ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μη Παραμετρικές Μέθοδοι Στατιστικής Ανάλυσης, ROC Ανάλυση και Εφαρμογές σε Πραγματικά Σεισμολογικά Δεδομένα

ΟΝΟΜΑΤΕΠΩΝΥΜΟ: ΠΑΡΑΣΚΕΥΟΠΟΥΛΟΥ ΠΗΝΕΛΟΠΗ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΧΡΗΣΤΟΣ ΚΟΥΚΟΥΒΙΝΟΣ

ΑΘΗΝΑ 2011

Ευχαριστίες

Για να ολοκληρωθεί η παρούσα διπλωματική εργασία ήταν απαραίτητη η συμβολή και βοήθεια ορισμένων ατόμων τους οποίους οφείλω να ευχαριστήσω. Καταρχήν τον Καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου κ. Χρήστο Κουκουβίνο καθώς μου έδωσε την ευκαιρία να ασχοληθώ με ένα εξαιρετικά ενδιαφέρον αντικείμενο. Επίσης θερμές ευχαριστίες θα ήθελα να εκφράσω στην υποψήφια διδάκτωρ Χριστίνα Παρπούλα για την συνεργασία και καθοδήγησή της. Τέλος θα ήθελα να ευχαριστήσω τη συμφοιτήτρια και φίλη Λένα Βαλαντή για τη συνεχή υποστήριξη και το ενδιαφέρον της καθ'όλη τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Περίληψη

Οι ROC καμπύλες αποτελούν μία χρήσιμη τεχνική για την οργάνωση, επιλογή και απεικόνιση ταξινομητών με βάση τη γραφική τους παράσταση. Η ROC ανάλυση έχει χρησιμοποιηθεί ευρέως στην απεικόνιση και ανάλυση της συμπεριφοράς διαγνωστικών συστημάτων σε διάφορους τομείς όπως η Ιατρική και πρόσφατα στον χώρο της Μηχανικής μάθησης καθώς και στην εξόρυξη δεδομένων. Οι ROC καμπύλες καθώς και ο υπολογισμός του εμβαδού κάτω από τις καμπύλες αυτές με τη χρήση παραμετρικών και μη παραμετρικών μεθόδων είναι το αντικείμενο μελέτης της παρούσας διπλωματικής εργασίας.

Στο πρώτο κεφάλαιο παρουσιάζονται αναλυτικά τρεις μη παραμετρικές στατιστικές μέθοδοι καθώς και ο τρόπος λειτουργία τους. Αυτές είναι οι Wilcoxon-Mann-Whitney Test, τεστ Kolmogorov-Smirnov και Kruskal-Walis τεστ..

Στο δεύτερο κεφάλαιο αναλύονται οι βασικές έννοιες της ROC ανάλυσης, εξετάζονται τα διάφορα σημεία του ROC χώρου και η σημασία τους καθώς και η τυχαία αναπαράσταση. Επιπλέον διερευνώνται οι μέθοδοι υπολογισμού μιας μέσης ROC καμπύλης καθώς και η μέθοδος του ROC convex hull για τη σύγκριση ενός πλήθους γνωστών ταξινομητών.

Το τρίτο κεφάλαιο πραγματεύεται τις διάφορες μεθόδους υπολογισμού του εμβαδού κάτω από μία ROC καμπύλη (AUC). Παρουσιάζονται δύο παραμετρικές και δύο μη παραμετρικές προσεγγίσεις για την εκτίμηση του εμβαδού AUC. Οι μη παραμετρικές προσεγγίσεις είναι η χρήση του Mann–Whitney στατιστικού (MW) και η προσαρμογή μιας λείας ROC καμπύλης χρησιμοποιώντας τη μέθοδο λείανσης με πυρήνες (K). Οι παραμετρικές προσεγγίσεις είναι πρώτον να υποθέσουμε ότι οι τιμές του δείκτη για τα παθολογικά αλλά και για τα υγιή περιστατικά ακολουθούν την κανονική κατανομή και έπειτα να υπολογίσουμε το AUC εμβαδόν χρησιμοποιώντας τυπικές παραμετρικές μεθόδους (N) και δεύτερον να εφαρμόσουμε ένα μετασχηματισμό τύπου Box–Cox και έπειτα έχοντας αποκτήσει τον κατάλληλο μετασχηματισμό να χρησιμοποιήσουμε τη θεωρία κανονικότητας (NT).

Στο τέταρτο και τελευταίο κεφάλαιο παρουσιάζεται μία εφαρμογή της ROC ανάλυσης σε ένα δείγμα τιμών που αναφέρονται στην ένταση σεισμικών δονήσεων. Στόχος της εφαρμογής αυτής ήταν η εύρεση και η περαιτέρω μελέτη των σημαντικότερων παραγόντων στην πρόβλεψη μιας ενδεχόμενης μελλοντικής ισχυρής δόνησης προκειμένου να ληφθούν τα αντίστοιχα μέτρα ασφάλειας.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ Ι

ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ ΣΤΗΝ ΑΝΑΛΥΣΗ ΚΑΤΗΓΟΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

1.1ΕΙΣΑΓΩΓΗ

Παραμετρικές μέθοδοι: Έγινε η υπόθεση ότι τα δεδομένα έρχονται από κάποια βαθύτερη κατανομή της οποίας η γενική μορφή είναι γνωστή. Οι στατιστικές μέθοδοι για αξιολόγηση και έλεγχο της υπόθεσης είναι τότε βασισμένες στην κατανομή. Ο στόχος μας βρίσκεται στις παραμέτρους και στον έλεγχο της υπόθεσης γύρω από αυτές.

Μη παραμετρικές μέθοδοι: Δεν γίνονται υποθέσεις πάνω στη βαθύτερη κατανομή.

- ♦ Το συμπέρασμα δεν επικεντρώνεται στις ειδικές πληθυσμιακές παραμέτρους.
- ♦ Ίσως λιγότερο δυνατές από τις αντίστοιχες παραμετρικές όταν οι υποθέσεις ικανοποιούνται.

1.2 WILCOXON-MANN-WHITNEY ΤΕΣΤ

■ Αυτό χρησιμοποιείται για τον έλεγχο της μηδενικής υπόθεσης κατά την οποία οι δύο πληθυσμοί έχουν ίδιες συναρτήσεις κατανομής έναντι της εναλλακτικής υπόθεσης κατά την οποία οι δύο συναρτήσεις κατανομής διαφέρουν μόνο ως προς την τοποθεσία (median),αν όχι καθόλου.

■ Δεν απαιτεί την προϋπόθεση οι διαφορές ανάμεσα στα δύο δείγματα να είναι ομαλά κατανεμημένες.

■ Το τεστ χρησιμοποιείται εις αντικατάσταση των δύο δειγματικών t-τεστ όπου η υπόθεση ομαλότητας είναι αμφίβολη.

■ Αυτό το τεστ μπορεί να εφαρμοστεί όταν οι παρατηρήσεις σε ένα δείγμα δεδομένων είναι βαθμιδωτές, δηλαδή ταξικά δεδομένα και όχι ευθείς μετρήσεις.

Το τεστ λειτουργεί με τη βαθμολόγηση του συνδυασμένου συνόλου δεδομένων, διαιρώντας τους βαθμούς σε δύο σύνολα σύμφωνα με το γκρουπ των μελών των αυθεντικών παρατηρήσεων, και υπολογίζοντας ένα δείγμα z στατιστικό, χρησιμοποιώντας τη συγκεντρωμένη εκτίμηση της ποικιλίας. Για μεγάλα δείγματα, το στατιστικό συγκρίνεται με τα ποσοστά επί τις εκατό της κανονικής κατανομής. Για μικρά δείγματα, το στατιστικό συγκρίνεται με το αποτέλεσμα που θα είχαμε αν τα δεδομένα συνδυάζονταν σε ένα μονό σύνολο δεδομένων και μεταβιβάζονταν στην τύχη σε δύο γκρουπ έχοντας τον ίδιο αριθμό παρατηρήσεων με αυτόν των αυθεντικών δειγμάτων.

1.2.1. ΕΦΑΡΜΟΓΗ ΤΟΥ WILCOXON-MANN-WHITNEY ΤΕΣΤ

Πρέπει να ελεγχθεί αν το δείγμα 1 και το δείγμα 2 έχουν επιλεχθεί τυχαία από την ίδια κατανομή. Τότε ακολουθείται η εξής διαδικασία:

1. Ταξινομούνται όλες οι παρατηρήσεις σε μία βαθμονομημένη σειρά. Αυτό θα πει ότι βαθμονομούνται οι παρατηρήσεις ανεξάρτητα από το γκρουπ στο οποίο ανήκει η κάθε μία
2. Αθροίζονται οι βαθμούς των παρατηρήσεων, που προέρχονται από κάθε δείγμα. Έστω $R_{1.}$ είναι το άθροισμα των βαθμών των παρατηρήσεων του δείγματος 1 και $R_{2.}$ το άθροισμα των βαθμών για το δείγμα 2.
3. Το U στατιστικό τότε δίνεται από τον τύπο:

$$U_1 = R_{1.} - \frac{n_1(n_1+1)}{2}$$

όπου n_1 είναι το μέγεθος του δείγματος 1 και R_1 είναι το άθροισμα των βαθμών για το δείγμα 1.

<u>Σημείωση</u>: Δεν υπάρχει διαφορά ως προς το ποιο δείγμα ονομάζουμε 1 ή 2.
Ένας ισοδύναμος τύπος για το U είναι :

$$U_2 = R_{2.} - \frac{n_2(n_2+1)}{2}$$

όπου n_2 είναι το μέγεθος του δείγματος 2 και R_2 είναι το άθροισμα των βαθμών για το δείγμα 2.

Από τα δύο στατιστικά U_1 και U_2 θα χρησιμοποιηθεί το μικρότερο, όταν θα ανατρέξουμε στους πίνακες σημαντικότητας. Το άθροισμα των δύο στατιστικών δίνει :

$$U_1 + U_2 = R_{1.} - \frac{n_1(n_1+1)}{2} + R_{2.} - \frac{n_2(n_2+1)}{2}$$

Γνωρίζοντας ότι $R_{1.} + R_{2.}$ = N(N+1)/2 και ότι N = $n_1 + n_2$ προκύπτει εύκολα μετά από πράξεις ότι :

$$U_1 + U_2 = n_1 n_2.$$

Για μεγάλα δείγματα το U στατιστικό είναι προσεγγιστικά κανονικά κατανεμημένο. Σε αυτή την περίπτωση έχουμε το στατιστικό :

$$z = \frac{U - m_U}{\sigma_U},$$

Όπου $m_U = \frac{n_1 n_2}{2}$ είναι ο μέσος

και $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$ είναι η τυπική απόκλιση του U

Σημείωση:Ηαπόλυτη τιμή του zστατιστικού θα είναι η ίδια οποιαδήποτε τιμή του U και αν χρησιμοποιηθεί.

1.3 KRUSKAL-WALIS ΤΕΣΤ

■ Χρησιμοποιείται για τη σύγκριση τριών ή περισσότερων δειγμάτων.
■ Χρησιμοποιείται για τον έλεγχο της μηδενικής υπόθεσης κατά την οποία όλοι οι πληθυσμοί έχουν ίδιες συναρτήσεις κατανομής έναντι της εναλλακτικής υπόθεσης κατά την οποία τουλάχιστον δύο από τα δείγματα διαφέρουν μόνο όσων αφορά την θέση.
■ Αποτελεί το ανάλογο του F-τεστ που χρησιμοποιείται στην ανάλυση διασποράς. Ενώ τα τεστ που χρησιμοποιούνται στην ανάλυση διασποράς εξαρτώνται από την υπόθεση ότι όλοι οι πληθυσμοί που βρίσκονται υπό σύγκριση είναι κανονικά κατανεμημένοι, το τεστ του Kruskal-Walis δε βάζει τέτοιους περιορισμούς στη σύγκριση.
■ Αποτελεί μία λογική επέκταση του Wilcoxon-Mann-Whitney τεστ σε περισσότερα από δύο γκρουπ.

1.4 ΕΦΑΡΜΟΓΗ ΤΟΥ KRUSKAL-WALIS ΤΕΣΤ

Έστω ότι υπάρχουν g δείγματα Πρέπει να ελεγχθεί αν τα δείγματα αυτά έχουν επιλεχθεί τυχαία από την ίδια κατανομή. Ακολουθούνται πάλι τα βήματα 1 και 2 που αναφέρθηκαν στην μέθοδο του Wilcoxon-Mann-Whitney τεστ. Έστω ότι συμβολίζεται με Ti το άθροισμα των n_i βαθμών στο δείγμαi(δηλαδή $T_i = R_{i.}$) και M_i το μέσο όρο των n_i βαθμών στο δείγμαi (δηλαδή $M_i = \overline{R}_{i.}$). Επίσης έστω ότι συμβολίζεται με T_{all} το άθροισμα των Νβαθμών σε όλα τα δείγματα συνολικά (δηλαδή $T_{all} = \sum_{i=1}^{g} R_{i.}$) και M_{all} τον αντίστοιχο μέσο όρο (δηλαδή $M_{all} = \frac{\sum_{i=1}^{g} R_{i.}}{N}$).

Ορίζουμε $SS_{bg(R)} = \sum n_i (M_i - M_{all})^2 = \sum \frac{T_i^2}{n_i} - \frac{T_{all}^2}{N}$, το στατιστικό του Kruskal-Walis είναι :

$$H = \frac{SS_{bg(R)}}{N(N+1)/12}$$

Κατά τη μηδενική υπόθεση προκύπτει:

$$H \sim X^2_{g-1}$$

1.4 ΤΕΣΤ KOLMOGOROV-SMIRNOV

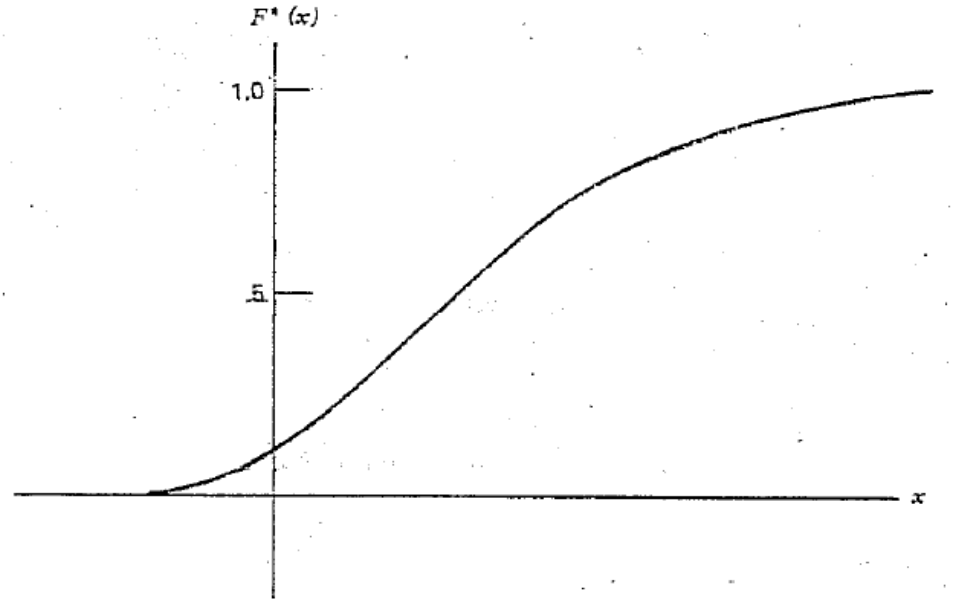
Η εμπειρική συνάρτηση κατανομής αναφέρθηκε ως η συνάρτηση, η βασισμένη σε ένα τυχαίο δείγμα το οποίο ίσως να χρησιμοποιήθηκε ακριβώς για να εκτιμηθεί η πραγματική συνάρτηση κατανομής του πληθυσμού. Αν θέλουμε να διαπιστώσουμε αν δύο ή περισσότερα δείγματα ακολουθούν την ίδια άγνωστη κατανομή είναι αναμενόμενο να συγκρίνουμε τις εμπειρικές συναρτήσεις κατανομής, αυτών των δειγμάτων, προκειμένου να ελέγξουμε τα κοινά χαρακτηριστικά που αυτές μπορεί να έχουν. Για να γίνουμε πιο ακριβείς, παρόλα αυτά, κάποια μέτρηση της ανομοιότητας μεταξύ ή ανάμεσα αυτών των συναρτήσεων είναι απαραίτητη. Ο Kolmogorovκαι oSmirnov ανέπτυξαν στατιστικές διαδικασίες οι οποίες χρησιμοποιούν την μέγιστη κάθετη απόσταση μεταξύ αυτών των συναρτήσεων ως ένα μέτρο για το πόσο πολύ μοιάζουν η μία στην άλλη. Οι μέθοδοί τους αυτές, παρουσιάζονται σε αυτήν την παράγραφο.

1.4.1 ΕΛΕΓΧΟΣ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΤΟΥ ΜΟΝΤΕΛΟΥ ΤΟΥ KOLMOGOROV

Αρχικά θα αναφερθεί ένας έλεγχος καταλληλότητας μοντέλου που εισάχθηκε από τον Kolmogorov (1933).Αυτό το τεστ είναι εξαιρετικής χρησιμότητας, πρώτον γιατί μας εφοδιάζει με μία εναλλακτική μέθοδο, σχεδιασμένη για ταξικά δεδομένα, ως προς αυτήν του chi-square τεστ καταλληλότητας μοντέλου που εισάχθηκε από τον Pearson, σχεδιασμένο όμως εκείνο για ονομαστικά δεδομένα και δεύτερον γιατί ο στατιστικός έλεγχος του Kolmogorov μας δίνει την δυνατότητα να σχηματίσουμε έναν "δεσμό εμπιστοσύνης" για την άγνωστη συνάρτηση κατανομής.

Ένας έλεγχος καταλληλότητας μοντέλου συνήθως περιλαμβάνει την εξέταση ενός τυχαίου δείγματος από κάποια άγνωστη κατανομή προκειμένου να ελέγξει τη μηδενική υπόθεση κατά την οποία η άγνωστη συνάρτηση κατανομής είναι στην πραγματικότητα μία γνωστή, συγκεκριμένη συνάρτηση. Αυτό σημαίνει ότι η μηδενική υπόθεση προσδιορίζει ακριβώς κάποια συνάρτηση κατανομής $F^*(x)$, ίσως γραφικά ,όπως στην εικόνα 1, ή ίσως σαν μία μαθηματική συνάρτηση η οποία μπορεί να σχεδιαστεί. Τότε παίρνουμε ένα τυχαίο δείγμα $X_1, X_2, ..., X_n$ από κάποιο

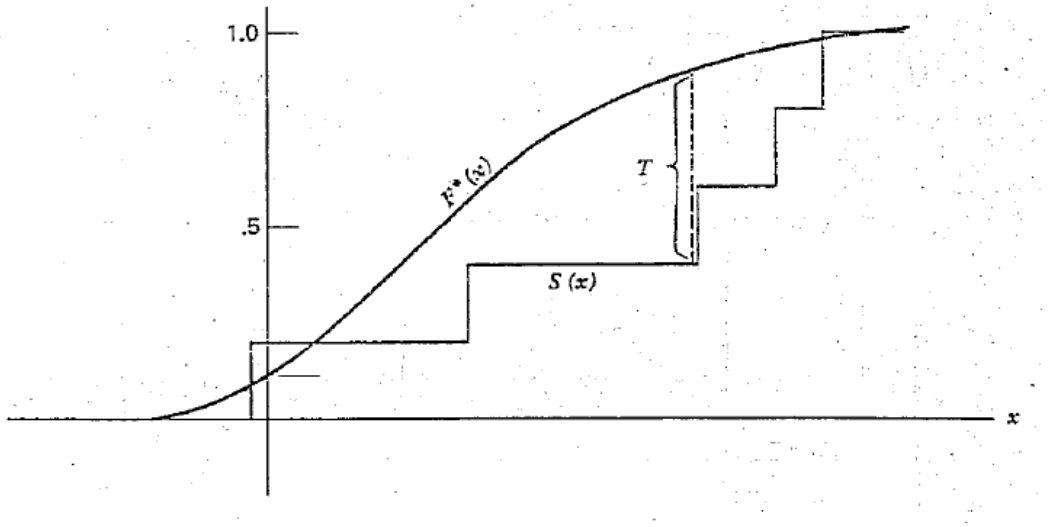
πληθυσμό και το συγκρίνουμε με την $F^*(x)$ με κάποιο τρόπο για να διαπιστώσουμε αν είναι αληθές, το να πούμε ότι η $F^*(x)$ είναι η πραγματική συνάρτηση κατανομής αυτού του τυχαίου δείγματος.



Εικόνα 1.1 Μία υποθετική συνάρτηση κατανομής.

Ένας λογικός τρόπος για να συγκριθεί το τυχαίο δείγμα με την $F^*(x)$ είναι με την μέθοδο της εμπειρικής συνάρτησης κατανομής S(x), η οποία ορίστηκε ως το κλάσμα των X_iς, που είναι μικρότερο ή ίσο του x, για κάθε x, $-\infty < x < +\infty$. Εμπειρική συνάρτηση κατανομής S(x)χρησιμοποιείται ως η εκτιμήτρια της $F(x)$, της άγνωστης δηλαδή συνάρτησης κατανομής των X_iς. Έτσι μπορούμε να συγκρίνουμε την εμπειρική συνάρτηση κατανομής S(x) με την υποθετική συνάρτηση κατανομής $F^*(x)$ προκειμένου να διαπιστώσουμε αν υπάρχει συμφωνία. Αν δεν υπάρχει συμφωνία τότε εμείς θα πρέπει να απορρίψουμε την μηδενική υπόθεση και να συμπεράνουμε ότι η πραγματική αλλά άγνωστη συνάρτηση κατανομής, $F(x)$, στην πραγματικότητα δεν δίνεται από την συνάρτηση $F^*(x)$ κατά την μηδενική υπόθεση. Αλλά τι είδους στατιστικό τεστ θα μπορούσαμε να χρησιμοποιήσουμε σαν μέτρο "διαφωνίας" (ανομοιογένειας) μεταξύ των S(x) και $F^*(x)$; Ένα από τα πιο απλά φανταστικά μέτρα είναι η μέγιστη απόσταση μεταξύ των δύο γραφικών παραστάσεων τωνS(x) και $F^*(x)$, μετρημένο σε κάθετη διεύθυνση. Αυτό είναι το στατιστικό το προτεινόμενο από τον Kolmogorov(1933). Αυτό σημαίνει ότι αν η $F^*(x)$ δίνεται από την εικόνα 1 και ένα τυχαίο δείγμα μεγέθους 5 παίρνεται από τον πληθυσμό, τότε η συνάρτηση κατανομής S(x) μπορεί να σχεδιαστεί στο ίδιο γράφημα κατά μήκος της $F^*(x)$ όπως φαίνεται στη εικόνα 2. Τώρα, αν οι S(x) και $F^*(x)$ έχουν μορφή όπως αυτή που δίνεται στην εικόνα 2, τότε η μέγιστη κάθετη απόσταση μεταξύ των δύο γραφικών παραστάσεων βρίσκεται μόλις πριν το τρίτο σκαλοπάτι της S(x). Αυτή η

απόσταση είναι περίπου 0.5 στην εικόνα 2. Παρόλα αυτά το στατιστικό T του Kolmogorov ισοδυναμεί με 0.5 σε αυτή την περίπτωση. Μεγάλες τιμές του Τα οδηγούν σε απόρριψη της $F^*(x)$ ως μία αξιόπιστη προσέγγιση της άγνωστης, πραγματικής συνάρτησης κατανομής $F(x)$.



Εικόνα 1.2.Η υποθετική συνάρτηση κατανομής $F^*(x)$, η εμπειρική συνάρτηση κατανομής S(x) και το T στατιστικό του Kolmogorov.

Το τεστ του Kolmogorov μπορεί να προτιμηθεί σε σχέση με το αντίστοιχο chi-squaretεστ, για την καταλληλότητα του μοντέλου, αν το μέγεθος του δείγματος είναι μικρό. Το τεστ του Kolmogorov είναι ακριβές ακόμα και για μικρά δείγματα ενώ το αντίστοιχο chi-squaretεστ προϋποθέτει ότι ο αριθμός των παρατηρήσεων είναι μεγάλος αρκετά ώστε ηchi-square κατανομή να δώσει μία καλή προσέγγιση όπως η κατανομή του στατιστικού τεστ. Υπάρχει μία διαφωνία ως προς το πιο τεστ είναι το πιο δυνατό, αλλά η γενική αίσθηση είναι ότι το τεστ του Kolmogorov είναι πιθανόν πιο δυνατό σε σχέση με αυτό του chi-square, στις περισσότερες περιπτώσεις.

<u>Έλεγχος καταλληλότητας μοντέλου του Kolmogorov</u>

ΔΕΔΟΜΕΝΑ. Τα δεδομένα αποτελούνται από ένα τυχαίο δείγμα $X_1, X_2, ..., X_n$, μεγέθους n, συνδεδεμένα με μία άγνωστη συνάρτηση κατανομής, δηλωμένη ως $F(x)$.

ΠΡΟΥΠΟΘΕΣΗ.
1.Το δείγμα είναι τυχαίο.

ΥΠΟΘΕΣΕΙΣ. Ας ορίσουμε $F^*(x)$, να είναι μία τελείως ειδική υποθετική συνάρτηση κατανομής.

Α. (Δίπλευρο τεστ)

$H_0 : F(x) = F^*(x)$ για όλα τα x από $-\infty$ έως $+\infty$

$H_1 : F(x) \neq F^*(x)$ για τουλάχιστον μία τιμή του x

Β. (Μονόπλευρο τεστ)

$H_0 : F(x) \geq F^*(x)$ για όλα τα x από $-\infty$ έως $+\infty$

$H_1 : F(x) < F^*(x)$ για τουλάχιστον μία τιμή του x

Γ. (Μονόπλευρο τεστ)

$H_0 : F(x) \leq F^*(x)$ για όλα τα x από $-\infty$ έως $+\infty$

$H_1 : F(x) > F^*(x)$ για τουλάχιστον μία τιμή του x

ΣΤΑΤΙΣΤΙΚΟ ΤΕΣΤ. Ας ορίσουμε S(x) να είναι η εμπειρική συνάρτηση κατανομής, βασισμένη στο τυχαίο δείγμα $X_1, X_2, ..., X_n$. Το στατιστικό τεστ ορίζεται διαφορετικά για τα τρία διαφορετικά σύνολα υποθέσεων Α, Β και Γ.

Α. (Δίπλευρο τεστ) Θέτω το στατιστικό ελέγχου Τ να είναι η μεγαλύτερη κάθετη απόσταση μεταξύ των S(x) και $F^*(x)$. Με σύμβολα έχουμε:

(1) $$T = \sup_x \left| F^*(x) - S(x) \right|$$

το οποίο διαβάζεται ‘‘ το Τ ισοδυναμεί με το supremum, κατά μήκος όλων των x, της απόλυτης τιμής της διαφοράς $F^*(x) - S(x).$’’

Β.(Μονόπλευρο τεστ) Δηλώνω αυτό το στατιστικό ελέγχου με T^+ και θέτω αυτό ίσο με την μεγαλύτερη κάθετη απόσταση που σχηματίζεται μεταξύ των S(x)και $F^*(x)$, όταν η $F^*(x)$ είναι πάνω από την S(x). Αυτό σημαίνει ότι,

(2) $$T^+ = \sup_x \left[F^*(x) - S(x) \right]$$

το οποίο είναι παρόμοιο με το Τ αν εξαιρέσουμε το γεγονός ότι εδώ εμείς υπολογίζουμε την μέγιστη απόσταση μόνο όπου η συνάρτηση $F^*(x)$ βρίσκεται πάνω από την συνάρτηση S(x).

Γ. (Μονόπλευρο τεστ) Για αυτό το τεστ χρησιμοποιούμε το στατιστικό ελέγχου T^-, ορισμένο ως την μεγαλύτερη κάθετη απόσταση που σχηματίζεται μεταξύ των S(x) και $F^*(x)$, όταν η S(x)είναι πάνω από την $F^*(x)$. Τυπικά, αυτό γίνεται

(3) $$T^- = \sup_x \left[S(x) - F^*(x) \right]$$

ΚΑΝΟΝΑΣ ΣΥΜΠΕΡΑΣΜΑΤΟΣ. Απορρίπτω την H_0, σε επίπεδο σημαντικότητας α, αν το κατάλληλο στατιστικό ελέγχου, $T, T^+ ή T^-$, υπερβαίνει το 1-α

Παράδειγμα1. Έχει επιλεχθεί τυχαίο δείγμα μεγέθους 10: $X_1 = 0.621, X_2 = 0.503$ $X_3 = 0.203, X_4 = 0.477, X_5 = 0.710, X_6 = 0.581, X_7 = 0.329, X_8 = 0.480, X_9 = 0.554, X_{10} = 0.382.$ Η μηδενική υπόθεση υποστηρίζει ότι η συνάρτηση κατανομής είναι η ομοιόμορφη συνάρτηση κατανομής της οποίας το γράφημα δίνεται στην εικόνα 3. Η μαθηματική έκφραση της υποθετικής συνάρτησης κατανομής είναι:

$F^*(x) = 0$ αν $x < 0$

(4) $= x$ αν $0 \le x < 1$

$= 1$ αν $1 \le x$

Τυπικά οι υποθέσεις δίνονται ως εξής:

$H_0 : F(x) = F^*(x)$ για όλα τα x

$H_1 : F(x) \neq F^*(x)$ για τουλάχιστον μία τιμή του x

όπου $F(x)$ είναι η άγνωστη συνάρτηση κατανομής, η οποία είναι κοινή για όλα τα X_i ς και $F^*(x)$ δίνεται από την εξίσωση 4.

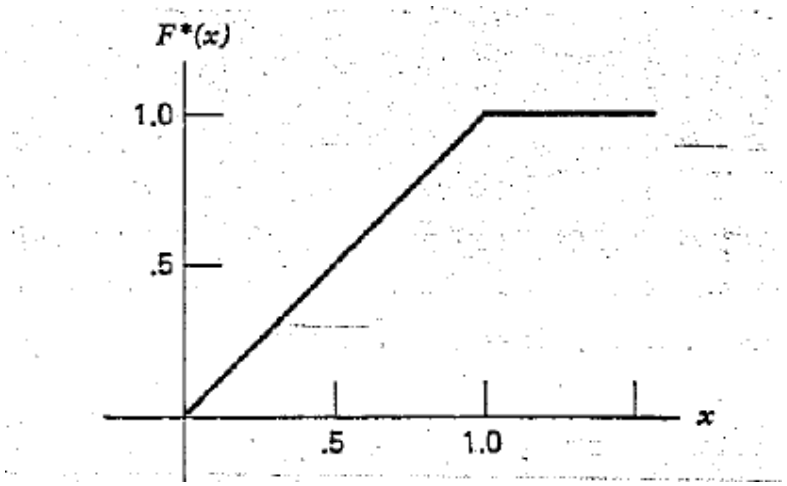
Το δίπλευρο Kolmogorov τεστ για την καταλληλότητα μοντέλου χρησιμοποιείται. Η κρίσιμη περιοχή, για το επίπεδο σημαντικότητας α=0.05, αντιστοιχεί σε τιμές του T μεγαλύτερες του 0.409. Η τιμή του T υπολογίζεται σχεδιάζοντας την γραφική παράσταση της εμπειρικής συνάρτησης κατανομήςS(x)στην κορυφή της υποθετικής συνάρτησης κατανομής $F^*(x)$ όπως φαίνεται στην εικόνα 4. Η μεγαλύτερη κάθετη απόσταση που διαχωρίζει τις δύο γραφικές παραστάσεις στην εικόνα 4 είναι 0.290, και αυτή βρίσκεται στο σημείο x = 0.710 επειδή S(0.710) = 1.000 και $F^*(0.710) = 0.710$. Με άλλα λόγια,

$$T = \sup_x \left| F^*(x) - S(x) \right|$$

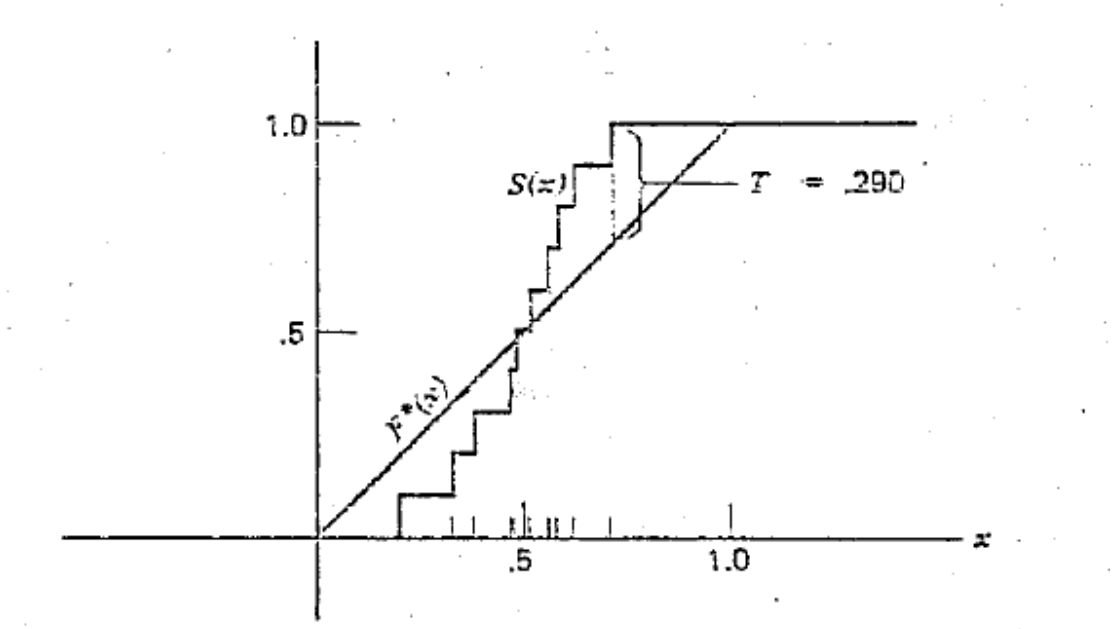
$$= \left| F^*(0.710) - S(0.710) \right|$$

$$= 0.290$$

Εφόσον T = 0.290 είναι μικρότερο από 0.409, η μηδενική υπόθεση γίνεται δεκτή.



Εικόνα 1.3. Η υποθετική συνάρτηση κατανομής



Εικόνα 1.4.Οι γραφικές παραστάσεις των $F^*(x)$ andS(x), με το T.

Αν έπρεπε να ελεγχθεί η εξής μηδενική υπόθεση
$H_0 : F(x) \geq F^*(x)$ για όλα τα x

έναντι της μονόπλευρης εναλλακτικής
$H_1 : F(x) < F^*(x)$ για κάποια x

το στατιστικός ελέγχου T^+ θα είχε χρησιμοποιηθεί. Ο κανόνας για το συμπέρασμα είναι ότι απορρίπτουμε την H_0, σε επίπεδο σημαντικότητας α = 0.05 αν το T^+

υπερβαίνει το 0.369. Η τιμή του T^+ σε αυτήν την περίπτωση υπολογίζεται μόλις στα αριστερά της δεύτερης αναπήδησης της S(x).

$$T^+ = \sup_x \left[F^*(x) - S(x)\right]$$

$$= F^*(0.3289) - S(0.3289)$$
$$= 0.3289 - 0.100$$
$$= 0.2289$$

Θα πρέπει να αναφερθεί ότι $T^+ = 0.228999\ldots$,το οποίο στρογγυλεύεται στο 0.229. Το τελικό αποτέλεσμα είναι το ίδιο.

Ένα μονόπλευρο τεστ, της άλλης κατεύθυνσης θα είχε σαν αποτέλεσμα

$$T^- = \sup_x \left[S(x) - F^*(x)\right]$$

$$= \mathrm{S}(0.710) - F^*(0.710)$$
$$= 1.000 - 0.710$$
$$= 0.290$$

Το δίπλευρο τεστ είναι το κατάλληλο για αυτή την κατάσταση. Τα μονόπλευρα τεστ παρουσιάστηκαν μόνο για να δείξουμε πως υπολογίζονται τα στατιστικά ελέγχου για καθένα από αυτά. Γενικά , βέβαια το δίπλευρο Τ στατιστικό ελέγχου ισοδυναμεί με το μεγαλύτερο από τα δύο μονόπλευρα στατιστικά ελέγχου T^+ και T^-.

1.4.2 ΜΙΑ ΜΕΘΟΔΟΣ ΓΙΑ ΤΗΝ ΑΠΟΚΤΗΣΗ ΤΟΥ ΑΚΡΙΒΟΥΣ ΚΡΙΣΙΜΟΥ ΕΠΙΠΕΔΟΥ ΟΤΑΝ Η $F^*(x)$ ΕΙΝΑΙ ΔΙΑΚΡΙΤΗ

Αν η υποθετική συνάρτηση κατανομής $F^*(x)$ είναι διακριτή και η "συντηρητική" προσέγγιση για το κρίσιμο επίπεδο που εξετάσαμε προηγουμένως δεν είναι ικανοποιητική, το ακριβές κρίσιμο επίπεδο μπορεί να αποκτηθεί για μία συγκεκριμένη παρατηρούμενη τιμή του στατιστικού ελέγχου. Αυτή η υπολογιζόμενη διαδικασία μπορεί να διεξαχθεί με το χέρι για δειγματικά μεγέθη ίσα ή μικρότερα του 5. Ένας υπολογιστής συνίσταται για μεγαλύτερα δειγματικά μεγέθη. Για μεγέθη δειγμάτων μεγαλύτερα από 30 ή 40 οι υπολογισμοί γίνονται περίπλοκοι ακόμα και σε έναν υπολογιστή.

Α. (Δίπλευρο τεστ) Έστω t, η παρατηρούμενη τιμή του Τ στατιστικού ελέγχου. Υπολογίζεται η $P(T^+ \geq t)$ και $P(T^- \geq t)$ όπως περιγράφεται στα μέρη Β και Γ που ακολουθούν, χρησιμοποιώντας t αντί για t^+ και t^-. Τότε

(5) $$\mathrm{P}(\mathrm{T} \geq t) = \mathrm{P}(T^+ \geq t) + \mathrm{P}(T^- \geq t)$$

είναι μία προσέγγιση, πολύ κοντά στο πραγματικό κρίσιμο επίπεδο στις περισσότερες περιπτώσεις, εκτός αν το t είναι πολύ μικρό. Το σφάλμα βρίσκεται στη συντηρητική πλευρά.

Β. (Μονόπλευρο τεστ) Ας θέσουμε t^+, να είναι η παρατηρούμενη τιμή του T^+.
Βήμα 1. Υπολογίζονται οι πιθανότητες f_i για $0 \le j < n(1-t^*)$,τραβώντας μία οριζόντια γραμμή με τεταγμένη $1-t^+-\frac{j}{n}$ σε ένα γράφημα της $F^*(x)$.Τότε $f_i = 1-t^*-\frac{j}{n}$ εκτός αν η οριζόντια γραμμή τέμνει την $F^*(x)$ με μία αναπήδηση, κατά την οποία η f_i ισοδυναμεί με το ύψος της $F^*(x)$ στην κορυφή της αναπήδησης. Μία από τις οριζόντιες γραμμές ίσως τέμνει την $F^*(x)$ ακριβώς στην κορυφή της αναπήδησης. Σε αυτή την περίπτωση η f_i ισοδυναμεί με την τεταγμένη της οριζόντιας γραμμής.
Βήμα 2. Υπολογίζουμε τις σταθερές $e_0, e_1, ...,$ από την σχέση $e_0 = 1$ και

(6) $$e_k = 1 - \sum_{j=0}^{k-1} \binom{k}{j} f_j^{k-j} e_j \quad k \ge 1$$

για όλα τα k, έτσι ώστε και f_k>0 στο Βήμα 1.Να σημειώσουμε ότι αυτές οι σταθερές είναι της μορφής:

$e_0 = 1$
$e_1 = 1 - f_0$
$e_2 = 1 - f_0^2 - 2f_1 e_1$
$e_3 = 1 - f_0^3 - 3f_1^2 e_1 - 3f_2 e_2$
$e_4 = 1 - f_0^4 - 4f_1^3 e_1 - 6f_2^2 e_2 - 4f_3 e_3$
$e_5 = 1 - f_0^5 - 5f_1^4 e_1 - 10f_2^3 e_2 - 10f_3^2 e_3 - 5f_4 e_4$
κ.λπ.

Βήμα 3. Υπολογίζω το κρίσιμο επίπεδο

(7) $$P(T^+ \ge t^+) = \sum_{j=0}^{\lceil n(1-t^+) \rceil} \binom{n}{j} f_j^{n-j} e_j$$

με τα f_j και e_i από τα βήματα 1 και 2.

Γ. (Μονόπλευρο τεστ) Έστω t^- να είναι η παρατηρούμενη τιμή του T^-.

Βήμα 1. Υπολογίζονται οι πιθανότητες c_j για $0 \le j < n(1-t^-)$ όπως ακολούθως. Φέρεται μία οριζόντια γραμμή με τεταγμένη $t^- + \frac{j}{n}$, πάνω στο γράφημα της $F^*(x)$. Τότε $c_j = 1 - t^- - \frac{j}{n}$ εφόσον η οριζόντια γραμμή δεν τέμνει $F^*(x)$ σε μία αναπήδηση της $F^*(x)$. Διαφορετικά c_j=1.0 μείον το ύψος της $F^*(x)$ στην κορυφή της αναπήδησης. Μία από τις οριζόντιες γραμμές ίσως να τέμνουν την $F^*(x)$ ακριβώς στην κορυφή της αναπήδησης. Σε αυτή την περίπτωση c_j=1.0 πλην την τεταγμένη αυτής της γραμμής.

Βήμα 2. Υπολογίζονται οι σταθερές $b_0, b_1, ...,$ από την σχέση $b_0 = 1$ και

$$b_k = 1 - \sum_{j=0}^{k-1} \binom{k}{j} c_{k-j} b_j \quad \text{k} \ge 1 \tag{8}$$

Για όλα τα kέτσι ώστε $c_k > 0$ στο βήμα 1.Αυτές οι σταθερές ακολουθούν το ίδια διαδικασία όπως και τα e_k ς, στο μέρος Β με τα f_j ς να έχουν αντικατασταθεί από τα c_j ς.

Βήμα 3. Υπολογίζουμε το κρίσιμο επίπεδο

$$P(T^- \ge t^-) = \sum_{j=0}^{\left[n\left(1-t^-\right)\right]} \binom{n}{j} c_j^{n-j} b_j \tag{9}$$

με τα c_j και b_j όπως στο βήματα 1 και 2.

Το ακόλουθο παράδειγμα απεικονίζει τη μέθοδο υπολογισμού του ακριβούς κρίσιμου επιπέδου όταν η $F^*(x)$ είναι διακριτή.

