Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Χημικών Μηχανικών
Τομέας ΙΙ: Ανάλυσης, Σχεδιασμού και
Ανάπτυξης Διεργασιών και Συστημάτων

Μελέτη και σχεδιασμός μεθόδων Εξόρυξης Δεδομένων

και εφαρμογές σε προβλήματα Μεταβολομικής

Διπλωματική εργασία
19 Φεβρουαρίου 2014
Επιβλέπων: Χ. Σαρίμβεης

Γεράσιμος Α. Χουρδάκης

ΠΕΡΙΛΗΨΗ

- Μελετήσαμε μερικούς από τους συνηθέστερους αλγορίθμους μηχανικής μάθησης και τους εφαρμόσαμε σε πραγματικά προβλήματα Μεταβολομικής.
- **Πρόβλημα 1:** Δημοσιευμένα δεδομένα σχετικά με την AKI (οξεία νεφρική βλάβη). 2 κλάσεις (υγιείς-ασθενείς), 106 δείγματα, 701 περιοχές NMR.
- **Πρόβλημα 2:** Δεδομένα του Τμήματος Φαρμακευτικής ΕΚΠΑ σχετικά με τη χορήγηση ολευρωπαΐνης για την αντιμετώπιση της καρδιοτοξικής δράσης της αδριαμυκίνης (DXR). 6 κλάσεις, 40 δείγματα, 38 περιοχές NMR.
- Πέρα από το ήδη διαθέσιμο λογισμικό, **αναπτύχθηκε ένας κώδικας για επιλογή μεταβλητών** με γενετική έρευνα και SVM, καθώς και ένα υπολογιστικό φύλλο για την αξιολόγηση συναινετικών μοντέλων.

ΠΕΡΙΛΗΨΗ

- Μελετήσαμε μερικούς από τους συνηθέστερους αλγορίθμους μηχανικής μάθησης και τους εφαρμόσαμε σε πραγματικά προβλήματα Μεταβολομικής.
- **Πρόβλημα 1:** Δημοσιευμένα δεδομένα σχετικά με την AKI (οξεία νεφρική βλάβη). 2 κλάσεις (υγιείς-ασθενείς), 106 δείγματα, 701 περιοχές NMR.
- **Πρόβλημα 2:** Δεδομένα του Τμήματος Φαρμακευτικής ΕΚΠΑ σχετικά με τη χορήγηση ολευρωπαΐνης για την αντιμετώπιση της καρδιοτοξικής δράσης της αδριαμυκίνης (DXR). 6 κλάσεις, 40 δείγματα, 38 περιοχές NMR.
- Πέρα από το ήδη διαθέσιμο λογισμικό, **αναπτύχθηκε ένας κώδικας για επιλογή μεταβλητών** με γενετική έρευνα και SVM, καθώς και ένα υπολογιστικό φύλλο για την αξιολόγηση συναινετικών μοντέλων.

ΠΕΡΙΛΗΨΗ

- Μελετήσαμε μερικούς από τους συνηθέστερους αλγορίθμους μηχανικής μάθησης και τους εφαρμόσαμε σε πραγματικά προβλήματα Μεταβολομικής.
- **Πρόβλημα 1:** Δημοσιευμένα δεδομένα σχετικά με την AKI (οξεία νεφρική βλάβη). 2 κλάσεις (υγιείς-ασθενείς), 106 δείγματα, 701 περιοχές NMR.
- **Πρόβλημα 2:** Δεδομένα του Τμήματος Φαρμακευτικής ΕΚΠΑ σχετικά με τη χορήγηση ολευρωπαΐνης για την αντιμετώπιση της καρδιοτοξικής δράσης της αδριαμυκίνης (DXR). 6 κλάσεις, 40 δείγματα, 38 περιοχές NMR.
- Πέρα από το ήδη διαθέσιμο λογισμικό, **αναπτύχθηκε ένας κώδικας για επιλογή μεταβλητών** με γενετική έρευνα και SVM, καθώς και ένα υπολογιστικό φύλλο για την αξιολόγηση συναινετικών μοντέλων.

ΠΕΡΙΛΗΨΗ

- Μελετήσαμε μερικούς από τους συνηθέστερους αλγορίθμους μηχανικής μάθησης και τους εφαρμόσαμε σε πραγματικά προβλήματα Μεταβολομικής.
- **Πρόβλημα 1:** Δημοσιευμένα δεδομένα σχετικά με την AKI (οξεία νεφρική βλάβη). 2 κλάσεις (υγιείς-ασθενείς), 106 δείγματα, 701 περιοχές NMR.
- **Πρόβλημα 2:** Δεδομένα του Τμήματος Φαρμακευτικής ΕΚΠΑ σχετικά με τη χορήγηση ολευρωπαΐνης για την αντιμετώπιση της καρδιοτοξικής δράσης της αδριαμυκίνης (DXR). 6 κλάσεις, 40 δείγματα, 38 περιοχές NMR.
- Πέρα από το ήδη διαθέσιμο λογισμικό, **αναπτύχθηκε ένας κώδικας για επιλογή μεταβλητών** με γενετική έρευνα και SVM, καθώς και ένα υπολογιστικό φύλλο για την αξιολόγηση συναινετικών μοντέλων.

ΑΓΝΩΣΤΕΣ ΛΕΞΕΙΣ;

ΜΕΤΑΒΟΛΟΜΙΚΗ

ΜΕΤΑΒΟΛΟΜΙΚΗ

Μπορούμε να εξάγουμε χρήσιμες πληροφορίες από τα προϊόντα μεταβολισμού κυττάρων και οργανισμών;

- Πλήρης ανάλυση μεταβολικών υγρών (NMR, MS, …).
- Φάσματα → δεκάδες ή **εκατοντάδες μεταβλητές**!
- Κάθε άτομο έχει το δικό του μεταβολισμό!
- Συνήθως εξετάζουμε δεκάδες διαφορετικά άτομα-δείγματα!
- Εφαρμόζουμε **μεθόδους πολυπαραμετρικής στατιστικής επεξεργασίας** (PCA, PLS-DA, ανάλυση συσχέτισης, …).

ΜΕΤΑΒΟΛΟΜΙΚΗ

Μπορούμε να εξάγουμε χρήσιμες πληροφορίες από τα προϊόντα μεταβολισμού κυττάρων και οργανισμών;

- Πλήρης ανάλυση μεταβολικών υγρών (NMR, MS, …).
- Φάσματα → δεκάδες ή **εκατοντάδες μεταβλητές**!
- Κάθε άτομο έχει το δικό του μεταβολισμό!
- Συνήθως εξετάζουμε δεκάδες διαφορετικά άτομα-δείγματα!
- Εφαρμόζουμε **μεθόδους πολυπαραμετρικής στατιστικής επεξεργασίας** (PCA, PLS-DA, ανάλυση συσχέτισης, …).

ΜΕΤΑΒΟΛΟΜΙΚΗ

Μπορούμε να εξάγουμε χρήσιμες πληροφορίες από τα προϊόντα μεταβολισμού κυττάρων και οργανισμών;

- Πλήρης ανάλυση μεταβολικών υγρών (NMR, MS, …).
- Φάσματα → δεκάδες ή **εκατοντάδες μεταβλητές**!
- Κάθε άτομο έχει το δικό του μεταβολισμό!
- Συνήθως εξετάζουμε δεκάδες διαφορετικά άτομα-δείγματα!
- Εφαρμόζουμε **μεθόδους πολυπαραμετρικής στατιστικής επεξεργασίας** (PCA, PLS-DA, ανάλυση συσχέτισης, …).

ΜΕΤΑΒΟΛΟΜΙΚΗ

Μπορούμε να εξάγουμε χρήσιμες πληροφορίες από τα προϊόντα μεταβολισμού κυττάρων και οργανισμών;

- Πλήρης ανάλυση μεταβολικών υγρών (NMR, MS, …).
- Φάσματα → δεκάδες ή **εκατοντάδες μεταβλητές**!
- Κάθε άτομο έχει το δικό του μεταβολισμό!
- Συνήθως εξετάζουμε δεκάδες διαφορετικά άτομα-δείγματα!
- Εφαρμόζουμε **μεθόδους πολυπαραμετρικής στατιστικής επεξεργασίας** (PCA, PLS-DA, ανάλυση συσχέτισης, …).

ΜΕΤΑΒΟΛΟΜΙΚΗ

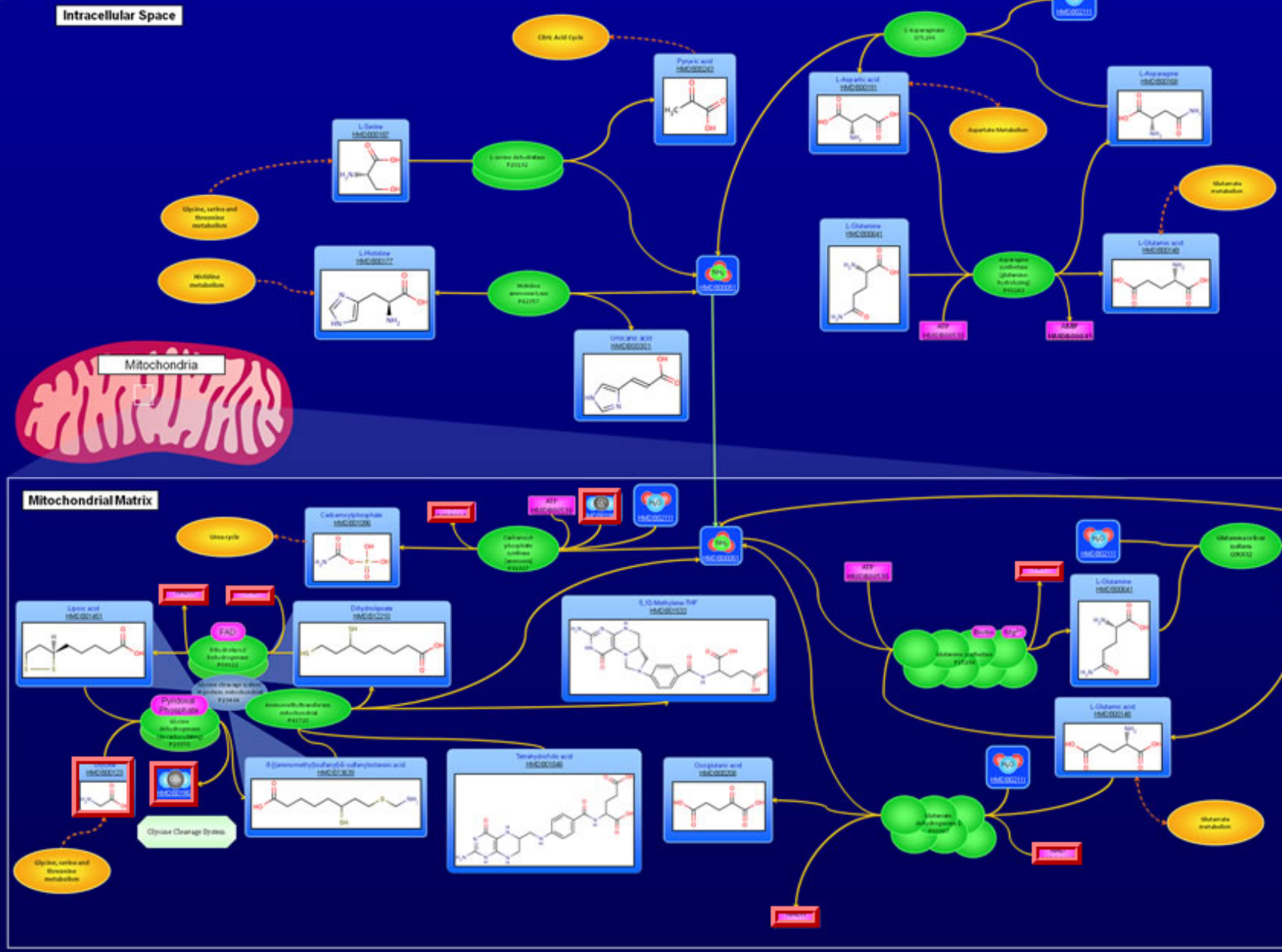
Μπορούμε να εξάγουμε χρήσιμες πληροφορίες από τα προϊόντα μεταβολισμού κυττάρων και οργανισμών;

- Πλήρης ανάλυση μεταβολικών υγρών (NMR, MS, …).
- Φάσματα → δεκάδες ή **εκατοντάδες μεταβλητές**!
- Κάθε άτομο έχει το δικό του μεταβολισμό!
- Συνήθως εξετάζουμε δεκάδες διαφορετικά άτομα-δείγματα!
- Εφαρμόζουμε **μεθόδους πολυπαραμετρικής στατιστικής επεξεργασίας** (PCA, PLS-DA, ανάλυση συσχέτισης, …).