Παράδειγμα 2.Θέτω $F^*(x)$ να είναι η διακριτή ομοιόμορφη κατανομή με ίσες πιθανότητες 1/5 στα πέντε σημείαx= 1,2,3,4,5. Υποθέτω ότι ένα τυχαίο δείγμα μεγέθους 10 με τις διατεταγμένες τιμές 1, 1,1, 2, 2,2, 3, 3,3,3, παίρνεται από κάποιο πληθυσμό και ότι η μηδενική υπόθεση υποστηρίζει ότι η $F^*(x)$ είναι η συνάρτηση κατανομής του πληθυσμού. Η μεγαλύτερη απόσταση μεταξύ των $F^*(x)$ και S(x) βρίσκεται στο σημείο x=3 (βλέπε εικόνα 5), έτσι το στατιστικό ελέγχου για το δίπλευρο τεστ του Kolmogorov γίνεται

$$T = \sup_x \left| F^*(x) - S(x) \right| = 0.4 = \text{t} \tag{10}$$

Για να βρεθεί το κρίσιμο επίπεδο που συνδέεται με t = 0.4 υπολογίζεται η πιθανότητα $P(T^+ \geq 0.4)$.

Βήμα 1 Επειδή n(1-t) = 10(0.6) = 6 ,οι πιθανότητες f_0 έως και f_5 πρέπει να υπολογιστούν. Η οριζόντια γραμμή με τεταγμένη 1-t = 0.6 τέμνει την $F^*(x)$ ακριβώς στην κορυφή της αναπήδησης στο x = 3 και έτσι η f_0 ισοδυναμεί με την τεταγμένη της οριζόντιας γραμμής : f_0=0.6. Για j=1, η οριζόντια γραμμή $1-t-1/10 = 0.5$ τέμνει την $F^*(x)$ κατά την μία αναπήδηση, έτσι ώστε f_1 να ισοδυναμεί με το ύψος της $F^*(x)$ στην κορυφή της αναπήδησης: f_1=0.4 .Με παρόμοιο τρόπο βρίσκουμε ότι $f_2 = 0.4, f_3 = 0.2, f_4 = 0.2$ και f_5=0.

Βήμα 2. Οι σταθερές e_0 έως και e_4 υπολογίζονται από την εξίσωση 6

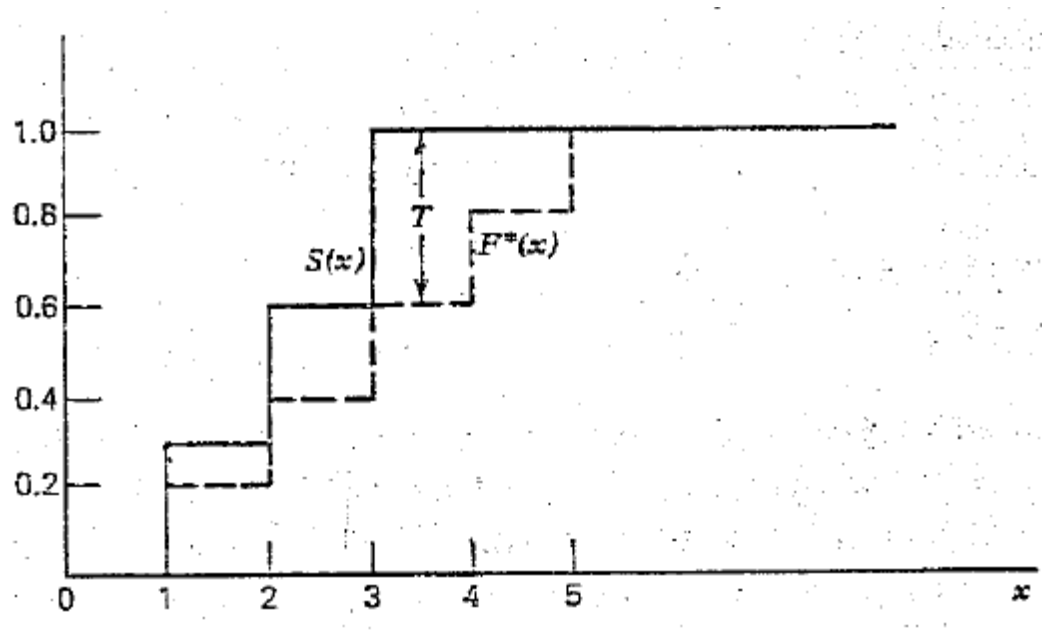
$e_0 = 1$

$e_1 = 1 - 0.6 = 0.4$

$e_2 = 1 - (0.6)^2 - 2(0.4)(0.4) = 0.32$

$e_3 = 1 - (0.6)^3 - 3(0.4)^2(0.4) - 3(0.4)(0.32) = 0.208$

$e_4 = 1 - (0.6)^4 - 4(0.4)^3(0.4) - 6(0.4)^2(0.32) - 4(0.4)(0.208) = 0.2944$



Εικόνα 1.5. Οι γραφικές παραστάσεις των $F^*(x)$ andS(x), με το T.

Βήμα 3. Το μονόπλευρο κρίσιμο επίπεδο $P(T^+ \geq t)$ υπολογίζεται από την εξίσωση 7.

$$(11) \qquad P(T^+ \geq t) = f_0^{10} + \binom{10}{1} f_1^9 e_1 + \binom{10}{2} f_2^8 e_2 + \binom{10}{3} f_3^7 e_3 + \binom{10}{4} f_4^6 e_4 = 0.2081$$

Επειδή η $F^*(x)$ είναι συμμετρική, ο υπολογισμός του άλλου μονόπλευρου κρίσιμου επιπέδου $P(T^- \geq 0.4)$ είναι ίδιος ακριβώς με τον προηγούμενο και έτσι $P(T^- \geq 0.4)$ = 0.02081 και το κρίσιμο επίπεδο για το δίπλευρο τεστ του Kolmogorov είναι προσεγγιστικά

(12) $P(T \geq 0.4) = 2(0.02081) = 0.04162$

Είναι ενδιαφέρον να σημειωθεί ότι αυτή η τιμή για το κρίσιμο επίπεδο δείχνει ότι η σωστή απόφαση είναι να απορρίψουμε τη μηδενική υπόθεση σε επίπεδο σημαντικότητας α = 0.05.

Σχόλιο. Ένα από τα πιο πολύτιμα χαρακτηριστικά του δίπλευρου στατιστικού ελέγχου του Kolmogorov είναι ότι το 1-α ποσοστημόριο του, w_{1-a} μπορεί να χρησιμοποιηθεί προκειμένου να διαμορφώσει ένα δεσμό εμπιστοσύνης για την πραγματική άγνωστη συνάρτηση κατανομής. Υπενθυμίζουμε ότι για να βρούμε ένα διάστημα εμπιστοσύνης για κάποια άγνωστη παράμετρο, εμείς πρώτα τραβήξαμε ένα τυχαίο δείγμα και τότε, από αυτό το δείγμα, υπολογίσαμε μία ανώτερη τιμή U και μία κατώτερη τιμή Loι οποίες περιείχαν την άγνωστη παράμετρο ανάμεσά τους με μία συγκεκριμένη πιθανότητα 1-α, ονομαζόμενη ως συντελεστής εμπιστοσύνης. Θα ήταν πολύ βολικό αν εμείς μπορούσαμε να κάνουμε το ίδιο πράγμα για να αποκτήσουμε ένα διάστημα εμπιστοσύνης μέσα στο οποίο ολόκληρη η άγνωστη συνάρτηση κατανομής θα κυμαίνονταν με πιθανότητα 1-α. Τότε θα μπορούσαμε να τραβήξουμε ένα τυχαίο δείγμα για κάποιο πληθυσμό του οποίου η συνάρτηση κατανομής είναι ολοκληρωτικά γνωστή και θα μπορούσαμε να τοποθετήσουμε κάποια όρια σε ένα γράφημα και να κάνουμε την δήλωση ότι η άγνωστη συνάρτηση κατανομής κυμαίνεται ολοκληρωτικά μέσα σε αυτά, με την πιθανότητα 1-α η δήλωση να είναι σωστή.

Διάστημα Εμπιστοσύνης Για Την Συνάρτηση Κατανομής

ΔΕΔΟΜΕΝΑ. Τα δεδομένα αποτελούνται από ένα τυχαίο δείγμα $X_1, X_2, ..., X_n$, μεγέθους n, συνδεδεμένα με μία άγνωστη συνάρτηση κατανομής, δηλωμένη ως $F(x)$.

ΠΡΟΥΠΟΘΕΣΗ.

1**.** Το δείγμα είναι τυχαίο.

2. Για να είναι ο συντελεστής εμπιστοσύνης ακριβής οι τυχαίες μεταβλητές θα πρέπει να είναι συνεχείς .Αν οι τυχαίες μεταβλητές είναι διακριτές το διάστημα εμπιστοσύνης είναι συντηρητικό. Αυτό σημαίνει ότι ο πραγματικός αλλά άγνωστος συντελεστής εμπιστοσύνης είναι μεγαλύτερος από αυτόν που δηλώσαμε εμείς.

ΜΕΘΟΔΟΣ. Σχεδιάζω την γραφική παράσταση της εμπειρικής συνάρτησης κατανομής S(x) βασισμένη στο τυχαίο δείγμα. Για να σχηματίσω ένα διάστημα εμπιστοσύνης με συντελεστή εμπιστοσύνης 1-α, βρίσκω το 1-α ποσοστημόριο του στατιστικού ελέγχου του Kolmogorov (από Πίνακα) για το δίπλευρο τεστ και για το κατάλληλο μέγεθος n. Θέτω ως w_{1-a} αυτό το ποσοστημόριο. Σχεδιάζω ένα γράφημα πάνω από αυτό της S(x) σε απόσταση w_{1-a} και ονομάζω το γράφημα αυτό U(x).Τότε σχεδιάζω ένα δεύτερο γράφημα σε απόσταση w_{1-a} κάτω από την S(x) και ονομάζω το δεύτερο αυτό γράφημα L(x). Τα γραφήματα των U(x) και S(x) αποτελούν τότε το ανώτερα και κατώτερα όρια ενός 1-α διαστήματος εμπιστοσύνης που περιέχει πλήρως την άγνωστη F(x) μέσα στα σύνορά του.
Δεν υπάρχει λόγος η U(x) να σχεδιαστεί πιο πάνω από 1.0 ακόμα και αν S(x) + w_{1-a} ξεπερνούν το 1.0, γιατί όπως γνωρίζουμε καμία συνάρτηση κατανομής δεν ξεπερνά το 1.0. Για τον ίδιο λόγο η L(x) δεν θα πρέπει να εκτείνεται κάτω από τον οριζόντιο άξονα. Οι τυπικοί μαθηματικοί ορισμοί των U(x) και L(x)είναι ως ακολούθως :

(13) $U(x) = S(x) + w_{1-a}$ αν $S(x) + w_{1-a} \le 1$

$U(x) = 1.0$ αν $S(x) + w_{1-a} > 1$

(14) $L(x) = S(x) - w_{1-\alpha}$ αν $S(x) - w_{1-\alpha} \ge 0$

$L(x) = 0$ αν $S(x) - w_{1-\alpha} < 0$

Η πιθανότητα να επαληθεύεται η υπόθεση:

(15) $P[L(x) \le F(x) \le U(x), \text{για όλα τα x}] \ge 1\text{-}\alpha$

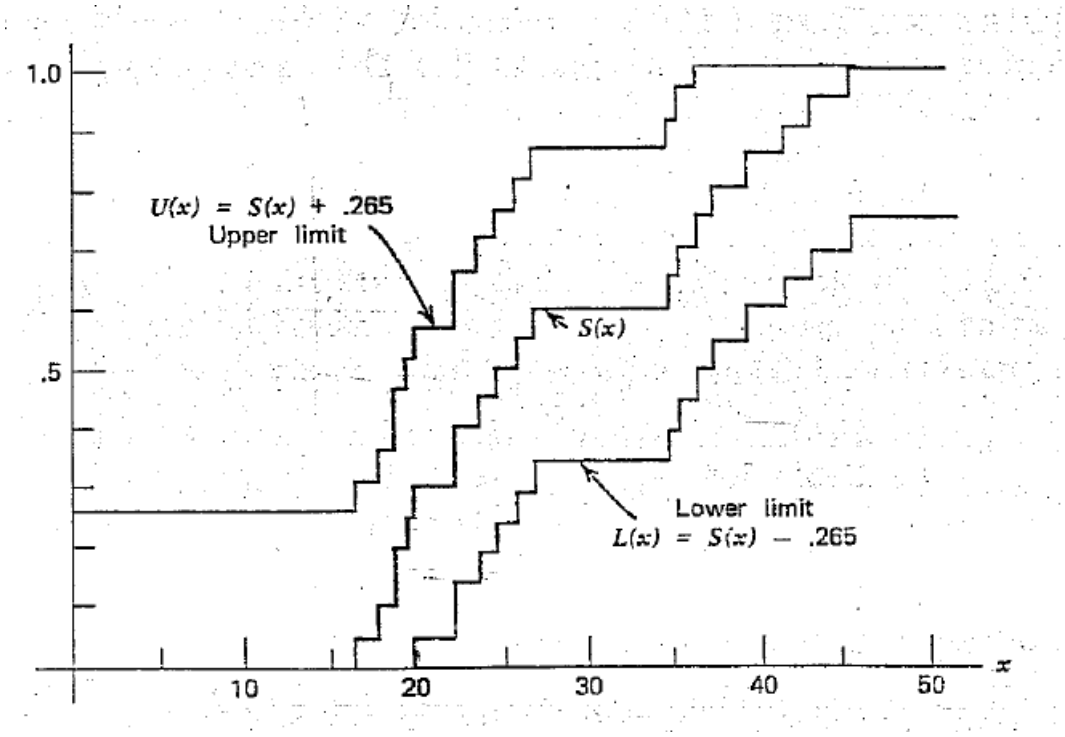
όπου η τελευταία ανισότητα εφαρμόζεται μόνο αν οι τυχαίες μεταβλητές είναι διακριτές.

Παράδειγμα 3. Έστω ότι προέκυπτε η ανάγκη να σχηματίστεί ένα 90% διάστημα εμπιστοσύνης για μία άγνωστη συνάρτηση κατανομής F(x). Ένα τυχαίο δείγμα μεγέθους 20 αποκτάται από κάποιο πληθυσμό με αυτή τη συνάρτηση κατανομής. Τα αποτελέσματα κατατάσσονται από το μικρότερο προς το μεγαλύτερο για μεγαλύτερη ευκολία.

16.7	17.4	18.1	18.2	18.8	19.3	22.4	22.4	24.0	24.7
25.9	27.0	25.1	35.8	36.5	37.6	39.8	42.1	43.2	46.2

Το 0.90 ποσοστημόριο ισούται με $w_{0.90}$ = 0.265 για n = 20.Το διάστημα εμπιστοσύνης είναι S(x) ± 0.265 εφόσον το διάστημα είναι μεταξύ 0 και 1. Η εικόνα

6 δείχνει τις S(x),U(x) και L(x). Η δήλωση "η F(x)κυμαίνεται ολοκληρωτικά ανάμεσα στιςU(x) και L(x)" επαληθεύεται με πιθανότητα 0.90.



Εικόνα 1.6. Ένα διάστημα εμπιστοσύνης για την $F(x)$.

Η ασυμπωτική κατανομή του δίπλευρου Τ στατιστικού βρέθηκε από τον Kolmogorov (1933) και καταχωρήθηκε σε πίνακα από τον Smirnov (1948). Οι ασυμπτωτικές κατανομές των μονόπλευρων στατιστικών T^+και T^-αποκτήθηκαν από τον Smirnov(1939). Η ακριβής κατανομή των στατιστικών τεστ για πεπερασμένα δειγματικά μεγέθη μελετήθηκε από τους Waldκαι Wolfowitz (1939) και καταχωρήθηκε σε πίνακα από τον Massey(1950a). Ησυνάρτηση κατανομής του T^- για πεπερασμένα δειγματικά μεγέθη, φτιάχτηκε από τους Birnbaum καιTingey (1951). Επιπλέον πραγματοποιήθηκαν συγκρίσεις μεταξύ των ακριβών ποσοστημορίων, που πήραμε από τις συναρτήσεις κατανομής και των ασυμπτωτικών ποσοστημορίων που δόθηκαν από τον Smirnov(1939, 1948). Διαπιστώθηκε ότι ασυμπτωτικά ποσοστημόρια οδηγούν σε ένα συντηρητικό τεστ.

Το δίπλευρο τεστ του Kolmogorovέχει την επιθυμητή ιδιότητα να διατηρεί την συνοχή του απέναντι σε όλες τις διαφορές ανάμεσα στις $F(x)$ και $F^*(x)$, την πραγματική και υποθετική συνάρτηση κατανομής. Ένα χαμηλότερο όριο για την δύναμη του δίπλευρου τεστ δίνεται από τον Massey(1950b). Το μεγαλύτερο, χαμηλότερο όριο για την δύναμη, κάτω από κάποιες εναλλακτικές υποθέσεις δόθηκε από τον Birnbaum (1953) και ένα δεύτερο από τα μεγαλύτερα χαμηλότερα όρια για την δύναμη , κάτω από άλλες εναλλακτικές υποθέσεις, δόθηκε από τον Lee (1966).

Ο Lee (1966) επίσης σύγκρινε την ακριβή δύναμη του τεστ του Kolmogorov με ένα τυπικό παραμετρικό τεστ. Κάποια από τα ευρήματά του θα τα παρουσιάσουμε εδώ. Ένα τυχαίο δείγμα μεγέθους 5 θεωρήθηκε ότι τραβήχτηκε από ένα πληθυσμό με

κανονική κατανομή, με μέση τιμή μ_1 και διασπορά σ^2. Η μηδενική υπόθεση υποστηρίζει ότι η κατανομή είναι κανονική με μέση τιμή μ_0, όχι μ_1, αλλά με την ίδια διασπορά. Η δύναμη του τεστ του Kolmogorov αποκτήθηκε από διαφορές μεταξύ των μ_0 και μ_1, σχετικές με το μέγεθος της τυπικής απόκλισης σ, και συγκρίθηκε με το "κανονικό τεστ" το οποίο είναι το πιο δυνατό παραμετρικό τεστ που υπάρχει για αυτήν την περίπτωση. Ακόμα και κάτω από αυτές τις καθόλου ευνοϊκές συνθήκες , η δύναμη του τεστ του Kolmogorov δεν είναι πολύ κατώτερη από αυτή του "κανονικού τεστ". Άλλες συγκρίσεις δύναμης έγιναν από τον Vander Waerden (1953), Suzuki(1968), Shapiro, Wilk, και Chen(1968), και Knott(1970). Να σημειώσουμε εδώ ότι αν μία παρέκκλιση από την υποθετική διασπορά υπάρξει έναντι μιας απόκλισης από τον υποθετικό μέσο, όπως πριν, το κανονικό τεστ είναι λιγότερο δυνατό στο να ανακαλύψει την διαφορά ενώ το τεστ του Kolmogorov είναι αντίστοιχα πιο δυνατό. Οι Govindarajulu και Klotz (1973) παρουσιάζουν μία σημείωση στην ασυμπτωτική κατανομή. Ο υπολογισμός και ο έλεγχος συμμετρικών κατανομών είναι το θέμα των σημειώσεων των Shcuster και Navarte(1973),Shcuster(1973), και Srinivasan και Godio(1974).

ΚΕΦΑΛΑΙΟ ΙΙ

ROC ΚΑΜΠΥΛΕΣ

2.1 ΕΙΣΑΓΩΓΗ ΣΤΙΣ ΚΑΜΠΥΛΕΣ ROC

Η σύλληψη των καμπύλων ROC(Receiver Operating Characteristic: Λειτουργικό Χαρακτηριστικό Δέκτη)έχει τις ρίζες τις στις αρχές της δεκαετίας του 1950 όταν προτάθηκαν από τους μεταπτυχιακούς φοιτητές WW Petersonκαι TG Birdsall του Τμήματος Ηλεκτρολόγων Μηχανικών του Πανεπιστημίου του Michigan, οι οποίοι εφάρμοσαν τη στατιστική θεωρία αποφάσεων σε προβλήματα της θεωρίας λήψης σημάτων (signal detection theory). Οι καμπύλες ROCπροτάθηκαν αρχικά ως γραφική μέθοδος μέτρησης της ποιότητας λήψης σήματος από ένα δέκτη σε ατελή διαγνωστικά συστήματα. Συγκεκριμένα, όταν για ένα δέκτη οι κατανομές του εξωτερικού θορύβου και του σήματος (στόχου της λήψης) συμπίπτουν σε κάποιο διάστημα (και έστω ότι χωρίς βλάβη της γενικότητας οι μετρήσεις σήματος έχουν υψηλότερες μετρήσεις γενικά από αυτές του θορύβου), η καμπύλη ROCπου αντιστοιχεί δείχνει τη συμμεταβολή των ποσοστών ‘ορθής λήψης σήματος’ από τις πραγματικές μετρήσεις σήματος και ‘εσφαλμένου συναγερμού’ από τις πραγματικές μετρήσεις θορύβου, συναρτήσει του σημείου απόφασης – δηλαδή της μέτρησης με βάση την οποία ο δέκτης θεωρεί ότι κάθε άλλη μέτρηση μεγαλύτερη αυτής αποτελεί σήμα.

Ένα ROC γράφημα αποτελεί χρήσιμη τεχνική για την οργάνωση, επιλογή και απεικόνιση ταξινομητών με βάση τη γραφική τους παράσταση. Η ROC ανάλυση έχει χρησιμοποιηθεί ευρέως στην απεικόνιση και ανάλυση της συμπεριφοράς διαγνωστικών συστημάτων. Η ιατρική κοινότητα λήψης αποφάσεων έχει ένα μακρύ ιστορικό πάνω στη χρήση των ROC γραφημάτων, στους διαγνωστικούς της ελέγχους. Οι Swets, Dawes και Monahan πρόσφατα έφεραν τις ROC καμπύλες στην προσοχή του ευρέος κοινού με το άρθρο τους *Scientific American.*

Ένας από τους πρώτους που υιοθέτησαν τα ROCγραφήματα στην μηχανική μάθηση ήταν ο Spackman (1989), ο οποίος ανέδειξε την αξία των ROCκαμπυλών στην αξιολόγηση και σύγκριση αλγορίθμων. Τα τελευταία χρόνια έχει παρατηρηθεί μία άνοδος στη χρήση των ROC γραφημάτων στην κοινότητα μηχανικής εκμάθησης. Δεν αποτελούν όμως μόνο μία χρήσιμη μέθοδο αναπαράστασης γραφημάτων αλλά έχουν και ιδιότητες που τις καθιστούν ιδιαίτερα χρήσιμες σε τομείς με ασύμμετρες κατανομές και άνισα ταξινομημένα σφάλματα.

Τα περισσότερα βιβλία εξόρυξης δεδομένων και μηχανικής εκμάθησης αν δεν αναφέρουν καθόλου τις ROCκαμπύλες, θα έχουν μόνο μία σύντομη περιγραφή της τεχνικής. Οι ROCκαμπύλες είναι εννοιολογικά απλές αλλά υπάρχουν κάποιες, όχι και τόσο εμφανείς περιπλοκές που εμφανίζονται όταν αυτές οι καμπύλες χρησιμοποιούνται στην έρευνα. Υπάρχουν επίσης συχνές παρανοήσεις και παγίδες όταν αυτές χρησιμοποιούνται στην πράξη.

2.2 ΒΑΣΙΚΟΙ ΟΡΙΣΜΟΙ-ΕΥΑΙΣΘΗΣΙΑ ΚΑΙ ΕΙΔΙΚΟΤΗΤΑ

Αρχικά θεωρήθηκαν προβλήματα ταξινόμησης που χρησιμοποιούν μόνο δύο κλάσεις(ROC Fawcett). Επίσης κάθε περίπτωση *I* αντιστοιχεί αυστηρά σε ένα μόνο στοιχείο από το ζεύγος {p,n}που αντιστοιχεί στις τιμές των κλάσεων των θετικών και αρνητικών αποτελεσμάτων. Ένα μοντέλο ταξινόμησης (ή ταξινομητής) είναι μία αντιστοίχηση από τα διάφορα περιστατικά στις αντίστοιχα προβλεπόμενες κλάσεις. Κάποια μοντέλα ταξινόμησης παράγουν ένα συνεχές αποτέλεσμα στο οποίο μπορούν να εφαρμοστούν διαφορετικά όρια προκειμένου να προβλεφθεί σε ποιά κλάση ανήκουν τα διάφορα περιστατικά. Άλλα μοντέλα παράγουν διακριτές τιμές κλάσεων υποδηλώνοντας μόνο την προβλεπόμενη κλάση του περιστατικού. Για να διακρίνουμε την πραγματική κλάση από την προβλεπόμενη χρησιμοποιούμε τις ενδείξεις {Y,N}για τις προβλέψεις κλάσης που παράγονται από ένα μοντέλο. Δοσμένων έναν ταξινομητή και ένα περιστατικό, υπάρχουν τέσσερα δυνατά αποτελέσματα. Αν το περιστατικό είναι θετικό και ταξινομείται ως θετικό, μετριέται ως αληθώς θετικό, ενώ αν αυτό ταξινομείται ως αρνητικό τότε θεωρείται ψευδώς αρνητικό. Αντίστοιχα αν το περιστατικό είναι αρνητικό και ταξινομείται ως αρνητικό τότε λογίζεται ως αληθώς αρνητικό, ενώ αν αυτό ταξινομείται ως θετικό τότε θεωρείται ψευδώς θετικό. Δοσμένων ένα ταξινομητή και ένα σύνολο από περιστατικά(το σύνολο των περιστατικών του τεστ) ένας, δύο επί δύο πίνακας ενδεχομένων μπορεί να κατασκευαστεί αντιπροσωπεύοντας όλες τις δυνατές περιπτώσεις. Αυτός ο πίνακας αποτελεί τη βάση για πολλές συνήθεις μετρικές. Ο πίνακας 1 δείχνει έναν πίνακα ενδεχομένων και εξισώσεις διαφόρων συνήθων μετρικών που μπορούν να υπολογιστούν από αυτό. Οι τιμές κατά μήκος της κυρίας διαγωνίου αντιπροσωπεύουν τις σωστές αποφάσεις ως προς τον χαρακτηρισμό του περιστατικού ως θετικό ή αρνητικό. Αντίθετα οι τιμές εκτός αυτής της διαγωνίου αντιπροσωπεύουν τις λανθασμένες αποφάσεις(ή διαφορετικά την σύγχυση ανάμεσα στις διάφορες κλάσεις). Το *ποσοστό των αληθώς θετικών αποτελεσμάτων,***TPR(** True Positive Rate) ενός ταξινομητή εκτιμάται ως εξής:

$$\mathbf{TPR} = \frac{TP}{P} = \frac{\boldsymbol{Θετικά Σωστά ταξινομημένα}}{\boldsymbol{Σύνολο Θετικών}}$$

Το *ποσοστό των ψευδών θετικών αποτελεσμάτων,***FPR**(False Positive Rate)ενός ταξινομητή εκτιμάται ως εξής:

$$\mathbf{FPR} = \frac{FP}{N} = \frac{\boldsymbol{Αρνητικά Λανθασμένα ταξινομημένα}}{\boldsymbol{Σύνολο\ Αρνητικών}}$$

Επιπρόσθετοι όροι συνδεδεμένοι με τις ROC καμπύλες είναι η **ευαισθησία** (**sensitivity**) και η **ειδικότητα** (**specificity**) και ορίζονται ως εξής:

Sensitivity=SE=TPR

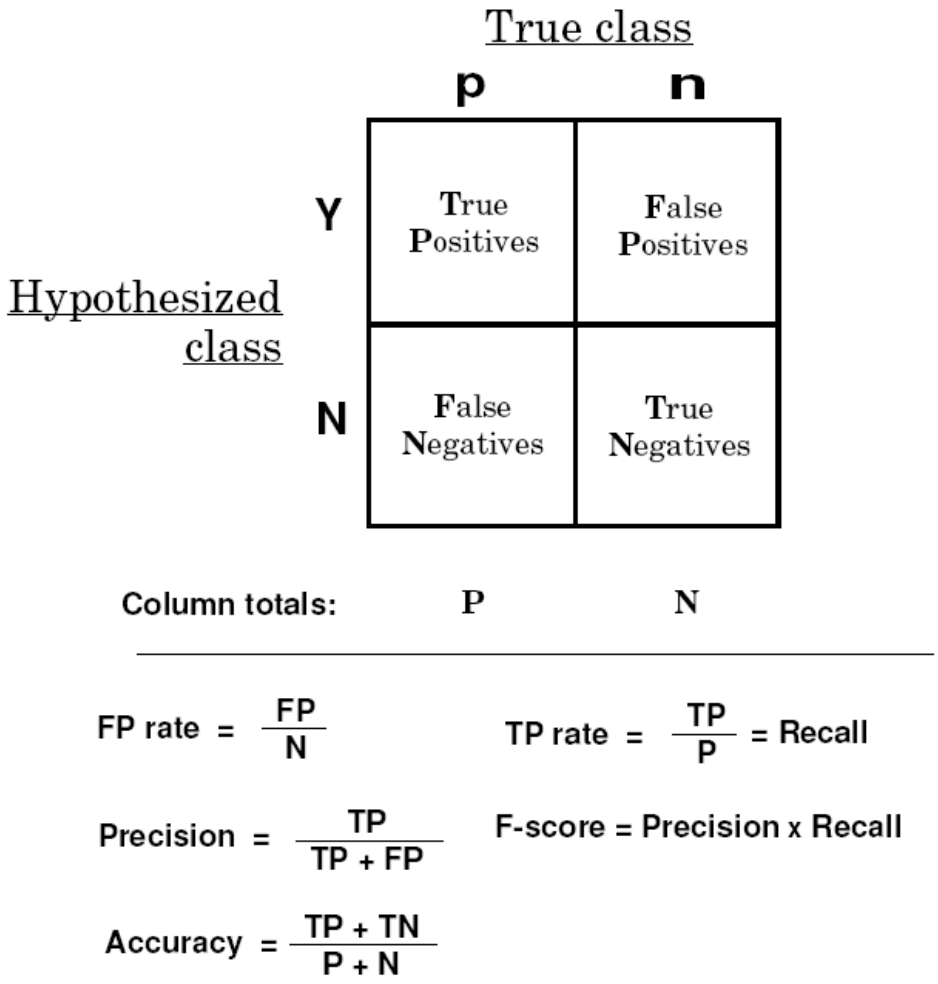
$$\mathbf{Specificity} = \frac{TN}{FP+TN}$$

$$=\frac{\boldsymbol{Αληθώς Αρνητικά}}{\boldsymbol{Ψευδώς Θετικά + Αληθώς Αρνητικά}}$$

=1-FPR

Ορίζονται εδώ και κάποιες άλλες πολύ χρήσιμες έννοιες που αφορούν τους διαγνωστικούς ελέγχους. Η ***θετική προβλεπόμενη τιμή*** (**PPV**) είναι μια από αυτές και ερμηνεύεται ως η πιθανότητα εμφάνισης θετικού περιστατικού μεταξύ όλων των θετικών προβλέψεων. Παρόμοια ορίζεται και η ***αρνητική προβλεπόμενη τιμή*** (**NPV**) ως η πιθανότητα εμφάνισης αρνητικού περιστατικού μεταξύ όλων των αρνητικών προβλέψεων. Οι παραπάνω υπολογίζονται ως εξής:

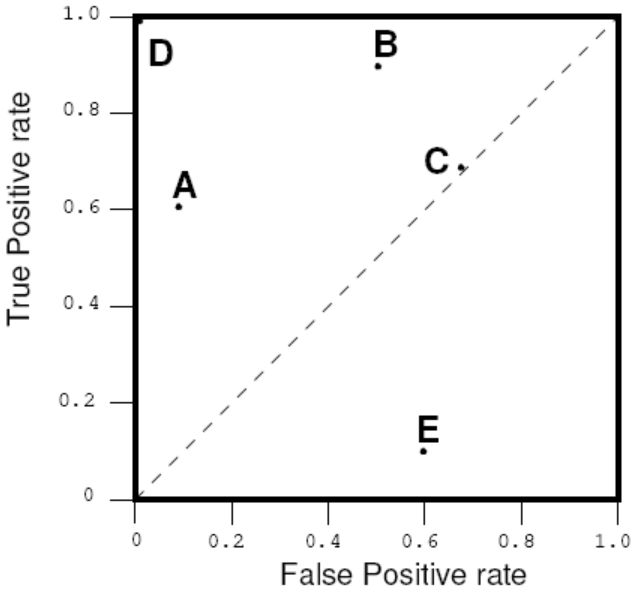
$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \textit{και} \quad \text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$



Εικόνα 2.1. Ένας πίνακας ενδεχομένων και διάφορες μετρικές αναπαράστασης που μπορούν να υπολογιστούν από αυτόν.

2.3 ROC ΧΩΡΟΣ

Τα ROCγραφήματα είναι δύο διαστάσεων γραφήματα στα οποία το TPRσχεδιάζεται στον Υάξονα και το FPRστον Χάξονα. Επιπλέον η ROC καμπύλη απεικονίζει τη συμμεταβολή του οφέλους(από το ποσοστό των αληθώς θετικών μετρήσεων) και του κόστους (από το ποσοστό των ψευδώς θετικών μετρήσεων). Η εικόνα 2.2 δείχνει ένα ROCγράφημα με πέντε ταξινομητές αριθμημένους από το Αέως το Ε. Ένας διακριτός ταξινομητής είναι αυτός που παράγει σαν αποτέλεσμα μόνο μία τιμή κλάσης. Κάθε διακριτός ταξινομητής παράγει ένα ζεύγος (FPR,TPR)το οποίο και αντιστοιχεί σε ένα μόνο σημείο πάνω στον ROCχώρο. Οι ταξινομητές στην εικόνα 2.2 είναι όλοι διακριτοί ταξινομητές.



Εικόνα 2.2.Ένα βασικό ROC γράφημα με πέντε διακριτούς ταξινομητές

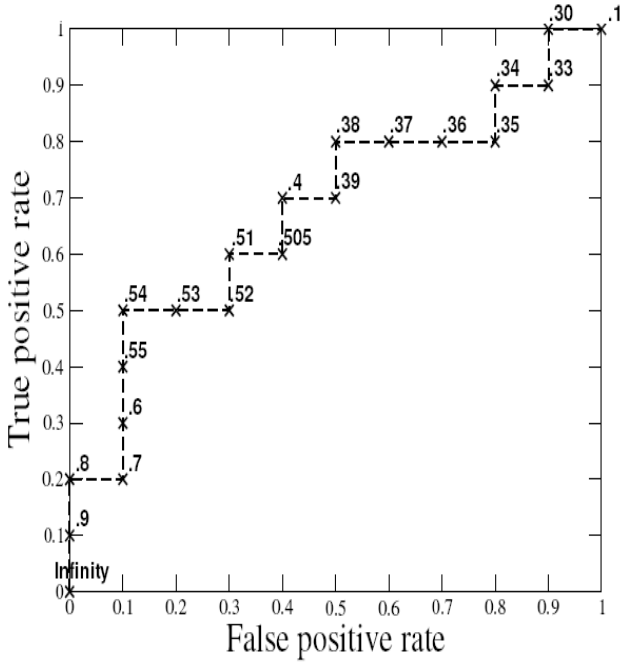
Κάποια σημεία πάνω στον ROCχώρο παρουσιάζουν ιδιαίτερο ενδιαφέρον. Αρχικά το χαμηλότερο σημείο από τα αριστερά, το (0,0) αντιπροσωπεύει την στρατηγική του να μην εκδίδεται ποτέ μία θετική ταξινόμηση. Αυτό σημαίνει ότι στην περίπτωση αυτή ένας ταξινομητής δεν παράγει ούτε ψευδώς θετικά αποτελέσματα αλλά και ούτε και αληθώς θετικά. Η αντίθετη περίπτωση, κατά την οποία, ανεξαρτήτως συνθηκών παράγονται μόνο θετικές ταξινομήσεις, αντιπροσωπεύεται από το ανώτερο από τα δεξιά σημείο, (1,1).

Το σημείο (0,1) αντιπροσωπεύει την τέλεια ταξινόμηση. Το σημείο Dστην παραπάνω γραφική παράσταση (Εικόνα 2) αντιστοιχεί στην ιδανική αυτή περίπτωση. Ανεπίσημα, ένα σημείο στον ROCχώρο είναι καλύτερο από ένα άλλο εφόσον βρίσκεται περισσότερο βορειοδυτικά από εκείνο, πράγμα που σημαίνει ότι (FPR χαμηλότερο,TPRυψηλότερο, ή και τα δύο). Οι ταξινομητές που εμφανίζονται στην αριστερή πλευρά του ROCγραφήματος, κοντά στον άξονα των X, μπορούν να θεωρηθούν πιο συντηρητικοί καθώς αυτοί πραγματοποιούν θετικές ταξινομήσεις μόνο όταν διαθέτουν ισχυρές ενδείξεις. Αυτό έχει σαν αποτέλεσμα να δίνουν λίγα ψευδώς θετικά αποτελέσματα. Η τακτική αυτή όμως έχει ως συνέπεια ένα χαμηλό TPR, δηλαδή περιορισμένο αριθμό αληθώς θετικών αποτελεσμάτων. Από την άλλη πλευρά, οι ταξινομητές που βρίσκονται στην άνω δεξιά πλευρά του ROCγραφήματος μπορούν να θεωρηθούν πιο φιλελεύθεροι. Αυτοί πραγματοποιούν θετικές ταξινομήσεις με ασθενέστερα στοιχεία με αποτέλεσμα να ταξινομούν σχεδόν όλα τα θετικά περιστατικά σωστά, αλλά συγχρόνως έχουν και υψηλό ποσοστό ψευδώς θετικών ταξινομήσεων. Στην Εικόνα 2.2, το σημείο A είναι περισσότερο συντηρητικό από το B. Επειδή σε πολλούς τομείς, στον πραγματικό κόσμο επικρατούν μεγάλοι αριθμοί αρνητικών περιστατικών, η περιοχή της αριστερής πλευράς του ROCγραφήματος παρουσιάζει ιδιαίτερο ενδιαφέρον.

2.3.1 ΤΥΧΑΙΑ ΑΝΑΠΑΡΑΣΤΑΣΗ

Η διαγώνιος, που αντιπροσωπεύεται από την εξίσωση y=x, αντιστοιχεί στην στρατηγική της τυχαίας πρόβλεψης μιας κλάσης. Για παράδειγμα, αν ένας ταξινομητής, τις μισές φορές τυχαία "μαντεύει" τη θετική κλάση, μπορεί να προβλεφτεί ότι θα ταξινομήσει τα μισά θετικά και τα μισά αρνητικά περιστατικά σωστά. Αυτή η περίπτωση αντιστοιχεί στο σημείο (0.5,0.5) του ROCχώρου. Αν αυτός "μαντεύει" τη θετική κλάση το 90% του χρόνου τότε μπορεί να προβλεφτεί ότι θα πάρει το 90% των θετικών σωστά αλλά και ότι το FPRθα αυξηθεί στο 90%, δίνοντας το σημείο (0.9,0.9) στο ROCχώρο. Έτσι ένας τυχαίος ταξινομητής θα δίνει ένα ROCσημείο που θα ολισθαίνει πότε μπροστά πότε πίσω πάνω στη διαγώνιο ανάλογα με την συχνότητα με την οποία ο ταξινομητής μαντεύει τη θετική κλάση. Προκειμένου να φύγουμε από την διαγώνιο και να μεταφερθούμε στον χώρο πάνω από αυτήν θα πρέπει ο ταξινομητής να εκμεταλλευτεί κάποιες πληροφορίες στα δεδομένα. Στην Εικόνα 2 ο ταξινομητής Cείναι σχεδόν τυχαίος. Δίνοντας το σημείο (0.7,0.7) μπορεί να προβλεφτεί ότι ο ταξινομητής C μαντεύει τη θετική κλάση το 70% των περιπτώσεων. Κάθε ταξινομητής που εμφανίζεται στη χαμηλότερη δεξιά γωνία αποδίδει χειρότερα από την τυχαία πρόβλεψη. Αυτή η γωνία είναι άλλωστε συνήθως άδεια στα ROCγραφήματα. Παρόλα αυτά αξίζει να σημειώσουμε ότι ο χώρος απόφασης εκτείνεται συμμετρικά γύρω από τη διαγώνιο, χωρίζοντας έτσι τις δύο γωνίες. Αν αναιρέσουμε ένα ταξινομητή αυτό σημαίνει αντιστροφή των αποφάσεων ταξινόμησής του, δηλαδή οι αρχικά αληθώς θετικές ταξινομήσεις του γίνονται λανθασμένα θετικές και οι λανθασμένα θετικές γίνονται αντίστοιχα αληθώς θετικές. Έτσι, κάθε ταξινομητής που δίνει ένα σημείο στη χαμηλότερη δεξιά γωνία του ROCχώρου μπορεί να αναιρεθεί και να δώσει έτσι ένα σημείο στην άνω αριστερή γωνία. Στην Εικόνα 2.2, ο ταξινομητής Εσυμπεριφέρεται πολύ χειρότερα από ότι ένας τυχαίος ταξινομητής και αποτελεί στην πραγματικότητα την αναίρεση του ταξινομητή Α.

Δοσμένου ενός ROC γραφήματος στο οποίο η αναπαράσταση του ταξινομητή φαίνεται ελαφρώς καλύτερη από τυχαία είναι εύλογη η ερώτηση αν αυτή η αναπαράσταση του συγκεκριμένου ταξινομητή είναι πραγματικά σημαντική ή απλά έτυχε να είναι καλύτερη από την τυχαία αναπαράσταση. Ενώ δεν υπάρχει συγκεκριμένο τεστ που να δίνει μία ικανοποιητική απάντηση σε αυτό ο Forman (2002) έδειξε μία μεθοδολογία η οποία κατευθύνεται στο να απαντήσει στο παραπάνω ερώτημα με τη χρήση ROC καμπυλών.



Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

Εικόνα 2.3. Η ROC καμπύλη που προήλθε από την οριοθέτηση των περιστατικών ενός τεστ. Ο πίνακας στα αριστερά περιέχει είκοσι δεδομένα και η τιμή που αντιστοιχεί στο καθένα από ένα ταξινομητή βαθμολόγησης. Το γράφημα στα αριστερά δείχνει την αντίστοιχη ROC καμπύλη και σε κάθε σημείο της υπάρχει η τιμή του ορίου από το οποίο προήλθε.