ΜΕΤΑΒΟΛΟΜΙΚΗ

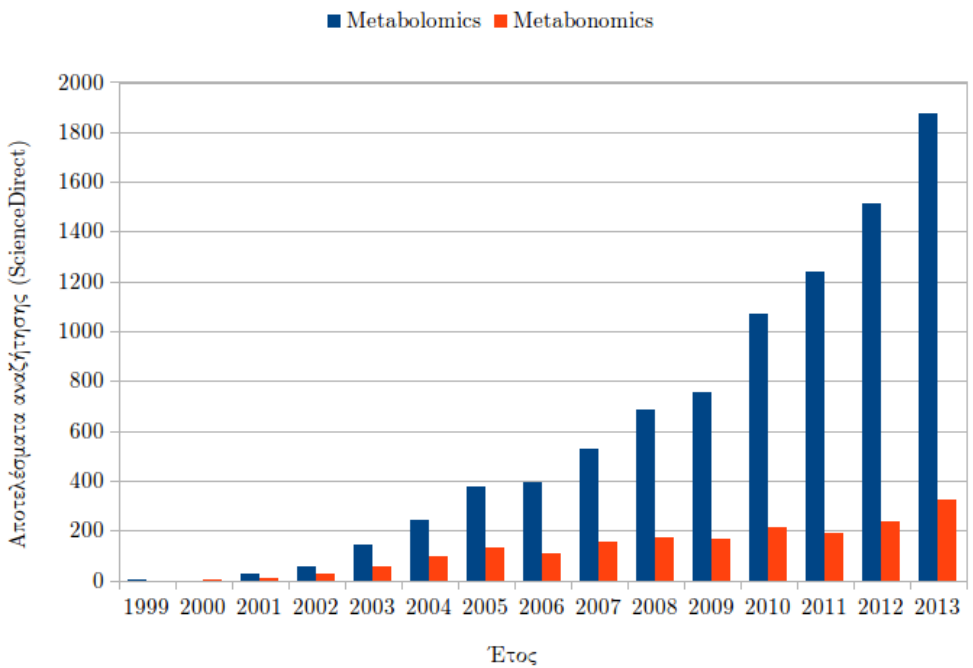
Μπορούμε να εξάγουμε χρήσιμες πληροφορίες από τα προϊόντα μεταβολισμού κυττάρων και οργανισμών;

- Πλήρης ανάλυση μεταβολικών υγρών (NMR, MS, …).
- Φάσματα → δεκάδες ή **εκατοντάδες μεταβλητές**!
- Κάθε άτομο έχει το δικό του μεταβολισμό!
- Συνήθως εξετάζουμε δεκάδες διαφορετικά άτομα-δείγματα!
- Εφαρμόζουμε **μεθόδους πολυπαραμετρικής στατιστικής επεξεργασίας** (PCA, PLS-DA, ανάλυση συσχέτισης, …).


Intracellular Space
Citric Acid Cycle
Pyruvic acid
L-Serine
Glycine, serine and threonine metabolism
L-Methionine
Methionine metabolism
Uricacic acid
L-Aspartic acid
Aspartate Metabolism
L-Asparagine
L-Glutamine
Glutamate metabolism
L-Glutamic acid
ATP
GTP
Mitochondria
Mitochondrial Matrix
Urea cycle
Carbamoylphosphate
Lipoic acid
Dihydrolipoate
5,10-Methylene-THF
FAD
Pyridoxal Phosphate
Glycine Cleavage System
Tetrahydrofolate acid
Oxoglutaric acid
Glutamate metabolism
Glycine, serine and threonine metabolism

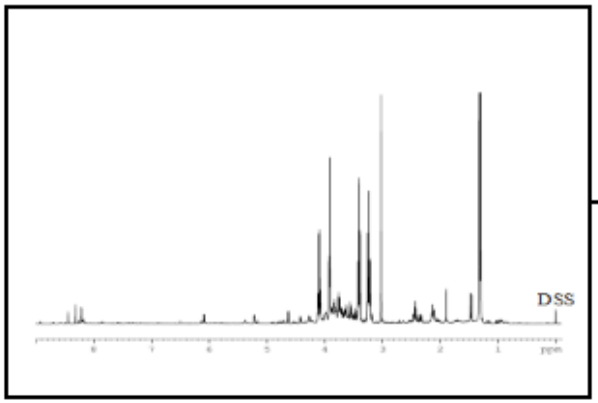
ΝΕΟΣ ΧΩΡΟΣ!

Ο όρος πρωτοεμφανίζεται το 1999 από το J. Nicholson και έκτοτε παρατηρείται ραγδαία αύξηση στο πλήθος των σχετικών δημοσιεύσεων.

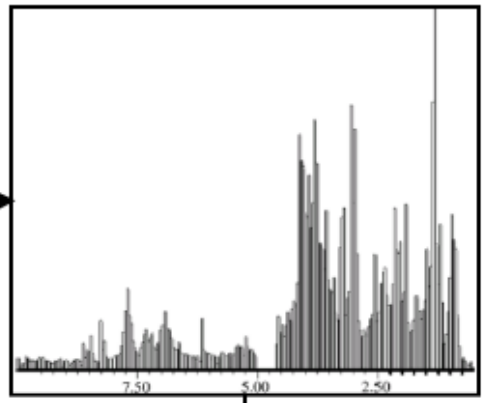


ΠΩΣ ΑΠΟΚΤΟΥΜΕ ΔΕΔΟΜΕΝΑ;

Φάσμα NMR



Διακριτοποίηση

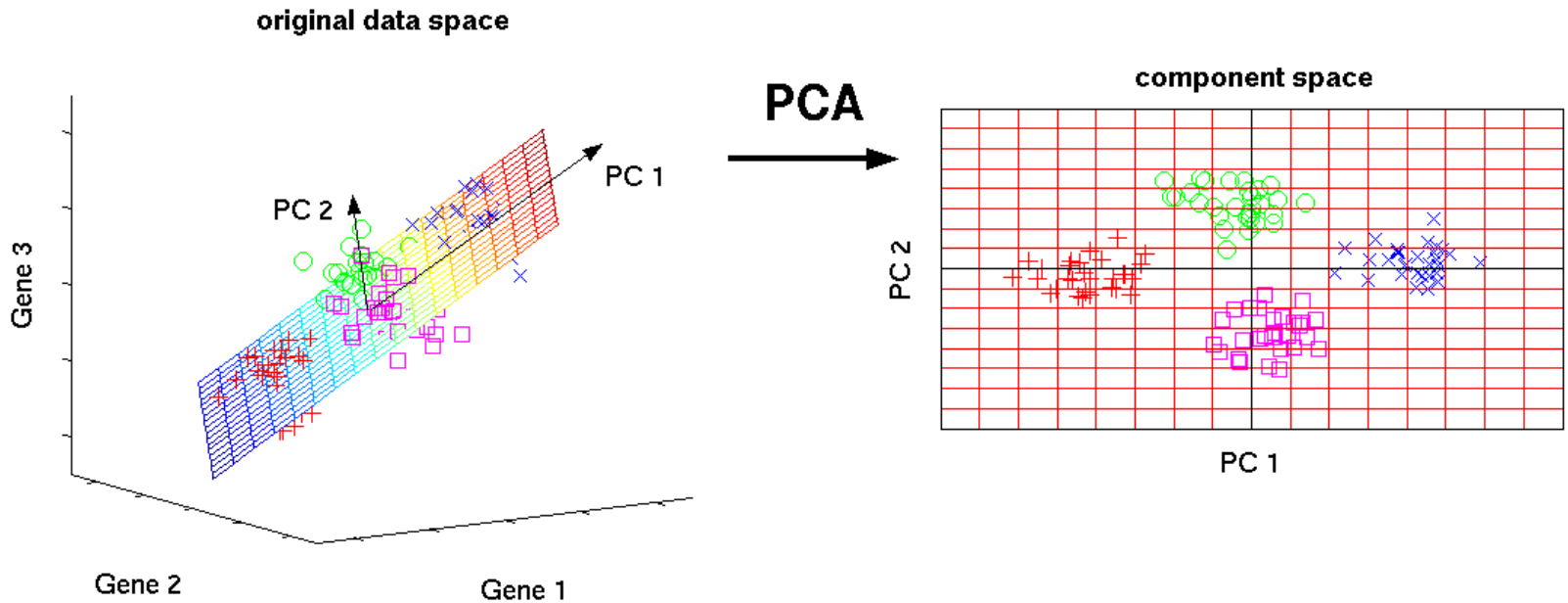


Πίνακας δεδομένων

patient	class	1	2	...	699	700	701
AKI_8_24_97_110812	Biopsy kidney normal	-0.002437107	-0.0035140763	...	-0.0006630864	-0.0005720405	-0.0006858139
AKI_8_24_100_110812	Biopsy kidney normal	-0.0013475575	-0.0035747009	...	-0.0006320954	-0.0006858139	-0.0008425529
AKI_8_24_101_110812	Biopsy kidney normal	-0.0060549417	-0.0046379988	...	-0.000791196	-0.0009159127	-0.0008425529
AKI_8_24_102_110812	Biopsy kidney normal	-0.0049196483	-0.0017621665	...	-0.0008894464	-0.0010515443	-0.0010613955
AKI_8_24_104_110812	Biopsy kidney normal	-0.0037098023	-0.0046379988	...	-0.0008737839	-0.0008425529	-0.0009611175
AKI_8_24_105_110812	Biopsy kidney normal	-0.0049196483	-0.0032075154	...	-0.0006630864	-0.0009376947	-0.0010613955
AKI_8_24_107_110812	Biopsy kidney normal	-0.0032075154	-0.0088158057	...	-0.0010150392	-0.0009611175	-0.0011037445
AKI_8_24_108_110812	Biopsy kidney normal	-0.0023361113	-0.0035747009	...	-0.0007549861	-0.0009749736	-0.0010515443
AKI_8_24_109_110812	Biopsy kidney normal	-0.0017972075	-0.0029727493	...	-0.000231685	-0.0006630864	-0.0004138545
AKI_8_24_110_110812	Biopsy kidney normal	-0.0040793693	-0.0026217018	...	-0.0006320954	-0.0008894464	-0.0008425529
...	...	...	...	...	...	...	...
AKI_8_24_02_110722	Acute Kidney Injury	-0.0060549417	-0.0066442452	...	-0.0007197942	-0.0007549861	-0.0008894464
AKI_8_24_04_110722	Acute Kidney Injury	-0.0028694225	-0.0049196483	...	-0.0009749736	-0.0011497981	-0.0014832325
AKI_8_24_05_110722	Acute Kidney Injury	-0.0030599754	-0.0017972075	...	-0.000791196	-0.0009611175	-0.0011037445
AKI_8_24_06_110722	Acute Kidney Injury	-0.0055588469	-0.0039890442	...	-0.0005504O9	-0.0006630864	-0.0006858139
AKI_8_24_08_110722	Acute Kidney Injury	-0.0032075154	-0.0035140763	...	0.0005251104	0.0003348619	0.0003175396

PCA: ΜΙΑ ΙΔΑΝΙΚΗ ΠΕΡΙΠΤΩΣΗ

Από τις διαθέσιμες μεταβλητές παράγουμε τα **κυρίαρχα συστατικά** (principal components) και απεικονίζουμε τα δεδομένα σε 2D ή 3D διαγράμματα προσπαθώντας να αναπαραστήσουμε το σημαντικότερο μέρος της διασποράς χρησιμοποιώντας λιγότερους άξονες.



ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

ΠΡΟΒΛΗΜΑ ΤΑΞΙΝΟΜΗΣΗΣ

Σε κάθε πρόβλημα ταξινόμησης (classification) έχουμε στη διάθεσή μας έναν πίνακα ο οποίος:

- έχει στις **γραμμές** του παραδείγματα εκπαίδευσης,
- έχει στις **στήλες** του τις τιμές μετρούμενων μεταβλητών για κάθε παράδειγμα,
- έχει σε μια επιπλέον στήλη τη γνωστή **κλάση** του κάθε δείγματος.

Σκοπός είναι να παράξουμε μοντέλα τα οποία να προβλέπουν τις κλάσεις νέων δειγμάτων.

ΠΡΟΒΛΗΜΑ ΤΑΞΙΝΟΜΗΣΗΣ

Σε κάθε πρόβλημα ταξινόμησης (classification) έχουμε στη διάθεσή μας έναν πίνακα ο οποίος:

- έχει στις **γραμμές** του παραδείγματα εκπαίδευσης,
- έχει στις **στήλες** του τις τιμές μετρούμενων μεταβλητών για κάθε παράδειγμα,
- έχει σε μια επιπλέον στήλη τη γνωστή **κλάση** του κάθε δείγματος.

Σκοπός είναι να παράξουμε μοντέλα τα οποία να προβλέπουν τις κλάσεις νέων δειγμάτων.

ΠΡΟΒΛΗΜΑ ΤΑΞΙΝΟΜΗΣΗΣ

Σε κάθε πρόβλημα ταξινόμησης (classification) έχουμε στη διάθεσή μας έναν πίνακα ο οποίος:

- έχει στις **γραμμές** του παραδείγματα εκπαίδευσης,
- έχει στις **στήλες** του τις τιμές μετρούμενων μεταβλητών για κάθε παράδειγμα,
- έχει σε μια επιπλέον στήλη τη γνωστή **κλάση** του κάθε δείγματος.

Σκοπός είναι να παράξουμε μοντέλα τα οποία να προβλέπουν τις κλάσεις νέων δειγμάτων.

ΠΡΟΒΛΗΜΑ ΤΑΞΙΝΟΜΗΣΗΣ

Σε κάθε πρόβλημα ταξινόμησης (classification) έχουμε στη διάθεσή μας έναν πίνακα ο οποίος:

- έχει στις **γραμμές** του παραδείγματα εκπαίδευσης,
- έχει στις **στήλες** του τις τιμές μετρούμενων μεταβλητών για κάθε παράδειγμα,
- έχει σε μια επιπλέον στήλη τη γνωστή **κλάση** του κάθε δείγματος.

Σκοπός είναι να παράξουμε μοντέλα τα οποία να προβλέπουν τις κλάσεις νέων δειγμάτων.

ΠΡΟΒΛΗΜΑ ΤΑΞΙΝΟΜΗΣΗΣ

Σε κάθε πρόβλημα ταξινόμησης (classification) έχουμε στη διάθεσή μας έναν πίνακα ο οποίος:

- έχει στις **γραμμές** του παραδείγματα εκπαίδευσης,
- έχει στις **στήλες** του τις τιμές μετρούμενων μεταβλητών για κάθε παράδειγμα,
- έχει σε μια επιπλέον στήλη τη γνωστή **κλάση** του κάθε δείγματος.

Σκοπός είναι να παράξουμε μοντέλα τα οποία να προβλέπουν τις κλάσεις νέων δειγμάτων.

ΠΡΟΒΛΗΜΑ ΤΑΞΙΝΟΜΗΣΗΣ

Παράδειγμα πίνακα δεδομένων:

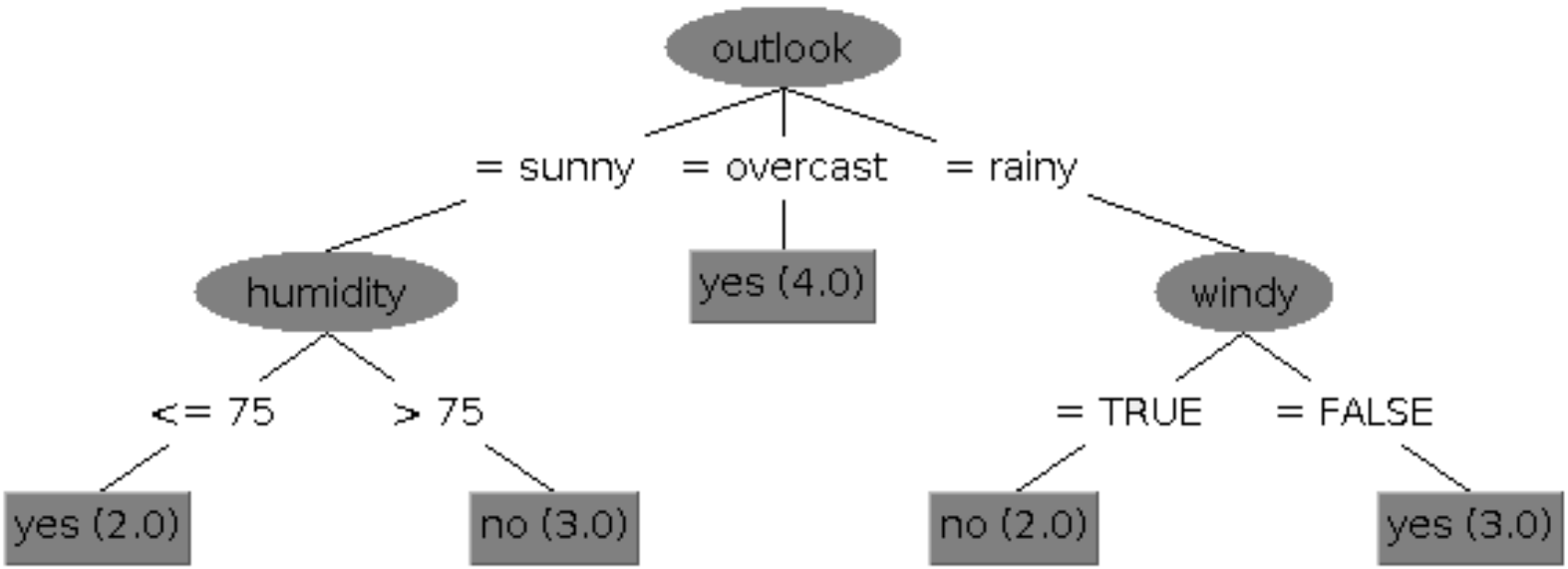
patient	class	1	2	...	699	700	701
AKI_8_24_97_110812	Biopsy kidney normal	-0.002437107	-0.0035140763	...	-0.0006630864	-0.0005720405	-0.0006858139
AKI_8_24_100_110812	Biopsy kidney normal	-0.0013475575	-0.0035747009	...	-0.0006320954	-0.0006858139	-0.0008425529
AKI_8_24_101_110812	Biopsy kidney normal	-0.0060549417	-0.0046379988	...	-0.000791196	-0.0009159127	-0.0008425529
AKI_8_24_102_110812	Biopsy kidney normal	-0.0049196483	-0.0017621665	...	-0.0008894464	-0.0010515443	-0.0010613955
AKI_8_24_104_110812	Biopsy kidney normal	-0.0037098023	-0.0046379988	...	-0.0008737839	-0.0008425529	-0.0009611175
AKI_8_24_105_110812	Biopsy kidney normal	-0.0049196483	-0.0032075154	...	-0.0006630864	-0.0009376947	-0.0010613955
AKI_8_24_107_110812	Biopsy kidney normal	-0.0032075154	-0.0088158057	...	-0.0010150392	-0.0009611175	-0.0011037445
AKI_8_24_108_110812	Biopsy kidney normal	-0.0023361113	-0.0035747009	...	-0.0007549861	-0.0009749736	-0.0010515443
AKI_8_24_109_110812	Biopsy kidney normal	-0.0017972075	-0.0029727493	...	-0.000231685	-0.0006630864	-0.0004138545
AKI_8_24_110_110812	Biopsy kidney normal	-0.0040793693	-0.0026217018	...	-0.0006320954	-0.0008894464	-0.0008425529
...	...	...	...	...	...	...	...
AKI_8_24_02_110722	Acute Kidney Injury	-0.0060549417	-0.0066442452	...	-0.0007197942	-0.0007549861	-0.0008894464
AKI_8_24_04_110722	Acute Kidney Injury	-0.0028694225	-0.0049196483	...	-0.0009749736	-0.0011497981	-0.0014832325
AKI_8_24_05_110722	Acute Kidney Injury	-0.0030599754	-0.0017972075	...	-0.000791196	-0.0009611175	-0.0011037445
AKI_8_24_06_110722	Acute Kidney Injury	-0.0055588469	-0.0039890442	...	-0.000550409	-0.0006630864	-0.0006858139
AKI_8_24_08_110722	Acute Kidney Injury	-0.0032075154	-0.0035140763	...	0.0005251104	0.0003348619	0.0003175396

ΤΙ ΠΡΟΣΠΑΘΟΥΜΕ ΝΑ ΠΑΡΑΞΟΥΜΕ;

Η γνώση που εξάγεται μπορεί να έχει τη μορφή:

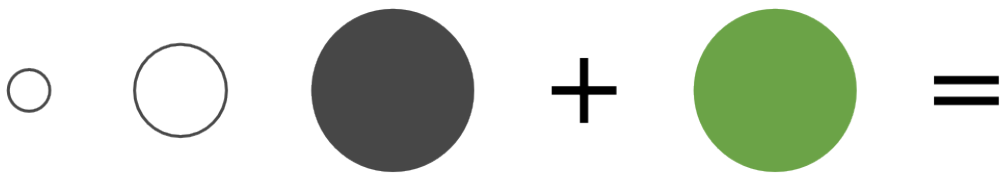
- Πίνακες με τα δεδομένα (δεν βοηθάνε ιδιαίτερα)
- Γραμμικές (ή μη) συναρτήσεις
- Δέντρα αποφάσεων
- Κανόνες κάλυψης
- Συστάδες

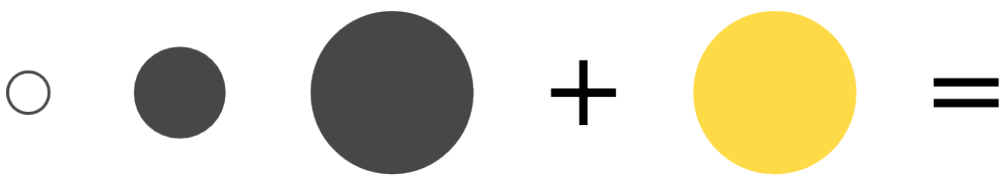
ΠΑΡΑΔΕΙΓΜΑ: ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ



ΠΑΡΑΔΕΙΓΜΑ: ΚΑΝΟΝΕΣ ΚΑΛΥΨΗΣ

Εάν συνθήκη1 [ΚΑΙ συνθήκη2 ΚΑΙ ...] τότε κλάση





ΣΤΑΔΙΑ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

1. **Συγκέντρωση** δεδομένων σε ένα αρχείο ή βάση δεδομένων.
2. **Προεπεξεργασία** (καθαρισμός, μείωση δεδομένων, κανονικοποίηση, …).
3. **Εκπαίδευση** μοντέλων μηχανικής μάθησης.
4. **Αξιολόγηση** μοντέλων.

Επιπλέον: **επιλογή μεταβλητών** και **συναινετική μάθηση**.