2.4 ΚΑΜΠΥΛΕΣ ΣΤΟΝ ROC ΧΩΡΟ

Πολλοί ταξινομητές όπως δέντρα απόφασης ή σύνολα κανόνων είναι σχεδιασμένοι έτσι ώστε να παράγουν μία μόνο απόφαση κλάσης, για παράδειγμα ένα Y(Yes) ήN(No) για την κάθε περίπτωση. Όταν ένας τέτοιος διακριτός ταξινομητής εφαρμόζεται σε ένα σύνολο δεδομένων ενός test, δίνει ένα πίνακα ενδεχομένων ο οποίος με τη σειρά του αντιστοιχεί σε ένα μόνο ROCσημείο. Έτσι, ένας διακριτός ταξινομητής παράγει ένα μόνο σημείο στον ROCχώρο. Κάποιοι ταξινομητές όπως ο Naïve Bayes ταξινομητής ή ένα νευρωνικό δίκτυο, δίνουν ένα περιστατικό πιθανότητας ή ένα σκορ ή μία αριθμητική τιμή που αντιπροσωπεύει το βαθμό κατά τον οποίο ένα περιστατικό αποτελεί μέλος μιας κλάσης. Τέτοιες τιμές μπορεί να είναι αυστηρές πιθανότητες κατά την οποία περίπτωση αυτές προσκολλώνται σε σταθερά θεωρήματα πιθανότητας. Επίσης οι τιμές αυτές θα μπορούσαν να είναι γενικά, μη βαθμονομημένα σκορ, κατά την οποία περίπτωση η μόνη ιδιότητα που διατηρούν είναι ότι ένα υψηλότερο σκορ υποδηλώνει μία μεγαλύτερη πιθανότητα. Εμείς θα επικαλεστούμε έναν πιθανολογικό ταξινομητή παρά το γεγονός ότι το αποτέλεσμα πιθανόν να μην είναι μία κατάλληλη πιθανότητα.

Ένας τέτοιος βαθμιδωτός ταξινομητής μπορεί να χρησιμοποιηθεί με ένα όριο(γνωστό και ως σημείο απόφασης) ώστε να παράγει ένα διακριτό (δυαδικό) ταξινομητή. Αν το αποτέλεσμα του ταξινομητή είναι πάνω από το όριο αυτό, ο ταξινομητής δίνει σαν αποτέλεσμα έναΥδιαφορετικά δίνει ένα N. Κάθε τιμή ορίου δίνει και ένα διαφορετικό σημείο πάνω στον ROC χώρο. Επομένως, εμείς μπορούμε να φανταστούμε ένα τέτοιο όριο-σημείο απόφασης να κυμαίνεται από το - ∞ έως το +∞ και να σχηματίζει έτσι

μία καμπύλη στον ROCχώρο. Ο αλγόριθμος 1(ο οποίος διατίθεται στο Παράρτημα Α) περιγράφει τη βασική ιδέα. Υπολογιστικά αυτός είναι ένας φτηνός τρόπος παραγωγής μιας ROCκαμπύλης και παρακάτω θα περιγραφεί μία πιο αποτελεσματική και προσεκτική μέθοδος. Η Εικόνα 2.3 δείχνει ένα παράδειγμα μιας ROCκαμπύλης από ένα σύνολο είκοσι περιστατικών ενός τεστ. Τα περιστατικά 10 θετικά και 10 αρνητικά φαίνονται στον πίνακα δίπλα από το γράφημα. Κάθε ROCκαμπύλη η οποία προέρχεται από ένα πεπερασμένο σύνολο περιστατικών είναι στην πραγματικότητα μία συνάρτηση βήματος η οποία προσεγγίζει μία πραγματική καμπύλη καθώς ο αριθμός των περιστατικών πλησιάζει το άπειρο. Η συνάρτηση βήματος στην Εικόνα 2.3 έχει προκύψει από ένα πολύ μικρό σύνολο περιστατικών και έτσι η προέλευση του κάθε σημείου μπορεί να γίνει εύκολα αντιληπτή. Στον πίνακα της Εικόνας 2.3, τα περιστατικά ταξινομούνται με βάση το σκορ τους και κάθε σημείο στο ROCγράφημα φέρει την τιμή του ορίου-σημείου απόφασης από το οποίο προέκυψε το σημείο αυτό. Το όριο $+\infty$ παράγει το σημείο (0,0). Καθώς εμείς κατεβάζουμε το όριο στο 0.9 το πρώτο θετικό περιστατικό ταξινομείται ως θετικό δίνοντας το σημείο (0,0.1). Καθώς το όριο περιορίζεται ακόμα, η καμπύλη σκαρφαλώνει προς τα πάνω και δεξιά καταλήγοντας στο σημείο (1,1) με μία τιμή ορίου 0.1. Αξίζει να σημειωθεί εδώ ότι, καθώς ρίχνουμε την τιμή του ορίου πιο χαμηλά αυτό αντιστοιχεί με μετακίνηση από τις πιο συντηρητικές περιοχές του γραφήματος στις περισσότερο φιλελεύθερες.

Αν και το σύνολο δεδομένων του τεστ είναι πολύ μικρό, είναι δυνατόν να γίνουν κάποιες δοκιμαστικές παρατηρήσεις για τον ταξινομητή. Αυτός εμφανίζεται να συμπεριφέρεται καλύτερα στην πιο συντηρητική περιοχή του γραφήματος. Το ROCσημείο (0.1,0.5) δίνει την μεγαλύτερη ακρίβεια(70%). Αυτό σημαίνει ότι ο ταξινομητής είναι καλύτερος στο να αναγνωρίζει πιθανά θετικά περιστατικά από ότι πιθανά αρνητικά περιστατικά. Αξίζει να σημειωθεί επίσης ότι η καλύτερη ακρίβεια του ταξινομητή επιτυγχάνεται στο σημείο απόφασης 0.54 αντί του σημείο 0.5 που ίσως να αναμέναμε(το οποίο αποδίδει 60%). Παρακάτω αναλύεται το φαινόμενο αυτό.

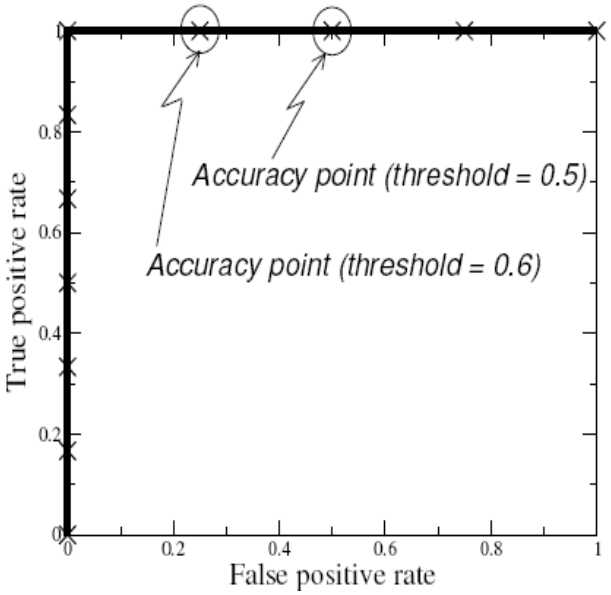
2.4.1 ΣΥΓΚΡΙΤΙΚΑ ΣΚΟΡ ΕΝΑΝΤΙ ΑΠΟΛΥΤΩΝ ΣΚΟΡ

Μία σημαντική ιδιότητα των ROCγραφημάτων είναι ότι μετρούν την ικανότητα ενός ταξινομητή να παράγει καλά συγκριτικά σκορ, των διαφόρων περιστατικών. Δεν είναι αναγκαίο και απαραίτητα χρήσιμο ένας ταξινομητής να παράγει ακριβείς, βαθμονομημένες εκτιμήσεις πιθανότητας. Αυτό που έχει κυρίως σημασία είναι να δίνει ακριβή συγκριτικά σκορ, γεγονός που θα εξυπηρετεί στην διάκριση μεταξύ θετικών και αρνητικών περιστατικών.

Έστω η απλή ένδειξη που φαίνεται στη Εικόνα 2.4, η οποία προήρθε από έναν NaiveBayes ταξινομητή. Συγκρίνοντας την υποθετική κλάση(η οποία είναι Y αν το σκορ>0.5, διαφορετικά N)με τις πραγματικές κλάσεις μπορούμε να παρατηρήσουμε ότι ο ταξινομητής ταξινομεί τα περιστατικά 7 και 8 ως λανθασμένα, με αποτέλεσμα να αποδίδει με 80% ακρίβεια. Παρόλα αυτά αν παρατηρήσουμε τη ROCκαμπύλη στην αριστερή πλευρά της Εικόνας 2.4 θα δούμε ότι η καμπύλη ανεβαίνει κάθετα από το (0,0) στο (0,1) και έπειτα συνεχίζει οριζόντια μέχρι το (1,1). Αυτό υποδηλώνει τέλεια αναπαράσταση ταξινόμησης για το σύνολο δεδομένων του συγκεκριμένου τεστ. Το ερώτημα λοιπόν είναι γιατί υπάρχει αυτή η διαφωνία. Η απάντηση δίνεται αν σκεφτούμε τι ακριβώς μετράμε στην κάθε περίπτωση. Η ROCκαμπύλη δείχνει την

ικανότητα ενός ταξινομητή να κατατάσσει τα θετικά περιστατικά συγκριτικά με τα αρνητικά και όπως είδαμε ανταπεξέρχεται τέλεια όσων αφορά αυτή τη λειτουργία. Η μετρική της ακρίβειας επιβάλει ένα όριο (σκορ>0.5) και μετρά τα αποτελέσματα των ταξινομήσεων με βάση τα σκορ. Το μέτρο της ακρίβειας θα ήταν τελείως κατάλληλο αν τα σκορ ήταν κατάλληλες πιθανότητες αλλά δεν είναι. Με άλλα λόγια, τα σκορ δεν είναι κατάλληλα βαθμονομημένα ενώ οι πραγματικές πιθανότητες είναι. ΣτονROC χώρο η επιβολή του 0.5 ως όριο έχει σαν αποτέλεσμα στην αναπαράσταση που έχει σχεδιαστεί στην Εικόνα 2.4, το κυκλωμένο σημείο ακρίβειας. Αυτό το λειτουργικό σημείο είναι υποδεέστερο από τέλειο. Θα μπορούσαμε επίσης να χρησιμοποιήσουμε το ίδιο σύνολο δεδομένων, για να πάρουμε ένα καλύτερο σημείο πάνω στη ROCκαμπύλη με το να επιλέξουμε ως όριο την τιμή της πιθανότητας p(P)= $=\frac{\theta\varepsilon\tau\iota\kappa\acute{\alpha}\ \Pi\varepsilon\rho\iota\sigma\tau\alpha\tau\iota\kappa\acute{\alpha}}{\Sigma\acute{\upsilon}\nu o\lambda o\ \Pi\varepsilon\rho\iota\sigma\tau\alpha\tau\iota\kappa\acute{\omega}\nu}=\frac{6}{10}=0.6$ αλλά ακόμα και σε αυτή την περίπτωση θα είχαμε μία όχι τέλεια αναπαράσταση(με ακρίβεια 90%). Ένας τρόπος για να εξαλείψουμε αυτό το φαινόμενο είναι να βαθμονομήσουμε τα σκορ του ταξινομητή. Υπάρχουν κάποιες μέθοδοι για να πραγματοποιήσουμε το παραπάνω (για παράδειγμα αυτές των Zadrozny & Elkan, 2001). Μια άλλη προσέγγιση είναι να χρησιμοποιήσουμε μία ROCμέθοδο η οποία επιλέγει λειτουργικά σημεία βασιζόμενη στην συγκριτική τους αναπαράσταση και υπάρχουν επίσης διάφορες μέθοδοι για αυτό(όπως αυτή των Provost & Fawcett, 1998,2001). Οι τελευταίες αυτές μέθοδοι θα συζητηθούν συνοπτικά στην παράγραφο 2.8.1).

Μία συνέπεια των συγκριτικών σκορ είναι ότι τα σκορ του ταξινομητή δεν πρέπει να συγκριθούν στις κλάσεις του μοντέλου. Μία κλάση μοντέλου μπορεί να είναι σχεδιασμένη να παράγει σκορ στο εύρος τιμών [0,1] ενώ μία άλλη να παράγει τιμές στο εύρος [-1,+1] ή [1,100]. Το να συγκρίνουμε την αναπαράσταση του μοντέλου με ένα κοινό όριο θα ήταν ανόητο.



Inst no.	Class True	Class Hyp	Score
1	p	Y	0.99999
2	p	Y	0.99999
3	p	Y	0.99993
4	p	Y	0.99986
5	p	Y	0.99964
6	p	Y	0.99955
7	n	Y	0.68139
8	n	Y	0.50961
9	n	N	0.48880
10	n	N	0.44951

Εικόνα 2.4. Τιμές και ταξινομήσεις δέκα περιστατικών καθώς και η ROC καμπύλη που προέρχεται από αυτά.

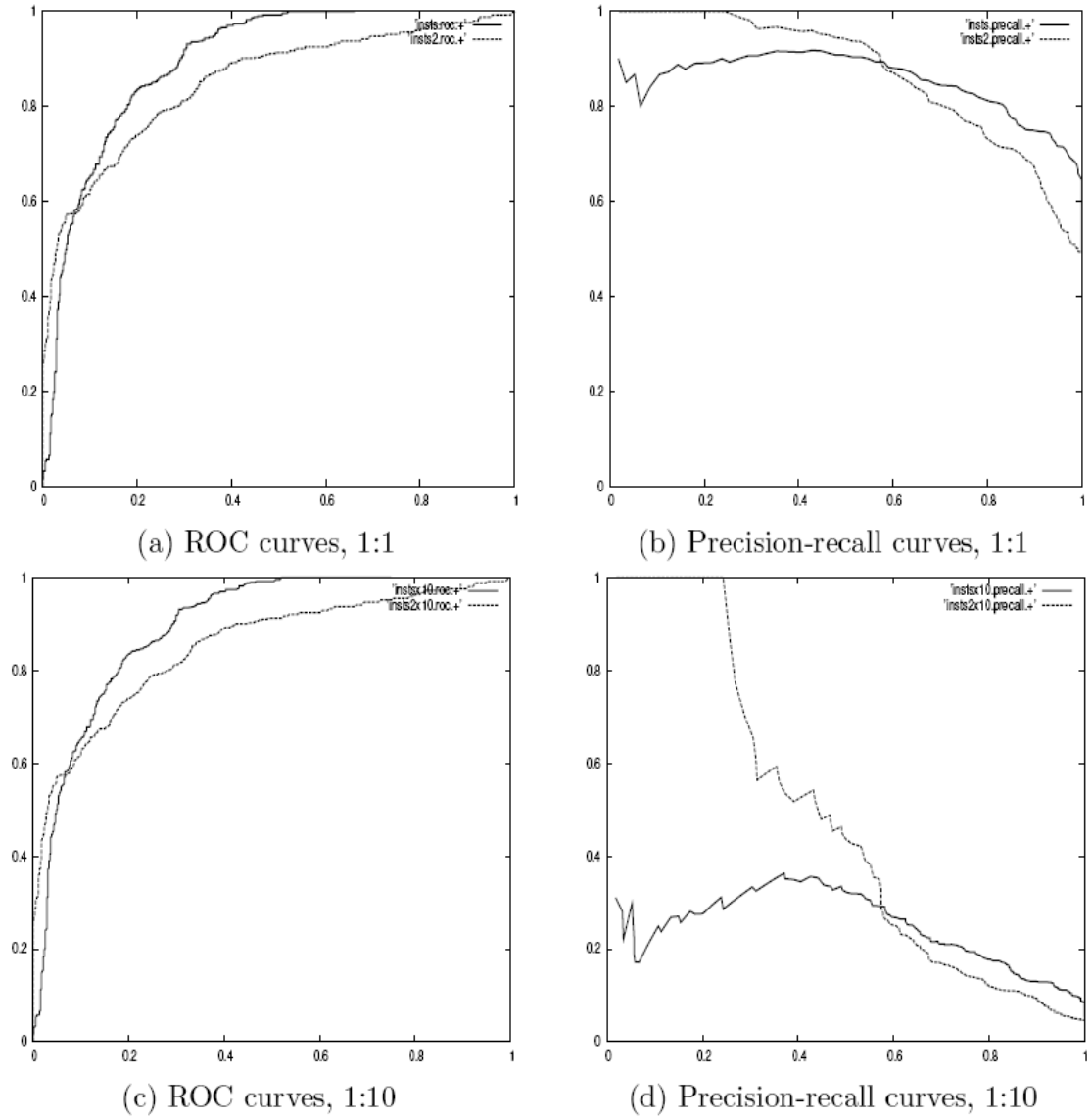
2.4.2 ΚΛΑΣΗ ΑΣΣΥΜΜΕΤΡΙΑΣ

Οι ROCκαμπύλες έχουν μία πολύ σημαντική ιδιότητα: δεν επηρεάζονται από αλλαγές στην κατανομή της κλάσης. Αυτό σημαίνει ότι αν η αναλογία των θετικών προς τα αρνητικά περιστατικά αλλάξει σε ένα σύνολο δεδομένων ενός τεστ οι ROCκαμπύλες θα παραμείνουν ανεπηρέαστες. Για να εξηγήσουμε γιατί συμβαίνει αυτό ας θεωρήσουμε τον πίνακα ενδεχομένων στην Εικόνα 2.1. Αξίζει εδώ να σημειωθεί ότι η κατανομή της κλάσης-η αναλογία δηλαδή των θετικών έναντι των αρνητικών περιστατικών-είναι η σχέση της αριστερής στήλης (-) με τη δεξιά στήλη των (+). Κάθε μετρική αναπαράστασης που χρησιμοποιεί τιμές και από τις δύο στήλες θα είναι εγγενώς ευαίσθητη σε ασυμμετρίες της κλάσης. Μετρικές όπως η ακρίβεια, η σαφήνεια και τα Fσκορ χρησιμοποιούν τιμές και από τις δύο στήλες του πίνακα ενδεχομένων. Καθώς η κατανομή της κλάσης αλλάζει, αυτές οι μετρήσεις θα αλλάζουν επίσης, ακόμα και αν η βασική αναπαράσταση του ταξινομητή δεν αλλάζει. Τα ROCγραφήματα είναι βασισμένα πάνω στο TPRκαι FPR στα οποία κάθε διάσταση είναι μία αυστηρή αναλογία στηλών επομένως δεν εξαρτώνται από τις κατανομές της κλάσης.

Για πολλούς ερευνητές, οι μεγάλες ασυμμετρίες κλάσης και οι μεγάλες αλλαγές στις κατανομές της κλάσης ίσως φαίνονται αφύσικες και μη ρεαλιστικές. Παρόλα αυτά ασυμμετρίες κλάσης, της τάξεως του10^1 και 10^2 είναι πολύ συνηθισμένες σε τομείς του πραγματικού κόσμου και ασυμμετρίες άνω της τάξεως του 10^6 έχουν παρατηρηθεί σε κάποιους τομείς (Clearwater & Stern, 1991; Fawcett & Provost, 1996; Kubat, Holte, & Matwin, 1998; Saitta & Neri, 1998). Μεγάλες αλλαγές στις κατανομές μιας κλάσης είναι επίσης μη ρεαλιστικές. Για παράδειγμα στη λήψη ιατρικών αποφάσεων κάποιες επιδημίες μπορεί να επιφέρουν αύξηση στα περιστατικά μιας ασθένειας με το χρόνο. Στην εσφαλμένη ανίχνευση, οι αναλογίες των εσφαλμένων διαφοροποιούνται σημαντικά από μήνα σε μήνα και από μέρος σε μέρος (Fawcett & Provost, 1997). Αλλαγές στη βιομηχανική πρακτική μπορούν να επιφέρουν αύξηση ή μείωση στην αναλογία των ελαττωματικών μονάδων που παράγονται από μία βιομηχανική γραμμή. Σε κάθε ένα από αυτά τα παραδείγματα, η επικράτηση μιας κλάσης μπορεί να αλλάξει δραστικά χωρίς να μεταβάλει το βασικό χαρακτηριστικό της κλάσης, όπως για παράδειγμα τον εννοιολογικό της στόχο.

Η ακρίβεια και η ευαισθησία είναι συνηθισμένες στην επανάκτηση πληροφορίας για την αξιολόγηση της αναπαράστασης ταξινόμησης(Lewis, 1990, 1991). Τα γραφήματα ακρίβειας-ευαισθησίας χρησιμοποιούνται ευρέως όταν κάποιες φορές μπορούμε να υποθέσουμε στατικά σύνολα δεδομένων. Παρόλα αυτά, αυτά χρησιμοποιούνται επίσης σε δυναμικά περιβάλλοντα όπως σε ιστοσελίδες επανάκτησης, όπου ο αριθμός των σελίδων που είναι άσχετες ως προς ένα ερώτημα(N) είναι πολλές τάξεις μεγέθους μεγαλύτερος από τον αριθμό των Pκαι πιθανώς αυξάνει σταθερά με το πέρασμα του χρόνου καθώς νέες ιστοσελίδες δημιουργούνται. Προκειμένου να παρατηρήσουμε το αποτέλεσμα της ασυμμετρίας μιας κλάσης, ας θεωρήσουμε τις καμπύλες στην Εικόνα 5, οι οποίες δείχνουν δύο ταξινομητές εκτιμημένους με τη χρήση ROC καμπυλών και καμπυλών ακρίβειας-ευαισθησίας. Στα γραφήματα 5ακαι b, το σύνολο δεδομένων του τεστ έχει μία ισορροπημένη 1:1 κατανομή κλάσης. Τα γραφήματα 5cκαι dδείχνουν τους ίδιους δύο ταξινομητές στον ίδιο τομέα αλλά ο αριθμός των αρνητικών περιστατικών έχει

αυξηθεί δεκαπλάσια. Αξίζει να σημειωθεί ότι οι ταξινομητές και το βαθύτερο σχέδιο δεν έχει αλλάξει, μόνο η κατανομή κλάσης είναι διαφορετική. Μπορούμε να παρατηρήσουμε ότι τα ROC γραφήματα στις εικόνες 5aκαι 5c είναι πανομοιότυπα ενώ αυτά στις εικόνες 5bκαι5d διαφέρουν δραματικά. Σε κάποιες περιπτώσεις το συμπέρασμα κατά το οποίο ένας ταξινομητής έχει ανώτερη αναπαράσταση μπορεί να αλλάξει από μία μετακίνηση κατανομής.



(a) ROC curves, 1:1 (b) Precision-recall curves, 1:1

(c) ROC curves, 1:10 (d) Precision-recall curves, 1:10

Εικόνα 2.5. Καμπύλες ROC και ακρίβειας-ευαισθησίας κάτω από την ασυμμετρία κλάσης

2.4.3 ΔΗΜΙΟΥΡΓΩΝΤΑΣ SCORING CLASSIFIERS

Πολλά μοντέλα ταξινομητών είναι διακριτά: είναι σχεδιασμένα να παράγουν μία μόνο τιμή κλάσης από κάθε περιστατικό του τεστ. Παρόλα αυτά συμβαίνει συχνά να θέλουμε να δημιουργήσουμε μία γεμάτη ROCκαμπύλη από έναν ταξινομητή και όχι μόνο απλά ένα σημείο. Με αυτόν τον σκοπό εμείς επιθυμούμε να δημιουργήσουμε

τιμές-αποτελέσματα από ένα ταξινομητή και όχι απλά και μόνο μία τιμή κλάσης. Υπάρχουν διάφοροι τρόποι για να παράγουμε τέτοιες τιμές.

Πολλά διακριτά μοντέλα ταξινομητών μπορούν εύκολα να μετατραπούν σε ταξινομητές βαθμολόγησης απλά με το να "κοιτάξουμε μέσα σε αυτούς" στα στατιστικά περιστατικών που αυτοί διατηρούν. Για παράδειγμα, ένα δέντρο απόφασης καθορίζει μία τιμή κλάσης από ένα κόμβο φύλλων από την αναλογία των περιστατικών στον κόμβο. Η κλάση απόφασης είναι απλά η πιο επικρατούσα κλάση. Αυτές οι αναλογίες κλάσης εξυπηρετούν και ως ένα αποτέλεσμα (Provost&Domingos, 2001). Ένας μαθητευόμενος κανόνων διατηρεί παρόμοια στατιστικά στην εχεμύθεια κανόνων και η εκμυστήρευση ενός κανόνα που ταιριάζει με ένα περιστατικό μπορεί ν χρησιμοποιηθεί ως αποτέλεσμα(Fawcett, 2001).

Ακόμα και αν ένας ταξινομητής παράγει απλά μία τιμή κλάσης, μία συσσώρευση αυτών μπορεί να χρησιμοποιηθεί για να παραχθεί ένα αποτέλεσμα. Η MetaCost χρησιμοποιεί το πακετάρισμα προκειμένου να δημιουργήσει ένα σύνολο από διακριτούς ταξινομητές, καθένας από τους οποίους να παράγει ένα ψήφο. Το σύνολο των ψήφων μπορεί να χρησιμοποιηθεί για να παραχθεί ένα αποτέλεσμα. Η MetaCost στην πραγματικότητα λειτουργεί προς την αντίθετη κατεύθυνση διότι ο στόχος της είναι να δημιουργήσει ένα διακριτό ταξινομητή. Αυτή αρχικά παράγει ένα πιθανολογικό ταξινομητή, μετά εφαρμόζει τη γνώση του κόστους των λαθών και των ασυμμετριών κλάσης ούτως ώστε να δώσει νέα τιμή στα περιστατικά και να βελτιστοποιήσει τις ταξινομήσεις τους. Τελικά αυτή μαθαίνει ένα συγκεκριμένο διακριτό ταξινομητή από αυτό το νέο σύνολο περιστατικών. Επομένως η MetaCost δεν είναι μία καλή μέθοδος για τη δημιουργία ενός καλού ταξινομητή βαθμολόγησης, αν και η μέθοδος πακεταρίσματος που διαθέτει ίσως να μπορεί να θεωρηθεί καλή.

Συμπερασματικά, κάποιοι συνδυασμοί αποτελεσμάτων και ψήφων μπορούν να χρησιμοποιηθούν. Για παράδειγμα κάποιοι κανόνες μπορούν να παρέχουν βασικές εκτιμήσεις πιθανότητας, οι οποίες μπορούν τότε να χρησιμοποιηθούν σε σταθμισμένη ψηφοφορία(Fawcett, 2001).

2.5 ΑΠΟΤΕΛΕΣΜΑΤΙΚΗ ΔΗΜΙΟΥΡΓΙΑ ROC ΚΑΜΠΥΛΩΝ

Δοσμένου ενός συνόλου δεδομένων από ένα τεστ, εμείς συχνά επιθυμούμε να δημιουργήσουμε αποτελεσματικά μία ROC καμπύλη από αυτό. Αν και κάποιοι ερευνητές έχουν εφαρμόσει μεθόδους όπως ο αλγόριθμος 1, αυτή η μέθοδος δεν είναι ούτε αποτελεσματική ούτε πρακτική: αυτή απαιτεί τη γνώση του ελάχιστου(min), του μέγιστου(max) και της προσαύξησης τα οποία πρέπει να εκτιμηθούν από το σύνολο δεδομένων του τεστ και τις ftιμές. Αυτή επίσης εμπλέκει και δύο ένθετους βρόγχους καθώς ο εξωτερικός βρόγχος πρέπει να προσαυξήσει το ttουλάχιστον nφορές και η περιπλοκότητα είναι της τάξεως $O(n^2)$ επί του αριθμού των περιστατικών του συνόλου του τεστ.

Ένας πολύ καλύτερος αλγόριθμος μπορεί να δημιουργηθεί αν εκμεταλλευτούμε την μονοτονία των ορίων ταξινόμησης. Κάθε περιστατικό που ταξινομείται ως θετικό με βάση το δοσμένο όριο θα εξακολουθεί να ταξινομείται ως θετικό για κάθε όριο μικρότερου του αρχικού. Έτσι μπορούμε απλά να ταξινομήσουμε τα περιστατικά των τεστ από τα *f*αποτελέσματα, από το υψηλότερο προς το χαμηλότερο και να

κατεβαίνουμε την λίστα, έχοντας ένα περιστατικό την κάθε φορά και ανανεώνοντας το TPR και FPR καθώς προχωρούμε. Κατά αυτόν τον τρόπο μπορούμε να δημιουργήσουμε ένα ROCγράφημα από μία γραμμική σάρωση.

Ο νέος αλγόριθμος φαίνεται στον Αλγόριθμο 2 του Παραρτήματος Α. Το TPRκαι το FPR ξεκινούν και τα δύο από το μηδέν. Με κάθε θετικό περιστατικό προσαυξάνουμε το FPRκαι με κάθε αρνητικό το FPR. Επιπλέον διατηρούμε ένα σωρό RROCσημείων προσθέτοντας ένα επιπλέον σημείο στο Rμετά από κάθε περιστατικό που επεξεργαζόμαστε. Το τελικό αποτέλεσμα είναι ο σωρός των Ro οποίος θα περιέχει σημεία της ROCκαμπύλης. Το Rξεκινά έχοντας ως πρώτο σημείο το (0,0) και όταν και το τελευταίο σημείο το έχουμε επεξεργαστεί το σύνολο Rκλείνει με το σημείο (1,1)

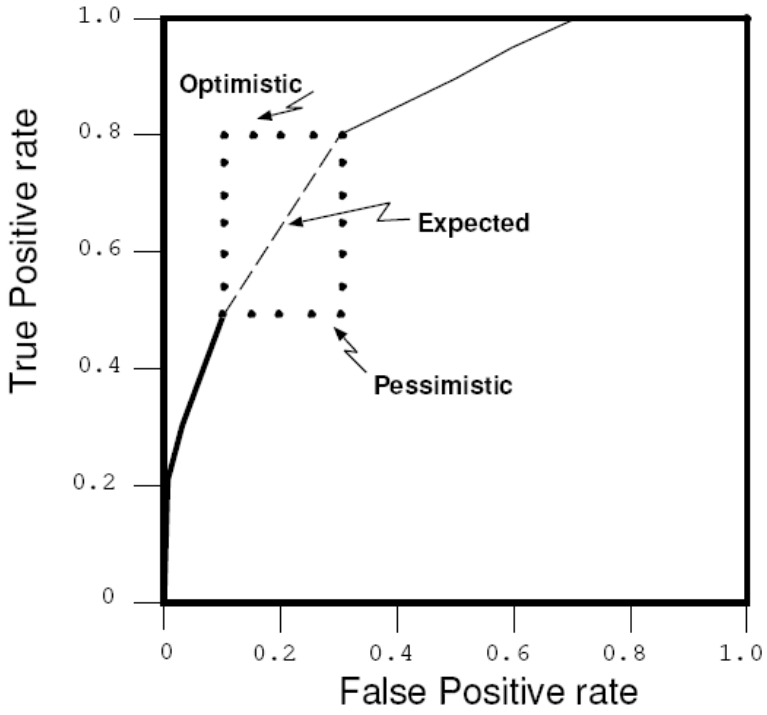
Έστω ητο πλήθος των σημείων που έχουν προκύψει από το σύνολο των περιστατικών του τεστ, αυτός ο αλγόριθμος απαιτεί μία $O(n\log n)$ ταξινόμηση ακολουθούμενη από μία $O(n)$σάρωση της λίστας και δίνει σαν αποτέλεσμα μία $O(n\log n)$ συνολική περιπλοκότητα.

2.5.1ΙΣΟΔΥΝΑΜΑ ΒΑΘΜΟΛΟΓΗΜΕΝΑ ΠΕΡΙΣΤΑΤΙΚΑ

Οι εντολές στις γραμμές 7 έως 9 του αλγορίθμου 2 χρειάζονται κάποια περαιτέρω ανάλυση. Αυτά τα βήματα είναι απαραίτητα προκειμένου να χειριστούμε σωστά σειρές από ισοδύναμα βαθμονομημένα περιστατικά. Ας θεωρήσουμε τη ROC καμπύλη που φαίνεται στην Εικόνα 6. Επιπλέον ας υποθέσουμε ότι έχουμε ένα σύνολο από δεδομένα στο οποίο υπάρχει μία σειρά από δεδομένα, τέσσερα αρνητικά και έξι θετικά, όλα βαθμολογημένα ισοδύναμα από την *f*. Η ταξινόμηση στη γραμμή 1 του αλγορίθμου 2 δεν επιβάλει καμία συγκεκριμένη διάταξη σε αυτά τα περιστατικά εφόσον τα*f*αποτελέσματα αυτών είναι ίσα. Το θέμα είναι τι συμβαίνει όταν δημιουργούμε μία ROCκαμπύλη. Στη μία ακραία περίπτωση, όλα τα θετικά καταλήγουν στην κορυφή της αλληλουχίας σχηματίζοντας έτσι το " αισιόδοξο" άνω τμήμα Lπου φαίνεται στην Εικόνα 2.6. Στην αντίθετη ακραία περίπτωση, όλα τα αρνητικά καταλήγουν στην κορυφή της αλληλουχίας δίνοντάς μας έτσι το "απαισιόδοξο" χαμηλότερο σημείο L,όπως φαίνεται στην Εικόνα 6. Κάθε αναμειγμένη διάταξη των περιστατικών θα δώσει ένα διαφορετικό σύνολο βημάτων-σημείων μέσα στο ορθογώνιο που σχηματίζεται από τις δύο παραπάνω ακραίες περιπτώσεις. Παρόλα αυτά εμείς θέλουμε η ROC καμπύλη να αντιπροσωπεύει την αναμενόμενη αναπαράσταση του ταξινομητή, η οποία, όταν δεν υπάρχει καμία άλλη πληροφορία είναι η μέση κατάσταση των απαισιόδοξων και αισιόδοξων σημείων. Αυτή η μέση κατάσταση των σημείων αντιπροσωπεύεται από τη διαγώνιο του ορθογωνίου και μπορεί να δημιουργηθεί στον αλγόριθμο της ROCκαμπύλης όταν δεν παράγεται ROCσημείο μέχρις ότου περάσουν όλα τα περιστατικά ίσων τιμών f. Αυτό ακριβώς υποδηλώνουν και η μεταβλητή f_{prev}με την εντολή **if** στη γραμμή 7 του αλγόριθμου.

Τα περιστατικά που έχουν βαθμολογηθεί ισοδύναμα μπορεί να θεωρούνται σπάνια αλλά με κάποια μοντέλα ταξινομητών είναι συνηθισμένα. Για παράδειγμα, αν χρησιμοποιούμε περιστατικά ως μετρητές, στους κόμβους των δέντρων απόφασης προκειμένου να βαθμολογήσουμε περιστατικά, ένα φύλλο κόμβος υψηλής εντροπίας μπορεί να παράγει πολλά ισόβαθμα περιστατικά. Αν από τέτοια περιστατικά δεν

παίρνουμε το μέσο όρο, οι ROCκαμπύλες που προκύπτουν θα είναι ευαίσθητες ως προς τη διάταξη του συνόλου δεδομένων του τεστ με αποτέλεσμα διαφορετικές διατάξεις να δίνουν παραπλανητικές καμπύλες. Το γεγονός αυτό μπορεί να κατακριθεί ιδιαίτερα στον υπολογισμό της περιοχής κάτω από μία ROCκαμπύλη(παράγραφος 2.6). Ας θεωρήσουμε ότι έχουμε ένα δέντρο απόφασης το οποίο περιέχει ένα κόμβο φύλλων από nθετικά και mαρνητικά περιστατικά. Σε κάθε περιστατικό που ταξινομείται σε αυτόν τον κόμβο θα δίνεται ο ίδιος βαθμός. Το ορθογώνιο της Εικόνας 6 θα έχει μέγεθος$\frac{nm}{PN}$ και αν αυτά τα περιστατικά δεν είναι υπολογισμένα κατά μέσο όρο, αυτό το ένα μόνο φύλλο μπορεί να είναι υπεύθυνο για λάθη στην περιοχή της ROCκαμπύλης μεγέθους έως και $\frac{mn}{2PN}$.



Εικόνα 2.6. Τα αισιόδοξα, απαισιόδοξα και αναμενόμενα ROC τμήματα προερχόμενα από μία σειρά από δέκα ισοδύναμα βαθμολογημένα περιστατικά.

2.5.2 ΔΗΜΙΟΥΡΓΩΝΤΑΣ ΚΥΡΤΕΣ ROCΚΑΜΠΥΛΕΣ

Για τον έλεγχο κατάλληλων στόχων, πρέπει να χρησιμοποιείται ο ROCαλγόριθμος 2 καθώς αυτός υπολογίζει μία μη τροποποιημένη ROCκαμπύλη. Οι καμπύλες που αυτός παράγει απεικονίζουν μία αναμενόμενη αναπαράσταση ταξινομητή στο σύνολο δεδομένων του τεστ.

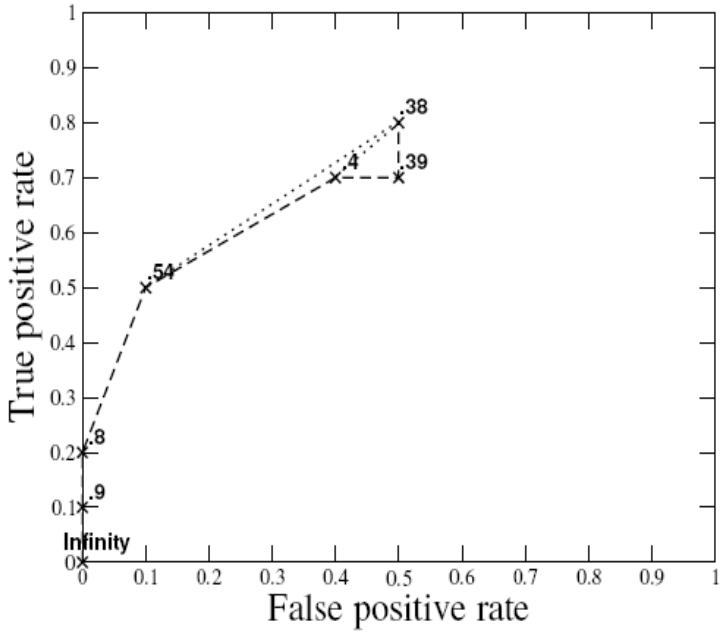
Παρόλα αυτά σε κάποιες περιπτώσεις, πιθανόν να επιθυμείται να παραχθεί μία κυρτή ROCκαμπύλη, μία δηλαδή με όλες τις κοιλότητες απομακρυσμένες. Μία κοιλότητα σε μία ROCκαμπύλη αντιπροσωπεύει μία ατέλεια στον ταξινομητή. Συγκεκριμένα μία κοιλότητα συμβαίνει όποτε ένα γραμμικό τμήμα κλίσης rσυνδέεται αριστερά ενός γραμμικού τμήματος κλίσης sκαι παράλληλα ισχύει s>r. Η κλίση μιας ROCκαμπύλης

αντιπροσωπεύει την αναλογία πιθανότητας της κλάσης. Μία κοιλότητα υποδηλώνει ότι το σύνολο των περιστατικών που παράγει την κλίση sέχει υψηλότερη μεταγενέστερη αναλογία κλάσης από ότι το σύνολο των περιστατικών που παράγει την κλίση r. Επιπλέον επειδή το sβρίσκεται δεξιά τουr, τα περιστατικά του rθα έπρεπε να έχουν βαθμολογηθεί πιο υψηλά από ότι αυτά του s, γεγονός που δεν ισχύει. Αυτό αποτελεί ατέλεια του ταξινομητή. Πρακτικά οι κοιλότητες στις ROCκαμπύλες παράχθηκαν από γνωστούς ταξινομητές, είτε λόγω ιδιοσυγκρασιών στην εκμάθηση είτε λόγω αποτελεσμάτων από μικρά σύνολα δεδομένων.

Ο αλγόριθμος 2 μπορεί να τροποποιηθεί ώστε να απομακρύνει κοιλότητες, απλά αντικαθιστώντας τη λειτουργία του ADD_POINT. Ο νέος ορισμός λειτουργίας του ADD_POINT φαίνεται στον αλγόριθμο 3. Στη νέα αυτή λειτουργία, η κλίση των τμημάτων του Rσυνόλου, εξετάζονται πριν προστεθεί ένα νέο σημείο. Μία κυρτή ROCκαμπύλη πρέπει να έχει μονότονα φθίνουσες κλήσεις. Αν λοιπόν προσθέτοντας ένα νέο σημείο, προκύπτει ένα γραμμικό τμήμα μεγαλύτερης κλίσης από αυτήν που προηγούνταν, τότε το ROCσημείο που αντιστοιχεί στο γραμμικό αυτό τμήμα που προηγούνταν διώχνεται. Το τεστ τότε επαναλαμβάνεται, εντοπίζοντας και διώχνοντας σημεία όταν είναι απαραίτητο από το Rσύνολο μέχρις ότου ένα τμήμα μεγαλύτερης κλίσης προσεγγίζεται (ή μέχρις ότου να έχει απομείνει ένα μόνο σημείο) . Αυτή η διαδικασία εξασφαλίζει μονοτονία κλίσης και έτσι κυρτότητα καμπύλης.
Ο Αλγόριθμος 3 μπορεί να βελτιωθεί αποθηκεύοντας στη μνήμη του, τις κλίσεις των γραμμικών τμημάτων κατά μήκος όλων των σημείων του Rσυνόλου. Με αυτόν τον τρόπο, αποφεύγονται οι μη απαραίτητες επαναλήψεις της συνάρτησης SLOPE. Για μεγαλύτερη διαύγεια η διαδικασία βελτιστοποίησης δεν φαίνεται.

Ο αλγόριθμος 3, υπολογιστικά δεν είναι περισσότερο περίπλοκος από ότι ο αλγόριθμος 2. Ένα κομμάτι μόνο προστίθεται σε ένα σημείο, στη γραμμή 11(η εντολή "push" στη γραμμή 10 εξυπηρετεί μόνο στο να αναιρεί την εντολή"pop" στη γραμμή 7).Ένα σημείο μπορεί να επισημανθεί το πολύ μία φορά. Η υπολογιστική περιπλοκότητα παραμένει $O(n\log n)$.

Η Εικόνα 2.7 απεικονίζει τα σημεία της Εικόνας 2.3 αλλά επεξεργασμένα στον αλγόριθμο 3 και όχι στον αλγόριθμο 2. Η Εικόνα 2.7a παρουσιάζει μία μεγέθυνση ενός ενδιάμεσου βήματος, στο οποίο το σημείο με βαθμό 0.4 προστίθεται στη καμπύλη. Οι κλίσεις στα προηγούμενα σημεία 0.505, 0.51 και 0.54 αναθεωρούνται και τα σημεία 0.505 και 0.51 απομακρύνονται.



Εικόνα 2.7. Τα σημεία της εικόνας 3 επεξεργασμένα από τον αλγόριθμο 3

2.6 ΕΜΒΑΔΟΝ ΤΗΣ ΠΕΡΙΟΧΗΣ ΚΑΤΩ ΑΠΟ ΜΙΑ ROC ΚΑΜΠΥΛΗ (AUC)

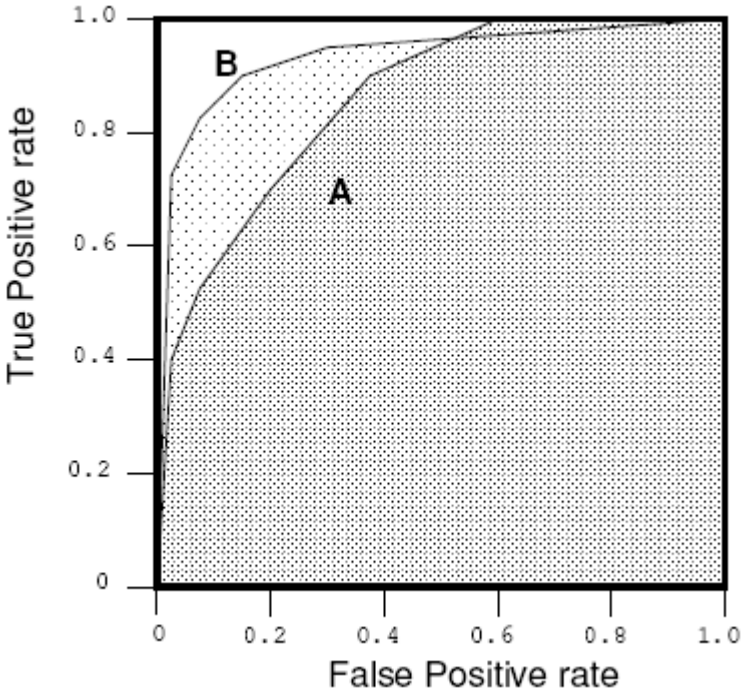
Η ROCκαμπύλη είναι μία δυσδιάστατη απεικόνιση της αναπαράστασης ενός ταξινομητή. Προκειμένου να συγκρίνουμε ταξινομητές, ίσως επιθυμούμε να περιορίσουμε τη ROCαναπαράσταση σε μία απλά βαθμιδωτή τιμή, η οποία θα αντιπροσωπεύει την αναμενόμενη αναπαράσταση. Μία συνηθισμένη μέθοδος είναι να υπολογίσουμε το εμβαδόν κάτω από τη ROCκαμπύλη, το οποίο συντομογραφικά είναι γνωστό ως AUC (AREAUNDERCURVE, Bradley, 1997; Hanley&McNeil, 1982). Επιπλέον, καθώς το AUCαποτελεί αναλογία του εμβαδού προς την τεραγονική μονάδα, η τιμή του θα κυμαίνεται πάντα μεταξύ του 0 και του 1.0. Παρόλα αυτά, καθώς με την τυχαία αναπαράσταση που είχαμε αναφέρει στη παράγραφο 2.3.1, σχηματίζεται η διαγώνιος γραμμή μεταξύ των σημείων (0,0) και (1,1), η οποία έχει εμβαδόν 0.5 κανένας ρεαλιστικός ταξινομητής δε μπορεί να έχει εμβαδό μικρότερο από αυτό.

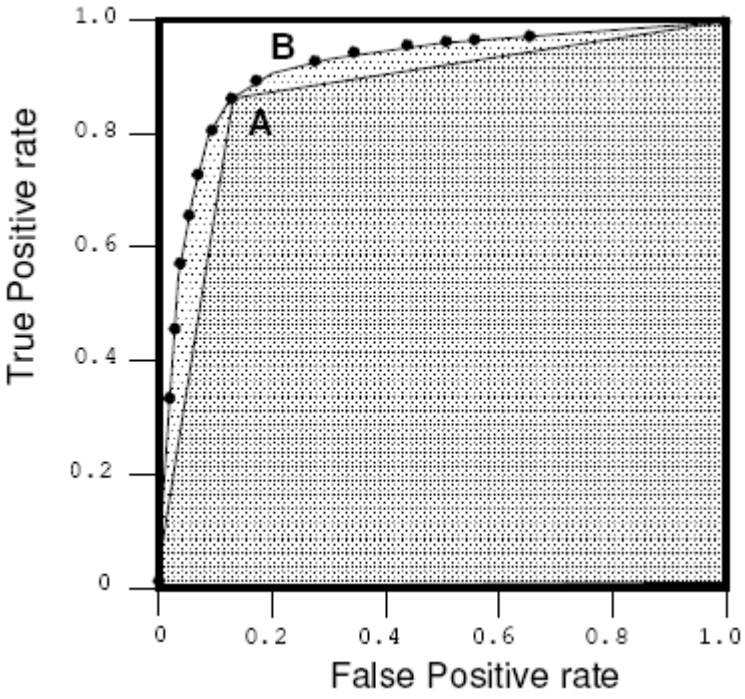
Το εμβαδόν AUCέχει μία πολύ σημαντική στατιστική ιδιότητα: το AUC ενός ταξινομητή είναι ισοδύναμο με την πιθανότητα ένας ταξινομητής να βαθμολογήσει ένα τυχαία επιλεγμένο θετικό περιστατικό υψηλότερα από ένα τυχαία επιλεγμένο αρνητικό περιστατικό. Αυτό είναι ισοδύναμο με το τεστ βαθμών του Wilcoxon (Hanley&McNeil, 1982). Επίσης το AUCείναι πολύ στενά συσχετισμένο με τον δείκτη Gini(Breiman, Friedman, Olshen, &Stone, 1984), ο οποίος ισοδυναμεί με δύο φορές το εμβαδόν ανάμεσα στη διαγώνιο και τη ROCκαμπύλη. Ο Handκαι oTill(2001) επεσήμαναν ότι **Gini + 1 = 2 ×AUC.**

Η εικόνα 2.8a δείχνει τα εμβαδά κάτω από δύο ROCκαμπύλες, τις A καιB. Παρατηρούμε ότι ο ταξινομητής B έχει μεγαλύτερο εμβαδόν και παρόλα αυτά καλύτερη μέση αναπαράσταση. Η εικόνα 2.8bδείχνει το εμβαδόν κάτω από τη

καμπύλη ενός δυαδικού ταξινομητή Ακαι ενός ταξινομητή βαθμολόγησης Β. Ο ταξινομητής Α αντιπροσωπεύει την αναπαράσταση του Β, όταν ο Β χρησιμοποιείται με ένα μονό, καθορισμένο όριο. Αν και η αναπαράσταση και των δύο είναι ισοδύναμη στο συγκεκριμένο καθορισμένο σημείο (όριο του Β ταξινομητή) η αναπαράσταση του Β γίνεται κατώτερη του Α πέραν αυτού του σημείου.

Είναι πιθανόν, για έναν, υψηλού εμβαδού AUC,ταξινομητή να συμπεριφέρεται χειρότερα σε μία συγκεκριμένη περιοχή του ROCχώρου από ότι ένας χαμηλού εμβαδούAUCταξινομητής. Η εικόνα 2.2a παρουσιάζει ένα παράδειγμα αυτής της περίπτωσης: ο ταξινομητής Βείναι γενικά καλύτερος από τον ταξινομητή Α εκτός από τον παράγοντα FPRγια τον οποίο ισχύει FPR> 0.6, στο σημείο αυτό ο Α ταξινομητής έχει ένα μικρό πλεονέκτημα. Παρόλα αυτά, στην πράξη το κριτήριο του AUCεμβαδού λειτουργεί πολύ καλά και χρησιμοποιείται συχνά όταν απαιτείται ένα γενικό μέτρο της ικανότητας πρόβλεψης. Το εμβαδόν AUCμπορεί να υπολογιστεί εύκολα χρησιμοποιώντας μία μικρή τροποποίηση του αλγορίθμου 2, η οποία φαίνεται στον αλγόριθμο 4. Αντί να συλλέγει ROCσημεία ο αλγόριθμος προσθέτει διαδοχικά εμβαδά με τον κανόνα του τραπεζίου. Στη συνέχεια διαιρεί το συνολικό εμβαδόν, για να υπολογίσει την τιμή στην τετραγωνική μονάδα.





Εικόνα 2.8. Δύο ROC γραφήματα. Το πάνω γράφημα δείχνει το εμβαδόν κάτω από τις δύο ROC καμπύλες. Το κάτω γράφημα δείχνει το εμβαδόν κάτω από τις καμπύλες ενός διακριτού ταξινομητή (A) και ενός πιθανολογικού ταξινομητή (B).

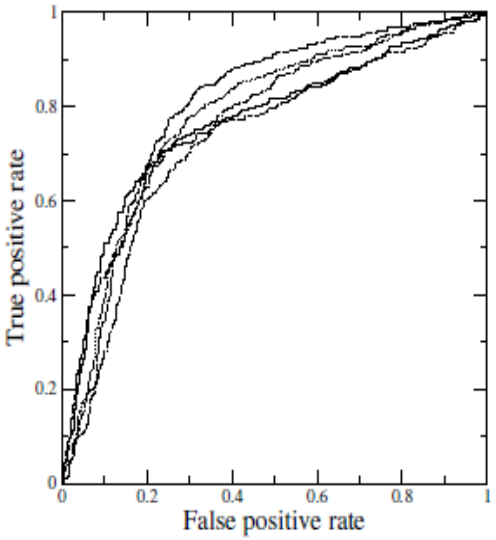
2.7 ΥΠΟΛΟΓΙΣΜΟΣ ΜΕΣΗΣ ΚΑΜΠΥΛΗΣ ROC

Αν και οι ROCκαμπύλες μπορούν να χρησιμοποιηθούν για την αξιολόγηση ταξινομητών πρέπει να υπάρξει προσοχή όταν αυτές χρησιμοποιούνται για την εξαγωγή συμπερασμάτων όσων αφορά την υπεροχή ταξινομητών. Κάποιοι ερευνητές έχουν υποθέσει ότι ένα ROCγράφημα μπορεί να χρησιμοποιηθεί προκειμένου να επιλεγούν οι καλύτεροι ταξινομητές απλά αναπαριστώντας τους σε ένα ROCχώρο και παρατηρώντας ποιος επικρατεί. Βέβαια, αυτό είναι παραπλανητικό. Είναι ανάλογο με το να πάρουμε το μέγιστο από ένα σύνολο γραφικών παραστάσεων ακρίβειας ενός συνόλου δεδομένων ενός μεμονωμένου τεστ. Χωρίς ένα μέτρο διασποράς δεν μπορούμε εύκολα να συγκρίνουμε τους ταξινομητές.

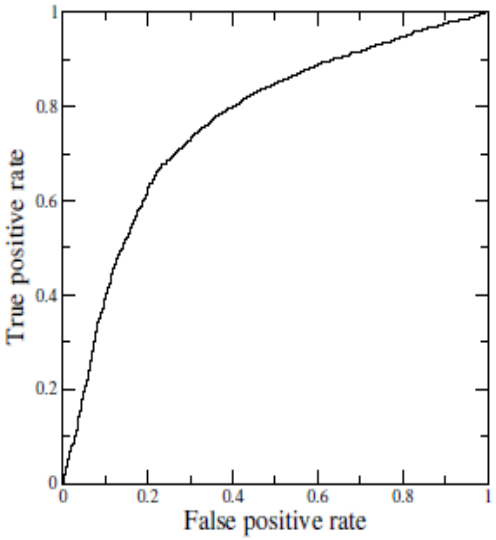
Ο υπολογισμός της μέσης ROCκαμπύλης αποτελεί εύκολη διαδικασία αν τα αυθεντικά περιστατικά είναι διαθέσιμα. Δοσμένου ενός συνόλου δεδομένων T_1, T_2,…,T_nτα οποία έχουν προέλθει από μία διασταυρωμένη επικύρωση ή μία μέθοδο bootstrap, εμείς μπορούμε απλά να ταξινομήσουμε με συγχώνευση τα περιστατικά με βάση την τιμή που τους αντιστοιχεί, σε ένα μεγάλο σύνολο T_M. Έπειτα, τρέχουμε έναν αλγόριθμο παραγωγής ROCκαμπύλης, όπως ο αλγόριθμος 2 και έχουμε τη γραφική παράσταση του αποτελέσματος. Αυτή η διαδικασία δίνει τη μέση αναμενόμενη ROCαναπαράσταση. Παρόλα αυτά ο πρωτεύον λόγος για τον οποίο χρησιμοποιούμε πολλαπλά σύνολα από τεστ είναι για να αντλήσουμε από αυτό ένα μέτρο διασποράς, την οποία αυτή η απλή μέθοδος συγχώνευσης δεν παρέχει. Για το

λόγο αυτό εμείς χρειαζόμαστε μία πιο περίπλοκη μέθοδο η οποία δοκιμάζει τις επιμέρους καμπύλες σε διαφορετικά σημεία.. Ο ROCχώρος είναι δύο διαστάσεων και κάθε μέσος όρος ενός μεγέθους είναι απαραιτήτως μίας διάστασης. Οι ROCκαμπύλες μπορούν να προβληθούν σε μία διάσταση και συμβατικά κατά μέσο όρο, αλλά αυτό οδηγεί στο ερώτημα αν η προβολή αυτή είναι κατάλληλη ή πιο συγκεκριμένα, αν με αυτή διατηρούνται κάποια χαρακτηριστικά της καμπύλης, που παρουσιάζουν ιδιαίτερο ενδιαφέρον. Η απάντηση εξαρτάται από το λόγο για τον οποίο επιθυμούμε να παράγουμε τη μέση καμπύλη. Σε αυτή την παράγραφο παρουσιάζονται δύο μέθοδοι για τον υπολογισμό του μέσου όρου των καμπυλών: η κάθετη μέθοδος και αυτή που βασίζεται στο όριο που επιλέγεται για τις τιμές των περιστατικών.

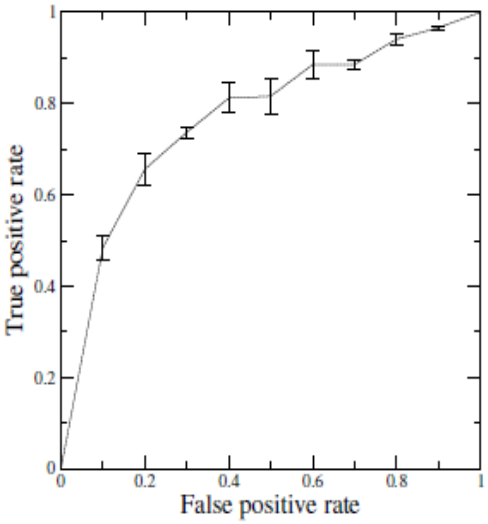
Η εικόνα 2.9απαρουσιάζει πέντε ROCκαμπύλες για τις οποίες επιθυμούμε να υπολογίσουμε τον μέσο όρο. Κάθε μία από αυτές περιέχει χιλιάδες σημεία και έχει κάποιες κοιλότητες. Η εικόνα 2.9b δείχνει την καμπύλη που σχηματίστηκε από τη συγχώνευση των πέντε συνόλων δεδομένων και από τον υπολογισμό της συνδυασμένης ROC καμπύλης τους. Οι εικόνες 2.9.b και 2.9.cδείχνουν μέσες καμπύλες οι οποίες σχηματίστηκαν από τις αρχικές πέντε ξεχωριστές καμπύλες. Οι μπάρες σφάλματος αντιστοιχούν σε 95% διάστημα εμπιστοσύνης.



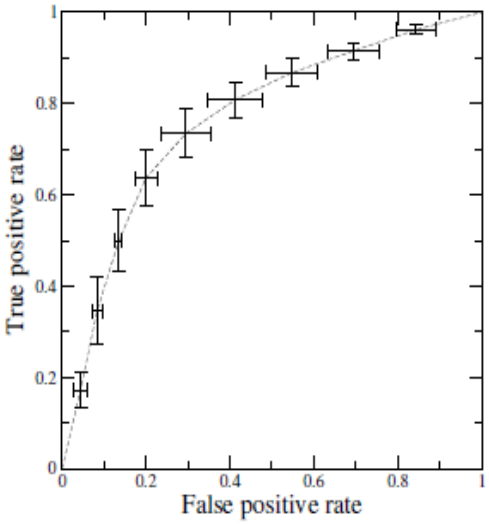
Εικόνα 2.9(α).ROC καμπύλες από πέντε δείγματα περιστατικών δειγμάτων



Εικόνα 2.9.(b)ROC καμπύλη σχηματισμένη από την ανάμειξη πέντε



Εικόνα 2.9.(c) ΜέσηROC καμπύλη υπολογισμένη με την κάθετη μέθοδο



Εικόνα 2.9.(d) Μέση ROC καμπύλη υπολογισμένη με τη μέθοδο του ορίου

2.7.1 ΥΠΟΛΟΓΙΣΜΟΣ ROC ΚΑΜΠΥΛΗΣ ΜΕ ΤΗΝ ΚΑΘΕΤΗ ΜΕΘΟΔΟ

Η κάθετη μέθοδος υπολογισμού ROCκαμπυλών παίρνει κάθετα δείγματα των ROCκαμπυλών για καθορισμένα FPR και υπολογίζει το μέσο όρο των αντίστοιχων TPR.Ένας τέτοιος υπολογισμός του μέσου όρου είναι κατάλληλος όταν το FPRμπορεί να είναι καθορισμένο από τον ερευνητή, ή όταν ζητείται ένα μονοδιάστατο μέτρο διασποράς. Οι Provost, FawcettκαιKohavi (1998) χρησιμοποίησαν αυτή τη μέθοδο στη δουλειά τους προκειμένου να υπολογίσουν τις μέσες ROC καμπύλες ενός ταξινομητή για κ φορές διασταυρωμένες επικυρώσεις. Σε αυτή τη μέθοδο κάθε ROCκαμπύλη χρησιμοποιείται σαν συνάρτηση, R_i,ούτως ώστε$TP = R_i(FP)$.Αυτό επιτυγχάνεται επιλέγοντας το μέγιστο TPRγια το κάθε FPRκαι παρεμβάλλοντας μεταξύ των σημείων όταν αυτό είναι απαραίτητο.

Η μέση ROCκαμπύλη είναι η συνάρτηση $\hat{R}(FP) = mean[R_i(FP)]$. Για να σχεδιάσουμε μία μέση ROCκαμπύλη, μπορούμε να δοκιμάσουμε σημεία από το $\hat{R}$
Τα οποία είναι τακτικά τοποθετημένα κατά μήκος τουFPRάξονα. Διαστήματα εμπιστοσύνης του μέσου των TPR υπολογίζονται χρησιμοποιώντας τη συνηθισμένη υπόθεση της διωνυμικής κατανομής. Ο αλγόριθμος 5 υπολογίζει τον κάθετο μέσο όρο ενός συνόλου ROCσημείων. Σε αυτόν οι μέση όροι αφήνονται στη διάταξη TPavg.

Πολλές επεκτάσεις έχουν παραληφθεί από τον αλγόριθμο για μεγαλύτερη διαύγεια. Ο αλγόριθμος μπορεί εύκολα να επεκταθεί για να υπολογιστούν οι τυπικές αποκλίσεις των δειγμάτων προκειμένου να σχεδιάσουμε τις μπάρες των διαστημάτων εμπιστοσύνης. Επίσης η συνάρτηση TP_FOR_FPμπορεί να βελτιωθεί κάπως. Επειδή αυτή καλείται μόνο σε μονότονα αύξουσες τιμές των FPR δεν χρειάζεται να εξετάζει κάθε ROCδιάταξη από την αρχή κάθε φορά. Θα μπορούσε να κρατήσει στη μνήμη το τελευταίο σημείο που εμφανίστηκε και να δώσει αρχική τιμή στο Iαπό αυτή τη διάταξη.

Η εικόνα 2.9.cδείχνει τον κάθετο υπολογισμό του μέσου όρου των πέντε καμπυλών της εικόνας 2.9.a. Οι κάθετες μπάρες στην καμπύλη παριστάνουν τα 95% διαστήματα

εμπιστοσύνης του ROCμέσου. Για αυτή τη μέση καμπύλη, οι αρχικές καμπύλες πήραν τιμές FPRαπό το 0 μέχρι το 1, με βήμα 0.1. Είναι πιθανό να πάρουμε καμπύλες πολύ πιο εύκολα αλλά οι μπάρες εμπιστοσύνης θα είναι πολύ πιο δυσδιάκριτες.

2.7.2 ΥΠΟΛΟΓΙΣΜΟΣ ΜΕΣΟΥ ΟΡΟΥ ΚΑΜΠΥΛΩΝ ΜΕ ΤΗ ΜΕΘΟΔΟ ΤΟΥ ΟΡΙΟΥ

Η κάθετη μέθοδος υπολογισμού της μέσης καμπύλης έχει ένα πλεονέκτημα: οι μέσες τιμές προκύπτουν από μία μόνο εξαρτημένη μεταβλητή το FPR. Αυτό απλοποιεί αρκετά τον υπολογισμό των διαστημάτων εμπιστοσύνης. Παρόλα αυτά oHolte (2002)έχει επισημάνει ότι το FPR, συχνά δε βρίσκεται υπό τον άμεσο έλεγχο του ερευνητή. Επομένως, ίσως να ήταν προτιμότερο, να υπολογίζουμε το μέσο όρο των ROCσημείων χρησιμοποιώντας μία ανεξάρτητη μεταβλητή της οποίας η τιμή να μπορεί να ελεγχθεί άμεσα όπως το όριο(σημείο απόφασης) των τιμών ενός ταξινομητή.

Η μέθοδος του ορίου επιτυγχάνει το παραπάνω. Αν και η μέθοδος αυτή δοκιμάζει σημεία με βάση τις θέσεις τους στο ROCχώρο, όπως ακριβώς πραγματοποιεί και η κάθετη μέθοδος, αυτή επιπλέον δοκιμάζει σημεία με βάση το όριο από το οποίο προήλθαν τα σημεία αυτά. Η μέθοδος αυτή πρέπει να παράγει ένα σύνολο ορίων και έπειτα για το καθένα από αυτά τα όρια να βρει το αντίστοιχο σημείο από κάθε ROCκαμπύλη και τέλος να υπολογίσει το μέσο όρο των σημείων αυτών.

Ο αλγόριθμος 6 δείχνει τη βασική μέθοδο για να πραγματοποιηθεί αυτό που μόλις περιγράψαμε. Κατά τον αλγόριθμο αυτό, δημιουργείται μία διάταξη T από τιμές ταξινομητών οι οποίες ταξινομούνται από τη μεγαλύτερη προς τη μικρότερη και χρησιμοποιείται ως σύνολο ορίων. Αυτά τα όρια δοκιμάζονται ανά καθορισμένα διαστήματα τα οποία αποφασίζονται με βάση τον αριθμό των δειγμάτων που επιθυμείται. Για ένα δοσμένο όριο, ο αλγόριθμος επιλέγει από κάθε ROC καμπύλη το σημείο με τη μεγαλύτερη τιμή η οποία όμως να είναι μικρότερη ή ίση ως προς την τιμή του ορίου(υποθέτουμε ότι τα ROCσημεία έχουν προέλθει από ένα αλγόριθμο όπως ο 2, ο οποίος χειρίζεται σωστά την περίπτωση των ισοδύναμα βαθμολογημένων περιστατικών). Από αυτά τα σημεία στη συνέχεια υπολογίζεται ο μέσος όρος ξεχωριστά κατά μήκος των X καιY αξόνων με το κεντρικό σημείο να επιστρέφει στη Avg διάταξη. Η εικόνα 2.9.dδείχνει το αποτέλεσμα του υπολογισμού του μέσου όρου των πέντε καμπυλών της εικόνας 2.9.a, με τη μέθοδο του ορίου. Η καμπύλη που προκύπτει αποτελείται από τα μέσα σημεία και μπάρες διαστημάτων εμπιστοσύνης στις Xκαι Y διευθύνσεις. Οι μπάρες που φαίνονται αντιστοιχούν σε 95% διάστημα εμπιστοσύνης.

Υπάρχουν κάποιοι μικροί περιορισμοί στη μέθοδο αυτή του ορίου σε σχέση με την αντίστοιχη κάθετη μέθοδο. Για να χρησιμοποιήσουμε τη μέθοδο αυτή χρειάζεται να γνωρίζουμε την τιμή του ταξινομητή που αντιστοιχεί στο κάθε σημείο. Επιπλέον στη παράγραφο 2.4.1 επισημάνθηκε ότι οι τιμές του ταξινομητή δεν πρέπει να συγκρίνονται κατά μήκος των κλάσεων του μοντέλου. Ως άμεση συνέπεια έχουμε ότι οι μέσες ROCκαμπύλες υπολογισμένες από διαφορετικές κλάσεις μοντέλων μπορεί να είναι παραπλανητικές καθώς οι τιμές είναι πιθανό να είναι δυσανάλογες.

2.8 ΕΠΙΠΡΟΣΘΕΤΑ ΘΕΜΑΤΑ

Οι προηγούμενες παράγραφοι στόχευαν στο να είναι αυτόνομες και στο να καλύψουν βασικά θέματα που ανακύπτουν από τη χρήση ROCκαμπυλών στην έρευνα εξόρυξης δεδομένων. Αυτή η παράγραφος πραγματεύεται επιπρόσθετα ελαφρώς πιο εσωτερικά θέματα.

2.8.1 ROC CONVEX HULL

Ένα πλεονέκτημα των ROC γραφημάτων είναι ότι καθιστούν εφικτή την απεικόνιση και οργάνωση της αναπαράστασης ταξινομητών (classifier performance)χωρίς να λαμβάνουν υπόψη τις κατανομές κλάσεων ή τα κόστη λαθών. Αυτή η ιδιότητα αποκτά ιδιαίτερη σημασία όταν διερευνάται η εκμάθηση με ασύμμετρες κατανομές ή η ευαίσθητη ως προς το κόστος μάθηση. Ένας ερευνητής μπορεί να σχεδιάσει την αναπαράσταση ενός συνόλου ταξινομητών και το γράφημα που θα προκύψει να παραμένει απαράλλακτο ως προς τα λειτουργικά χαρακτηριστικά (ασυμμετρία κλάσης και κόστος σφαλμάτων). Αν τα χαρακτηριστικά αυτά αλλάξουν, η περιοχή ενδιαφέροντος ίσως αλλάξει, αλλά το καθεαυτό γράφημα όχι. Γενικά ένας ταξινομητής θεωρείται βέλτιστος αν βρίσκεται πάνω στη κυρτή θήκη(Barber, Dobkin, & Huhdanpaa, 1993)του συνόλου των σημείων στο ROCχώρο. Την κυρτή αυτή θήκη την ονομάζουμε ROC convexhull (ROCCH) του αντίστοιχου συνόλου ταξινομητών.

Η μέθοδος του ROC convexhull βασίζεται στο συνδυασμό της ανάλυσης αποφάσεων και της ROCανάλυσης και τις υιοθετεί για *τη σύγκριση ενός πλήθους γνωστών ταξινομητών*. Η μέθοδος βασίζεται σε τρεις ύψιστης σημασίας αρχές:

1. Πρώτον, οROC χώρος χρησιμοποιείται για να αξιολογήσει τις αναπαραστάσεις ταξινομητών (classification performance) βάση πληροφοριών για την κατανομή της κλάσης και του κόστους.
2. Δεύτερον, η πληροφορία για την λήψη απόφασης απεικονίζεται μέσω της ROC καμπύλης.
3. Τρίτον, χρησιμοποιούμε την κυρτή θήκη (convex hull) του ROC χώρου προκειμένου να προσδιοριστεί το υποσύνολο των βέλτιστων μεθόδων.

Ορίζονται παρακάτω κάποιες βασικές έννοιες οι οποίες θα χρειαστούν παρακάτω. Έστω λοιπόν cμια κλάση-ταξινόμηση για την οποία ορίζονται τα εξής:

1. Το κόστος από το σφάλμα ενός ψευδώς θετικού αποτελέσματος συμβολίζεται ως $c(Y, n)$
2. Το κόστος από το σφάλμα ενός ψευδώς αρνητικού αποτελέσματος συμβολίζεται ως $c(N, p)$

Εάν ένας ταξινομητής παράγει πιθανότητες που προκύπτουν εκ των υστέρων, η ανάλυση αποφάσεων εξασφαλίζει costsensitive ταξινομήσεις. Έστω ένα γεγονός I, για την απόφαση να πραγματοποιηθεί θετική ταξινόμηση ισχύει η παρακάτω σχέση:

$$[1-p(p|I)]\cdot c(Y,n) < p(p|I)\cdot c(N,p)$$

Όπου $p(p|I)$ η εκ των υστέρων πιθανότητα το γεγονός I να είναι θετικό.

Άσχετα από το είδος των ταξινομήσεων, το κανονικοποιημένο κόστος ενός τεστ μπορεί να εκτιμηθεί εμπειρικά ως εξής:

$$Cost = FP\cdot c(Y,n) + FN\cdot c(N,p)$$

Έστω λοιπόν $p(p)$ η εκ των προτέρων πιθανότητα ενός θετικού παραδείγματος και $p(n)=1-p(p)$ η πιθανότητα ενός αρνητικού αποτελέσματος. Το κόστος των ψευδώς θετικών και ψευδώς αρνητικών σφαλμάτων δίνονται από τις $c(Y,n)$ και $c(N,p)$ αντίστοιχα.

Το αναμενόμενο κόστος της ταξινόμησης ενός ταξινομητή που αντιπροσωπεύεται από ένα σημείο (TP,FP) στο ROCχώρο δίνεται από τον παρακάτω τύπο:

$$p(p)\cdot(1-TP)\cdot c(N,p) + p(p)\cdot FP\cdot c(Y,n)$$

Συνεπώς δύο σημεία $(TP_1,FP_1))$ και (TP_2,FP_2) έχουν την ίδια αναπαράσταση όταν:

$$\frac{TP_2-TP_1}{FP_2-FP_1} = \frac{p(n)\cdot c(Y,n)}{p(p)\cdot c(N,p)}$$

Αυτή η εξίσωση ορίζει την κλίση μιας iso-performance καμπύλης. Όλοι οι ταξινομητές που αντιστοιχούν σε σημεία πάνω σε αυτήν την καμπύλη έχουν το ίδιο αναμενόμενο κόστος. Κάθε σύνολο από κατανομές κλάσεων και κόστους ορίζει μια οικογένεια από ισομορφικές καμπύλες. Οι καμπύλες που βρίσκονται πιο βόρεια - έχουν μεγαλύτερο TPδείκτη και θεωρούνται καλύτερες καθώς αντιστοιχούν σε ταξινομητές με χαμηλότερο αναμενόμενο κόστος.

Συγκεκριμένα, το αναμενόμενο κόστος από την εφαρμογή ενός ταξινομητή αντιστοιχεί σε ένα σημείο (FPR,TPR) στο ROCχώρο και δίνεται από την παρακάτω σχέση:

$$p(\mathbf{p})\cdot(1-TP)\cdot c(\mathbf{N},\mathbf{p}) + p(\mathbf{n})\cdot FP\cdot c(\mathbf{Y},\mathbf{n})$$

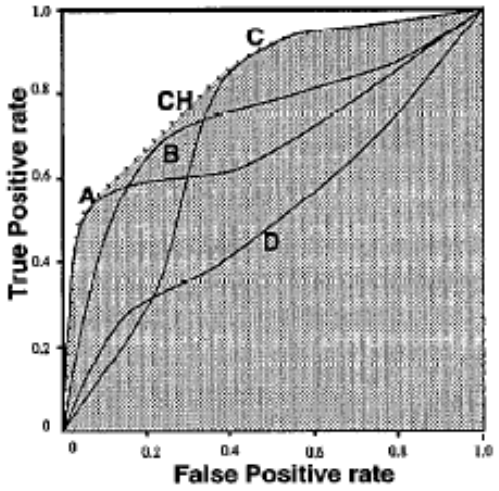
Γεννάται λοιπόν το ερώτημα πώς παράγεται αυτή η μέθοδος σύγκρισης ταξινομητών. Τα βήματα που ακολουθούνται είναι τα παρακάτω:

1. Για κάθε ταξινομητή κάνουμε τη γραφική παράσταση των TP και FP επιλέγοντας κατάλληλο διαχωριστικό όριο.
2. Βρίσκουμε την κυρτή θήκη του συνόλου των σημείων που αντιστοιχούν στους ταξινομητές που μας ενδιαφέρουν. Για n ταξινομητές αυτό μπορεί να γίνει σε χρονικό διάστημα ίσο με O(nlog(n)) μέσω του QuickHull αλγόριθμου.
3. Για κάθε κατανομή που μας ενδιαφέρει βρίσκουμε την κλιση των αντίστοιχων iso-performance καμπυλών.

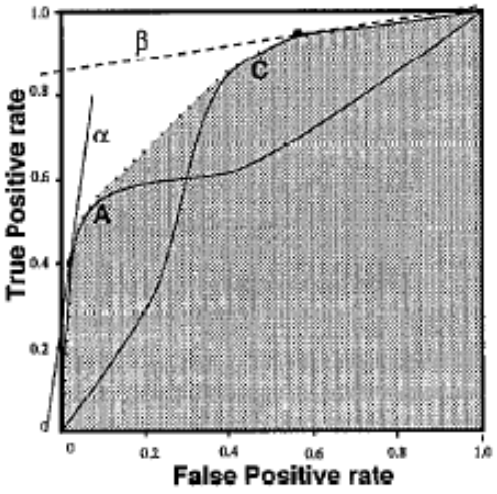
4. Για κάθε σύνολο κατανομών ο βέλτιστος ταξινομητής θα είναι το σημείο της κυρτής θήκης που τέμνει την ισομορφική καμπύλη με το μεγαλύτερο TP. Οι διακυμάνσεις των κλίσεων προσδιορίζουν hull τμήματα.

2.8.2 ΧΡΗΣΗΤΟΥROCCONVEXHULL

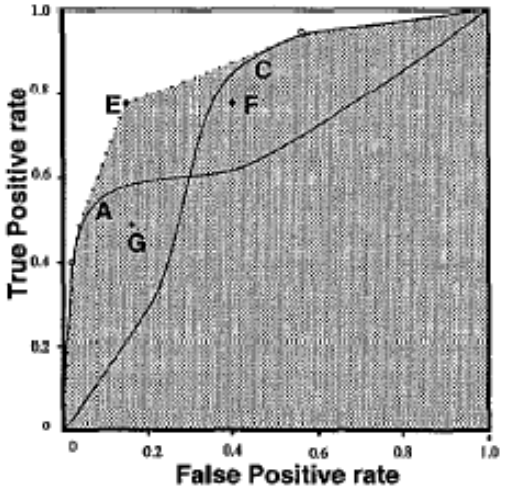
Η μέθοδος του ***ROC convex hull*** εφαρμόζεται και σε δίτιμους αλλά και σε συνεχείς ταξινομητές. Οι δίτιμοι ταξινομητές αναπαριστώνται από χωριστά σημεία στο ROCχώρο. Οι συνεχείς ταξινομητές παράγουν αριθμητικά αποτελέσματα στα οποία εφαρμόζεται ένα διαχωριστικό όριο και έτσι παράγεται μια σειρά από ζεύγη τιμών (FP, TP) τα οποία συνθέτουν τηROCκαμπύλη. Στο σχήμα 4 απεικονίζονται οι δίτιμοι εκτιμητές E, Fκαι G. Ο E μπορεί να είναι βέλτιστος κάτω από συγκεκριμένες συνθήκες καθώς ξεπερνάει την κυρτή θήκη ενώ οι Fκαι G δε μπορούν να είναι καθώς δεν την ξεπερνούν.



Εικόνα 2.8.1: Η ROCκυρτή θήκη αναδεικνύει τους πιθανά βέλτιστους ταξινομητές



Εικόνα 2.8.2: Οι ευθείες α και β προσδιορίζουν τους βέλτιστους ταξινομητές υπό διαφορετικές συνθήκες



Εικόνα 2.8.3: Ο ταξινομητής Ε μπορεί να είναι βέλτιστος επειδή βρίσκεται πάνω από την κυρτή θήκη σε αντίθεση με τους F και G

Νέοι ταξινομητές μπορεί να προστεθούν στην ανάλυση όπως απεικονίζεται στην εικόνα 2.8.3. Κάθε νέος ταξινομητής είτε υπερέχει της κυρτής θήκης είτε όχι. Στην πρώτη περίπτωση η θήκη πρέπει να υπολογιστεί πάλι λαμβάνοντας υπόψη τον νέο ταξινομητή, ενώ στη δεύτερη ο νέος ταξινομητής μπορεί να αγνοηθεί.

Συμπερασματικά, η μέθοδος του ***ROC convex hull***-όταν τίθεται ζήτημα σύγκρισης κάτω από νέα κατανομή- απαιτεί μόνο τον υπολογισμό των κλίσεων των αντιστοίχων ισομορφικών καμπύλων και την τομή αυτών με τη θήκη, όπως φαίνεται στην εικόνα 2.8.2.

Επιπλέον η συγκεκριμένη μέθοδος εξασφαλίζει ακρίβεια οποιουδήποτε βαθμού στον καθορισμό των κατανομών κλάσεων και κόστους. Ακόμα και στην περίπτωση που τίποτα δεν είναι γνωστό για μια κατανομή η μέθοδος του ***ROC convex hull*** εξασφαλίζει τους βέλτιστους ταξινομητές κάτω από οποιεσδήποτε συνθήκες. Η εικόνα 2.8.1 δείχνει ότι δεδομένων των ταξινομητών Α, Β, Cκαι D,μόνο ο Α και οCμπορούν να θεωρηθούν βέλτιστοι.
Επιπρόσθετα παρατηρείται ότι ταξινομητής Α είναι βέλτιστος με δεδομένο το σενάριο ***A*** και ο Cυπό το σενάριο ***B.***

2.8.3 ΠΡΟΒΛΗΜΑΤΑ ΑΠΟΦΑΣΗΣ ΜΕ ΠΕΡΙΣΣΟΤΕΡΕΣ ΑΠΟ ΔΥΟ ΚΛΑΣΕΙΣ

Όλες οι περιπτώσεις μέχρι αυτό το σημείο, είχαν να αντιμετωπίσουν δύο μόνο κλάσεις και το μεγαλύτερο τμήμα της ιστορίας των ROC καμπυλών διατηρεί αυτή την υπόθεση. Η ROCανάλυση εφαρμόζεται συχνά στη λήψη ιατρικών αποφάσεων κατά την οποία δύο κλάσεις διάγνωσης παρουσία ή απουσία μίας παθολογικής κατάστασης είναι απόλυτα συνηθισμένο φαινόμενο. Οι δύο άξονες αντιπροσωπεύουν τη συσχέτιση μεταξύ σφαλμάτων (λανθασμένα θετικά αποτελέσματα) και οφελών (σωστά θετικά αποτελέσματα) που ένας ταξινομητής πραγματοποιεί μεταξύ δύο κλάσεων. Το μεγαλύτερο μέρος της ανάλυσης διενεργείται χωρίς προβλήματα χάρη

στη συμμετρία που υπάρχει στα προβλήματα δύο κλάσεων. Η αναπαράσταση που προκύπτει μπορεί να σχεδιαστεί σε δύο διαστάσεις, πράγμα που είναι εύκολο να απεικονιστεί.

2.8.4 ROC ΓΡΑΦΗΜΑΤΑ ΠΟΛΛΩΝ ΚΛΑΣΕΩΝ

Με περισσότερες από δύο κλάσεις η κατάσταση γίνεται πολύ πιο περίπλοκη, αν πρέπει να διαχειριστούμε ολόκληρο το χώρο. Με nκλάσεις ο πίνακας ενδεχομένων γίνεται έναςn × n πίνακας ο οποίος περιέχει τιςnσωστές ταξινομήσεις(οι οποίες αντιστοιχούν σε σημεία της κυρίας διαγωνίου) και$n^2 - n$ πιθανά λανθασμένες ταξινομήσεις (ο οποίες αντιστοιχούν σε όλες τις άλλες θέσεις πλην της κυρίας διαγωνίου). Εκτός από μία εικόνα της συσχέτισης μεταξύ του TP και FP, εμείς παίρνουμε και ένα σύνολο από nοφέλη και $n^2 - n$ σφάλματα. Με τρείς μόνο κλάσεις θα προκύψουν $3^2 - 3$ πιθανά λανθασμένα αποτελέσματα. OLane (2000) έχει ασχοληθεί με το να σκιαγραφήσει τα θέματα που περικλείονται στα ROCγραφήματα πολλών κλάσεων και με τις προοπτικές διευθέτησής τους. Ο Srinivasan (1999) έδειξε ότι η ανάλυση πίσω από την ROCκυρτή θήκη προεκτείνεται και σε πολλαπλές κλάσεις και πολυδιάστατες κυρτές κλάσεις.

Μία μέθοδος για να χειριστούν οι n κλάσεις είναι να παραχθούν n διαφορετικά ROCγραφήματα ένα για την κάθε κλάση. Η παραπάνω διαδικασία λέγεται classreference σχηματισμός. Συγκεκριμένα, αν C είναι το σύνολο όλων των κλάσεων, το iROC γράφημα απεικονίζει την αναπαράσταση ταξινόμησης χρησιμοποιώντας την κλάση c_iσαν θετική κλάση και όλες τις άλλες κλάσεις σαν την αρνητική κλάση:

$$P_i = c_i \qquad (1)$$

$$N_i = \bigcup_{j \neq i} c_j \in C \qquad (2)$$

Αν και ο παραπάνω είναι ένας βολικός τύπος, παράλληλα διακυβεύει μία από τις πιο ελκυστικές ιδιότητες των ROCγραφημάτων, το γεγονός ότι αυτά είναι ανεπηρέαστα ως προς την ασυμμετρία κλάσης (βλέπε παράγραφο 2.4.2). Το παραπάνω συμβαίνει γιατί κάθε N_iσυμπεριλαμβάνει την ένωση $n-1$κλάσεων και αλλαγές στην ιεράρχηση μέσα σε αυτές τις κλάσεις πιθανόν να αλλάξουν το ROCγράφημα της c_i κλάσης. Για παράδειγμα αν υποθέσουμε ότι κάποια κλάση $c_k \in N$ να πιστοποιηθεί. Ένας ταξινομητής για την κλάση c_i , $i \neq k$ίσως εκμεταλλευτεί κάποια χαρακτηριστικά της κλάσης c_k προκειμένου να παράγει χαμηλές τιμές για τα περιστατικά της c_k $κλάσης$. Αυξάνοντας το προβάδισμα της c_k κλάσης πιθανόν να μετατραπεί η αναπαράσταση του ταξινομητή πράγμα ισοδύναμο με το να αλλάξει ο ουσιαστικός στόχος με την αύξηση της κυριαρχίας μιας εκ των ζεύξεών του. Το γεγονός αυτό με τη σειρά του θα αλλάξει τη ROCκαμπύλη. Παρόλα αυτά, αυτή η μέθοδος μπορεί να δουλέψει καλά στην πράξη και να παρέχει μια λογική ευελιξία στην αξιολόγηση.