ΑΛΓΟΡΙΘΜΟΙ ΜΑΘΗΣΗΣ

Οι συνηθέστεροι αλγόριθμοι ταξινόμησης χωρίζονται στις εξής κατηγορίες:

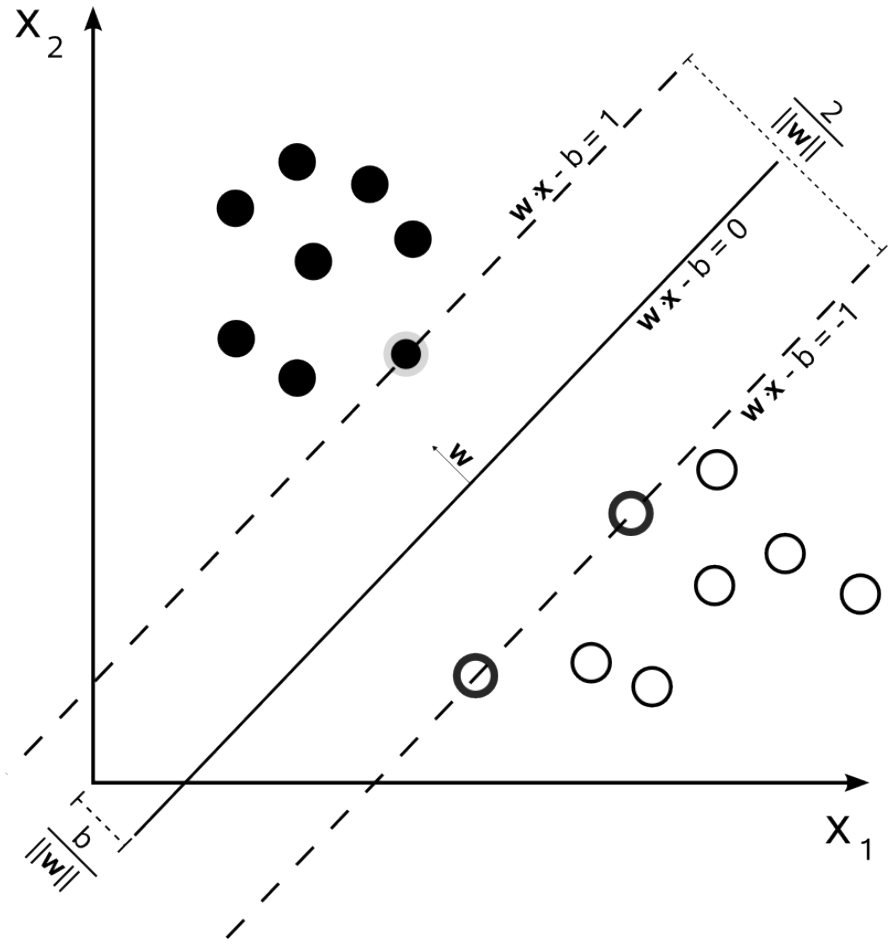
- Απλοί κανόνες (π.χ. `1R`, `ZeroR`).
- Στατιστική μοντελοποίηση (π.χ. `Bayes`).
- Δέντρα αποφάσεων (π.χ. `DecisionStump`, `REPTree`).
- Κανόνες κάλυψης (π.χ. `Jrip`).
- Γραμμικές (ή μη) συναρτήσεις (π.χ. Νευρωνικά Δίκτυα, Support Vector Machines).

Μελετήσαμε και εφαρμόσαμε πολλούς διαφορετικούς αλγορίθμους. Ιδιαίτερο ενδιαφέρον παρουσίασε ο αλγόριθμος `LibSVM`.

ΠΑΡΑΔΕΙΓΜΑ: SVM

Σε 2 διαστάσεις ψάχνουμε την καλύτερη δυνατή ευθεία που διαχωρίζει τα δεδομένα.

ΠΑΡΑΔΕΙΓΜΑ: SVM



ΠΑΡΑΔΕΙΓΜΑ: SVM

- Τα σημεία που βρίσκονται πιο κοντά σε αυτό ονομάζονται «διανύσματα υποστήριξης» (support vectors).
- Στόχος είναι η μεγιστοποίηση της απόστασης του υπερεπιπέδου από τα διανύσματα υποστήριξης.

ΠΑΡΑΔΕΙΓΜΑ: SVM

Τα δεδομένα μπορεί να μην είναι τελείως γραμμικώς διαχωρίσιμα: δεχόμαστε μεμονωμένα λάθη, με μια παράμετρο ποινής `C` (`C-SVC`).

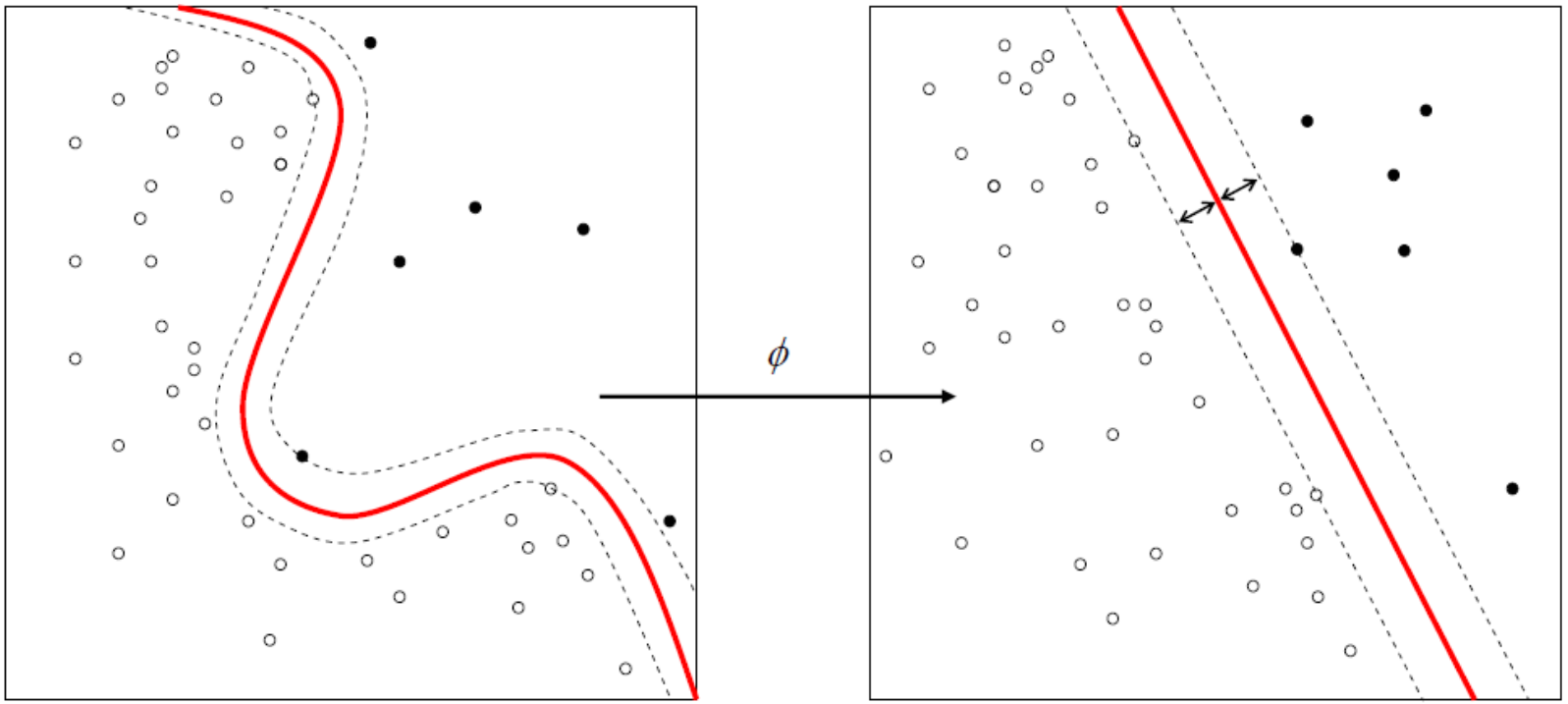
Δεδομένα που δεν είναι γραμμικώς διαχωρίσιμα ενδέχεται, αν εφαρμόσουμε ένα μη-γραμμικό μετασχηματισμό (kernel functions), να γίνουν τελικά γραμμικώς διαχωρίσιμα.

ΠΑΡΑΔΕΙΓΜΑ: SVM

Τα δεδομένα μπορεί να μην είναι τελείως γραμμικώς διαχωρίσιμα: δεχόμαστε μεμονωμένα λάθη, με μια παράμετρο ποινής `C` (`C-SVC`).

Δεδομένα που δεν είναι γραμμικώς διαχωρίσιμα ενδέχεται, αν εφαρμόσουμε ένα μη-γραμμικό μετασχηματισμό (kernel functions), να γίνουν τελικά γραμμικώς διαχωρίσιμα.

ΠΑΡΑΔΕΙΓΜΑ: SVM



ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΩΝ

Κάθε μοντέλο που παράγεται πρέπει να αξιολογηθεί ως προς την ικανότητά του να προβλέπει σωστά τις κλάσεις **αγνώστων** δειγμάτων.

Για τον έλεγχο **δεν** πρέπει να χρησιμοποιείται το ίδιο το **σύνολο εκπαίδευσης**! (υπερπροσαρμογή)

ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΩΝ

Χρησιμοποιούμε τη μέθοδο n-fold **Cross-Validation** για να αποφύγουμε το overfitting:

1. Χωρίζουμε τα δεδομένα σε n περίπου ίσα κομμάτια.
2. Εκπαιδεύουμε διαδοχικά n μοντέλα.
3. Σε κάθε μοντέλο, τα n-1 κομμάτια χρησιμοποιούνται για την εκπαίδευση και το 1 για τον έλεγχο του μοντέλου.

ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΩΝ

Δεν αρκεί μόνο η συνολική επιτυχία! Σε ένα σύνολο με 95% υγιή άτομα και 5% ασθενή, ένα μοντέλο που προβλέπει ότι «όλοι είναι υγιείς» έχει 95% επιτυχία!

Εξετάζουμε την επιτυχία ως προς κάθε κλάση ξεχωριστά (confusion matrix).

ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΩΝ

Δεν αρκεί μόνο η συνολική επιτυχία! Σε ένα σύνολο με 95% υγιή άτομα και 5% ασθενή, ένα μοντέλο που προβλέπει ότι «όλοι είναι υγιείς» έχει 95% επιτυχία!

Εξετάζουμε την επιτυχία ως προς κάθε κλάση ξεχωριστά (confusion matrix).

ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΩΝ

Πρόβλεψη	**A**	**B**	**Επιτυχία**
Πραγματικά A	90	10	90% (δείγματα A)
Πραγματικά B	30	70	70% (δείγματα B)
Όλα τα δείγματα			80%