2.8.5 ΕΜΒΑΔΟΝ AUC ΣΤΗΝ ΠΕΡΙΠΤΩΣΗ ΠΟΛΛΩΝ ΚΛΑΣΕΩΝ

Το εμβαδόν AUCείναι ένα μέτρο της ικανότητας διαχωρισμού σε ένα ζεύγος κλάσεων. Σε ένα πρόβλημα δύο κλάσεων, το εμβαδόν AUCείναι μία τιμή μιας μόνο κλίμακας, αλλά ένα πρόβλημα πολλών κλάσεων εισάγει το ζήτημα του συνδυασμού πολλαπλών τιμών διαχωρισμού. Ο αναγνώστης μπορεί να ανατρέξει στο άρθρο των Hand και Till's (2001)για μία τέλεια ανάπτυξη των παραπάνω θεμάτων.

Μία προσέγγιση για τον υπολογισμό του εμβαδού AUCστην περίπτωση πολλών κλάσεων έδωσαν οι Provost και Domingos (2001) πάνω στη δουλειά τους για την πιθανολογική εκτίμηση δέντρων. Αυτοί υπολόγισαν το εμβαδόν αυτό, δημιουργώντας τη ROCκαμπύλη κάθε κλάσης, έπειτα μετρώντας το εμβαδόν κάτω από τη καμπύλη και τέλος αθροίζοντας τα AUC εμβαδά αυτά πολλαπλασιασμένα με τον αντίστοιχο συντελεστή βαρύτητας των κλάσεων. Πιο συγκεκριμένα αυτοί υπολογίζουν το AUCεμβαδό με τον παρακάτω τύπο:

$$AUC_{total} = \sum_{c_i \in C} AUC(c_i) \cdot p(c_i)$$

Όπου το $AUC(c_i)$ είναι το εμβαδόν κάτω από τη class reference ROCκαμπύλη της κλάσης c_i όπως και στη εξίσωση 2. Αυτός ο ορισμός απαιτεί μόνο $|C|$υπολογισμούς εμβαδών AUC επομένως η συνολική του περιπλοκότητα είναι $O(|C|n\log n)$.

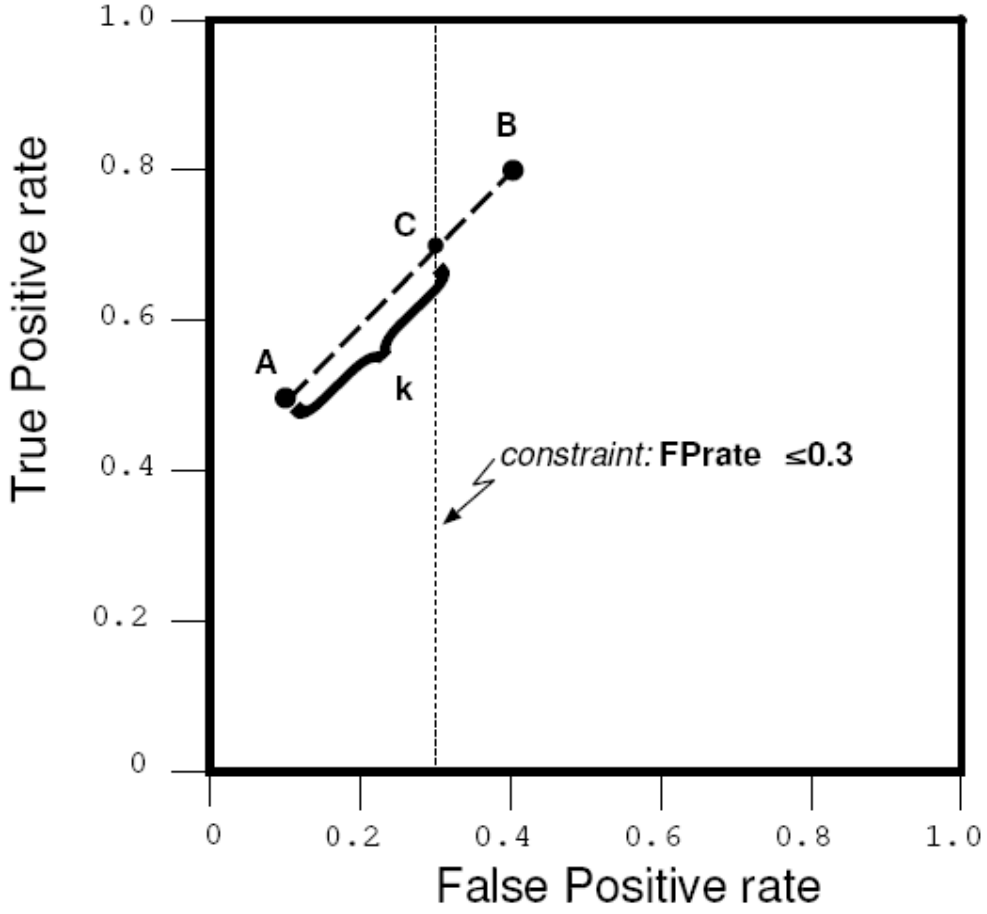
Το πλεονέκτημα του τύπου των Provost και Domingos για το εμβαδό AUCείναι ότι τοAUC_{total} προέρχεται απευθείας από τη classreference ROC καμπύλη και αυτές οι καμπύλες μπορούν να δημιουργηθούν και να απεικονιστούν εύκολα. Το μειονέκτημα του τύπου αυτού είναι ότι η ROCκαμπύλη της κλάσης αναφοράς είναι ευαίσθητη ως προς τις κατανομές κλάσης και τα κόστη σφαλμάτων, με αποτέλεσμα να γίνεται το ίδιο ευαίσθητος και ο τύπος υπολογισμού του AUC_{total}.

Οι Hand και Till (2001) έδωσαν μια διαφορετική προσέγγιση για τον υπολογισμό του εμβαδού AUCπολλών κλάσεων. Αρχικά επεδίωξαν την εύρεση ενός μέτρου που θα παραμένει ανεπηρέαστο ως προς τις κατανομές κλάσης και τα κόστη σφαλμάτων. Η εξόρυξη του τύπου αυτού είναι πολύ περίπλοκη ώστε να παρουσιαστεί εδώ, αλλά βασίζεται στο γεγονός ότι το εμβαδόν AUCείναι ισοδύναμο με την πιθανότητα ένας ταξινομητής να βαθμολογήσει ένα τυχαία επιλεγμένο θετικό περιστατικό, υψηλότερα από ότι ένα τυχαία επιλεγμένο αρνητικό περιστατικό. Από αυτό το πιθανολογικό σχήμα, αυτοί αντλούν ένα τύπο που μετρά την αμερόληπτη pairwise ικανότητα διαχωρισμού των κλάσεων. Το μέτρο τους αυτό, γνωστό ως Μυπολογίζεται ως εξής:

$$AUC_{total} = \frac{2}{|C|(|C|-1)} \sum_{\{c_i, c_j \in C\}} AUC(c_i, c_j)$$

Όπουn ο αριθμός των κλάσεων και $AUC(c_i, c_j)$ το εμβαδόν κάτω από την δύο κλάσεων ROCκαμπύλη, η οποία συμπεριλαμβάνει και τις δύο κλάσεις c_i $και$ c_j. Το άθροισμα υπολογίζεται πάνω σε όλα τα ζευγάρια των διακριτών κλάσεων άσχετα με τη σειρά διάταξής τους. Υπάρχουν $|C|(|C|-1)$ τέτοια ζευγάρια επομένως η τάξη περιπλοκότητας του μέτρου αυτού είναι $O(|C|^2 n\log n)$. Ενώ ο τύπος των Hand και Till είναι άρτια δικαιολογημένος και ανεπηρέαστος ως προς αλλαγές στην κατανομή

κλάσης δεν υπάρχει εύκολος τρόπος να απεικονιστεί η επιφάνεια της οποίας το εμβαδόν, ο τύπος αυτός υπολογίζει.



Εικόνα 2.10 Παρεμβάλλοντας ταξινομητές

2.8.6 ΣΥΝΔΥΑΣΜΟΙ ΤΑΞΙΝΟΜΗΤΩΝ

Ενώ οι ROCκαμπύλες χρησιμοποιούνται κυρίως για την απεικόνιση και αξιολόγηση ανεξάρτητων ταξινομητών ο ROCχώρος μπορεί επίσης να χρησιμοποιηθεί για να εκτιμήσει την αναπαράσταση συνδυασμών ταξινομητών.

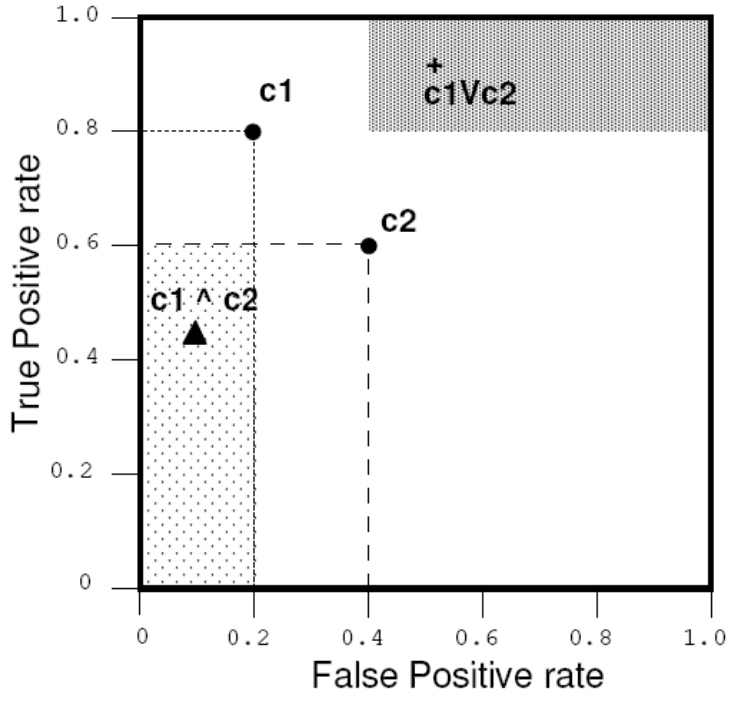
2.8.7 ΠΑΡΕΜΒΑΛΛΟΝΤΑΣ ΤΑΞΙΝΟΜΗΤΕΣ

Είναι πιθανό να παρουσιαστεί κάποια στιγμή η περίπτωση κατά την οποία η αναπαράσταση που επιθυμείται από ένα ταξινομητή να μην αντιπροσωπεύεται ακριβώς από κάποιον διαθέσιμο ταξινομητή αλλά να κυμαίνεται μεταξύ δύο ταξινομητών. Η επιθυμητή αναπαράσταση ταξινόμησης μπορεί να αποκτηθεί δοκιμάζοντας τις αποφάσεις κάθε ταξινομητή. Η αναλογία της δειγματοληψίας θα καθορίσει που θα κυμαίνεται η προκύπτουσα αναπαράσταση ταξινόμησης.

Η εικόνα 2.10 απεικονίζει την παραπάνω ιδέα. Αν γίνει η υπόθεση ότι έχουν δημιουργηθεί δύο ταξινομητές οι Α και Β των οποίων η αναπαράσταση φαίνεται ο Α έχει FPR = 0.1 και ο Β έχει FPR = 0.4. Επιπλέον, στην περίπτωση που απαιτείται ένας ταξινομητής με FPR όχι μεγαλύτερο από 0.3 (η κάθετη διακεκομμένη γραμμή στην εικόνα 2.10) μια λύση είναι απλά να επιλεγεί ο Α ταξινομητής καθώς ο Β δεν ικανοποιεί το κριτήριο. Παρόλα αυτά υπάρχει καλύτερη λύση. Μπορεί να δημιουργηθεί ένας ταξινομητής C ο οποίος θα παρεμβάλλεται μεταξύ των ταξινομητών Α και Β. Για το επιθυμητό FPRτου 0.3, η γραμμική παρεμβολή δίνει :

$$k = \frac{0.3 - 0.1}{0.4 - 0.1} = 2/3$$

Αν δοκιμαστούν τα αποτελέσματα του Β ταξινομητή κατά ένα ποσοστό 2/3 και τα αποτελέσματα του Α ταξινομητή κατά ένα ποσοστό 1-2/3=1/3 θα πρέπει να προκύψει η αναπαράσταση του Cταξινομητή. Στην πράξη αυτός ο κλασματικός δειγματισμός μπορεί να γίνει δοκιμάζοντας τυχαία αποτελέσματα από τον κάθε ένα ταξινομητή: για το κάθε περιστατικό παίρνουμε τυχαία έναν αριθμό ανάμεσα στο 0 και το 1. Αν ο τυχαίος αριθμός είναι μεγαλύτερος από τον k εφαρμόζεται ο ταξινομητής Α στο περιστατικό και διατηρείται το αποτέλεσμά του, διαφορετικά εφαρμόζεται ο ταξινομητής Β στο περιστατικό αυτό.



Εικόνα 2.11Οι αναμενόμενες θέσεις των boolean συνδυασμών των c_1 και c_2 κλάσεων. Η c_1 κλάση τοποθετείται στο σημείο (0.2,0.8), ενώ ηc_2 στο σημείο (0.4,0.6). Η τομή $c_1 \wedge c_2$ θα κυμαίνεται κάπου στο σκιαγραφημένο κομμάτι στην κάτω αριστερή περιοχή. Κάτω από συνθήκες ανεξαρτησίας ταξινομητών, η τομή των παραπάνω κλάσεων θα τοποθετείται στο σημείο (0.08,0.42). Από την άλλη πλευρά η ένωση $c_1 \vee c_2$θα τοποθετηθεί κάπου εντός της σκιαγραφημένου κομματιού στην άνω δεξιά περιοχή. Κάτω από συνθήκες ανεξαρτησίας ταξινομητών, η ένωση των παραπάνω ταξινομητών θα τοποθετηθεί στο σημείο (0.45,9.0).

2.8.8 ΛΟΓΙΚΑ ΣΥΝΔΥΑΖΟΜΕΝΟΙ ΤΑΞΙΝΟΜΗΤΕΣ

Με δύο κλάσεις ένας ταξινομητής εμπορεί να θεωρηθεί ως υποστηρικτής για ένα περιστατικό I όπου c(I)= trueαν και μόνο αν c(I)=Y. Τότε μπορούμε να μιλάμε για Boolean συνδυασμούς ταξινομητών και ένα ROCγράφημα θα παρέχει ένα τρόπο απεικόνισης της αναπαράστασης τέτοιων συνδυασμών. Αυτό μπορεί να συμβάλλει στην απεικόνιση της περιοχής οριοθέτησης του νέου ταξινομητή καθώς και της αναμενόμενης θέσης του.

Αν δύο ταξινομητές c_1 και c_2 συνδυάζονται προκειμένου να δημιουργήσουν τον ταξινομητή $c_3 = c_1 \wedge c_2$ προκύπτει το ερώτημα, που θα τοποθετηθεί ο c_3στον ROCχώρο. Έστω TPR_3 και FPR_3οι ROCσυντεταγμένες του c_3. Ο ελάχιστος αριθμός περιστατικών του c_3 ταξινομητή που μπορεί να ταιριάζουν είναι μηδέν. Ο μέγιστος περιορίζεται από την τομή των θετικών συνόλων. Καθώς ένα νέο περιστατικό μπορεί να ικανοποιεί και τους δύο ταξινομητές c_1 και c_2 μπορούμε να οριοθετήσουμε τη θέση του c_3 ταξινομητή με την παρακάτω σχέση:

$$0 \leq TPR_3 \leq \min(TPR_1, TPR_2)$$

$$0 \leq TPR_3 \leq \min(FPR_1, FPR_2)$$

Η εικόνα 2.11 παρουσιάζει αυτό το ορθογώνιο οριοθέτησης για τους δύο ταξινομητές $c_1 \wedge c_2$, το οποίο αποτελεί το σκιαγραφημένο ορθογώνιο στην κάτω αριστερή γωνία. Το ερώτημα που προκύπτει είναι που ακριβώς μέσα σε αυτό το ορθογώνιο αναμένεται να τοποθετηθεί ο ταξινομητής c_3 . Έστω x ένα περιστατικό εντός του συνόλου των αληθώς θετικών αποτελεσμάτων TP_3του ταξινομητή c_3 . Τότε ισχύει η παρακάτω σχέση :

$$TPR_3 \approx p(x \in TP_3)$$

$$\approx p(x \in TP_1 \wedge TP_2)$$

Υποθέτοντας ανεξαρτησία των δύο ταξινομητών c_1 και c_2, προκύπτει το παρακάτω:

$$TPR_3 \approx p(x \in TP_1) \cdot p(x \in TP_2)$$

$$\approx \frac{|TP_1|}{|P|} \cdot \frac{|TP_2|}{|P|}$$

$$\approx TPR_1 \cdot TPR_2$$

Μία αντίστοιχη σχέση προκύπτει και για το FPR_3 , συγκεκριμένα θα έχουμε$FPR_3 = FPR_1 \cdot \mathrm{FPR}_2$. Έτσι η τομή των δύο ταξινομητών μπορεί να αναμένεται να τοποθετηθεί στο παρακάτω σημείο του ROCχώρου:

$$(FPR_1 \cdot FPR_2, TPR_1 \cdot TPR_2)$$

Το σημείο αυτό βρίσκεται ακριβώς στο τρίγωνο της εικόνας 2.11, με συντεταγμένες (0.08,0.42). Δε θα πρέπει να ξεχνάμε ότι η παραπάνω εκτίμηση προϋποθέτει ανεξαρτησία των ταξινομητών. Αλληλεπιδράσεις μεταξύ των ταξινομητών c_1 και c_2 μπορεί να έχουν ως αποτέλεσμα η θέση του ταξινομητή c_3 στον ROCχώρο να διαφέρει από την θέση που εκτιμήσαμε προηγουμένως.

Μπορούμε να αντλήσουμε παρόμοιες εκφράσεις και για την ένωση $c_4 = c_1 \vee c_2$.Σε αυτή την περίπτωση οι αναλογίες οριοθετούνται ως εξής:

$$max(TPR_1, TPR_2) \leq TPR_4 \leq \min(1, TPR_1 + TPR_2)$$

$$max(FPR_1, FPR_2) \leq FPR_4 \leq \min(1, FPR_1 + FPR_2)$$

Αυτή η οριοθετημένη περιοχή υποδηλώνεται στην εικόνα 2.11 με το σκιαγραφημένο ορθογώνιο στην πάνω δεξιά πλευρά του ROCγραφήματος. Η αναμενόμενη, υποθέτοντας πάντοτε ανεξαρτησία των ταξινομητών δίνεται παρακάτω:

$$TPR_4 = 1 - [1 - TPR_1 - TPR_2 + TPR_1 \cdot TPR_2$$

$$FPR_4 = 1 - [1 - FPR_1 - FPR_2 + FPR_1 \cdot FPR_2$$

Αυτό το σημείο υποδηλώνεται με το σύμβολο + μέσα στο ορθογώνιο οριοθέτησης.

Αυτές οι εξισώσεις επιτρέπουν περιορισμένη απεικόνιση των αποτελεσμάτων συνδυασμών ταξινομητών στον ROCχώρο. Αυτοί θα μπορούσαν επίσης να χρησιμοποιηθούν προκειμένου να κατευθύνουν ένα συμπέρασμα ή ένα χαρακτηριστικό κατασκευής ενός αλγορίθμου. Για παράδειγμα, υπολογίζοντας την αναμενόμενη θέση κάθε συνδυασμού και συγκρίνοντάς την με την ROC convex hull, μία μέθοδος θα μπορούσε ολίγα υποσχόμενους νέους ταξινομητές, πριν ακόμα αυτοί παραχθούν.

2.8.9 ΑΛΥΣΙΔΩΤΟΙ ΤΑΞΙΝΟΜΗΤΕΣ

Σε προηγούμενη παράγραφο αναφέρθηκε ότι ταξινομητές στην αριστερή πλευρά ενός ROCγραφήματος κοντά X=0 μπορούν να θεωρηθούν ως "συντηρητικοί" ενώ αυτοί στην πάνω πλευρά κοντά στο Y=1 μπορούν να θεωρηθούν ως "φιλελεύθεροι". Με τους παραπάνω χαρακτηρισμούς θα ήταν δελεαστικό να επινοηθεί ένα σύνθετο σχέδιο το οποίο θα εφαρμόζει ταξινομητές διαδοχικά σαν έναν κατάλογο κανόνων. Μία τέτοια τεχνική θα μπορούσε να λειτουργεί ως ακολούθως: Δοσμένων των ταξινομητών στο ROC convex hull, ένα περιστατικό δίνεται στον πιο συντηρητικό (περισσότερο στα αριστερά) ταξινομητή. Αν αυτός ο ταξινομητής δώσει σαν αποτέλεσμα Υτότε ο σύνθετος ταξινομητής θα δώσει και αυτός σαν αποτέλεσμαΥ. Διαφορετικά ο δεύτερος πιο συντηρητικός ταξινομητής δοκιμάζεται και πάει λέγοντας. Η διαδικασία αυτή τερματίζει όταν κάποιος ταξινομητής δίνει μία Υταξινόμηση ή όταν οι ταξινομητές φτάσουν ένα μέγιστο αναμενόμενο κόστος, τέτοιο ώστε να μπορεί να προσδιοριστεί από μία ισομορφική καμπύλη. Ο ταξινομητής που προκύπτει είναι ο $c_1 \vee c_2 \vee \ldots \vee c_k$όπου ο c_kέχει το υψηλότερο αναμενόμενο ανεχτό κόστος.

Δυστυχώς αυτή η αλυσίδα των ταξινομητών πιθανόν να μην λειτουργήσει όπως επιθυμείται. Οι θέσεις των ταξινομητών στον ROCχώρο, βασίζονται άμεσα στην υπόθεση της ανεξάρτητης αναπαράστασης. Όταν οι ταξινομητές εφαρμόζονται στην σειρά κατά αυτόν τον τρόπο δεν χρησιμοποιούνται ανεξάρτητα αλλά αντί αυτού εφαρμόζονται σε περιστατικά τα οποία πιο συντηρητικοί ταξινομητές έχουν ήδη ταξινομήσει ως αρνητικά. Εξαιτίας των αλληλεπιδράσεων των ταξινομητών(τομές

μεταξύ των TPκαι FP συνόλων των ταξινομητών), ο ταξινομητής που προκύπτει πιθανόν να έχει πολύ διαφορετικά χαρακτηριστικά αναπαράστασης από ότι οι συνθετικοί ταξινομητές. Παρά το γεγονός ότι σε προηγούμενη παράγραφο εισήχθηκε μία υπόθεση ανεξαρτησίας, η οποία μπορεί πράγματι να είναι δικαιολογημένη στον συνδυασμό δύο ταξινομητών, αυτή η υπόθεση διατηρείται όλο και λιγότερο καθώς μεγαλύτερες αλυσίδες ταξινομητών δημιουργούνται.

2.8.10 Η ΣΗΜΑΣΙΑ ΤΗΣ ΤΕΛΙΚΗΣ ΕΠΑΛΗΘΕΥΣΗΣ

Προκειμένου να κλείσει αυτή η παράγραφος για τον συνδυασμό ταξινομητών, θα δωθεί έμφαση σε ένα βασικό σημείο που είναι εύκολο να ξεχαστεί. Τα ROCγραφήματα χρησιμοποιούνται ευρέως στην αξιολόγηση ταξινομητών και δημιουργούνται από ένα τελικό σύνολο δεδομένων. Αν παρόλα αυτά ένα ROCγράφημα χρησιμοποιηθεί για την επιλογή ή τον συνδυασμό ταξινομητών αυτό πρέπει να θεωρηθεί ότι αποτελεί μέρος της φάσης κατάρτισης. Ένα ξεχωριστό held-out σύνολο επαλήθευσης πρέπει να χρησιμοποιηθεί προκειμένου να εκτιμήσει την αναμενόμενη αναπαράσταση του ταξινομητή(ή των ταξινομητών). Το παραπάνω ισχύει ακόμα και όταν οι ROCκαμπύλες χρησιμοποιούνται για να παράγουν ένα ROC convexhull.

2.9 ΕΝΑΛΛΑΚΤΙΚΕΣ ΤΩΝ ROC ΓΡΑΦΗΜΑΤΩΝ

Πρόσφατα, διάφορες εναλλακτικές των ROCγραφημάτων έχουν προταθεί. Θα τις παρουσιάσουμε συνοπτικά παρακάτω.

2.9.1 ΚΑΜΠΥΛΕΣ DET

ΤαDETγραφήματα (Martin, Doddington, Kamm, OrdowskikaιPrzybocki, 1997) δεν μπορούν να θεωρηθούν τόσο, εναλλακτικές των ROCκαμπυλών όσο εναλλακτικοί τρόποι παρουσίασής τους. Υπάρχουν δύο διαφορές. Αρχικά τα DETγραφήματα σχεδιάζουν τα λανθασμένα αρνητικά αποτελέσματα στον άξονα των Υέναντι των αληθώς θετικών αποτελεσμάτων, με αποτέλεσμα να σχεδιάζουν ένα τύπο σφάλματος έναντι ενός άλλου. Δεύτερον τα DETγραφήματα, είναι βαθμολογημένα σε logκλίμακα και στους δύο άξονες ούτως ώστε η περιοχή στο χαμηλότερο αριστερά μέρος της καμπύλης (η οποία αντιστοιχεί στην υψηλότερη περιοχή ενός ROCγραφήματος) επεκτείνεται. Martinetal. (1997) ισχυρίστηκε ότι οι ταξινομητές που αναπαριστώνται καλώς με χαμηλά FPR(ισοδύναμα χαμηλά false negative rates) τείνουν να συσσωρεύονται στη χαμηλότερη αριστερή περιοχή ενός ROCγραφήματος. Η βαθμολόγηση σelogκλίμακα δίνει σε ένα DETγράφημα μεγαλύτερο εμβαδόν και επιτρέπει σε αυτούς τους ταξινομητές να συγκριθούν πιο εύκολα.

2.9.2 ΚΑΜΠΥΛΕΣ ΚΟΣΤΟΥΣ

Στη παράγραφο 2.8.1 έγινε εκτενής περιγραφή για το πως οι πληροφορίες για τις αναλογίες κλάσης και κοστών σφαλμάτων μπορούν να συνδυαστούν προκειμένου να καθορίσουν την κλίση μιας ισομορφικής γραμμής. Μια τέτοια γραμμή μπορεί να

τοποθετηθεί πάνω σε μία ROCκαμπύλη και να χρησιμοποιηθεί προκειμένου να αναγνωρίσει ποιοι ταξινομητές συμπεριφέρονται καλύτερα υπό τις συνθήκες που μας ενδιαφέρουν. Σε πολλά σενάρια ελαχιστοποίησης κόστους αυτό απαιτεί να ελέγξουμε τις καμπύλες και να κρίνουμε την εφαπτομένη για την οποία ένας ταξινομητής επικρατεί.

Οι Drummond και Holte (2000, 2002) επεσήμαναν ότι το να μετρήσουμε τις κλήσεις της εφαπτομένης, πιθανόν να είναι δύσκολο να πραγματοποιηθεί. Καθώς καθορίζουν τις περιοχές επικράτησης και το μέγεθος πάνω από το οποίο ένας ταξινομητής θεωρείται ανώτερος από έναν άλλο, είναι δελεαστικό το γεγονός ότι οι γραμμές σύγκρισης είναι εφαπτομένες και όχι απλές κάθετες γραμμές. Οι Drummond και Holte εξηγούν ότι αν η πρωταρχική χρήση μιας καμπύλης είναι να συγκρίνει τα σχετικά κόστη τότε τα γραφήματα πρέπει να αναπαριστούν αυτά τα κόστη με μεγάλη ευκρίνεια. Αυτοί προτείνουν τις cost curves (καμπύλες κόστους ως εναλλακτική των ROCκαμπυλών). Σε μία καμπύλη κόστους ο Χάξονας κυμαίνεται από το 0 μέχρι το 1 και μετρά την αναλογία των θετικών στην κατανομή. Ο Υάξονας, ο οποίος επίσης κυμαίνεται από το 0 μέχρι το 1, αντιπροσωπεύει το σχετικό αναμενόμενο κόστος της λανθασμένης ταξινόμησης. Ένας τέλειος ταξινομητής αποτελεί μία οριζόντια γραμμή από το (0,0) έως το (0,1). Οι καμπύλες κόστους είναι μία γραμμή-σημείο η οποία προέρχεται από ROC καμπύλες: ένα σημείο (δηλαδή ένας διακριτός ταξινομητής) στο ROC χώρο αντιπροσωπεύεται από μία γραμμή στον χώρο των καμπυλών κόστους, με τη γραμμή να σχεδιάζει το σχετικό αναμενόμενο κόστος του ταξινομητή. Για κάθε Χ σημείο, τα αντίστοιχα Υσημεία αντιπροσωπεύουν τα αναμενόμενα κόστη των ταξινομητών. Έτσι ενώ στον ROCχώρο, η ROC convex hull περιέχει το σύνολο των ταξινομητών χαμηλότερου κόστους, στον χώρο κόστους ο χαμηλότερος φάκελος αντιπροσωπεύει αυτό το σύνολο.

2.9.3 ΣΧΕΤΙΚΗ ΑΝΩΤΕΡΟΤΗΤΑ ΓΡΑΦΗΜΑΤΩΝ ΚΑΙ Ο LC ΔΕΙΚΤΗΣ

Όπως και οι καμπύλες κόστους έτσι και ο LCδείκτης (Adams και Hand, 1999) αποτελεί έναν μετασχηματισμό των ROCκαμπυλών, ο οποίος διευκολύνει τη σύγκριση ταξινομητών βάση του κόστους. Οι Adams και Hand ισχυρίζονται ότι συγκεκριμένη πληροφορία κόστους είναι σπάνια, αλλά κάποιες πληροφορίες σχετικά με τα κόστη είναι πάντα διαθέσιμες, επομένως το εμβαδόν AUCείναι πλέον ακατάλληλο μέτρο σύγκρισης αναπαράστασης ταξινομητών. Ένας ειδικός πιθανόν να μην είναι σε θέση να προσδιορίσει ακριβώς ποιο θα πρέπει να είναι το κόστος ενός λανθασμένα θετικού και ενός λανθασμένα αρνητικού αποτελέσματος αλλά συνήθως αυτός έχει μία ιδέα για το πόσο ποιο ακριβό είναι ένα κόστος από ένα άλλο. Αυτό μπορεί να εκφραστεί ως ένα εύρος τιμών πάνω στο οποίο θα κυμαίνεται η αναλογία κόστους σφάλματος.

Η μέθοδος των Adams και Hand, απεικονίζει την αναλογία του κόστους σφάλματος πάνω στο διάστημα (0,1). Αυτή η μέθοδος στη συνέχεια μετασχηματίζει ένα σύνολο από ROCκαμπύλες σε ένα σύνολο από παράλληλες γραμμές οι οποίες δείχνουν ποιος ταξινομητής επικρατεί και σε ποια περιοχή, μέσα στο παραπάνω διάστημα. Ένας ειδικός παρέχει ένα υποσύνολο του (0,1) μέσα στο οποίο η παραπάνω αναλογία κόστους αναμένεται να μειωθεί όπως επίσης και την πιο πιθανή τιμή της αναλογίας αυτής. Αυτό εξυπηρετεί στο να επικεντρώσουμε την προσοχή μας στο μεσοδιάστημα αυτό μεγαλύτερου ενδιαφέροντος. Πάνω σε αυτά τα σχετικής "ανωτερότητας

γραφήματα" ένα μέτρο εμπιστοσύνης- ο LCδείκτης-μπορεί να οριστεί ως ο δείκτης για το πόσο πιθανόν είναι ένας ταξινομητής να είναι ανώτερος από έναν άλλο μέσα σε αυτό το διάστημα.
Τα σχετικής ανωτερότητας γραφήματα μπορούν να θεωρηθούν ως μία δυαδική εκδοχή των καμπυλών κόστους, στην οποία εμείς ενδιαφερόμαστε μόνο για το ποιος ταξινομητής είναι ανώτερος. Ο LCδείκτης (για την σύγκριση κόστους) αποτελεί κατά αυτόν τον τόπο ένα μέτρο εμπιστοσύνης περισσότερο για την ανωτερότητα ταξινομητών παρά για την διαφορά κόστους.

2.10 ΣΥΜΠΕΡΑΣΜΑ

Τα ROCγραφήματα είναι ένα πολύ χρήσιμο εργαλείο για την απεικόνιση και την αξιολόγηση ταξινομητών. Αυτά παρέχουν τη δυνατότητα ενός πολύτιμου μέτρου για την αναπαράσταση ταξινομητών σε σχέση με την τιμή του σφάλματος ή της ακρίβειας και διατηρούν κάποια πλεονεκτήματα, έναντι άλλων μέτρων αξιολόγησης, όπως γραφήματα ακρίβειας-ευαισθησίας και lift curves. Παρόλα αυτά όπως και με κάθε άλλη μετρική αξιολόγησης, προκειμένου να τις χρησιμοποιήσουμε σωστά απαιτείται γνώση των ιδιαίτερων χαρακτηριστικών και των περιορισμών τους.

ΚΕΦΑΛΑΙΟ ΙΙΙ

ΜΕΘΟΔΟΙ ΥΠΟΛΟΓΙΣΜΟΥ ΠΕΡΙΟΧΗΣ ΚΑΤΩ ΑΠΟ THROC ΚΑΜΠΥΛΗ

3.1 ΕΙΣΑΓΩΓΗ

Όπως ήδη αναφέρθηκε σε προηγούμενες παραγράφους η ROCκαμπύλη (receiver operating characteristic curve) χρησιμοποιείται συχνά σαν μέτρο αποτελεσματικότητας διαγνωστικών δεικτών. Στο κεφάλαιο αυτό θα αναλύσουμε και θα συγκρίνουμε μεθόδους εκτίμησης της περιοχής κάτω από αυτήν την καμπύλη. Αυτές είναι βασισμένες i)στο Mann–Whitney στατιστικό τεστ,ii) στη λείανση της καμπύλης ROC με τη χρήση πυρήνων, iii)στις υποθέσεις κανονικότητας και iv) στον εμπειρικό μετασχηματισμό σε συνθήκες κανονικότητας. Αυτές οι μέθοδοι συγκρίνονται με βάση την προτίμηση αλλά και το μέσο τετραγωνικό σφάλμα σε μία ευρεία κλίμακα συνθηκών, μέσα από μία εκτεταμένη μελέτη προσομοίωσης. Τέλος καταλήγουμε στο ότι ο μετασχηματισμός σε συνθήκες κανονικότητας πρέπει να προτιμάται με εξαίρεση τις bimodalπεριπτώσεις όπου η μέθοδος των πυρήνων μπορεί να γίνει αποτελεσματική.

Οι μέθοδοι εκτίμησης του εμβαδού κάτω από μία ROCκαμπύλη, οι οποίες παρουσιάζονται και συγκρίνονται σε αυτό το κεφάλαιο θα αφορούν συνεχείς διαγνωστικούς δείκτες.

Ο σχεδιασμός της ROCκαμπύλης είναι ένας ευρέως διαδεδομένος τρόπος παρουσίασης της ακρίβειας διαχωρισμού ενός διαγνωστικού τεστ το οποίο έχει ως σκοπό να ανιχνεύσει αν ένας ασθενής πάσχει ή όχι από μία ασθένεια ή αν ανήκει σε κάποια δεδομένη κατηγορία ή όχι. Η ROCμεθοδολογία προέρχεται από μία θεωρία ανίχνευσης σήματος η οποία χρησιμοποιείται προκειμένου να καθοριστεί αν ένας ηλεκτρονικός δέκτης είναι ικανός να διακρίνει ικανοποιητικά ένα σήμα από ένα θόρυβο. Αυτή έχει χρησιμοποιηθεί πέραν της ιατρικής, στη ραδιολογία, στη ψυχιατρική, στον μη καταστρεπτικό έλεγχο και στην βιομηχανία συστημάτων επιθεώρησης..

Πρόσφατα έχει παρατηρηθεί μία αυξημένη χρήση των ROC καμπυλών για την εκτίμηση της αποτελεσματικότητας των συνεχών διαγνωστικών δεικτών στην διάκριση μεταξύ ασθενών και υγειών ατόμων. Ένα άτομο εκτιμάται ως ασθενής (θετικό αποτέλεσμα) ή υγιής (αρνητικό αποτέλεσμα) ανάλογα με το αν η τιμή του αντίστοιχου δείκτη είναι μεγαλύτερη ή μικρότερη ή ίση από μία δοσμένη τιμή ορίου(γνωστή και ως σημείο απόφασης). Συνδεδεμένη με την εκάστοτε τιμή του ορίου είναι και η πιθανότητα ενός αληθώς θετικού αποτελέσματος(ευαισθησία) και η πιθανότητα ενός αληθώς αρνητικού αποτελέσματος(ειδικότητα). Η θεωρητική ROCκαμπύλη είναι μία γραφική παράσταση του q = ευαισθησία(sensitivity) έναντι του p=1-ειδικότητα(specificity) για όλες τις πιθανές τιμές του ορίου. Οι ROCκαμπύλες μπορούν να εκτιμηθούν κάτω από παραμετρικές ή μη παραμετρικές υποθέσεις.

Ο πιο ευρέως χρησιμοποιούμενος δείκτης, παγκοσμίως, της διαγνωστικής ακρίβειας είναι το εμβαδόν κάτω από τη ROCκαμπύλη (AUC). Έστω ότι τα Χκαι Υ δηλώνουν τις μετρήσεις του διαγνωστικού δείκτη για τα παθολογικά και υγιή περιστατικά

αντίστοιχα. Ο Bamber έδειξε ότι AUC= Prob(X>Y). Αυτό μπορεί να μεταφραστεί ως η πιθανότητα, από ένα τυχαία επιλεγμένο ζεύγος ενός ασθενή και ενός υγιούς ατόμου, η τιμή του διαγνωστικού δείκτη να είναι υψηλότερη για τον ασθενή. Οι τιμές του εμβαδού AUCκοντά στο 1.0 υποδηλώνουν ότι ο δείκτης έχει υψηλή διαγνωστική ακρίβεια. Αυτή η ένδειξη εμφανίζεται σε πολλά προβλήματα, όχι συνδεδεμένα με διαγνωστικούς δείκτες. Για παράδειγμα, αν θεωρήσουμε ότι η μεταβλητή X αντιπροσωπεύει τη δύναμη ενός μηχανικού συστήματος, στο οποίο εφαρμόζεται πίεση Y, τότε η πιθανότητα Prob(X>Y)εκφράζει την αξιοπιστία του συστήματος αυτού. Οι Wolfe και Hogg συστήνουν τη χρήση του δείκτη αυτού ως γενικό μέτρο των διαφορών μεταξύ δύο κατανομών. Σε αυτό το κεφάλαιο θα παρουσιάσουμε δύο παραμετρικές και δύο μη παραμετρικές προσεγγίσεις για την εκτίμηση του εμβαδού AUC. Οι μη παραμετρικές προσεγγίσεις είναι i) η χρήση του Mann–Whitney στατιστικού(MW) και ii) η προσαρμογή μιας λείας ROCκαμπύλης χρησιμοποιώντας τη μέθοδο λείανσης με πυρήνες και έπειτα εκτιμώντας το εμβαδόν AUCμε ολοκλήρωση(K). Οι παραμετρικές προσεγγίσεις είναι i) να υποθέσουμε ότι οι τιμές του δείκτη και για τα παθολογικά αλλά και για τα υγιή περιστατικά ακολουθούν την κανονική κατανομή και έπειτα να υπολογίσουμε το AUC εμβαδόν χρησιμοποιώντας τυπικές παραμετρικές μεθόδους(N)καιi) να εφαρμόσουμε ένα μετασχηματισμό τύπου Box–Cox και έπειτα έχοντας αποκτήσει τον κατάλληλο μετασχηματισμό να χρησιμοποιήσουμε τη θεωρία κανονικότητας(NT). Για τη μέθοδο των πυρήνων εμείς χρησιμοποιούμε τον τυπικό κανονικό πυρήνα και θεωρούμε δύο διαφορετικές προσεγγίσεις προκειμένου να θέσουμε την παράμετρο. Επιπλέον εξετάζουμε τη χρήση μετασχηματισμού δεδομένων πριν εφαρμόσουμε τον πυρήνα.

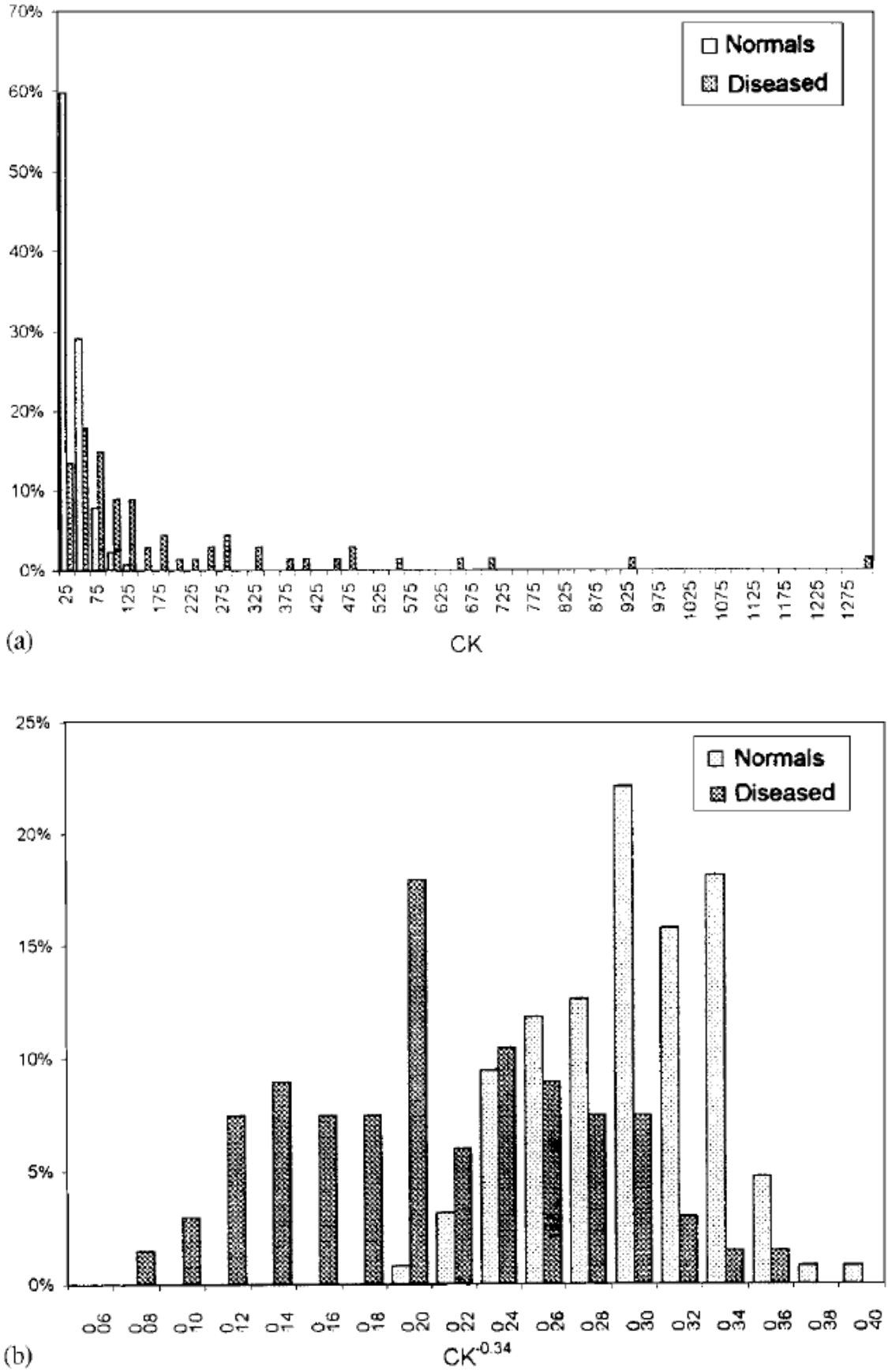
Διαφορετικές μέθοδοι υπολογισμού θα παρέχουν όπως είναι φυσικό διαφορετικές τιμές για το εμβαδόν AUC. Οι Goddard και Hinberg εξετάζουν δεδομένα για διάφορους δείκτες πάνω στον καρκίνο του προστάτη. Αυτοί υπολόγισαν το AUCεμβαδόν χρησιμοποιώντας μεθόδους MWκαιN καθώς επίσης και logμετασχηματισμούς δεδομένων και έπειτα εφαρμόζοντας την Νμία ειδική εκδοχή της NT. Αυτοί ανακάλυψαν ότι οι μέθοδοι MWκαιN, σε ορισμένες περιπτώσεις διαφέρουν σημαντικά, ενώ εφαρμόζοντας τη N μετά από logμετασχηματισμό πήραν αποτελέσματα κοντά σε αυτά της MW. Ένα επιπλέον παράδειγμα παρουσιάζεται στην επόμενη παράγραφο 3.2. Καθώς λοιπόν οι διαφορετικές μέθοδοι υπολογισμού μπορεί να παρέχουν ένα εύρος τιμών για το εμβαδόν AUC, πάνω στο ίδιο σύνολο δεδομένων, χρειάστηκε να εξεταστούν πάλι οι ιδιότητές τους προκειμένου να καταλήξουμε ως προς ποια προσέγγιση πρέπει να προτιμάται.