ΣΥΝΑΙΝΕΤΙΚΗ ΜΑΘΗΣΗ

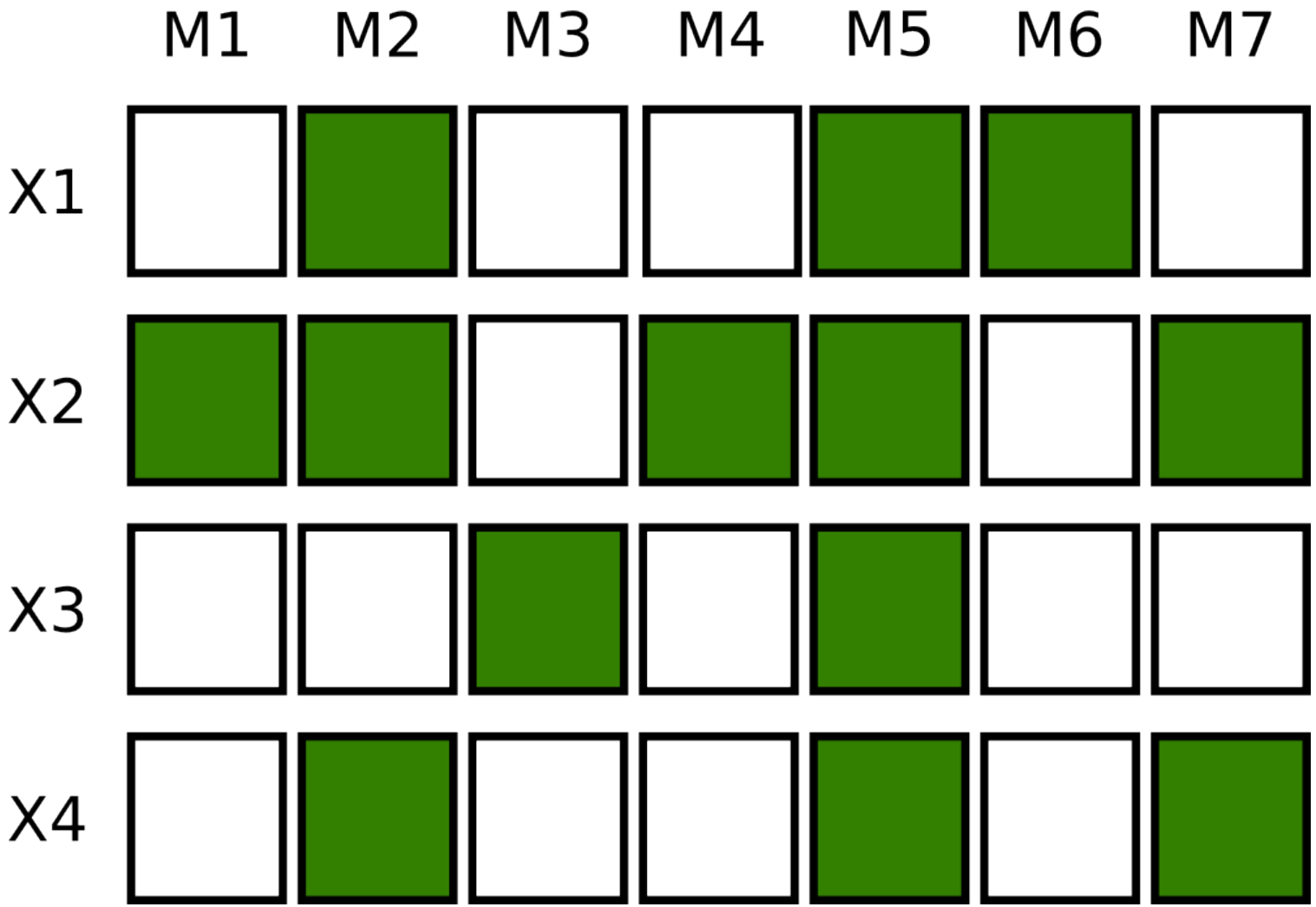
Συνδυάζοντας περισσότερα του ενός μοντέλα μπορούμε να δημιουργήσουμε consensus models.

Η τελική απόφαση προκύπτει από συμψηφισμό (απλό ή με συντελεστές βαρύτητας) των αποφάσεων κάθε μεμονωμένου μοντέλου (voting).

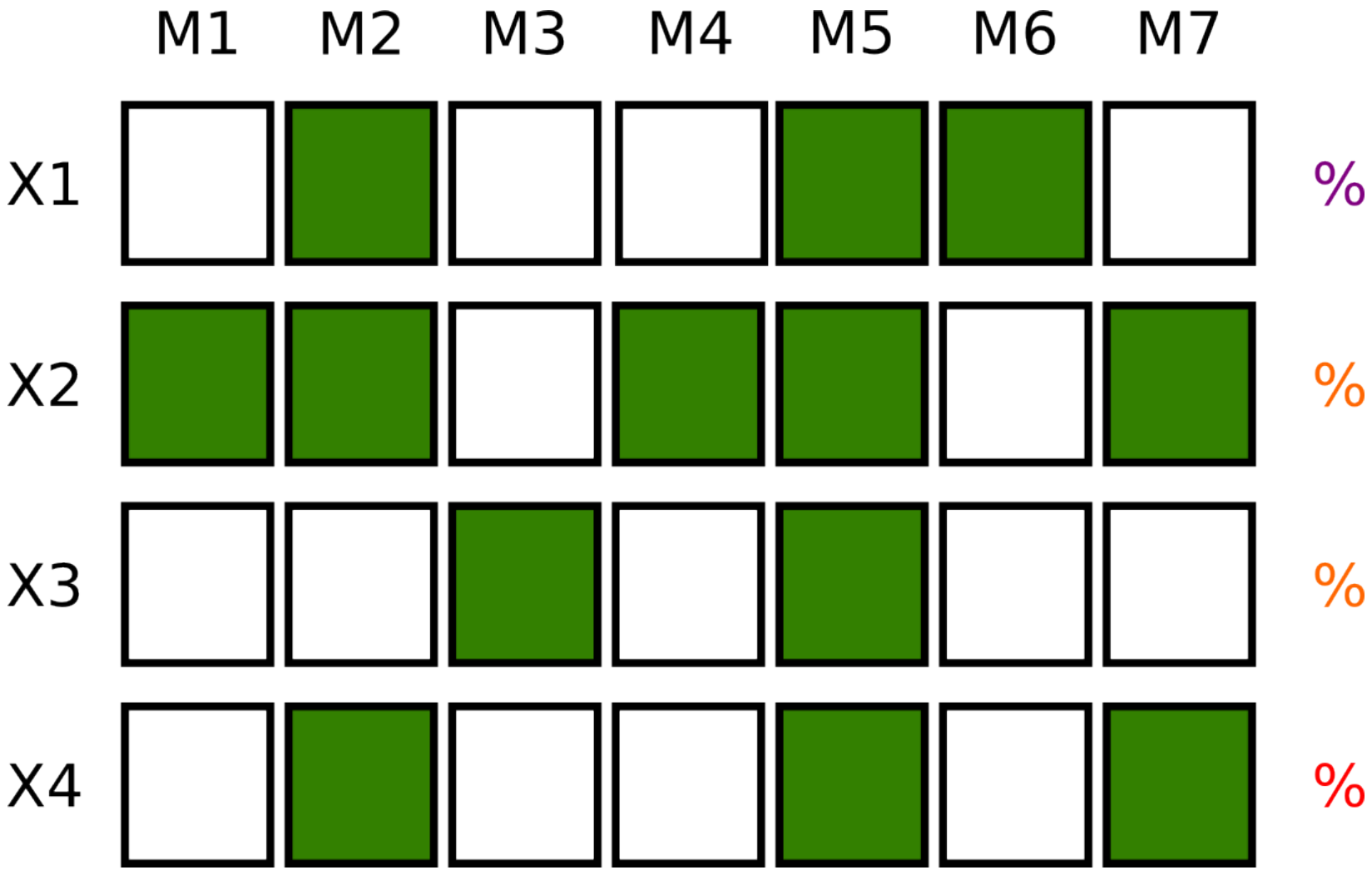
ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ

- Δεν συνεισφέρουν όλες οι διαθέσιμες μεταβλητές το ίδιο στις αποφάσεις.
- Ορισμένες από αυτές ενδέχεται να δυσκολεύουν τις προβλέψεις.
- Μεταβολομική: επιλογή μεταβλητών → επιλογή **μεταβολιτών** → **μεταβολικά μονοπάτια** → **μηχανισμοί**!

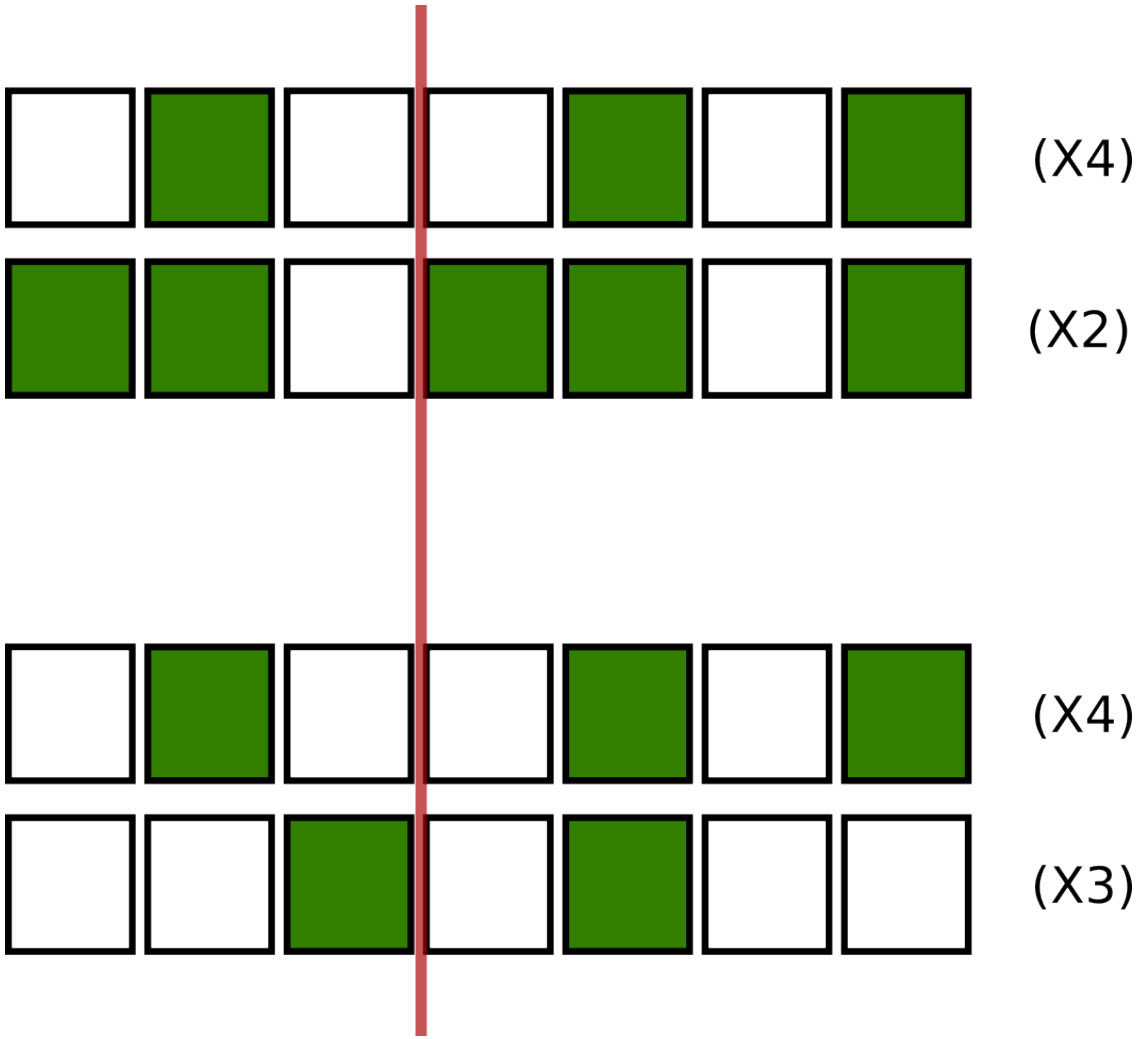
ΓΕΝΕΤΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ: ΑΡΧΙΚΟΣ ΠΛΗΘΥΣΜΟΣ