Ο υπολογισμός του AUCεμβαδού με τη χρήση της MWπροσέγγισης ακολουθεί φυσικά την εκτίμηση της ROCκαμπύλης ως μία συνάρτηση βήματος βασισμένη σε εμπειρικές αθροιστικές συναρτήσεις κατανομών. Οι άλλες προσεγγίσεις αποκτώνται με τον υπολογισμό των ROCκαμπυλών ως λεία συνάρτηση και την εκτίμηση του AUCεμβαδού ως την περιοχή κάτω από αυτή τη λεία καμπύλη. Έχει αποδειχθεί ότι υπάρχουν αρκετά πλεονεκτήματα στη χρήση μιας λείας ROCκαμπύλης στις εκτιμήσεις μας. Όταν αυτή χρησιμοποιείται είναι φυσικό να χρησιμοποιείται και ο αντίστοιχος AUCεκτιμητής. Αν και ο MWεκτιμητής είναι γνωστός ως αμερόληπτος μπορούμε να ελπίζουμε ότι κάποιες από τις εναλλακτικές προσεγγίσεις μπορεί να είναι αρκετά αμερόληπτες και να παρέχουν κάποιο κέρδος στην αποτελεσματικότητα καθώς μετρούμε με τη ρίζα του μέσου τετραγωνικού σφάλματος. Συνεπώς οι προσεγγίσεις MW, K,N και NT συγκρίνονται βάση της προτίμησης και της ρίζας του μέσου τετραγωνικού σφάλματος (RMSE) μέσω μιας εκτεταμένης μελέτης

προσομοίωσης η οποία αναλύεται στην παράγραφο 3.4. Μια άλλη προσέγγιση, η (RC) από τον Metz. Υποθέτει ότι κάποιος άγνωστος μονότονος μετασχηματισμός των τιμών του δείκτη και για τους ασθενείς αλλά και για τους υγιείς πληθυσμούς έχει σαν αποτέλεσμα κανονικά κατανεμημένες μεταβλητές. Αυτοί παρέχουν ένα σύνθετο αλγόριθμο ο οποίος αρχικά διαχωρίζει τα συνεχή δεδομένα σε Ικατηγορίες(με το Ιαρκετά μεγάλο) και στη συνέχεια αριθμητικά βελτιστοποιούν μία συνάρτηση πιθανότητας πάνω σε I+1 παραμέτρους. Ένα υπολογιστικό πρόγραμμα (ROCKIT)το οποίο διεξάγει αυτή τη διαδικασία μπορεί να αποκτηθεί στην διεύθυνση http://www-radiology.uchicago.edu/krl/toppage11.htm#software. Αυτή η μέθοδος συγκρίνεται με τις υπόλοιπες μεθόδους στη παράγραφο 3.4.4. Το παράδειγμα επανεξετάζεται στην παράγραφο 3.5 ενώ η παράγραφος 3.6 περιλαμβάνει μία συμπερασματολογική συζήτηση.

3.2 ΠΑΡΑΔΕΙΓΜΑ:ΔΕΔΟΜΕΝΑ ΠΑΝΩ ΣΤΗ DUCHENNE ΔΥΣΤΡΟΦΙΑ ΜΥΩΝ

Η Duchenne muscular dystrophy (DMD) είναι μία εξελικτική κληρονομική διαταραχή που περνά από τη μητέρα στο παιδί. Με την απουσία μιας αποτελεσματικής θεραπείας, ο προληπτικός έλεγχος των γυναικών ως ενδεχόμενες φορείς της ασθένειας είναι υψίστης σημασίας. Ο Percy επεξεργάζεται δεδομένα συγκεντρωμένα σε τέσσερις διαφορετικούς δείκτες σε ένα μέρος ενός προγράμματος για την ανάπτυξη μιας αποτελεσματικής μεθόδου ελέγχου. Ολοκληρωμένα δεδομένα είναι διαθέσιμα από 127 δείγματα ορού από ελέγχους υγιών γυναικών(Y) και 67 αντίστοιχα δείγματα από φορείς(X). Εμείς θεωρούμε για λόγους απεικόνισης μόνο τις μετρήσεις στον ορό αίματος κρεατινικής κινάσας (CK). Η εικόνα 3.1.απαρουσιάζει ένα ιστόγραμμα των δεδομένων αυτών. Η έλλειψη κανονικότητας στα δεδομένα είναι εμφανής. Οι τιμές της CKδοσμένες στη δύναμη -0.34 φαίνονται στην εικόνα 3.1.b. Αυτή η δύναμη προέκυψε από την εφαρμογή της μεθόδου Box-Coxεκτίμησης μετασχηματισμών (βλέπε παράγραφο 3.3.2). Είναι φανερό ότι τα μετασχηματισμένα δεδομένα εμφανίζονται περισσότερο συμμετρικά κατανεμημένα. Εφαρμόζοντας τις μεθόδους εκτίμησης του AUCεμβαδού, MW, N, NT, K1, K1T, K2, K2T και RC στα δεδομένα της CK παίρνουμε σαν αποτέλεσμα τις τιμές 0.870, 0.740, 0.873, 0.787, 0.852, 0.843, 0.857 και 0.875 αντίστοιχα. Οι K1 και K2 δηλώνουν την μέθοδο των πυρήνων χρησιμοποιώντας δύο διαφορετικούς υπολογισμούς παραμέτρων. Όταν αυτές οι μέθοδοι πυρήνων εφαρμόζονται στα μετασχηματισμένα CK δεδομένα αναφέρονται ωςK1T και K2T, αντίστοιχα. Υπάρχουν μεγάλες διαφορές μεταξύ των δύο αυτών μεθόδων. Προκειμένου να κατανοήσουμε καλύτερα τις διαφορές αυτές και να βοηθηθούμε στο να επιλέξουμε σωστά μεταξύ αυτών διεξαγάγαμε μία μελέτη προσομοίωσης η οποία παρουσιάζεται στην παράγραφο 3.4.



Εικόνα 3.1 Ιστογράμματα της CK:a)Πριν τον μετασχηματισμό,b) Μετά τον μετασχηματισμό

3.3 ΕΚΤΙΜΗΣΗ AUCΕΜΒΑΔΟΥ

Έγινε η υπόθεση ότι τα αποτελέσματα από διαγνωστικά τεστ, $x_1, \ldots, x_m$ και $y_1, \ldots, y_n$ είναι διαθέσιμα και από τον ασθενή και από τον υγιή πληθυσμό έχοντας αθροιστικές συναρτήσεις κατανομής F και G, αντίστοιχα. Τότε στο όριο c,$q = 1 - F(c)$ και $p = 1 - G(c)$. Η θεωρητική ROC καμπύλη είναι μία γραφική παράσταση των σημείων (1 − G(c),1 − F(c)) για όλες τις πιθανές τιμές του c, ή ισοδύναμα μία γραφική παράσταση των σημείων (p,q) όπου το p κυμαίνεται μεταξύ του 0 και του 1

$$q = 1 - F(G^{-1}(1-p)) \qquad (1)$$

Το εμβαδόν AUCαντιπροσωπεύει την περιοχή κάτω από αυτή τη καμπύλη. Διαφορετικές εκτιμήσεις για το AUCεμβαδόν προκύπτουν από διαφορετικές προσεγγίσεις για τον υπολογισμό της ROCκαμπύλης. Παρακάτω παρουσιάζουμε διαφορετικούς εκτιμητές για το AUCεμβαδόν.

3.3.1. ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ

Η απλούστερη μη παραμετρική μέθοδος εκτίμησης της ROCκαμπύλης περιλαμβάνει αντικατάσταση του Fκαι του G στην σχέση (1) από τις εμπειρικές τους συναρτήσεις κατανομής $\hat{F}_m(t)$και $\hat{G}_n(t)$, αντίστοιχα. Η εμπειρική αθροιστική συνάρτηση κατανομής ορίζεται ,για κάθε δοσμένη τιμή t, ως το παρατηρούμενο ποσοστό των δειγματικών τιμών που είναι μικρότερες ή ίσες του t. Η ROCκαμπύλη που προκύπτει είναι μία συνάρτηση αυξανόμενου βήματος στο μοναδιαίο τετράγωνο, και η μορφή της μπορεί να είναι αρκετάjagged(ακανόνιστη). Το εμβαδόν κάτω από αυτήν την καμπύλη ισοδυναμεί με το Mann–Whitney U στατιστικό και παρέχει έναν απροκατάληπτο μη παραμετρικό εκτιμητή για το AUCεμβαδόν. Θα δηλώνουμε από εδώ και στο εξής τον εκτιμητή αυτόν ως MW.

Διάφοροι συγγραφείς συζητούν τη βελτίωση της μη παραμετρικής προσέγγισης προκειμένου να παρέχουν μία λεία ROC καμπύλη με τη χρήση της μεθόδου των πυρήνων. Σύμφωνα με τον Zou επιλέγουμε τον Γκαουσιανό πυρήνα και υπολογίζουμε τη συνάρτηση πυκνότητας πιθανότητας (PDF), *f(t)=F'(t)*

$$\hat{f}(t) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{h_x}\varphi\left(\frac{t-x_i}{h_x}\right) \qquad (2)$$

όπου φ είναι η συνάρτηση πυκνότητας πιθανότητας της τυπικής κανονικής κατανομής. Όσον αφορά τώρα το h_x, αποτελεί την **παράμετρο λείανσης** και εκφράζει την τυπική απόκλιση σύμφωνα με την οποία εκτείνονται οι τιμές του πυρήνα γύρω από την τιμή x_i .Τα βήματα για την κατάλληλη επιλογή της αναλύονται παρακάτω. Η συνάρτηση πυκνότητας πιθανότητας $g(t) = G'(t)$ υπολογίζεται ανάλογα. Από την (2) προκύπτει με ολοκλήρωση ότι ο εκτιμητής της Feίναι

$$\hat{F}(t) = \frac{1}{m}\sum_{i=1}^{m}\Phi\left(\frac{t-x_i}{h_x}\right) \qquad (3)$$

Όπου Φ είναι η συνάρτηση της τυπικής αθροιστικής κανονικής κατανομής. Η $\hat{G}$ υπολογίζεται ανάλογα.

Χρησιμοποιώντας τώρα τις $\hat{G}$ και $\hat{F}$ στην (1) προκύπτει ένας λείος εκτιμητής της καμπύλης ROC. Τέλος, ο Zou αποδεικνύει ότι η εκτίμηση του εμβαδού κάτω από την προσαρμοσμένη καμπύλη με τη μέθοδο των πυρήνων δίνεται από τον τύπο:

$$K = \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}\Phi\left(\frac{x_i - y_i}{\sqrt{\left(h_x^2 + h_y^2\right)}}\right) \quad (4)$$

Όπου ο Zou χρησιμοποιεί για την παράμετρο λείανσης την τιμή

$$h_x = 0.9 \cdot \min(s_x, iqr_x/1.34) \cdot m^{-1.5} (5)$$

όπου s_xκαι iqr_xείναι η τυπική απόκλιση και το quartile του εύρους, αντίστοιχα των αποτελεσμάτων του m τεστ πάνω στο δείγμα των ασθενών. Το h_yυπολογίζεται ανάλογα. Αυτή η επιλογή των παραμέτρων έχει διαπιστωθεί πρόσφατα ότι είναι αποτελεσματική στον υπολογισμό της overlapping coefficient, η οποία αποτελεί ένα εναλλακτικό μέτρο της διαφοράς ανάμεσα στις δύο κατανομές. Ο Silverman επίσης ισχυρίζεται ότι αυτό συμπεριφέρεται πολύ καλά για ένα ευρύ φάσμα πυκνοτήτων. Να διευκρινίσουμε, εδώ ότι ο υπολογισμός του AUCβρέθηκε χρησιμοποιώντας τη σχέση (5) στη σχέση (4), ως K1 μέθοδος.

Οι Lloyd και Yong περιγράφουν μία πιο σύνθετη επιλογή για την παράμετρο λείανσης κατά την διαδικασία λείανσης ROCκαμπυλών. Η μέθοδός τους είναι μία επέκταση της ”δύο σταδίων plug-in” διαδικασίας των Wand και Jones. Να υπενθυμίσουμε ότι ο υπολογισμός του AUCεμβαδού αποκτήθηκε από τον αλγόριθμό τους, ως K2 μέθοδος.

Κάποιες φορές, πριν εφαρμοστεί ο πυρήνας, τα δεδομένα μετασχηματίζονται προκειμένου να γίνουν πιο συμμετρικά. Επομένως, εξετάζουμε την χρήση του μετασχηματισμού Box–Cox πριν εφαρμόσουμε τη μέθοδο των πυρήνων, και δηλώνουμε τα αποτελέσματα υπολογισμού του AUCεμβαδού ως K1T και K2T.

3.3.2 ΠΑΡΑΜΕΤΡΙΚΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ

Μία απλή παραμετρική προσέγγιση είναι να γίνει η υπόθεση ότι ο *X* και ο *Y*είναι ανεξάρτητες κανονικές μεταβλητές που ακολουθούν κανονική κατανομή. Έτσι,

$$X \sim \mathrm{N}(\mu_X, \sigma_X^2) \quad \kappa\alpha\iota \quad Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$$

Συνεπώς, έχουμε:

$$q = \Phi\left(\frac{\mu_X - c}{\sigma_X}\right), \quad p = \Phi\left(\frac{\mu_Y - c}{\sigma_Y}\right)$$

Τέλος η ROCκαμπύλη υπολογίζεται από το γράφημα των ρκαι qγια όλες τις πιθανές τιμές της μεταβλητής c,με τις άγνωστες παραμέτρους να αντικαθίστανται από τις συνήθεις εκτιμήσεις τους. Επιπλέον

$$AUC = \Phi\left(\frac{\mu_X - \mu_Y}{\sqrt{\sigma_X^2 + \sigma_Y^2}}\right) \quad (6)$$

η οποία μπορεί να υπολογιστεί αντικαθιστώντας τους μέσους και τις τυπικές αποκλίσεις στην παραπάνω σχέση. Θα δηλώνουμε στο εξής τον εκτιμητή αυτόν ως N. Κάποιοι εναλλακτικοί εκτιμητές προτείνονται από τους Reiser και Guttman κάτω από υποθέσεις κανονικότητας, οι οποίοι όμως αποδεικνύονται κατώτεροι.

Στην περίπτωση που η ανάλυση δεδομένων υποδεικνύει ότι η υπόθεση κανονικότητας είναι αστήρικτη, προτείνεται συχνά ένας adhoc μετασχηματισμός όπως ο λογάριθμος. Μέχρι πρόσφατα διάφοροι συγγραφείς έχουν προτείνει την αξιοποίηση των δεδομένων για την προσαρμογή ενός δυναμικού μετασχηματισμού του Box-Coxμοντέλου και έπειτα την εφαρμογή του N εκτιμητή στα μετασχηματισμένα δεδομένα. Σε αυτήν την κατάσταση, θεωρείται ότι οι X και Y μετρήσεις του διαγνωστικού δείκτη πάνω στα παθολογικά και υγιή περιστατικά, δεν ακολουθούν κανονική κατανομή αλλά οι μετασχηματισμένες εκδοχές αυτών. Πιο συγκεκριμένα ορίζουμε

$$X^{(\lambda)} = \begin{cases} \dfrac{X^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \log(X) & \lambda = 0 \end{cases} \quad (7)$$

όπου υποτέθηκε ότι:

$$X^{(\lambda)} \sim \mathrm{N}(\mu_1, \sigma_1^2) \quad \kappa\alpha\iota \quad Y^{(\lambda)} \sim \mathrm{N}(\mu_2, \sigma_2^2)$$

Παρατηρούμε ότι:

$$AUC = P(X > Y) = P(X^{(\lambda)} > Y^{(\lambda)}) \quad (8)$$

Με βάση τις παρατηρήσεις των υγιών και των ασθενών, μπορεί να κατασκευαστεί κατάλληλη συνάρτηση πιθανοφάνειας με τη μεγιστοποίηση της οποίας προκύπτει το $\hat{\lambda}$, η μέγιστη πιθανοφάνεια εκτίμησης του λ. Από την (8) προκύπτει ότι το εμβαδόνAUCμπορεί να εκτιμηθεί από τον μετασχηματισμό των δεδομένων χρησιμοποιώντας το $\hat{\lambda}$ και έπειτα εφαρμόζοντας τη μέθοδο εκτίμησης N (σχέση 6). Δηλώνουμε τον εκτιμητή αυτόν ως NT.

3.3.3 ΠΡΟΣΕΓΓΙΣΗ ΜΕ ΕΚΘΕΤΙΚΟ ΜΟΝΤΕΛΟ ΚΑΙ ΜΟΝΤΕΛΟ PARETO

Για το εκθετικό μοντέλο ισχύει, $\hat{\theta} = \dfrac{E}{E+1}$ όπου E η παράμετρος ασυμμετρίας της προσαρμοσμένης καμπύλης ROC.

Ενώ για το μοντέλο Paretoμε παραμέτρους k_1, k_2, θ_2 είναι:

$$\widehat{\theta}_p = \frac{k_1}{k_1 + k_2} F_1^2(k_2, 1; k_1 + k_2 + 1; 1 - \theta_2)$$

όπου F_1^2 η υπεργεωμετρική συνάρτηση.

Σημειώνεται εδώ ότι για όλα τα παραπάνω μοντέλα μπορεί να χρησιμοποιηθεί ο τύπος της διακύμανσης που προαναφέρθηκε:

$$Var\left(\hat{\theta}\right) = \frac{\hat{\theta}\left(1-\hat{\theta}\right) + (m-1)\left(Q_1 - \hat{\theta}^2\right) + (n-1)\left(Q_2 - \hat{\theta}^2\right)}{mn} \qquad (2)$$

Επιπλέον κάνοντας χρήση της ασυμπτωτικής κανονικότητας είναι δυνατόν να ελεγχθούν οι υποθέσεις που αφορούν το εμβαδόν κάτω από την καμπύλη ROC. Συγκεκριμένα για τον έλεγχο της σημαντικότητας ενός ιατρικού διαγνωστικού ελέγχου εξετάζεται η ισχύς της
$H_0 : \theta = 0.5$ έναντι της $H_A : \theta \neq 0.5$. Για το λόγο ότι θεωρούμε ότι οι υγιείς έχουν μικρότερες μετρήσεις από τους ασθενείς.

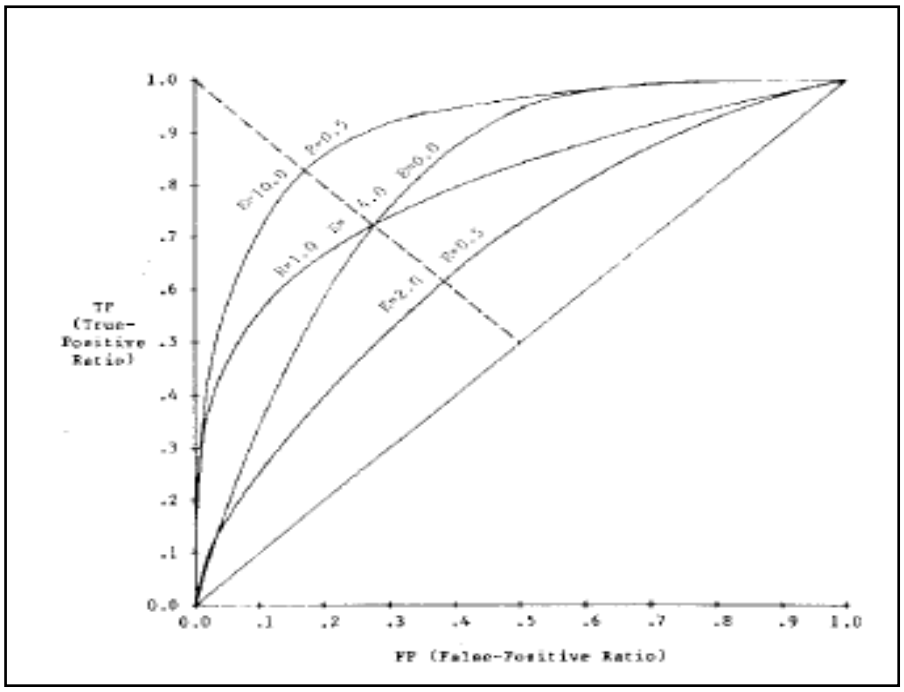
3.3.4 ΆΛΛΑ ΠΑΡΑΜΕΤΡΙΚΑ ΜΟΝΤΕΛΑ ΠΡΟΣΑΡΜΟΓΗΣ ΛΕΙΑΣ ΚΑΜΠΥΛΗΣ ROC

Το δικανονικό μοντέλο αν και είναι αυτό που χρησιμοποιείται συχνότερα στις εφαρμογές των καμπύλων ROC δεν έχει απλή κλειστή έκφραση που να περιγράφει την καμπύλη (δεδομένου ότι η παραμετρική της μορφή είναι $TP=1-F_1(F_0^{-1}(1-FP))$, όταν οι F_0, F_1 είναι κανονικές η έκφραση αυτή γίνεται περίπλοκη). Για το λόγο αυτό έχουν αναπτυχθεί μοντέλα που δίνουν απλή κλειστή έκφραση για την καμπύλη και έχει δειχθεί με διαδικασίες MonteCarlo ότι προσαρμόζονται στα δεδομένα εξίσου ικανοποιητικά με το δικανονικό μοντέλο.

Το μοντέλο που εξετάζουμε προτάθηκε από τον England (1988) και στηρίζεται στην παρατήρηση του Egan (1975) ότι όταν οι κατανομές υγιών – ασθενών είναι εκθετικές με $\mu_0<\mu_1$ η καμπύλη ROC υπακούει σε μια εξίσωση της μορφής $TP=FP^k$ (όπου $k=\mu_0/\mu_1$), ενώ για $\mu_0>\mu_1$ ισχύει $TP=1-(1-FP)^k$. Συνδυάζοντας τις δύο προηγούμενες εξισώσεις το βέλτιστο μοντέλο –που προέκυψε πειραματικά– είναι το:

$$TP = R \cdot FP^{1/E} + (1 - R)(1 - (1 - FP)^{E}) \qquad (6)$$

όπου R η παράμετρος που σχετίζεται με το σημείοτομής της καμπύλης ROC με την αρνητική διαγώνιο του μοναδιαίου τετραγώνου και E η παράμετρος που σχετίζεται με την ασυμμετρία (skewness) της καμπύλης ROC ως προς την αρνητική διαγώνιο του μοναδιαίου τετραγώνου. Στην εικόνα παρακάτω φαίνονται τα διάφορα σχήματα που παίρνει η καμπύλη ROC ανάλογα με τις τιμές των παραμέτρων E και R.



Εικόνα 1.18: Μεγαλύτερες τιμές για το E συνεπάγονται καμπύλη ROC με μεγάλο εμβαδόν κάτω απ' την καμπύλη, για R=0.5 η καμπύλη είναι συμμετρική ως προς την αρνητική διαγώνιο.

Το λεγόμενο *εκθετικό μοντέλο* που μόλις είδαμε έχει τις ιδιότητες ότι προσαρμόζεται απ' ευθείας στα σημεία (FP, TP) που έχουν υπολογιστεί οπότε είναι ανεξάρτητο από το αν τα αρχικά δεδομένα είναι συνεχή ή διακριτά, ενώ για το εμβαδόν κάτω από την καμπύλη με βάση αυτό το μοντέλο η έκφραση είναι πολύ απλή:

$$\hat{\theta}_E = \int_0^1 \left\{ R \cdot FP^{1/E} + (1-R)(1-(1-FP)^E) \right\} dFP = \frac{E}{E+1}$$

Το εκθετικό μοντέλο έχει επίσης την ιδιότητα ότι το βέλτιστο σημείο απόφασης πάνω στην καμπύλη ROC υπολογίζεται πολύ εύκολα με βάση τις σχέσεις:

$$\frac{dTP}{dFP} = \frac{R}{E} FP^{\frac{1}{E}-1} + (1-R)E(1-FP)^{E-1}$$

$$\frac{dTP}{dFP} = \frac{1-\Pr ev}{Prev} \cdot \frac{B_{TN} - C_{FP}}{B_{TP} - C_{FN}}$$

Οι τελευταίες λύνονται ως προς FP και στη συνέχεια υπολογίζεται το TP που αντιστοιχεί.

Οι παράμετροι του μοντέλου (6) υπολογίζονται με αριθμητικές μεθόδους με βάση την αρχή ελαχίστων τετραγώνων (με τη χρήση επαναληπτικού αλγορίθμου) και σαν

αρχικές τιμές προτείνονται οι E = $\frac{\hat{\vartheta}}{1-\hat{\vartheta}}$ και R=0.5, όπου θ το εμβαδόν κάτω από την καμπύλη ROC υπολογισμένο με τον κανόνα του τραπεζίου.

3.4 ΜΕΛΕΤΗ ΠΡΟΣΟΜΟΙΩΣΗΣ ΓΙΑ ΤΗ ΣΥΓΚΡΙΣΗ ΤΩΝ AUC ΕΚΤΙΜΗΤΩΝ

Συγκρίνονται οι εκτιμητές του AUC, που αναφέρθηκαν και παραπάνω, συγκεκριμένα τους MW, N, NT, K1, K1T, K2 και K2T, με βάση την προτίμηση και της ρίζας του μέσου τετραγωνικού τους σφάλματος (RMSE)) μέσω μιας εκτεταμένης μελέτης προσομοίωσης. Η μέθοδος RCαναπτύσσεται στην παράγραφο 3.4.4.

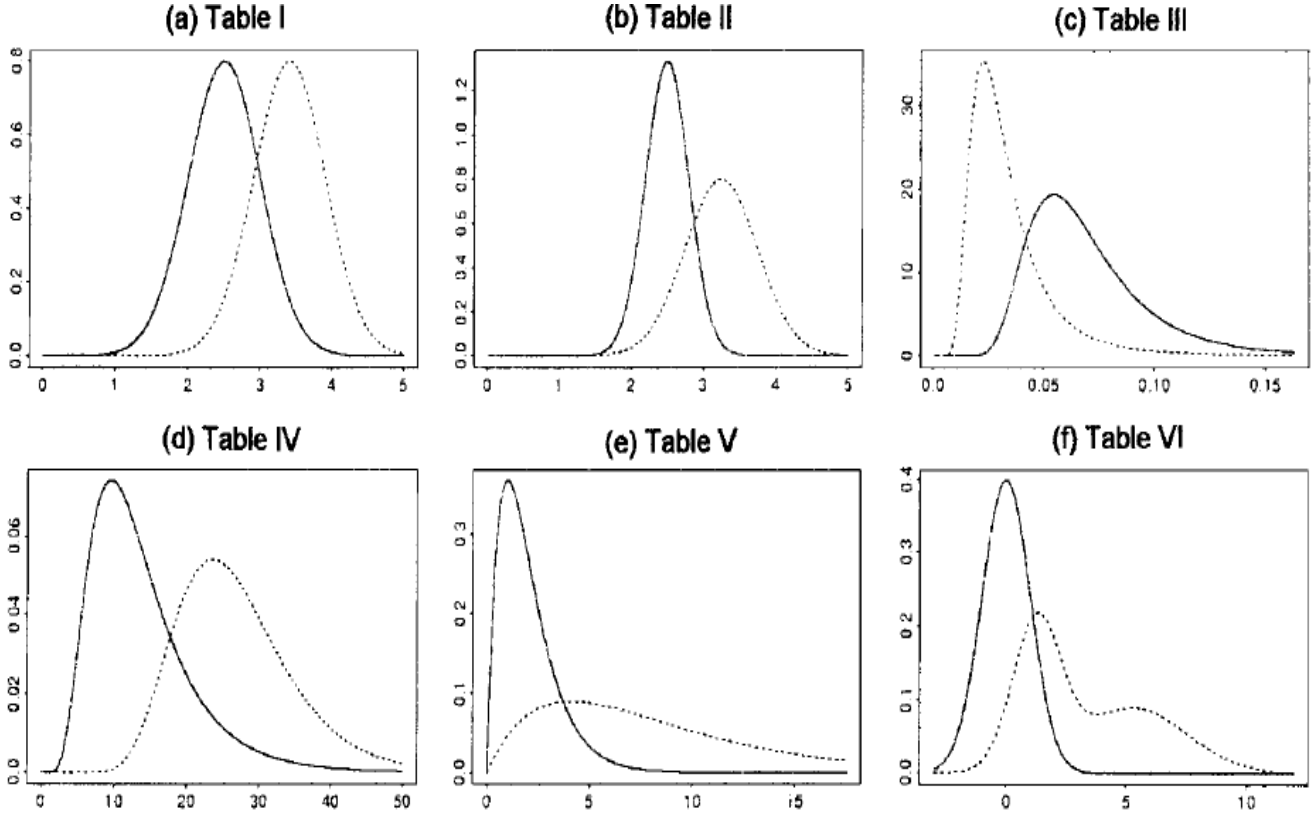
Οι προσομοιώσεις , καλύπτουν ένα ευρύ φάσμα διαφορετικών μορφών κατανομής, ένα δείγμα των οποίων παρουσιάζεται στην εικόνα 3.2. Στην μελέτη αυτή λήφθηκαν υπόψη πολλοί διαφορετικοί συνδυασμοί κατανομών, ο καθένας για διαφορετικές επιλογές του εμβαδού AUC και με μέγεθος δείγματος m=n=20,50,100.
Με στόχο τον υπολογισμό του RMSE και της προτίμησης, υπολογίστηκαν 1000 προσομοιώσεις για το κάθε σενάριο. Τα αποτελέσματα συνοψίζονται στους πίνακες I–IX. Κάθε κελί των πινάκων αυτών παρουσιάζει πρώτα την προτίμηση και έπειτα το RMSE.

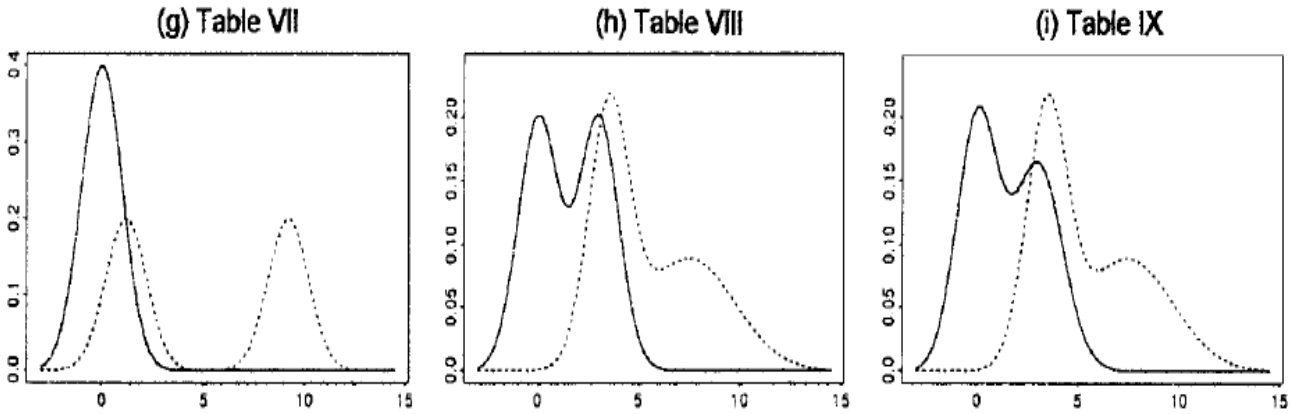
Οι κατανομές στην εικόνα 3.2 σταθεροποιούνται προκειμένου να δώσουν εμβαδόν AUC=0.90. Σε άλλες περιπτώσεις οι κατανομές για τους υγιείς και παθολογικούς πληθυσμούς θα κινούνται πιο κοντά η μία ως προς την άλλη. Το εύρος των περιπτώσεων που απεικονίζεται στην εικόνα 3.2 καλύπτει τις συμμετρικές , ασύμμετρες και bimodalσυνθήκες οποίες εμφανίζονται συχνά στα πραγματικά δεδομένα και είναι παρόμοιες στις μορφές κατανομών που παρουσιάζονται στην μελέτη προσομοίωσης του Hajian-Tilaki. Ένα πρόγραμμα γράφτηκε σε Γκάους προκειμένου να υπολογίσει τους διάφορους εκτιμητές και να διεξάγει τις προσομοιώσεις. Εάν, κατά τη διαδικασία υπολογισμού του NT εκτιμητή κάποιο από τα προσομοιωμένα δεδομένα βρεθεί αρνητικό τότε και οι τιμές των υγιών τμημάτων αλλά και των παθολογικών ανυψώνονται ισοδύναμα ούτως ώστε η ελάχιστη τιμή να είναι ελαφρά μεγαλύτερη του μηδενός, προτού η παράμετρος μετασχηματισμού λ εκτιμηθεί.

3.4.1.ΠΡΟΣΟΜΟΙΩΣΕΙΣ ΜΕ ΚΑΝΟΝΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

Αρχικά θεωρούνται δείγματα από κανονικούς πληθυσμούς ίσων διακυμάνσεων (ΠίνακαςI) και έπειτα άνισων διακυμάνσεων (ΠίνακαςII).Για αυτούς τους πίνακες το μ_χ, η αναμενόμενη τιμή των ασθενών πληθυσμών επιλέγεται έτσι ώστε να ανταποκρίνεται στις τιμές του AUCεμβαδού. Εκτός από τα μεγέθη των ταξινομημένων δειγμάτων(20,100) εμείς θεωρήσαμε και το ενδιάμεσο μέγεθος δείγματος το 50 το οποίο έδωσε ενδιάμεσα αποτελέσματα και δεν κατατάχτηκε στον πίνακα για συντομία. Επίσης διεξήχθησαν υπολογισμοί και για άλλες επιλογές του μέσου και της διασποράς αλλά καθώς δίνουν παρόμοια αποτελέσματα, δεν αναφέρονται. Αν και η κατανομή για τον υγιή πληθυσμό είναι η ίδια για τους πίνακες I και II, οι προσομοιώσεις διεξήχθησαν ανεξάρτητα ούτως ώστε οι δύο πίνακες να

είναι ανεξάρτητοι. Οι πίνακες Ικαι II δείχνουν, όπως αναμενόταν ότι ο N εκτιμητής συνήθως έχει το μικρότερο RMSE όταν τα δεδομένα είναι κανονικά αλλά αυτό του το πλεονέκτημα θεωρείται μικρό αν και αυτός μπορεί να τρέξει μέχρι και 6% υψηλότερα σε σχέση με τον μη παραμετρικό MWεκτιμητή. Μία εξαίρεση αποτελεί η περίπτωση κατά την οποία το AUC=0.7 και n=m=20 για τα οποία οι μέθοδοι των πυρήνων έχουν το μικρότερο RMSE. OMWεκτιμητής είναι γνωστός ως απροκατάληπτος και οι μικρές εκτιμήσεις προτίμησης που εμφανίζονται στους πίνακες είναι αποτέλεσμα του θορύβου προσομοίωσης. Αυτή η παρατήρηση επίσης εφαρμόζεται σε όλες τις προσομοιώσεις που συζητούνται παρακάτω. Καθώς οι παρατηρούμενες προτιμήσεις για της εκτιμήσεις του N είναι της ίδιας διάταξης με αυτές του MWεκτιμητή, η μέθοδος N μπορεί να θεωρηθεί αρκετά απροκατάληπτη για κανονικά δεδομένα. Η NT διαδικασία δίνει αποτελέσματα ουσιαστικά ισοδύναμα με αυτά της N διαδικασίας. Στην πραγματικότητα, ο αρχικός μετασχηματισμός των δεδομένων δεν βελτιώνει ούτε τις διαδικασίες των πυρήνων. Οι διαδικασίες των πυρήνων είναι κατά κάποιο τρόπο αρνητικά προκατειλημμένες. Οι K1 και K2 διαδικασίες συμπεριφέρονται παρόμοια αν και η K2 εμφανίζεται να είναι με συνέπεια, ελαφρώς λιγότερο προκατειλημμένη. Η περισσότερο σύνθετη K2 διαδικασία είναι καθαρά καλύτερη από την K1 μόνο για AUC=0.9 και n=m=100. Εκτός από την περίπτωση κατά την οποία AUC=0.7 και n=m=20 οι μέθοδοι των πυρήνων δεν βελτιώνονται στον MWεκτιμητή και για μεγαλύτερο AUCσυμπεριφέρονται χειρότερα. Τέλος τα αποτελέσματα για ίσες ή άνισες διασπορές είναι αρκετά παρόμοια.





Εικόνα 3.1Κατανομές που χρησιμοποιήθηκαν στη μελέτη προσομοίωσης με τιμή εμβαδού AUC=0.9, υποθέσεις ίδιες με αυτές των πινάκων I-IX.

Methods	$n=m=20$			$n=m=100$		
	AUC = 0.7	AUC = 0.8	AUC = 0.9	AUC = 0.7	AUC = 0.8	AUC = 0.9
MW	−0.001 0.081	−0.000 0.068	0.001 0.048	−0.001 0.036	−0.001 0.031	−0.001 0.021
N	−0.002 0.078	−0.003 0.066	−0.002 0.047	−0.000 0.035	−0.001 0.030	−0.001 0.021
NT	−0.000 0.078	−0.000 0.066	−0.000 0.046	−0.000 0.035	−0.001 0.030	−0.001 0.020
K1	−0.017 0.076	−0.021 0.068	−0.019 0.053	−0.010 0.036	−0.014 0.033	−0.014 0.026
K1T	−0.016 0.076	−0.021 0.068	−0.019 0.053	−0.010 0.036	−0.014 0.033	−0.014 0.026
K2	−0.015 0.076	−0.018 0.068	−0.016 0.052	−0.006 0.036	−0.008 0.031	−0.007 0.023
K2T	−0.014 0.076	−0.018 0.068	−0.017 0.052	−0.006 0.036	−0.008 0.031	−0.007 0.023

Πίνακας I. Προσομοιωμένη κλίση και RMSE για τους εκτιμητές του AUC εμβαδού: Y~N(2.5,0.25), X~N(μ_X,0.25)

Table II. Simulated bias and RMSE for AUC estimators: $Y \sim \mathrm{N}(2.5, 0.09)$, $X \sim \mathrm{N}(\mu_X, 0.25)$.

Methods	$n=m=20$			$n=m=100$		
	AUC = 0.7	AUC = 0.8	AUC = 0.9	AUC = 0.7	AUC = 0.8	AUC = 0.9
MW	0.005 0.081	−0.001 0.073	−0.002 0.052	−0.000 0.039	0.000 0.031	−0.000 0.023
N	0.002 0.078	−0.002 0.070	−0.005 0.049	−0.000 0.037	−0.000 0.030	−0.000 0.022
NT	0.005 0.080	0.001 0.070	−0.002 0.048	−0.000 0.037	0.000 0.030	−0.000 0.022
K1	−0.010 0.075	−0.021 0.072	−0.023 0.057	−0.010 0.038	−0.012 0.033	−0.013 0.027
K1T	−0.010 0.076	−0.020 0.072	−0.023 0.057	−0.010 0.038	−0.012 0.033	−0.012 0.027
K2	−0.008 0.076	−0.018 0.072	−0.020 0.056	−0.005 0.038	−0.006 0.031	−0.006 0.024
K2T	−0.008 0.076	−0.018 0.072	−0.020 0.056	−0.005 0.038	−0.006 0.032	−0.006 0.024

Πίνακας II. Προσομοιωμένη κλίση και RMSE για τους εκτιμητές του AUC εμβαδού: Y~N(2.5,0.09), X~N(μ_X,0.25)

Table III. Simulated bias and RMSE for AUC estimators: $Y^{1/3} \sim \mathrm{N}(2.5, 0.09)$, $X^{1/3} \sim \mathrm{N}(\mu_X, 0.25)$.

Methods	$n=m=20$						$n=m=100$					
	AUC = 0.7		AUC = 0.8		AUC = 0.9		AUC = 0.7		AUC = 0.8		AUC = 0.9	
MW	−0.000	0.083	0.000	0.073	0.001	0.049	0.000	0.038	−0.001	0.031	−0.000	0.021
N	−0.064	0.109	−0.056	0.109	−0.045	0.082	−0.090	0.103	−0.075	0.091	−0.053	0.064
NT	−0.000	0.081	0.002	0.070	0.000	0.047	0.000	0.038	−0.000	0.030	−0.000	0.020
K1	−0.025	0.083	−0.026	0.077	−0.025	0.057	−0.016	0.041	−0.019	0.037	−0.016	0.027
K1T	−0.015	0.078	−0.018	0.072	−0.019	0.053	−0.009	0.038	−0.014	0.034	−0.013	0.025
K2	−0.017	0.081	−0.017	0.074	−0.016	0.052	−0.005	0.039	−0.007	0.032	−0.005	0.022
K2T	−0.013	0.078	−0.016	0.072	−0.016	0.052	−0.004	0.038	−0.007	0.032	−0.006	0.023

Πίνακας ΙΙΙ. Προσομοιωμένη κλίση και RMSE για τους εκτιμητές του AUC εμβαδού: $Y^{1/3} \sim \mathrm{N}(2.5, 0.09)$, $X^{1/3} \sim \mathrm{N}(\mu_X, 0.25)$

3.4.2 ΠΡΟΣΟΜΟΙΩΣΕΙΣ ΜΕ ΑΣΥΜΜΕΤΡΕΣ ΚΑΤΑΝΟΜΕΣ

Δημιουργήθηκαν ασύμμετρα δεδομένα αρχικά παράγοντας κανονικές ποικιλίες τις οποίες υψώθηκαν στη δύναμη του (-3). Αυτή η δύναμη επιλέχθηκε προκειμένου να ανταποκρίνεται στη δύναμη του μετασχηματισμού($-0.34 \approx -1/3$) η οποία βρέθηκε ότι είναι καλύτερη για τα CKδεδομένα. Η σχετική συνάρτηση πυκνότητας πιθανότητας φαίνεται στην εικόνα 3.2.c. Η κλίση και το RMSEδίνονται στον πίνακα ΙΙΙ. Αναμενόμενα λοιπόν, η N διαδικασία η οποία προϋποθέτει κανονικότητα, συμπεριφέρεται χειρότερα εδώ, όσων αφορά και την κλίση αλλά και το RMSE. Η κλίση δεν μειώνεται με ένα μεγαλύτερο μέγεθος δείγματος και είναι μεγαλύτερη για AUC=0.7. Η NT παρέχει τα καλύτερα αποτελέσματα καθώς είναι αρκετά απροκατάληπτη και διατηρεί σταθερά το χαμηλότερο RMSE με εξαίρεση την περίπτωση που έχουμε AUC=0.7 και n=m=20 όπου οι K1T και K2T έχουν χαμηλότερο RMSE. Πρέπει να σημειωθεί ότι η μη παραμετρική μέθοδος MW δίνει παρόμοιες αλλά παρόλα αυτά ελαφρά υψηλότερες τιμές RMSE με μία διαφορά του 4% για ένα μέγεθος δείγματος 20. Οι διαδικασίες των πυρήνων είναι επίσης επηρεασμένες, αλλά σε αντίθεση με τα αποτελέσματα στην παράγραφο 3.4.1, εδώ ο μετασχηματισμός των δεδομένων πριν την εφαρμογή της μεθόδου λείανσης με τους πυρήνες γενικά βελτιώνει και τις τιμές του RMSEκαι της κλίσης. Αυτή η βελτίωση είναι πιο ουσιαστική για την K1. Καθώς τα δεδομένα μετασχηματίζονται υπάρχει μικρή διαφορά ανάμεσα στις δύο ,διαφορετικού εύρους ζώνης, μεθόδους επιλογής. Για την τιμή AUC=0.7 και n=m=20, οι K1T καιK2T βελτιώνονται κάπως και στην NTμέθοδο και στην MW, με βάση το RMSE. Παρόλα αυτά αυτό το πλεονέκτημα εξασθενεί καθώς αυξάνεται η τιμή του AUCεμβαδού, και για την τιμή AUC=0,9 αυτές είναι σημαντικά χειρότερες από την NTμέθοδο και κάπως χειρότερες από την MW.

Επιπλέον εξετάσθηκαν τα ασύμμετρα δεδομένα στον πίνακα IV ο οποίος συνοψίζει τα αποτελέσματα για την κανονική λογαριθμική κατανομή. Τα δεδομένα προέκυψαν, αρχικά προσομοιώνοντας κανονικές κατανομές και έπειτα αντιστοιχίζοντας τις τιμές αυτές. Οι πίνακες ΙΙΙ και IV είναι κατασκευασμένοι ανεξάρτητα. Εξετάζοντας τον πίνακα IVοδηγούμαστε στα ίδια συμπεράσματα με αυτά που δίνονται στον πίνακα ΙΙΙ. Εξετάσθηκαν άλλοι μετασχηματισμοί στην οικογένεια των δυνάμεων και

προέκυψε το συμπέρασμα ότι αν ο μετασχηματισμός είχε σαν αποτέλεσμα κατανομές αρκετά ασύμμετρες(skewed), τα αποτελέσματα παραλληλίζονται με εκείνα των πινάκων IIIκαι IV ενώ για μετασχηματισμούς που έδωσαν σχεδόν συμμετρικές (bellshaped)κατανομές, τα αποτελέσματα ήταν παρόμοια με αυτά των πινάκων Ικαι II.

Καθώς οι κατανομές που χρησιμοποιήθηκαν προέκυψαν από την εφαρμογή μετασχηματισμού δύναμης σε κανονικά δεδομένα και αυτές εντάσσονται στην Box-Coxοικογένεια μετασχηματισμών, δεν μας εκπλήσσει μάλλον το γεγονός ότι η NTμέθοδος έχει καλύτερα αποτελέσματα από ότι η μέθοδος των πυρήνων. Στον πίνακα Vεμείς θεωρούμε μία προσομοίωση η οποία δεν ανήκει σε αυτή την οικογένεια: την γάμμα κατανομή. Ορίζοντας την γάμμα συνάρτηση πυκνότητας ως$e^{-Z/\lambda}Z^{p-1}/(\lambda^p\Gamma(p))$, με δείκτες Χκαι Y στις παραμέτρους προκειμένου να ξεχωρίζουμε τους δύο πληθυσμούς, δημιουργήσαμε δεδομένα για $p_X = p_Y = 2$, $\lambda_Y = 0.5\ και\ 1.0\ και\ \lambda_X$ επιλεγμένο ώστε να δώσει AUC=0.9 . Η γραφική παράσταση της κατανομής που παίρνουμε σαν αποτέλεσμα, φαίνεται στην εικόνα 2(e).

Τα συμπεράσματά είναι παρόμοια με αυτά που βρέθηκαν στους πίνακες III και IV για κατανομές στην ίδια οικογένεια δυνάμεων. Η μέθοδος N συμπεριφέρεται χειρότερα και με βάση το RMSEαλλά και με βάση την κλίση, ενώ η NT μέθοδος συμπεριφέρεται καλύτερα από όλες. Στις ασύμμετρες κατανομές η N μέθοδος κλίνει προς τα κάτω. Από την άλλη όμως η NT μέθοδος παραμένει αποτελεσματικά ανεπηρέαστη και βελτιώνεται συγκριτικά με την MWμέθοδο, όσων αφορά το RMSE, σε περιπτώσεις μικρού μεγέθους δείγματος, μέχρι και 8%. Η K1 συμπεριφέρεται το ίδιο άσχημα με την N μέθοδο για μικρά μεγέθη δείγματος. Η K2 είναι καλύτερη από τη K1 και όσων αφορά το RMSEαλλά και την κλίση. Μετασχηματίζοντας τα δεδομένα πριν εφαρμόσουμε τη μέθοδο των πυρήνων, η K1 βελτιώνεται σημαντικά ενώ η K2 μόνο οριακά, αλλά παρόλα αυτά, αυτές συμπεριφέρονται χειρότερα από τις MWκαιNT. Για τις κανονικές και ασύμμετρες κατανομές που συζητήθηκαν παραπάνω, οι μέθοδοι των πυρήνων έχουν σαν αποτέλεσμα εκτιμήσεις που τείνουν να είναι πολύ κατώτερες.

ΕΚΤΙΜΗΣΗ ΤΟΥ ΕΜΒΑΔΟΥ ΚΑΤΩ ΑΠΟ ΤΗ ROC ΚΑΜΠΥΛΗ

Table IV. Simulated bias and RMSE for AUC estimators: $\log(Y) \sim \mathrm{N}(2.5, 0.25)$, $\log(X) \sim \mathrm{N}(\mu_X, 0.09)$.

Methods	$n = m = 20$						$n = m = 100$					
	AUC = 0.7		AUC = 0.8		AUC = 0.9		AUC = 0.7		AUC = 0.8		AUC = 0.9	
MW	−0.001	0.084	−0.003	0.072	−0.002	0.050	−0.001	0.037	−0.000	0.032	−0.000	0.022
N	−0.038	0.095	−0.035	0.091	−0.020	0.062	−0.049	0.065	−0.034	0.053	−0.019	0.033
NT	0.001	0.082	−0.001	0.069	−0.002	0.046	−0.000	0.036	0.000	0.031	−0.000	0.021
K1	−0.021	0.082	−0.027	0.076	−0.023	0.057	−0.015	0.039	−0.015	0.036	−0.013	0.026
K1T	−0.015	0.079	−0.023	0.072	−0.023	0.055	−0.011	0.037	−0.013	0.034	−0.013	0.026
K2	−0.016	0.082	−0.020	0.074	−0.017	0.054	−0.007	0.037	−0.006	0.033	−0.005	0.022
K2T	−0.013	0.080	−0.021	0.072	−0.020	0.054	−0.006	0.037	−0.006	0.033	−0.007	0.023

Πίνακας IV. Προσομοιωμένη κλίση και RMSE για τους εκτιμητές του AUC εμβαδού: $\log Y$~N(2.5,0.25), $\log X$~N(μ_X,0.09)

Table V. Simulated bias and RMSE for AUC estimators: Gamma AUC = 0.9, $p = 2$, $p1 = 2$.

Methods	$n = m = 20$				$n = m = 100$			
	$\lambda_Y = 0.5$		$\lambda_Y = 1$		$\lambda_Y = 0.5$		$\lambda_Y = 1$	
MW	0.001	0.048	0.001	0.050	0.000	0.022	−0.001	0.021
N	−0.041	0.061	−0.041	0.061	−0.047	0.052	−0.048	0.053
NT	0.001	0.045	0.001	0.046	0.001	0.021	−0.000	0.020
K1	−0.040	0.061	−0.040	0.063	−0.029	0.037	−0.031	0.038
K1T	−0.019	0.052	−0.018	0.054	−0.012	0.025	−0.014	0.026
K2	−0.028	0.054	−0.028	0.055	−0.010	0.024	−0.012	0.024
K2T	−0.017	0.051	−0.016	0.053	−0.005	0.023	−0.008	0.023

Πίνακας V. Προσομοιωμένη κλίση και RMSE για τους εκτιμητές του AUC εμβαδού: Gammma AUC = 0.9, p = 2, p_1=2.

Table VI. Simulated bias and RMSE for AUC estimators: $Y \sim \mathrm{N}(0,1)$, $X \sim 0.5\mathrm{N}(\mu,1) + 0.5\mathrm{N}(\mu+4,5)$.

Methods	$n = m = 20$						$n = m = 100$					
	AUC = 0.7		AUC = 0.8		AUC = 0.9		AUC = 0.7		AUC = 0.8		AUC = 0.9	
MW	0.004	0.081	−0.001	0.069	0.000	0.049	0.000	0.038	−0.000	0.031	0.000	0.021
N	0.041	0.076	0.006	0.052	−0.024	0.048	0.043	0.052	0.009	0.026	−0.023	0.029
NT	0.027	0.081	0.008	0.059	−0.012	0.043	0.032	0.047	0.012	0.029	−0.012	0.021
K1	−0.003	0.069	−0.026	0.064	−0.038	0.060	−0.002	0.034	−0.017	0.033	−0.026	0.033
K1T	−0.004	0.073	−0.022	0.065	−0.030	0.056	−0.003	0.035	−0.014	0.032	−0.020	0.029
K2	−0.001	0.072	−0.019	0.063	−0.028	0.054	−0.000	0.036	−0.006	0.031	−0.008	0.022
K2T	−0.003	0.074	−0.018	0.064	−0.024	0.052	−0.000	0.037	−0.006	0.031	−0.007	0.022

Πίνακας VI. Προσομοιωμένη κλίση και RMSE για τους εκτιμητές του AUC εμβαδού: Y~N(0,1), X~0.5N(μ ,1) + 0.5N(μ + 4.5)

Table VII. Simulated bias and RMSE for AUC estimators: $Y \sim \mathrm{N}(0,1)$, $X \sim 0.5\mathrm{N}(\mu,1) + 0.5\mathrm{N}(\mu+8,5)$.

Methods	$n = m = 20$						$n = m = 100$					
	AUC = 0.7		AUC = 0.8		AUC = 0.9		AUC = 0.7		AUC = 0.8		AUC = 0.9	
MW	0.001	0.082	−0.003	0.073	−0.000	0.049	0.000	0.039	0.001	0.031	−0.000	0.020
N	0.099	0.116	0.040	0.069	−0.017	0.048	0.103	0.107	0.048	0.054	−0.011	0.022
NT	0.077	0.111	0.037	0.073	−0.006	0.042	0.089	0.096	0.047	0.054	−0.004	0.017
K1	0.021	0.070	−0.022	0.064	−0.058	0.074	0.015	0.036	−0.013	0.030	−0.041	0.045
K1T	0.015	0.075	−0.018	0.064	−0.044	0.062	0.011	0.037	−0.010	0.028	−0.032	0.037
K2	0.016	0.073	−0.017	0.065	−0.040	0.061	0.003	0.037	−0.001	0.030	−0.008	0.022
K2T	0.012	0.076	−0.013	0.066	−0.030	0.054	0.002	0.038	−0.000	0.030	−0.006	0.021

Πίνακας VII. Προσομοιωμένη κλίση και RMSE για τους εκτιμητές του AUC εμβαδού:

Y~N(0,1), X~0.5N(μ ,1) + 0.5N(μ + 8.5)

3.4.3 ΠΡΟΣΟΜΟΙΩΣΕΙΣ ΜΕ ΜΕΙΓΜΑΤΑ ΑΠΟ ΚΑΝΟΝΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

Προκειμένου να μελετηθούν επιπρόσθετες περιπτώσεις οι οποίες δεν προκύπτουν από την οικογένεια μετασχηματισμού δύναμης κανονικών κατανομών, εξετάσθηκαν μείγματα κανονικών κατανομών. Αυτά μπορούν να παράγουν bimodal σχηματισμούς. Στους πίνακες VIκαιVII, θεωρούμε ότι το X δεδομένο (το παθολογικό) προέρχεται από ένα μείγμα ενώ το Υδεδομένο (το υγιές) ακολουθεί μία κανονική κατανομή. Στους πίνακες VIIIκαι IX και οι δύο πληθυσμοί θεωρούνται μείγματα. Οι συγκεκριμένες κατανομές χρησιμοποιούνται για τις προσομοιώσεις που δίνονται στους πίνακες με την παράμετρο μ να ποικίλει σύμφωνα με τις καθορισμένες AUCτιμές. Αυτές οι κατανομές αναπαριστώνται στις εικόνες 2(f)-(i). Όπως μπορεί να δει κανείς από τα γραφήματα τα αποτελέσματα στον πίνακα VIIείναι βασισμένα σε μία πιο ισχυρή bimodalityστον ασθενή πληθυσμό σε σύγκριση με τα αποτελέσματα στον πίνακα VI. Στον πίνακα IXο υγιής πληθυσμός εκθέτει μία πιο ασθενή bimodality απ'ότι στον πίνακα VIII. Από τους πίνακες VIκαι VII μπορούμε να παρατηρήσουμε ότι όπως και στις περιπτώσεις ασύμμετρων κατανομών που μελετήθηκαν προηγουμένως, η N μέθοδος είναι επηρεασμένη, με κλίση η οποία δεν επηρεάζεται από την αύξηση του μεγέθους δείγματος αλλά σημειώνει την μέγιστη τιμή της για AUC= 0.7. Όσων αφορά το RMSE, η N διαδικασία παρουσιάζει διακυμάνσεις, δίνοντας κάποιες φορές τη χαμηλότερη τιμή για το RMSE και άλλες την υψηλότερη. Εξαιτίας της υψηλήςbimodality (όπως φαίνεται στον πίνακα VII) η N διαδικασία είναι ιδιαίτερα ακατάλληλη για μικρές τιμές του AUC. ΗΝΤμέθοδος, από την άλλη, δεν παρέχει καμία σταθερή βελτίωση εκτός για την τιμή AUC=0.9. Επιπλέον μετασχηματίζοντας τα δεδομένα πριν την εφαρμογή της προσέγγισης των πυρήνων, βελτιώνει κάπως τα αποτελέσματα για τις τιμές των RMSEκαι της κλίσης εκτός από την περίπτωση όπου έχουμε AUC=0.7 καιn=m=20. Η K2 μέθοδος είναι πάλι ανώτερη ή ίδια ως προς την K1, εκτός από την περίπτωση όπου AUC=0.7. Για μικρά μεγέθη δείγματος, η μέθοδος των πυρήνων μπορεί να βελτιώσει σημαντικά την MWμέθοδο όσων αφορά το RMSE (μέχρι και 8%) εκτός από την περίπτωση όπου AUC=0.9, όπου η μέθοδος των πυρήνων φέρνει αντίθετα αποτελέσματα.

Στους πίνακες VIIIκαι IX, η N διαδικασία δίνει πολύ καλύτερα αποτελέσματα. Αν και είναι ολοφάνερα επηρεασμένη, έχει συχνά το χαμηλότερο RMSE. Τα αποτελέσματα της NT μεθόδου είναι αρκετά παρόμοια με αυτά της N. Όποτε είχαμε την τιμή AUC=0.9, η NT μέθοδος παρέμενε αποτελεσματικά ανεπηρέαστη και είχε την χαμηλότερη τιμή RMSE. Στις μεθόδους των πυρήνων, ο μετασχηματισμός δεν οδηγεί σε κάποια βελτίωση. Οι K1 και K2 συμπεριφέρονται παρόμοια εκτός από την περίπτωση όπουAUC=0.9 όπου η K2 μέθοδος είναι ολοφάνερα ανώτερη. Για την τιμή AUC=0.7 ή 0.8 η προσέγγιση των πυρήνων παρουσιάζει τιμή RMSEτο ίδιο καλή ή και καλύτερη από τη MWμέθοδο.

Table VIII. Simulated bias and RMSE for AUC estimators: $Y \sim 0.5\mathrm{N}(0,1) + 0.5\mathrm{N}(3,1)$, $X \sim 0.5\mathrm{N}(\mu,1) + 0.5\mathrm{N}(\mu+4,5)$.

Methods	$n=m=20$						$n=m=100$					
	AUC = 0.7		AUC = 0.8		AUC = 0.9		AUC = 0.7		AUC = 0.8		AUC = 0.9	
MW	0.000	0.079	0.002	0.068	−0.000	0.047	−0.000	0.036	0.001	0.029	−0.000	0.020
N	0.021	0.071	0.015	0.057	−0.008	0.039	0.024	0.039	0.016	0.029	−0.006	0.018
NT	0.023	0.073	0.020	0.058	−0.005	0.037	0.024	0.040	0.021	0.032	−0.001	0.016
K1	−0.012	0.071	−0.015	0.062	−0.027	0.051	−0.008	0.035	−0.008	0.028	−0.016	0.025
K1T	−0.009	0.071	−0.012	0.061	−0.027	0.051	−0.006	0.034	−0.006	0.027	−0.015	0.025
K2	−0.009	0.072	−0.010	0.062	−0.020	0.048	−0.003	0.035	−0.001	0.028	−0.006	0.020
K2T	−0.007	0.072	−0.009	0.062	−0.022	0.048	−0.002	0.035	−0.001	0.028	−0.006	0.020

Πίνακας VIII. Προσομοιωμένη κλίση και RMSE για τους εκτιμητές του AUC εμβαδού: Y~0.5N(0,1) + 0.5N(3,1) , X~0.5N(μ ,1) + 0.5N(μ + 4.5)

Table IX. Simulated bias and RMSE for AUC estimators: $Y \sim 0.5\mathrm{N}(0,1) + 0.5\mathrm{N}(3,1.5)$, $X \sim 0.5\mathrm{N}(\mu,1) + 0.5\mathrm{N}(\mu+4,5)$.

Methods	$n=m=20$						$n=m=100$					
	AUC = 0.7		AUC = 0.8		AUC = 0.9		AUC = 0.7		AUC = 0.8		AUC = 0.9	
MW	0.003	0.081	0.001	0.064	−0.000	0.047	−0.000	0.036	−0.000	0.029	−0.000	0.020
N	0.018	0.073	0.010	0.054	−0.007	0.041	0.017	0.036	0.011	0.027	−0.005	0.017
NT	0.020	0.075	0.016	0.055	−0.003	0.039	0.018	0.037	0.016	0.029	−0.000	0.016
K1	−0.013	0.073	−0.017	0.060	−0.025	0.052	−0.010	0.035	−0.012	0.030	−0.016	0.025
K1T	−0.009	0.073	−0.014	0.059	−0.025	0.051	−0.008	0.034	−0.009	0.029	−0.015	0.024
K2	−0.009	0.074	−0.012	0.060	−0.019	0.049	−0.004	0.035	−0.004	0.028	−0.006	0.020
K2T	−0.007	0.074	−0.011	0.059	−0.021	0.049	−0.004	0.035	−0.003	0.028	−0.007	0.020

Πίνακας VIII. Προσομοιωμένη κλίση και RMSE για τους εκτιμητές του AUC εμβαδού: Y~0.5N(0,1) + 0.5N(3,1.5) , X~0.5N(μ ,1) + 0.5N(μ + 4.5)

Table X. Simulated bias and RMSE for RC estimates.

Scenario	Bias	RMSE
Table I, AUC = 0.9, $n=m=100$	0.000	0.020
Table III, AUC = 0.9, $n=m=100$	0.002	0.024
Table V, AUC = 0.9, $\lambda_Y=1$, $n=m=100$	−0.001	0.023
Table VII, AUC = 0.7, $n=m=100$	0.038	0.050

Πίνακας VIII. Προσομοιωμένη κλίση και RMSE για τους RC εκτιμητές.

3.4.4 Η RC ΜΕΘΟΔΟΣ

Η RCμέθοδος προτάθηκε επίσης προκειμένου να μελετηθεί. Οι Hajian και Tilaki πραγματοποίησαν μία μελέτη προσομοίωσης η οποία σύγκρινε τις RCκαιMW μεθόδους με βάση την κλίση και την διακύμανση, για διάφορες κατανομές. Αυτοί συμπέραναν ότι οι δύο προσεγγίσεις έδιναν παρόμοια αποτελέσματα. Συνεπώς οι συγκρίσεις της MW και άλλων μεθόδων πρέπει να μεταφερθούν στην RC.

Προκειμένου να εξετασθεί περαιτέρω το παραπάνω, διεξήχθηκε μία μικρή προσομοίωση της RCμεθόδου (χρησιμοποιώντας 100 προσομοιώσεις) για μερικές από τις περιπτώσεις που περιγράφηκαν παραπάνω. Αυτές δίνονται στον πίνακα Χ. Στις τρείς πρώτες περιπτώσεις, η RCμέθοδος συμπεριφέρεται παρόμοια με τη MWκαι τη NT ενώ στην τέταρτη περίπτωση συμπεριφέρεται αρκετά διαφορετικά. Αυτή η περίπτωση αντιμετωπίζει μία ισχυρή bimodalκατανομή (δες εικόνα 2(g))για τον άρρωστο πληθυσμό. Αν και οι Hajian και Tilaki μελέτησαν bimodalδείγματα δεν ασχολήθηκαν με κάποια περίπτωση τόσο ισχυρά bimodalόπως αυτή. Βλέπουμε μία ολοφάνερη κλήση στην RC. Για την τέταρτη περίπτωση η RCείναι σίγουρα προτιμότερη έναντι της NT και με βάση την τιμή RMSEαλλά και την κλίση αλλά ξεπεράστηκε σημαντικά από την μέθοδο MWκαι τις μεθόδους των πυρήνων(δες πίνακα VII).

3.5 ΕΠΑΝΕΞΕΤΑΖΟΝΤΑΣ ΤΟ ΠΑΡΑΔΕΙΓΜΑ

Η εικόνα 1(a)δείχνει ότι τα δεδομένα είναι αρκετά ασύμμετρα. Η υπόθεση κανονικότητας είναι αστήρικτη. Δεν μας εκπλήσσει το γεγονός ότι η παραμετρική μέθοδος εκτίμησης, η οποία υπέθετε κανονικότητα δίνει μία εκτίμηση αρκετά διαφορετική από τις άλλες. Το γεγονός ότι αυτή η εκτίμηση είναι αρκετά χαμηλότερη από αυτήν της αμερόληπτης MWμεθόδου, οφείλεται στις αρνητικές κλήσεις που σημειώθηκαν στις προσομοιώσεις μας, στις ασύμμετρες κατανομές. Οι προσομοιώσεις μας δείχνουν ότι η NT μέθοδος πρέπει να είναι κατάλληλη σε αυτή την περίπτωση. Τα μετασχηματισμένα δεδομένα (εικόνα1(b)) είναι πιο "κοντά" στην κανονικότητα. Οι μέθοδοι NT, MW καιRC είναι αρκετά παρόμοιες όσων αφορά τις προσομοιώσεις μας. Επιπλέον, οι προσομοιώσεις αυτές (δες πίνακες III, IVκαιV) μας οδηγούν στο να συμπεράνουμε ότι για τις ασύμμετρες κατανομές, όλες οι εκτιμήσεις των πυρήνων θα τείνουν να είναι πολύ χαμηλές με την K1 να είναι η χειρότερη ενώ οι K1T και K2T θα είναι παρόμοιες η μία στην άλλη και πιο κοντά στο να γίνουν αμερόληπτες. Αυτό φαίνεται να συμφωνεί αρκετά με ότι εμείς παρατηρούμε για τα CKδεδομένα.

3.6. ΣΥΖΗΤΗΣΗ

Οι RC και NT προσεγγίσεις είναι και οι δύο βασισμένες σε αρκετά παρόμοιες υποθέσεις. Η RCμέθοδος υποθέτει ότι κάποιοι μονοτονικοί μετασχηματισμοί των δεδομένων θα έχουν σαν αποτέλεσμα την κανονικότητα ενώ η NT μέθοδος είναι λιγότερο γενική σε αυτό. Αυτή, από την άλλη πλευρά, υποθέτει ότι ο μετασχηματισμός ανήκει σε μία συγκεκριμένη οικογένεια, η οποία είναι η οικογένεια μετασχηματισμών δύναμης. Έτσι δεν μας εκπλήσσει το γεγονός ότι οι δύο αυτές μέθοδοι συνήθως παράγουν παρόμοια αποτελέσματα. Αυτές, μπορεί παρόλα αυτά να

διαφέρουν ουσιαστικά για ισχυρές bimodal κατανομές, όπου η RC είναι καλύτερη αλλά καμία από τις δύο δεν είναι κατάλληλη. Θα πρέπει να σημειωθεί επίσης ότι η NT είναι υπολογιστικά πολύ πιο απλή μέθοδος από ότι η RC.

Από τα παραπάνω αποτελέσματα είναι φανερό ότι η NT μέθοδος είναι η προτιμότερη εφόσον υπάρχει η υποψία πως οι κατανομές του δείκτη για τους υγιείς και ασθενείς πληθυσμούς είναι μίγματα. Πρέπει επίσης να σημειωθεί πως αν και η MWμέθοδος συνήθως δεν είναι καλύτερη όσων αφορά την RMSEείναι πολύ κοντά στο να είναι η καλύτερη. Στις μονοτροπικές καταστάσεις(unimodal situations) θεωρήσαμε πως η NT μέθοδος παρέμενε αποτελεσματικά ανεπηρέαστη και σε σύγκριση με την RMSE λειτούργησε κατά κάποιον τρόπο καλύτερα από τη μη παραμετρική MW προσέγγιση, κάποιες φορές έως και 8% καλύτερα. Η NT προσέγγιση έχει την επιπρόσθετη χρήσιμη ιδιότητα να δίνει μία συνεχή ROCκαμπύλη. Για unimodal κατανομές διαφόρων μορφών, δεν βρήκαμε την μέθοδο των πυρήνων χρήσιμη αν και για μικρά AUCεμβαδά συνοδευόμενα από μικρά μεγέθη δείγματος αυτοί οδήγησαν σε μία μικρή βελτίωση του RMSEεις βάρος μιας μεγαλύτερης κλίσης. Αντιμετωπίζοντας μίγματα ανακαλύψαμε πώς αν οι δύο πληθυσμοί είναι καλά διαχωρισμένοι (Εμβαδόν AUC=0.9)η NT μέθοδος εξακολουθεί να συμπεριφέρεται καλύτερα. Προφανώς σε μία τέτοια κατάσταση οι πραγματικές λεπτομέρειες των μορφών της κατανομής δεν έχουν μεγάλη σημασία. Για λιγότερο διαχωρισμένους πληθυσμούς (εμβαδό AUC=0.7 ,0.8) βασισμένοι σε μίγματα δεν υπήρχε μέθοδος που να υπερέχει σημαντικά. Σε αυτή τη περίπτωση η προσέγγιση των πυρήνων αποδείχθηκε ανώτερη της MW μεθόδου και συγκρινόμενη με την N ή την NT μέθοδο η μέθοδος των πυρήνων κυμαινόταν από πολύ καλύτερη(για παράδειγμα στον πίνακαVII, n=m=20, AUC=0.7) έως ελαφρώς κατώτερη(για παράδειγμα πίνακας VIIIn=m=20, AUC=0.8). Για τις κατανομές μιγμάτων με μέτριο AUC υπήρχε πολύ μικρή διαφορά ανάμεσα στις διάφορες προσεγγίσεις πυρήνων και την K1 η οποία θα μπορούσε να προταθεί χάρη στην απλότητά της.

Επιπλέον προτάθηκε τα τυπικά σφάλματα των προσομοιωμένων δεικτών να υπολογίζονται προκειμένου να ερμηνεύονται πιο εύκολα οι πίνακες. Αυτό θα απαιτούσε το να επαναλάβουμε το κάθε ένα από τα σύνολα προσομοίωσης τα οποία είναι μεγέθους 1000 το καθένα πολλές φορές, και ως εκ τούτου κάτι τέτοιο θα ήταν εξαιρετικά χρονοβόρο. Εμείς πραγματοποιήσαμε ένα περιορισμένο αριθμό επαναλήψεων σε μερικές περιπτώσεις και βρήκαμε ότι το τυπικό σφάλμα για το RMSE είναι περίπου 2.5% της υπολογισμένης του τιμής. Αυτό ενισχύει τις παραπάνω ερμηνείες των αποτελεσμάτων προσομοίωσης.

ΚΕΦΑΛΑΙΟ IV

ΕΦΑΡΜΟΓΗ ΣΕ ΠΡΑΓΜΑΤΙΚΑ ΣΕΙΣΜΟΛΟΓΙΚΑ ΔΕΔΟΜΕΝΑ

4.1 ΕΙΣΑΓΩΓΗ ΣΤΟ ΠΡΟΒΛΗΜΑ

Πρόκειται για μία βάση δεδομένων με 10.333 (n) εγγραφές και 11 επεξηγηματικές μεταβλητές.
Παρακάτω παραθέτουμε επιγραμματικά τις 11 αυτές επεξηγηματικές μεταβλητές που μελετήθηκαν :

X_1	**Χρόνια (years)**
X_2	**Νομός (1-54)**
X_3	**Γεωγραφικό μήκος (longitude)**
X_4	**Γεωγραφικό πλάτος (latitude)**
X_5	**Ένταση (1-12)**
X_6	**Απόσταση από το σεισμό (km)**
X_7	**Hyper distance (degrees)**
X_8	**Αζιμούθιο (degrees)**
X_9	**Επίκεντρο, στον άξονα x (τεταγμένη)**
X_{10}	**Επίκεντρο, στον άξονα y (τετμημένη)**
X_{11}	**Βάθος (0-700 km)**
Y	**Magnitude (0(<6.5), 1(>6.5))**

Αυτές τις χωρίζουμε ανάλογα σε συνεχείς (range) και κατηγορικές(ordinal).

- Συνεχείς (Range)

X_1	**Χρόνια (years)**
X_3	**Γεωγραφικό μήκος (longitude)**
X_4	**Γεωγραφικό πλάτος (latitude)**
X_6	**Απόσταση από το σεισμό (km)**
X_7	**Hyperdistance (degrees)**
X_8	**Αζιμούθιο (degrees)**
X_9	**Επίκεντρο, στον άξονα x (τεταγμένη)**
X_{10}	**Επίκεντρο, στον άξονα y (τετμημένη)**
X_{11}	**Βάθος (0-700 km)**

- Κατηγορικές (ordinal)

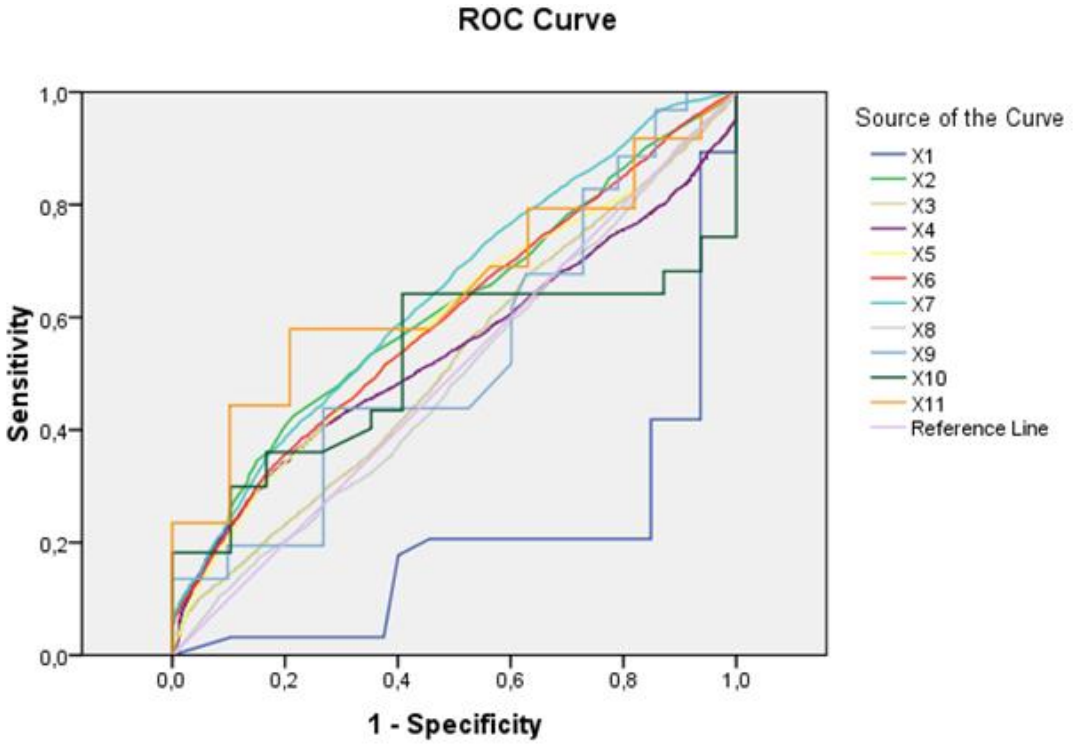
X_2	**Νομός (1-54)**
X_5	**Ένταση (1-12)**

Σκοπός μας είναι η μελέτη των παραπάνω 11 παραγόντων που κρίνεται ότι επηρεάζουν την ένταση των σεισμικών δονήσεων σε μία περιοχή. Ως εξαρτημένη μεταβλητή Y ορίζεται η ένταση της σεισμικής δόνησης. Η εξαρτημένη αυτή μεταβλητή είναι δίτιμη και παίρνει τις τιμές Y=1 (όταν η ένταση >6.5) και Y=0 (όταν η ένταση < 6.5).

4.2 ΜΕΛΕΤΗ ΤΟΥ ΠΛΗΡΟΥΣ ΜΟΝΤΕΛΟΥ

Αρχικά θα μελετήσουμε το πλήρες μοντέλο το οποίο περιλαμβάνει όλες τις μεταβλητές και στη συνέχεια θα εξετάσουμε παραδειγματικά κάποιες από αυτές επικεντρώνοντας το ενδιαφέρον μας στις πιο σημαντικές.

Θέτοντας ως μεταβλητή αναφοράς την ένταση για την τιμή Y=1, και εφαρμόζουμε ROC ανάλυση στα δεδομένα μας. Το στατιστικό πακέτο SPSS 16.0 δίνει το παρακάτω γράφημα (Εικόνα 4.1) που απεικονίζει τις ROC καμπύλες για την κάθε μία από τις προαναφερθείσες μεταβλητές καθώς και τον πίνακα 4.1, στις στήλες του οποίου εμφανίζονται η κάθε ανεξάρτητη μεταβλητή με το αντίστοιχο εμβαδόν κάτω από την καμπύλη ROC (AUC), το τυπικό σφάλμα και το επίπεδο σημαντικότητας p-value. Όλες οι στατιστικές επεξεργασίες πραγματοποιήθηκαν σε στάθμη εμπιστοσύνης 95% και ως βαθμός σημαντικότητας ορίστηκε $p< 0.05$.



Εικόνα 4.1.ROC καμπύλες για όλες τις μεταβλητές

Area Under the Curve

Test Result Variable...	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
X1	,198	,005	,000	,189	,208
X2	,613	,006	,000	,602	,624
X3	,520	,006	,001	,508	,533
X4	,540	,006	,000	,529	,551
X5	,587	,006	,000	,575	,598
X6	,597	,006	,000	,585	,608
X7	,640	,006	,000	,629	,652
X8	,493	,006	,287	,481	,506
X9	,532	,006	,000	,520	,545
X10	,520	,006	,002	,508	,531
X11	,648	,005	,000	,638	,659

The test result variable(s): X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

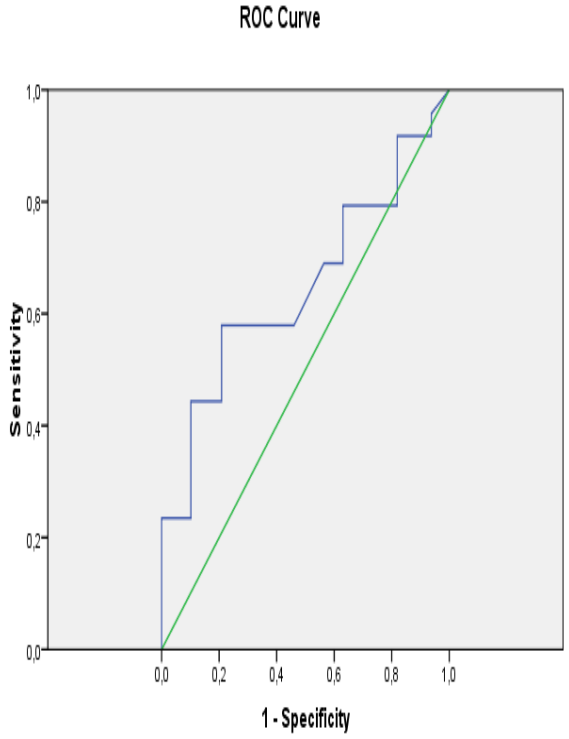
a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

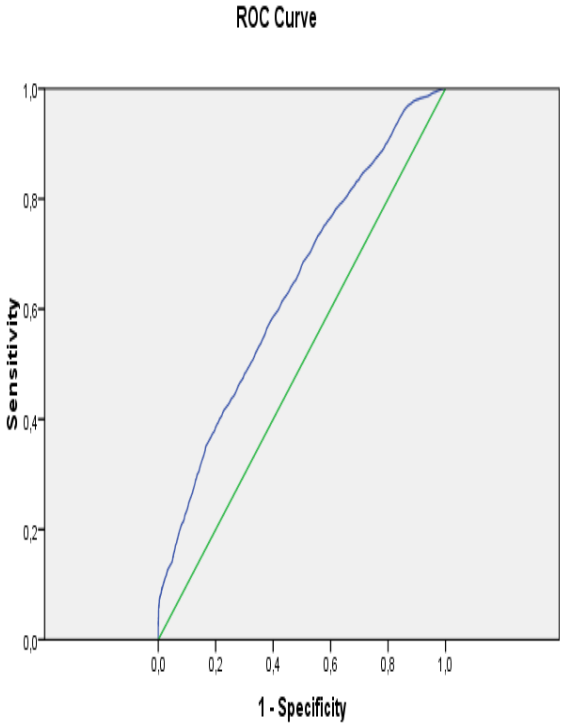
Πίνακας 4.1.Αποτελέσματα ROCανάλυσης στο πλήρες μοντέλο

Από τον πίνακα 4.1 μπορούμε να συμπεράνουμε ότι οι μεταβλητές που παίζουν σημαντικότερο ρόλο στη πρόβλεψη της έντασης μιας σεισμικής δόνησης είναι ο X11 (ο οποίος δηλώνει το βάθος από την επιφάνεια του εδάφους που έχει το επίκεντρο), ο X7 (hyper distance), ο X2 (νομός) και ο X6 (απόσταση από το σεισμό). Τις μεταβλητές αυτές τις διακρίναμε ως πιο σημαντικές καθώς το αντίστοιχο εμβαδόν κάτω από την ROC καμπύλη είναι αρκετά μεγαλύτερο από την τιμή 0.5 (τιμή που αντιστοιχεί στην τυχαία αναπαράσταση). Όπως είχαμε προαναφέρει, το εμβαδόν της περιοχής κάτω από την καμπύλη ROC αποτελεί ένα ισχυρό μέτρο του συνολικού πληροφοριακού περιεχομένου της διαγνωστικής δοκιμασίας και εκφράζει την πιθανότητα ένα ζεύγος μετρήσεων, μιας από τις μεγάλης έντασης σεισμικές δονήσεις και μιας από τις χαμηλής έντασης, να ταξινομηθεί με τη σωστή σειρά, δηλαδή η μέτρηση της επικίνδυνης (δυνατής έντασης) δόνησης να είναι υψηλότερη από αυτή της ασθενέστερης δόνησης.