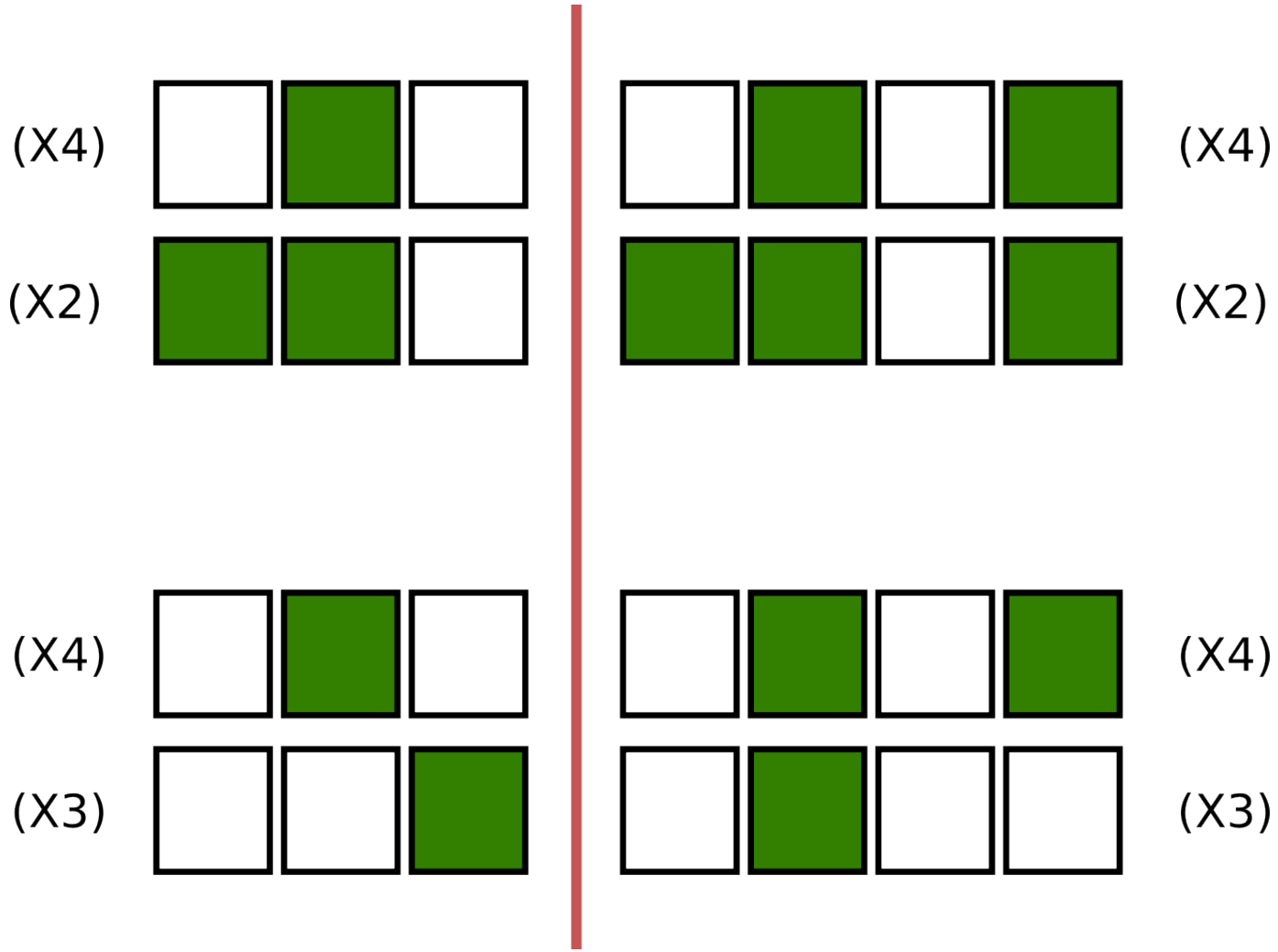
ΓΑ: ΑΞΙΟΛΟΓΗΣΗ ΧΡΩΜΟΣΩΜΑΤΩΝ



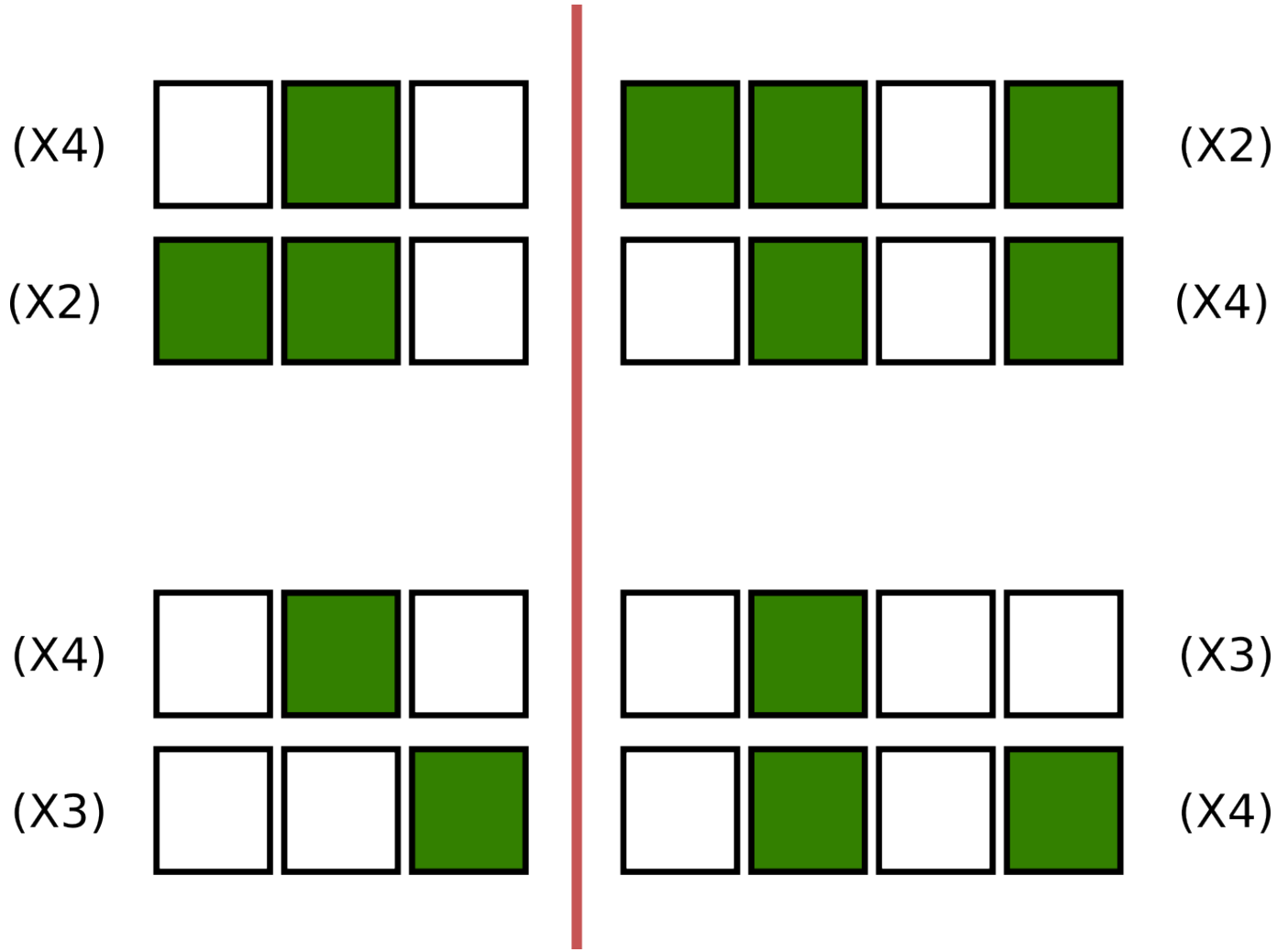
ΓΑ: ΕΠΙΛΟΓΗ ΧΡΩΜΟΣΩΜΑΤΩΝ



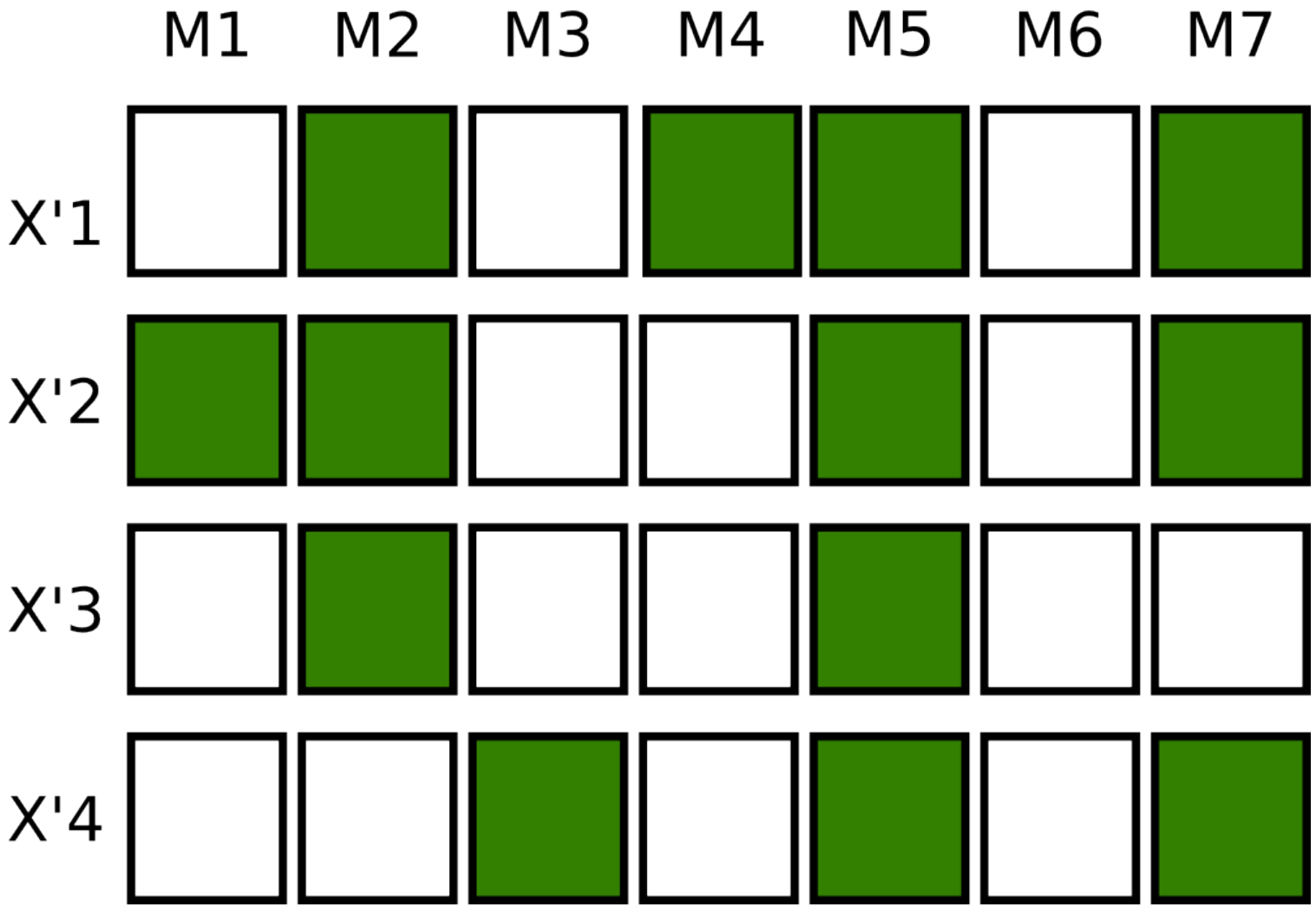
ΓΑ: ΤΟΜΗ ΧΡΩΜΟΣΩΜΑΤΩΝ



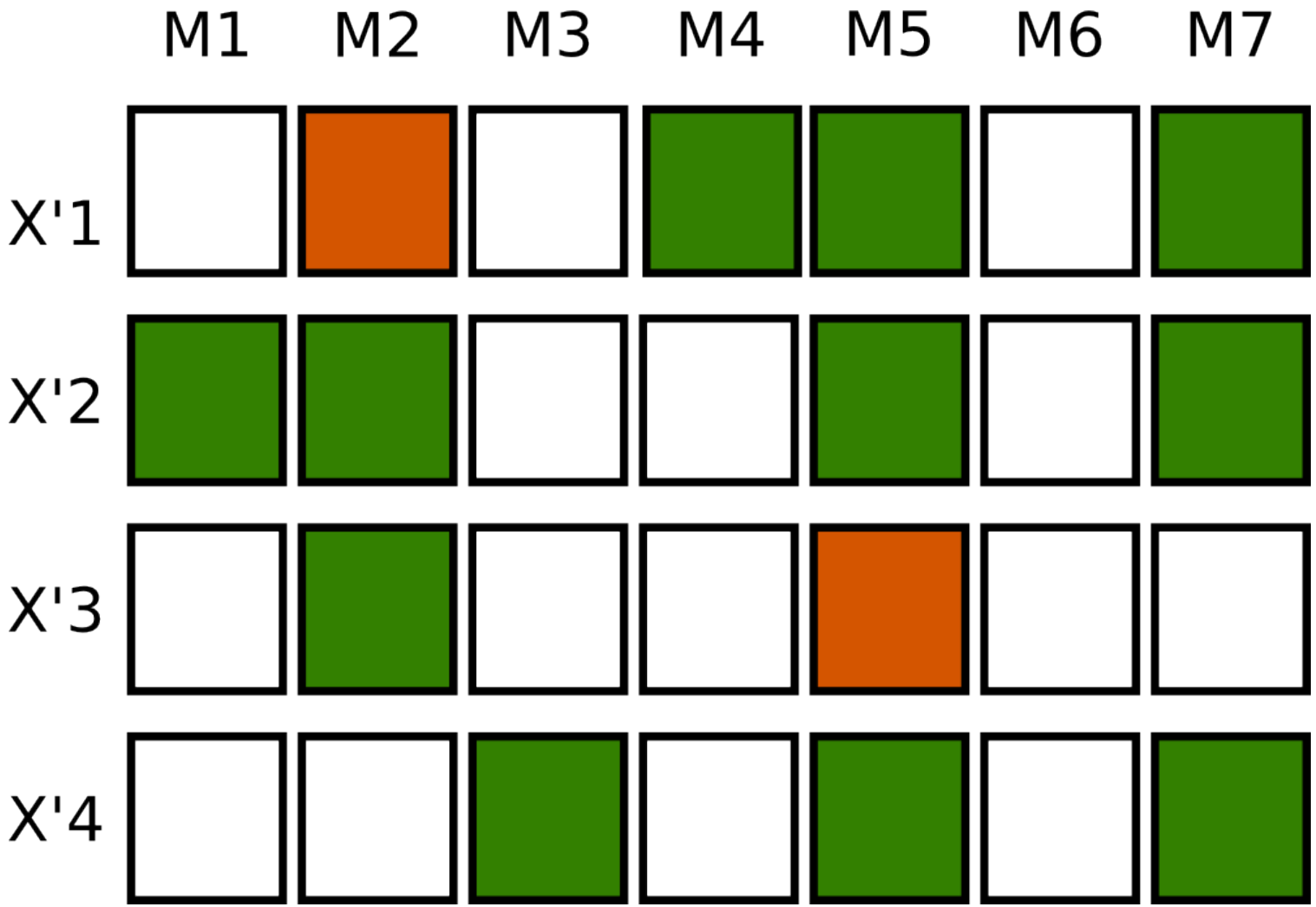