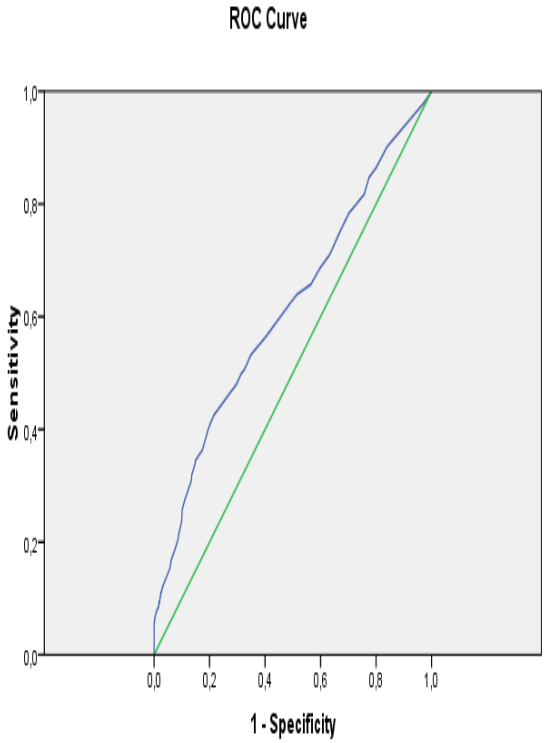
ΣΗΜΑΝΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ ΜΕ ΥΨΗΛΟ ΔΕΙΚΤΗ AUC



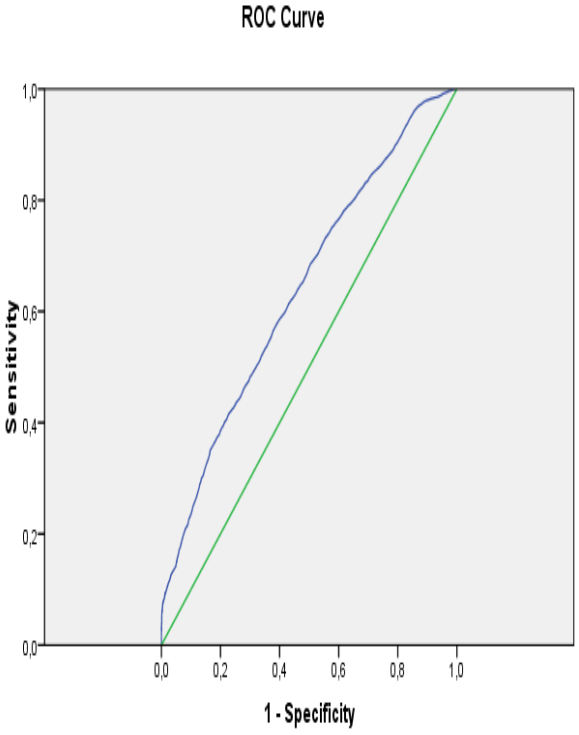
Εικόνα 4.3 ROCκαμπύλη για μεταβλητή X11 (βάθος)



Εικόνα 4.2 ROCκαμπύλη για μεταβλητή X7 (hyper distance)

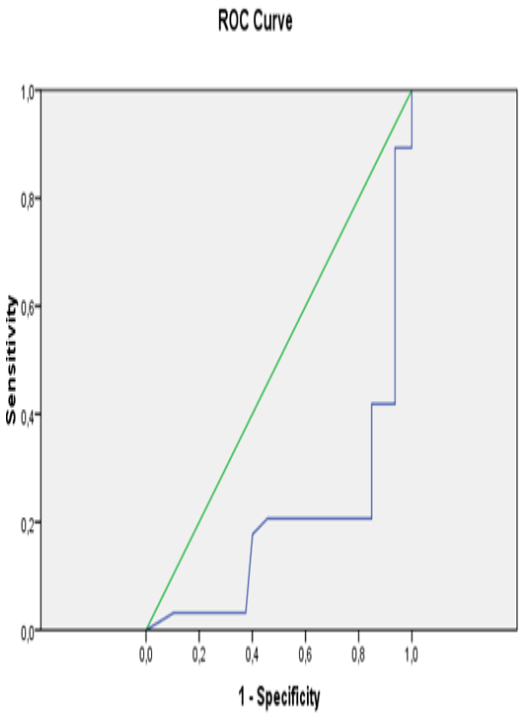


Εικόνα 4.4ROCκαμπύλη για μεταβλητή X2 (νομός)

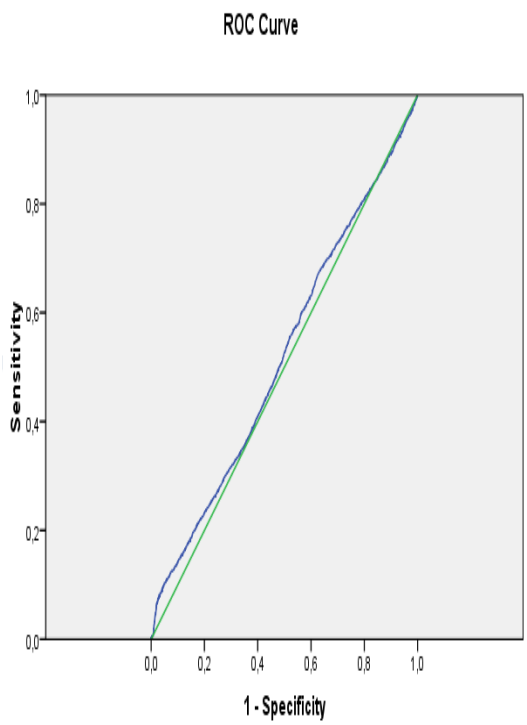


Εικόνα 4.5 ROCκαμπύλη για μεταβλητή X6 (απόσταση από επίκεντρο)

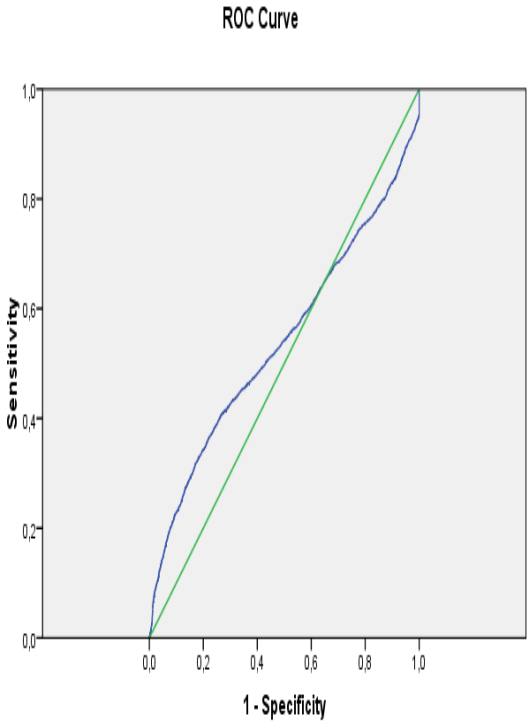
ΜΗ ΣΗΜΑΝΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ ΜΕ ΧΑΜΗΛΟ ΔΕΙΚΤΗ AUC



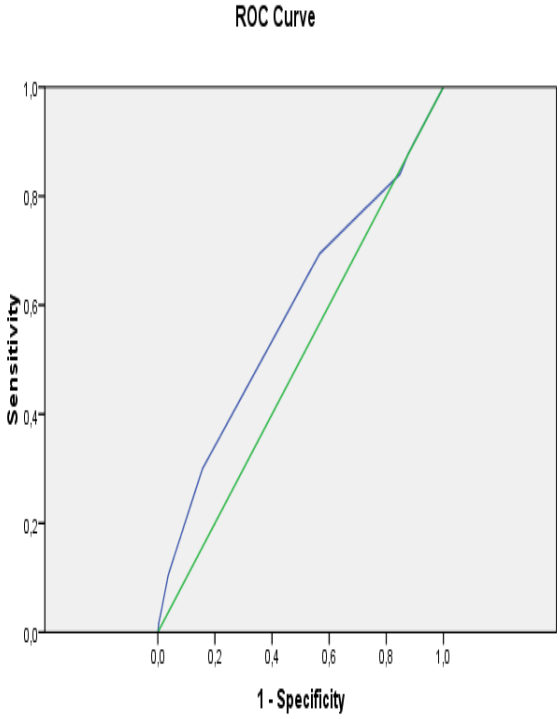
Εικόνα 4.6. ROCκαμπύλη για μεταβλητή X1 (χρόνια)



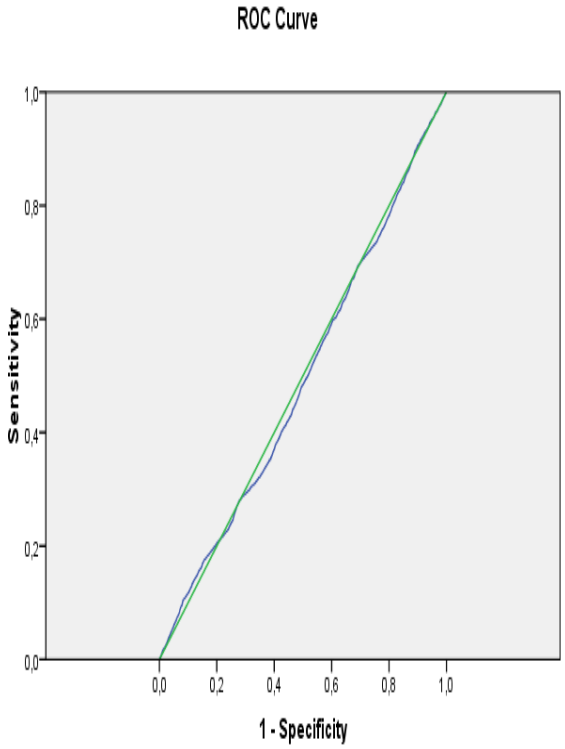
Εικόνα 4.7 ROCκαμπύλη για μεταβλητή X3 (γεωγραφικό μήκος)



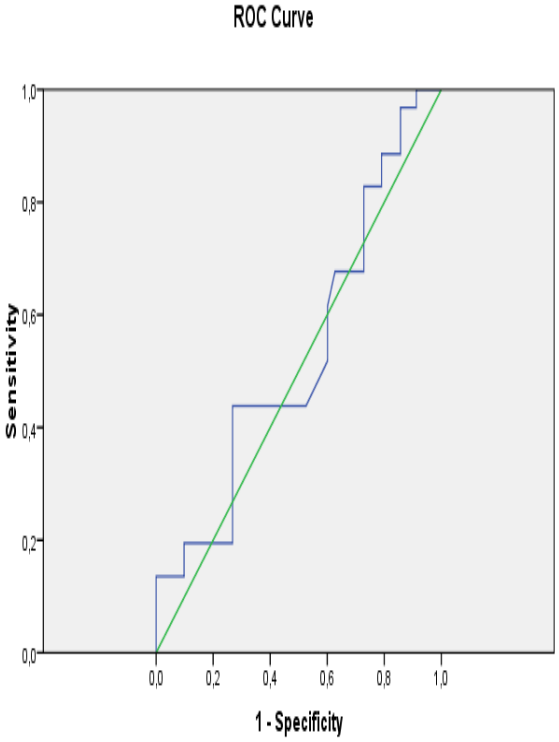
Εικόνα 4.9 ROCκαμπύλη για μεταβλητή X4 (γεωγραφικό πλάτος)



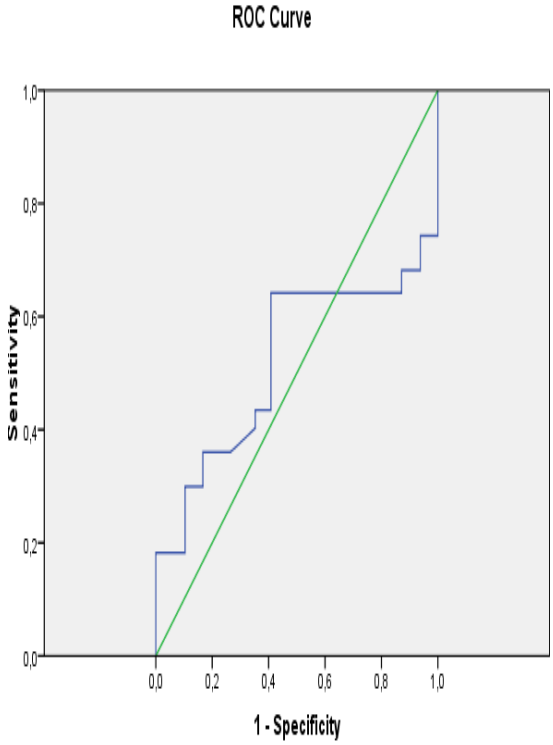
Εικόνα 4.8 ROCκαμπύλη για μεταβλητή X5 (ένταση)



Εικόνα4.10ROCκαμπύλη για μεταβλητή X8 (αζιμούθιο)



Εικόνα 4.11ROCκαμπύλη για μεταβλητή X9 (επίκεντρο, στον άξονα x)



Εικόνα 4.12 ROCκαμπύλη για μεταβλητή X10 (επίκεντρο, στον άξονα y)

4.3 ΕΥΡΕΣΗ ΒΕΛΤΙΣΤΟΥ ΟΡΙΟΥ

Το διαχωριστικό όριο είναι μία συγκεκριμένη τιμή της διαχωρίζουσας μεταβλητής που επιλέγεται στην κλίμακα μέτρησής της, πέραν της οποίας οι τιμές τις μεταβλητής (τα αποτελέσματα της δοκιμασίας) θεωρούνται «θετικές» (στην περίπτωσή μας, υψηλής έντασης) και κάτω της οποίας «αρνητικές» (χαμηλής έντασης).

Το βέλτιστο διαχωριστικό όριο παρέχει τη μέγιστη πληροφορία για το λόγο αυτό είναι υψίστης σημασίας η επιτυχής εύρεσή του. Το βέλτιστο ΔΟ δίνει το μέγιστο άθροισμα αληθώς θετικών και αληθώς αρνητικών αποτελεσμάτων και επομένως το ελάχιστο άθροισμα ψευδώς θετικών και ψευδώς αρνητικών αποτελεσμάτων. Είναι το σημείο που μεγιστοποιεί την τιμή TPR+TNR ή Sensitivity + Specificity. Τέλος για την εύρεση του βέλτιστου ορίου θα πρέπει η τιμή Sensitivity να είναι όσο το δυνατόν πλησιέστερα στη μονάδα ενώ η τιμή 1-Specificity πλησιέστερα στο μηδέν.

Παρακάτω θα περιγράψουμε τη διαδικασία επιλογής του ορίου αυτού παραδειγματικά για την επεξηγηματική μεταβλητή X11 (βάθος του επίκεντρου). Εφαρμόζοντας τη ROC ανάλυση στο SPSS παίρνουμε τον παρακάτω πίνακα με τα διάφορα διαχωριστικά όρια για την υπό εξέταση μεταβλητή καθώς και τις αντίστοιχες τιμές ευαισθησίας και ειδικότητας. Επιπλέον ταξινομούμε τα δεδομένα μας ως προς τη στήλη Sensitivity+Specificity κατά φθίνουσα σειρά. Το πρώτο στοιχείο του πίνακα που προκύπτει είναι και το βέλτιστο όριο.

Coordinates of the curve		
Test result variable(s):X11		
Positive if Greater Than or Equal TO	Sensitivity	1-Specificity
0	1	1
2,5	0,958	0,938
4,5	0,918	0,938
7,5	0,918	0,882
10,5	0,918	0,819
11,5	0,825	0,819
12,5	0,793	0,819
14	0,793	0,631
15,5	0,751	0,631
16,5	0,69	0,631
17,5	0,69	0,564
18,5	0,579	0,46
21	0,579	0,362
24,5	0,579	0,209
27	0,523	0,209
29	0,443	0,209
32	0,443	0,102
36,5	0,325	0,102
43	0,235	0,102
49	0,235	0,076
64,5	0,235	0
79,5	0,135	0
88	0,107	0
96	0	0

Πίνακας 4.2:Διαχωριστικά όρια για τη μεταβλητή X11

Positive if Greater Than or Equal TO	Sensitivity	1-Specificity	Sensitivity+Specificity
24,5	**0,579**	**0,209**	**1,37**
32	0,443	0,102	1,341
27	0,523	0,209	1,314
64,5	0,235	0	1,235
29	0,443	0,209	1,234
36,5	0,325	0,102	1,223
21	0,579	0,362	1,217
14	0,793	0,631	1,162
49	0,235	0,076	1,159
79,5	0,135	0	1,135
43	0,235	0,102	1,133
17,5	0,69	0,564	1,126
15,5	0,751	0,631	1,12
18,5	0,579	0,46	1,119
88	0,107	0	1,107
10,5	0,918	0,819	1,099
16,5	0,69	0,631	1,059
7,5	0,918	0,882	1,036
2,5	0,958	0,938	1,02
11,5	0,825	0,819	1,006
0	1	1	1
96	0	0	1
4,5	0,918	0,938	0,98
12,5	0,793	0,819	0,974

Πίνακας 4.3: Εύρεση διαχωριστικού ορίου για τη μεταβλητή X11

Επομένως το βέλτιστο διαχωριστικό όριο που προέκυψε με την παραπάνω διαδικασία για τη μεταβλητή X11 είναι το 24,5. Στις διπλανές στήλες φαίνονται και οι τιμές της ευαισθησίας και ειδικότητας που του αντιστοιχούν.

Στον Πίνακα 4.4 που ακολουθεί παραθέτονται συγκεντρωτικά τα βέλτιστα διαχωριστικά όρια, η ευαισθησία και η ειδικότητα ενδεικτικά κάποιων από τις μεταβλητές.

Μεταβλητή	Βέλτιστο ΔΟ	Ευαισθησία	Ειδικότητα
X1(Χρόνια)	1964.5	0.893	0.063
X2(Νομός)	20.5	0.426	0.784
X5(Ένταση)	4.5	0.299	0.844
X7(Hyper distance)	107.5	0.571	0.617
X9(Επίκεντρο, στον άξονα x)	22.375	0.438	0.732
X10(Επίκεντρο, στον άξονα y)	38.32	0.642	0.592
X11(Βάθος)	24.5	0.579	0.791

Πίνακας 4.4 Τα βέλτιστα διαχωριστικά όρια των διαδικασιών

Έχοντας υπολογίσει τα βέλτιστα διαχωριστικά όρια των μεταβλητών μπορούμε να προχωρήσουμε σε επιπλέον ανάλυση των αποτελεσμάτων. Είναι πλέον εύκολος ο υπολογισμός των Αληθώς Θετικών, Ψευδώς Θετικών, Ψευδώς Αρνητικών και Αληθώς Αρνητικών αποτελεσμάτων καθώς και της Ακρίβειας της κάθε δοκιμασίας:

X1

TPR	FPR	FNR	TNR
89,3%	93,7%	10,7%	6,3%

Πραγματική τιμή

Αποτέλεσμα Πρόβλεψης		**p**	**n**
	p'	6526	2835
	n'	782	191

Accuracy =65%

Οι τιμές της μεταβλητής X1 που είναι μεγαλύτερες του ΔΟ 1964.5 θεωρούνται «επικίνδυνες» (υψηλής έντασης), ενώ αυτές που είναι μικρότερες του, θεωρούνται «ακίνδυνες» (χαμηλής έντασης). Για το συγκεκριμένο διαχωριστικό όριο, για το 89,3% (TPR) των υψηλών δονήσεων το αποτέλεσμα της δοκιμασίας θα είναι θετικό, ενώ για το 10,7% (FNR) αρνητικό. Επιπλέον για το 93,7% (FPR) των χαμηλών δονήσεων το αποτέλεσμα θα είναι θετικό ενώ για το 6,3% (TNR) αρνητικό. Η ακρίβεια της δοκιμασίας προέκυψε 65%.

X2

TPR	FPR	FNR	TNR
42,6%	21,6%	57,4%	78,4%

		Πραγματική τιμή	
Αποτέλεσμα Πρόβλεψης		**p**	**n**
	p'	3113	654
	n'	4195	2372

Accuracy =53%

Οι τιμές της μεταβλητής X2 που είναι μεγαλύτερες του ΔΟ 20,5 θεωρούνται «επικίνδυνες» (υψηλής έντασης), ενώ αυτές που είναι μικρότερες του, θεωρούνται «ακίνδυνες» (χαμηλής έντασης). Για το συγκεκριμένο διαχωριστικό όριο, για το 42,6% (TPR) των υψηλών δονήσεων το αποτέλεσμα της δοκιμασίας θα είναι θετικό, ενώ για το 57,4% (FNR) αρνητικό. Επιπλέον για το 21,6% (FPR) των χαμηλών δονήσεων το αποτέλεσμα θα είναι θετικό ενώ για το 78,4% (TNR) αρνητικό. Η ακρίβεια της δοκιμασίας προέκυψε 53%.

X5

TPR	FPR	FNR	TNR
29,9%	15,6%	70,1%	84,4%

		Πραγματική τιμή	
Αποτέλεσμα Πρόβλεψης		**p**	**n**
	p'	2185	472
	n'	5123	2554

Accuracy =46%

Οι τιμές της μεταβλητής X5 που είναι μεγαλύτερες του ΔΟ 4,5 θεωρούνται «επικίνδυνες» (υψηλής έντασης), ενώ αυτές που είναι μικρότερες του, θεωρούνται «ακίνδυνες» (χαμηλής έντασης). Για το συγκεκριμένο διαχωριστικό όριο, για το 29,9% (TPR) των υψηλών δονήσεων το αποτέλεσμα της δοκιμασίας θα είναι θετικό, ενώ για το 70,1% (FNR) αρνητικό. Επιπλέον για το 15,6% (FPR) των χαμηλών δονήσεων το αποτέλεσμα θα είναι θετικό ενώ για το 84,4% (TNR) αρνητικό. Η ακρίβεια της δοκιμασίας προέκυψε 46%.

X7

TPR	FPR	FNR	TNR
57,1%	38,3%	42,9%	61,7%

Πραγματική τιμή

Αποτέλεσμα Πρόβλεψης		**p**	**n**
	p'	4173	1159
	n'	3135	1867

Accuracy =58%

Οι τιμές της μεταβλητής X7 που είναι μεγαλύτερες του ΔΟ 107,5 θεωρούνται «επικίνδυνες» (υψηλής έντασης), ενώ αυτές που είναι μικρότερες του, θεωρούνται «ακίνδυνες» (χαμηλής έντασης). Για το συγκεκριμένο διαχωριστικό όριο, για το 57,1% (TPR) των υψηλών δονήσεων το αποτέλεσμα της δοκιμασίας θα είναι θετικό, ενώ για το 42,9% (FNR) αρνητικό. Επιπλέον για το 38,3% (FPR) των χαμηλών δονήσεων το αποτέλεσμα θα είναι θετικό ενώ για το 61,7% (TNR) αρνητικό. Η ακρίβεια της δοκιμασίας προέκυψε 58%.

X9

TPR	FPR	FNR	TNR
43,8%	26,8%	56,2%	73,2%

Πραγματική τιμή

Αποτέλεσμα Πρόβλεψης		**p**	**n**
	p'	3201	811
	n'	4107	2215

Accuracy =52%

Οι τιμές της μεταβλητής X9 που είναι μεγαλύτερες του ΔΟ 22,375 θεωρούνται «επικίνδυνες» (υψηλής έντασης), ενώ αυτές που είναι μικρότερες του, θεωρούνται «ακίνδυνες» (χαμηλής έντασης). Για το συγκεκριμένο διαχωριστικό όριο, για το 43,8% (TPR) των υψηλών δονήσεων το αποτέλεσμα της δοκιμασίας θα είναι θετικό, ενώ για το 56,2% (FNR) αρνητικό. Επιπλέον για το 26,8% (FPR) των χαμηλών δονήσεων το αποτέλεσμα θα είναι θετικό ενώ για το 73,2% (TNR) αρνητικό. Η ακρίβεια της δοκιμασίας προέκυψε 52%.

X10

TPR	FPR	FNR	TNR
64,2%	40,8%	35,8%	59,2%

Πραγματική τιμή

Αποτέλεσμα Πρόβλεψης		**p**	**n**
	p'	4692	1235
	n'	2616	1791

Accuracy =63%

Οι τιμές της μεταβλητής X10 που είναι μεγαλύτερες του ΔΟ 38,32 θεωρούνται «επικίνδυνες» (υψηλής έντασης), ενώ αυτές που είναι μικρότερες του, θεωρούνται «ακίνδυνες» (χαμηλής έντασης). Για το συγκεκριμένο διαχωριστικό όριο, για το 64,2% (TPR) των υψηλών δονήσεων το αποτέλεσμα της δοκιμασίας θα είναι θετικό, ενώ για το 35,8% (FNR) αρνητικό. Επιπλέον για το 40,8% (FPR) των χαμηλών δονήσεων το αποτέλεσμα θα είναι θετικό ενώ για το 59,2% (TNR) αρνητικό. Η ακρίβεια της δοκιμασίας προέκυψε 63%.

X11

TPR	FPR	FNR	TNR
57,9%	20,9%	42.1%	79,1%

Πραγματική τιμή

Αποτέλεσμα Πρόβλεψης		**p**	**n**
	p'	4231	632
	n'	3077	2394

Accuracy =64%

Οι τιμές της μεταβλητής X11 που είναι μεγαλύτερες του ΔΟ 24,5 θεωρούνται «επικίνδυνες» (υψηλής έντασης), ενώ αυτές που είναι μικρότερες του, θεωρούνται «ακίνδυνες» (χαμηλής έντασης). Για το συγκεκριμένο διαχωριστικό όριο, για το 57,9% (TPR) των υψηλών δονήσεων το αποτέλεσμα της δοκιμασίας θα είναι θετικό, ενώ για το 42,1% (FNR) αρνητικό. Επιπλέον για το 20,9% (FPR) των χαμηλών δονήσεων το αποτέλεσμα θα είναι θετικό ενώ για το 79,1% (TNR) αρνητικό. Η ακρίβεια της δοκιμασίας προέκυψε 64%.

ΣΥΜΠΕΡΑΣΜΑ

Από τα παραπάνω αποτελέσματα μπορούμε να συμπεράνουμε ότι οι μεταβλητές που χαρακτηρίζονται με υψηλότερη ακρίβεια είναι οι: X1, X11 και X10. Αυτοί που

χαρακτηρίζονται με υψηλότερη ευαισθησία είναι οι : X1 και X10 και με υψηλότερη ειδικότητα είναι οι : X5, X11, X2 και X9.

ΠΑΡΑΡΤΗΜΑ Α

Algorithm 1 Conceptual method for calculating an ROC curve. See algorithm 2 for a practical method.

Inputs: L, the set of test instances; $f(i)$, the probabilistic classifier's estimate that instance i is positive; *min* and *max*, the smallest and largest values returned by f; *increment*, the smallest difference between any two f values.

```
for t = min to max by increment do
  FP ⇐ 0
  TP ⇐ 0
  for i ∈ L do
    if f(i) ≥ t then                                   /* This example is over threshold */
      if i is a positive example then
        TP ⇐ TP + 1
      else                                             /* i is a negative example, so this is a false positive */
        FP ⇐ FP + 1
  Add point (FP/N, TP/P) to ROC curve
end
```

ΑΛΓΟΡΙΘΜΟΣ 1.Ενοιολογική μέθοδος για τον υπολογισμό μιας ROC καμπύλης

Algorithm 2 Practical method for calculating an ROC curve from a test set

Inputs: L, the set of test instances; $f(i)$, the probabilistic classifier's estimate that instance i is positive.
Outputs: R, a list of ROC points from (0,0) to (1,1)

```
L_sorted ⇐ L sorted decreasing by f scores
FP ⇐ 0
TP ⇐ 0
R ⇐ ⟨⟩
f_prev ⇐ −∞
for i ∈ L_sorted do
  if f(i) ≠ f_prev then
    ADD_POINT((FP/N, TP/P), R)
    f_prev ⇐ f(i)
  if i is a positive example then
    TP ⇐ TP + 1
  else                                                 /* i is a negative example, so this is a false positive */
    FP ⇐ FP + 1
ADD_POINT((FP/N, TP/P), R)
end
```

```
subroutine ADD_POINT(P, R)
push P onto R
end subroutine
```

ΑΛΓΟΡΙΘΜΟΣ 2.Πρακτική μέθοδος για τον υπολογισμό μιας ROC καμπύλης από ένα σύνολο δεδομένων ενός τεστ.

Algorithm 3 Modifications to algorithm 2 to avoid introducing concavities.

```
1: subroutine ADD_POINT(P, R)
2: loop
3:   if |R| < 2 then
4:     push P onto R
5:     return
6:   else
7:     T ⇐ pop(R)
8:     T2 ⇐ top_of_stack(R)
9:     if SLOPE(T2, T) < SLOPE(T, P) then
10:      push T onto R
11:      push P onto R
12:      return
13: end subroutine
```

ΑΛΓΟΡΙΘΜΟΣ 3.Τροποποιήσεις στον αλγόριθμο 2 προκειμένου να αποφευχθεί η εισαγωγή κοιλότητας.

Algorithm 4 Calculating the area under an ROC curve

Inputs: L, the set of test instances; $f(i)$, the probabilistic classifier's estimate that instance i is positive.
Outputs: A, the area under the ROC curve.

```
1: L_sorted ⇐ L sorted decreasing by f scores
2: FP ⇐ TP ⇐ 0
3: FP_prev ⇐ TP_prev ⇐ 0
4: A ⇐ 0
5: f_prev ⇐ −∞
6: for i ∈ L_sorted do
7:   if f(i) ≠ f_prev then
8:     A ⇐ A + TRAP_AREA(FP, FP_prev, TP, TP_prev)        /* See A.3 for TRAP_AREA */
9:     f_prev ⇐ f(i)
10:    FP_prev ⇐ FP
11:    TP_prev ⇐ TP
12:  if i is a positive example then
13:    TP ⇐ TP + 1
14:  else
15:    FP ⇐ FP + 1
16: A ⇐ A + TRAP_AREA(1, FP_prev, 1, TP_prev)
17: A ⇐ A/(P · N)                                          /* scale from P × N onto the unit square */
18: end
```

ΑΛΓΟΡΙΘΜΟΣ 4. Υπολογισμός περιοχής κάτω από μία ROC καμπύλη.

Algorithm 5 Vertical averaging of ROC curves.

Inputs: *samples*, the number of FP samples; *nrocs*, the number of ROC curves to be sampled, $ROCS[nrocs]$, an array of *nrocs* ROC curves; $npts[m]$, the number of points in ROC curve m. Each ROC point is a structure of two members, FP and TP, whose values are referenced by subscripts here.

Output: Array $TPavg$, containing the vertical (TP) averages.

$s \Leftarrow 1$
for $FP_{sample} = 0$ **to** 1 **by** $1/samples$ **do**
  $TPsum \Leftarrow 0$
  for $i = 1$ **to** nrocs **do**
    $TPsum \Leftarrow TPsum +$ TP_FOR_FP$(FP_{sample}, ROCS[i], npts[i])$
  $TPavg[s] \Leftarrow TPsum/i$
  $s \Leftarrow s + 1$
end

function TP_FOR_FP$(FP_{sample}, ROC, npts)$
$i \Leftarrow 1$
while $i < npts$ **and** $ROC[i+1]_{FP} \leq FP_{sample}$ **do**
  $i \Leftarrow i + 1$
if $ROC[i]_{FP} = FP_{sample}$ **then**
  return $ROC[i]_{TP}$
else if $ROC[i]_{FP} < FP_{sample}$ **then**
  return INTERPOLATE$(ROC[i], ROC[i+1], FP_{sample})$
end function

ΑΛΓΟΡΙΘΜΟΣ 5.Κάθετη μέθοδος υπολογισμού μέσου όρουROC καμπυλών.

Algorithm 6 Threshold averaging of ROC curves.

Inputs: *samples*, the number of threshold samples; *nrocs*, the number of ROC curves to be sampled; $ROCS[nrocs]$, an array of *nrocs* ROC curves; $npts[m]$, the number of points in ROC curve m. Each ROC point is a structure of
ep instructions for common features
P and Score, whose values are referenced by subscripts here.

Output: *Avg*, an array of (X,Y) points constituting the average ROC curve.

$T \Leftarrow$ all Scores of all ROC points
sort T in descending order
$s \Leftarrow 1$
for $tidx = 1$ **to** $length(T)$ **by** $int(length(T)/samples)$ **do**
  $FPsum \Leftarrow 0$
  $TPsum \Leftarrow 0$
  for $i = 1$ **to** nrocs **do**
    $p \Leftarrow$ POINT_AT_THRESH$(ROCS[i], npts[i], T[tidx])$
    $FPsum \Leftarrow FPsum + p_{FP}$
    $TPsum \Leftarrow TPsum + p_{TP}$
  $Avg[s] \Leftarrow (FPsum/i\ , TPsum/i)$
  $s \Leftarrow s + 1$
end

function POINT_AT_THRESH$(ROC, npts, thresh)$
$i \Leftarrow 1$
while $i < npts$ **and** $ROC[i]_{Score} > thresh$ **do**
  $i \Leftarrow i + 1$
return $ROC[i]$
end function

ΑΛΓΟΡΙΘΜΟΣ 6. Υπολογισμός μέσου όρου ROC καμπυλών με τη μέθοδο του ορίου.

ΠΑΡΑΡΤΗΜΑ Β

ΒΑΣΙΚΕΣ ΣΥΝΑΡΤΗΣΕΙΣ ΓΕΩΜΕΤΡΙΑΣ

Οι παραπάνω αλγόριθμοι κάνουν χρήση των ακόλουθων συναρτήσεων γεωμετρίας:

Β.1 ΚΛΙΣΗ ΓΡΑΜΜΗΣ

Η κλίση μιας γραμμής ορίζεται ως η αναλογία(ποσοστό) της μεταβολής του y ως προς τη μεταβολή του x (διαφορετικά γνωστή ως "riseoverrun"). Αυτή μπορεί να υπολογιστεί απευθείας με εξαίρεση την περίπτωση των άπειρων κλίσεων.

1: **function** SLOPE(P, Q)
2: **if** $P_x = Q_x$ **then**
3: return ∞
4: **else**
5: return $\frac{P_y - Q_y}{P_x - Q_x}$
6: **end function**

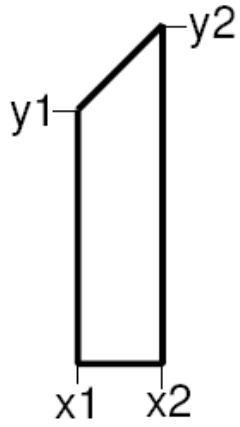
Β.2 ΓΡΑΜΜΙΚΗ ΠΑΡΕΜΒΟΛΗ ΑΝΑΜΕΣΑ ΣΕ ΔΥΟ ΣΗΜΕΙΑ

Δοσμένων δύο σημείων, P1 και P2 και μιας τιμής X, παρεμβάλλουμε την τιμή Υπου αντιστοιχεί στο X. Έτσι εξασφαλίζεται ότι η τιμή Χθα τοποθετηθεί μεταξύ των σημείων P1 και P2 και ότι θα ισχύει P1x ≠P2x.

1: **function** INTERPOLATE$(P1, P2, X)$
2: $\Delta x = P2_X - P1_X$
3: $\Delta y = P2_Y - P1_Y$
4: $m = \Delta y / \Delta x$
5: **return** $P1_Y + m \cdot (X - P1_X)$
6: **end function**

Β.3ΕΜΒΑΔΟΝ ΕΝΟΣ ΤΡΑΠΕΖΙΟΥ

Το εμβαδόν είναι απλά το γινόμενο του μέσου ύψους επί του πλάτους της βάσης.



1: **function** TRAP_AREA$(X1, X2, Y1, Y2)$
2: $Base \Leftarrow |X1 - X2|$
3: $Height_{avg} \Leftarrow (Y1 + Y2)/2$
4: **return** $Base \times Height_{avg}$
5: **end function**

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Adams, N. M., & Hand, D. J. (1999). Comparing classifiers when the misallocations costs are uncertain. *Pattern Recognition*, 32, 1139-1147.
2. Box G.E.P, Cox D.R, 1964. An analysis of transformations. Journal *of the Royal Statistical Society*, Series B; 26, 211–243.
3. Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30 (7), 1145-1159.
4. Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Wadsworth International Group, Belmont, CA.
5. Barber, C., Dobkin, D., & Huhdanpaa, H. (1993). The quickhull algorithm for convex hull. Tech. rep. GCG53, University of Minnesota. Available: ftp://geom.umn.edu/pub/software/qhull.tar.Z.
6. Bamber DC, (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*; 12, 387–415.
7. Clearwater, S., & Stern, E. (1991). A rule-learning program in high energy physics event classification. *Comp Physics Comm*, 67, 159-182.
8. Domingos, P. (1999). MetaCost: A general method for making classifiers cost-sensitive. *In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155-164.
9. Dreiseitl, S., Ohno-Machado, L., & Binder, M. (2000). Comparing three-class diagnostic tests by three-way ROC analysis. *Medical Decision Making*, 20, 323-331.
10. Drummond, C., & Holte, R. C. (2000). Explicitly representing expected cost: An alternative to ROC representation. In Ramakrishnan, R., &Stolfo, S. (Eds.), *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 198-207.ACM Press.
11. Egan, J. P. (1975). Signal Detection Theory and ROC Analysis. Series in Cognitition and Perception. Academic Press, New York.
12. Fawcett, T. (2001). Using rule sets to maximize ROC performance. In *Proceedings of the IEEE International Conference on Data Mining (ICDM-2001)*, pp. 131-138.
13. Fawcett, T., & Provost, F. (1996).Combining data mining and machine learning for effective user profiling. *In Simoudis, Han, & Fayyad (Eds.), Proceedings on the Second International Conference on Knowledge Discovery and Data Minin*g, pp. 8-13 Menlo Park, CA. AAAI Press.
14. Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. *DataMining and Knowledge Discovery*, 1 (3), 291-316.
15. Forman, G. (2002). A method for discovering the insignificance of one's best classifier and the unlearnability of a classification task. *In Lavrac, Motoda, & Fawcett (Eds.), Proceedings of th First International Workshop on Data Mining Lessons Learned (DMLL-2002*).
16. Green DM, Swets JA., (1966). Signal Detection Theory and Psychophysics. Wiley: New York,
17. Goddard MJ, Hinberg I. (1990). Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. *Statistics in Medicine*; 9, 325–337.

18. Hanley JA, McNeil BJ. (1982). The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology*, 143, 29–36.
19. Hajian-Tilaki KO, Hanley JA, Joseph L, Collet J. (2002). A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. *Statist. Med.*, 21, 3093–3106
20. Hsiao JK, Bartko JJ, Potter WZ. (1989). Diagnosing diagnoses receiver operating characteristic methods and psychiatry. *Archives of General Psychiatry*; 46, 664–667.
21. Hand, D. J., & Till, R. J. (2001). A simple generalization of the area under the ROC curve to multiple class classification problems. *Machine Learning*, 45 (2), 171-186.
22. Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29-36.
23. Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30 (2-3), 195–215.
24. Kramar A, Faraggi D, Ychou M, Reiser B, Grenier J. Criteres, (1999). ROC generalises pour l'evaluation de plusieurs marqueurstumoraux. *Revue d'Epidemiologieei de SantePublique*; 47, 217–226.
25. Lloyd CJ. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*; 93,1356–1364.
26. Lloyd CJ, Yong Z. (1999). Kernel estimators of the ROC curves are better than empirical. *Statistics and Probability Letters*, 44, 221–228.
27. Lewis, D. (1990). Representation quality in text classification: An introduction and experiment. *In Proceedings of Workshop on Speech and Natural Language*, pp. 288-295 Hidden Valley, PA. Morgan Kaufmann.
28. Lewis, D. (1991). Evaluating text categorization. *In Proceedings of Speech and Natural Language Workshop*, pp. 312-318. Morgan Kaufmann.
29. Lane, T. (2000). Extensions of ROC analysis to multi-class domains. In Dietterich, T., Margineantu, D., Provost, F., & Turney, P. (Eds.), *ICML-2000 Workshop on Cost-Sensitive Learning*.
30. Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). *The DET curve in assessment of detection task performance. In Proc. Eurospeech '97*, pp. 1895-1898 Rhodes, Greece.
31. Mossman, D. (1999). Three-way ROCs. *Medical Decision Making*, 19, 78-89.
32. Metz CE. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigation Radiology*, 24, 234–245.
33. McCool JI. (1991). Inference on P(Y˝X) in the Weibull case. *Communications in Statistics – Simulation and Computation*, 20, 129–148.
34. Metz CE, Herman BA, Shen J. (1998). Maximum likelihood estimator of receiver operator characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, 17, 1033–1053.
35. Nockemann C, Heidt H, Thomsen N. (1991). Reliability in NDT: ROC study of radiographic weld inspections. *Nondestructive Testing and Evaluation International*, 24, 235–245.
36. Percy ME, Andrews DF, Thompson MW. (1982). Duchene muscular dystrophy carrier detection using logistic discrimination: serum creatine kinase, hemopexin, pyravite kinase and lactate dehydrogeneze in combination. *American Journal of Medical Genetics*; 13, 27–38.

37. Provost, F., & Domingos, P. (2001). Well-trained PETs: Improving probability estimation trees. *CeDER working paper #IS-00-04*, Stern School of Business, New York University, NY, NY 10012.
38. Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, 42 (3), 203-231.
39. Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *In Shavlik, J. (Ed.), Proceedings of the Fifteenth International Conference on Machine Learning*,pp. 445-453 San Francisco, CA. Morgan Kaufmann.
40. Reiser B, Faraggi D. (2002). Confidence intervals for the generalized ROC criterion. *Statist. Med*., 21:3093–3106.
41. Stine RA, Heyse JF. (2001). Nonparametric estimates of overlap. *Statistics in Medicine*, 20, 215–236.
42. Schisterman E. (1999). Lipid peroxidation and antioxidant biomarkers and biomarker disease. PhD thesis, State University of New York, Buffalo.
43. Silverman BW. (1986). Density Estimator for Statistics and Data Analysis. Chapman and Hall: London.
44. Shapiro DE. (1999). The interpretation of diagnostic tests. *Statistical Methods in Medical Research*, 8, 113–134.
45. Spackman, K. A. (1989). Signal detection theory: Valuable tools for evaluating inductive learning. *In Proceedings of the Sixth International Workshop on Machine Learning*, pp. 160-163 San Mateo, CA. Morgan Kaufman.
46. Srinivasan, A. (1999). Note on the location of optimal classifiers in n-dimensional ROC space. *Technical report PRG-TR-2-99*, Oxford University Computing Laboratory, Oxford, England.
47. Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240, 1285-1293.
48. Swets, J. A., Dawes, R. M., & Monahan, J. (2000a). Better decisions through science. *Scientific American*, 283, 82-87. Available:http://www.psychologicalscience.org/newsresearch/publications/journals/%siam.pdf.
49. Swets, J. A., Dawes, R. M., & Monahan, J. (2000b). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1 (1), 1-26.
50. Wand MP, Jones MC. (1995). Kernel Smoothing. Chapman and Hall: London.
51. Wolfe DA, Hogg RV. (1971). On constructing statistics and reporting data. *American Statistician*; 25, 27–30.
52. Zou KH, Hall WJ, Shapiro DE. (1997). Smooth non-parametric receiver-operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, 16, 2143–2156.
53. Zou KH, Tempany CM, Fielding JR, Silverman SG. (1998). Original smooth receiver operating characteristic curves estimation from continuous data: statistical methods for analyzing the predictive value of spiral CT of ureteral stones. *Academic Radiology*, 5, 680–687.
54. Zou, K. H. (2002). Receiver operating characteristic (ROC) literature research. On-line bibliography available from http://splweb.bwh.harvard.edu:8000/pages/ppl/zou/roc.html.