ΓΑ: ΔΙΑΣΤΑΥΡΩΣΗ ΧΡΩΜΟΣΩΜΑΤΩΝ



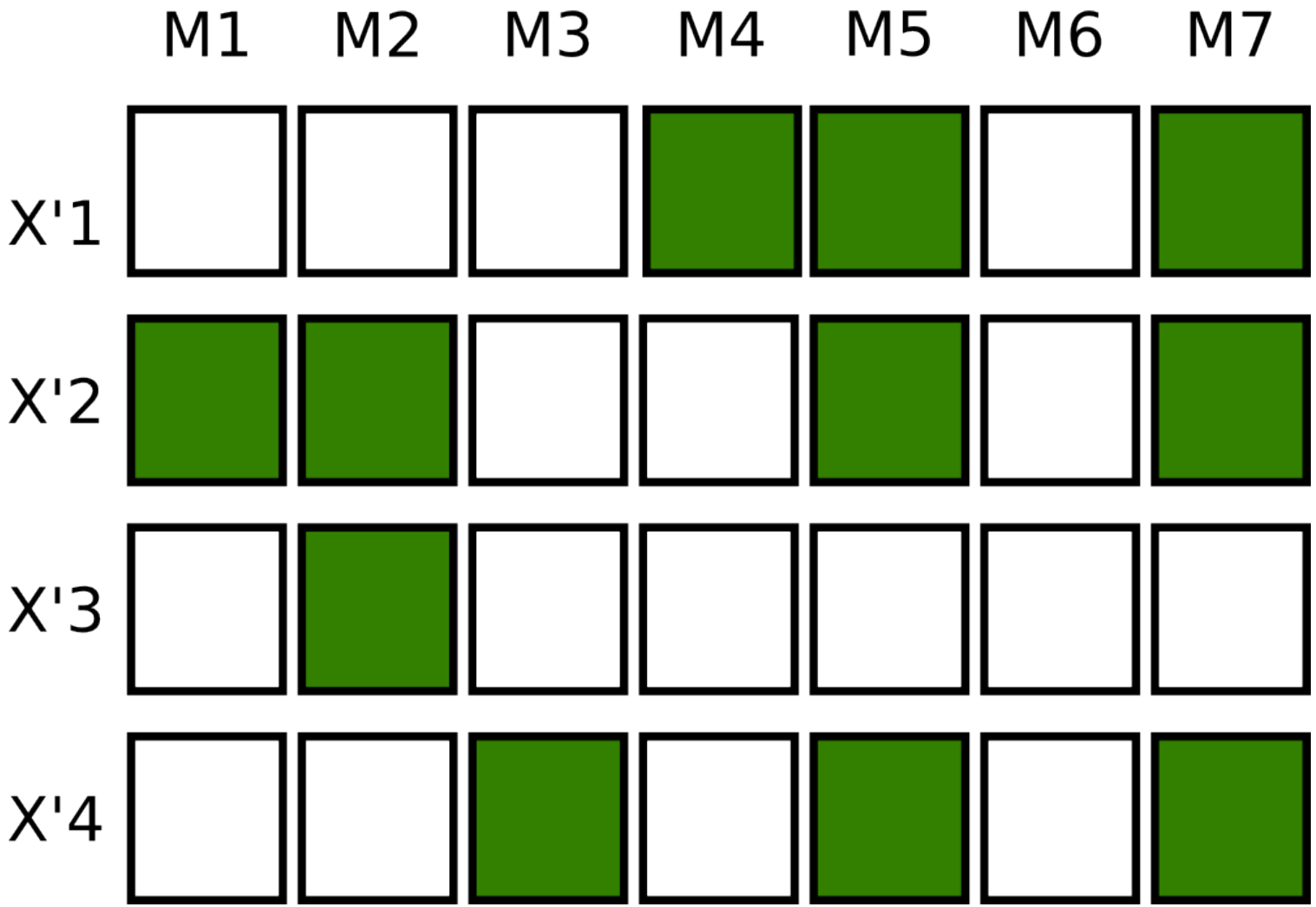
ΓΑ: ΝΕΟΣ ΠΛΗΘΥΣΜΟΣ



ΓΑ: ΜΕΤΑΛΛΑΞΗ

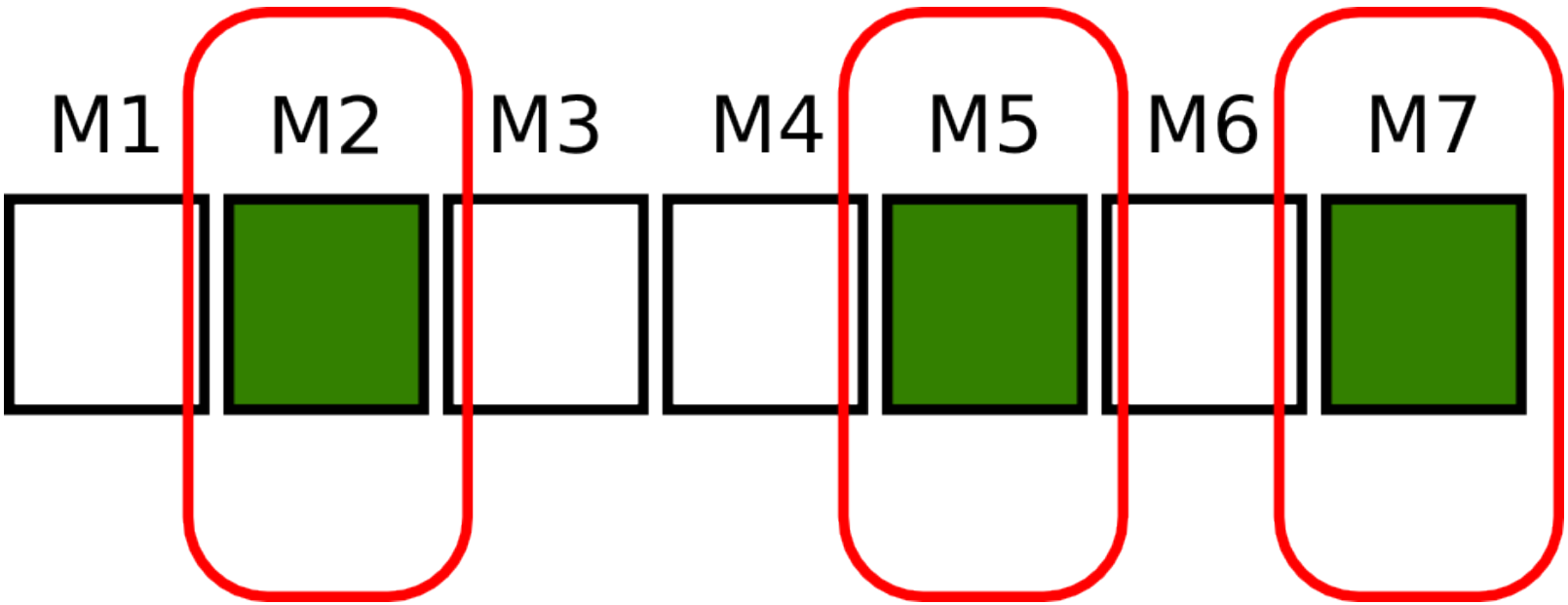


ΓΑ: ΜΕΤΑΛΛΑΞΗ



ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΓΕΝΕΤΙΚΗ ΕΡΕΥΝΑ

Τάση για εκδήλωση ή μη κάθε γονιδίου:

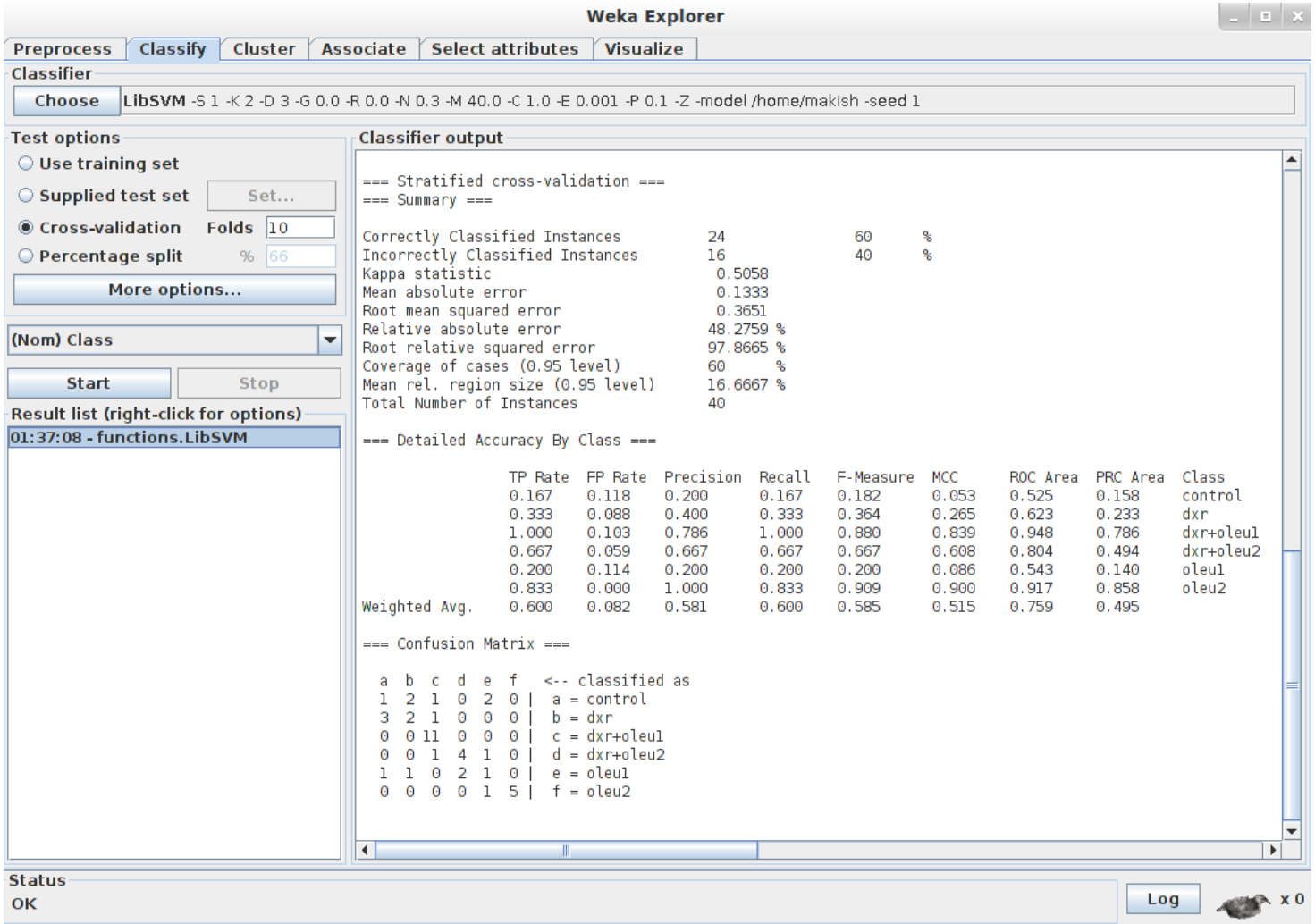


ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΓΕΝΕΤΙΚΗ ΕΡΕΥΝΑ

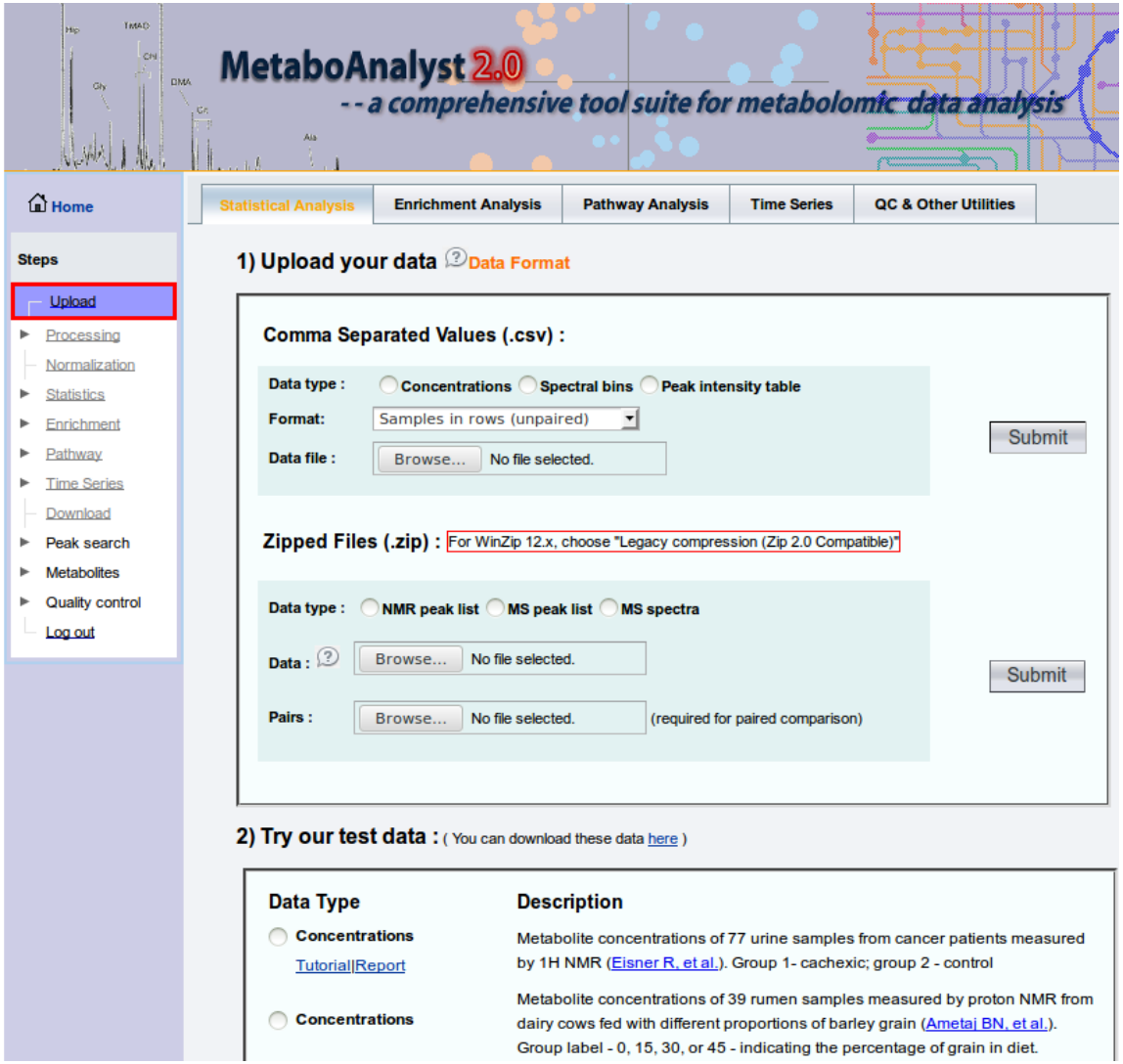
- Η επιλογή των γεννητόρων γίνεται με πιθανότητα ανάλογη της σχετικής τους καταλληλότητας (π.χ. «τροχός ρουλέτας»).
- Η αναπαραγωγή τους γίνεται διασταυρώνοντάς τα σε ένα σημείο (ανταλλαγή γενετικού υλικού).
- Τα γονίδια μεταλλάσσονται με μία πιθανότητα.
- Μετά τη δημιουργία μιας πλήρους νέας γενιάς, η προηγούμενη καταστρέφεται.
- Ο αλγόριθμος τερματίζεται μετά από ένα πλήθος γενεών ή αν συγκλίνουν τα χρωμοσώματα.
- Λόγω της στοχαστικής του φύσης επιβάλλονται πολλαπλές επανεκκινήσεις.

ΛΟΓΙΣΜΙΚΟ

WEKA


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
LibSVM -S 1 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.3 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -Z -model /home/makish -seed 1
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Class
Start
Stop
Result list (right-click for options)
01:37:08 - functions.LibSVM
Classifier output
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 24 60 %
Incorrectly Classified Instances 16 40 %
Kappa statistic 0.5058
Mean absolute error 0.1333
Root mean squared error 0.3651
Relative absolute error 48.2759 %
Root relative squared error 97.8665 %
Coverage of cases (0.95 level) 60 %
Mean rel. region size (0.95 level) 16.6667 %
Total Number of Instances 40
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.167 0.118 0.200 0.167 0.182 0.053 0.525 0.158 control
0.333 0.088 0.400 0.333 0.364 0.265 0.623 0.233 dxr
1.000 0.103 0.786 1.000 0.880 0.839 0.948 0.786 dxr+oleu1
0.667 0.059 0.667 0.667 0.667 0.608 0.804 0.494 dxr+oleu2
0.200 0.114 0.200 0.200 0.200 0.086 0.543 0.140 oleu1
0.833 0.000 1.000 0.833 0.909 0.900 0.917 0.858 oleu2
Weighted Avg. 0.600 0.082 0.581 0.600 0.585 0.515 0.759 0.495
=== Confusion Matrix ===
a b c d e f <-- classified as
1 2 1 0 2 0 | a = control
3 2 1 0 0 0 | b = dxr
0 0 11 0 0 0 | c = dxr+oleu1
0 0 1 4 1 0 | d = dxr+oleu2
1 1 0 2 1 0 | e = oleu1
0 0 0 0 1 5 | f = oleu2
Status
OK
Log
x 0

METABOANALYST



LIBSVM

```
makish@makish-nautilus ~/Desktop/Thesis/SVM/libsvm-3.17 $ ./svm-train --help
Usage: svm-train [options] training_set_file [model_file]
options:
-s svm_type : set type of SVM (default 0)
        0 -- C-SVC              (multi-class classification)
        1 -- nu-SVC             (multi-class classification)
        2 -- one-class SVM
        3 -- epsilon-SVR        (regression)
        4 -- nu-SVR             (regression)
-t kernel_type : set type of kernel function (default 2)
        0 -- linear: u'*v
        1 -- polynomial: (gamma*u'*v + coef0)^degree
        2 -- radial basis function: exp(-gamma*|u-v|^2)
        3 -- sigmoid: tanh(gamma*u'*v + coef0)
        4 -- precomputed kernel (kernel values in training_set_file)
-d degree : set degree in kernel function (default 3)
-g gamma : set gamma in kernel function (default 1/num_features)
-r coef0 : set coef0 in kernel function (default 0)
-c cost : set the parameter C of C-SVC, epsilon-SVR, and nu-SVR (default 1)
-n nu : set the parameter nu of nu-SVC, one-class SVM, and nu-SVR (default 0.5)
-p epsilon : set the epsilon in loss function of epsilon-SVR (default 0.1)
-m cachesize : set cache memory size in MB (default 100)
-e epsilon : set tolerance of termination criterion (default 0.001)
-h shrinking : whether to use the shrinking heuristics, 0 or 1 (default 1)
-b probability_estimates : whether to train a SVC or SVR model for probability estimates, 0 or 1 (default 0)
-wi weight : set the parameter C of class i to weight*C, for C-SVC (default 1)
-v n: n-fold cross validation mode
-q : quiet mode (no outputs)
```

ΛΟΓΙΣΜΙΚΟ

Χρησιμοποιήσαμε:

WEKA	Πλήρης σουίτα εργαλείων εξόρυξης δεδομένων, με πολλούς έτοιμους αλγορίθμους μηχανικής μάθησης.
MetaboAnalyst	Εξειδικευμένη, on the cloud συλλογή εργαλείων μεταβολομικής ανάλυσης.
LibSVM	Αυτόνομη υλοποίηση SVM η οποία συνεργάζεται και με GNU Octave/MATLAB™.

Επίσης, δημιουργήσαμε έναν κώδικα επιλογής μεταβλητών με γενετική έρευνα και SVM.

Όλα τα πακέτα είναι ελεύθερα διαθέσιμα και ανοιχτού κώδικα.

ΚΩΔΙΚΑΣ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

Αναπτύξαμε έναν γενετικό αλγόριθμο με συνάρτηση καταλληλότητας τη συνολική ακρίβεια του μοντέλου που προκύπτει από τη LibSVM.

- Ένα βασικό GNU Octave script, τέσσερις συναρτήσεις και ένα script επανεκκινήσεων.
- Δέχεται ως εισόδους τον πίνακα δεδομένων, τις παραμέτρους της LibSVM και τις παραμέτρους του γενετικού αλγορίθμου.
- Ο βασικός αλγόριθμος εκτελείται πολλές φορές και επιστρέφει ένα χρωμόσωμα με τις συχνότερα εμφανιζόμενες τιμές των γονιδίων.

ΚΩΔΙΚΑΣ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

Αναπτύξαμε έναν γενετικό αλγόριθμο με συνάρτηση καταλληλότητας τη συνολική ακρίβεια του μοντέλου που προκύπτει από τη LibSVM.

- Ένα βασικό GNU Octave script, τέσσερις συναρτήσεις και ένα script επανεκκινήσεων.
- Δέχεται ως εισόδους τον πίνακα δεδομένων, τις παραμέτρους της LibSVM και τις παραμέτρους του γενετικού αλγορίθμου.
- Ο βασικός αλγόριθμος εκτελείται πολλές φορές και επιστρέφει ένα χρωμόσωμα με τις συχνότερα εμφανιζόμενες τιμές των γονιδίων.

ΚΩΔΙΚΑΣ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

Αναπτύξαμε έναν γενετικό αλγόριθμο με συνάρτηση καταλληλότητας τη συνολική ακρίβεια του μοντέλου που προκύπτει από τη LibSVM.

- Ένα βασικό GNU Octave script, τέσσερις συναρτήσεις και ένα script επανεκκινήσεων.
- Δέχεται ως εισόδους τον πίνακα δεδομένων, τις παραμέτρους της LibSVM και τις παραμέτρους του γενετικού αλγορίθμου.
- Ο βασικός αλγόριθμος εκτελείται πολλές φορές και επιστρέφει ένα χρωμόσωμα με τις συχνότερα εμφανιζόμενες τιμές των γονιδίων.

ΚΩΔΙΚΑΣ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

Αναπτύξαμε έναν γενετικό αλγόριθμο με συνάρτηση καταλληλότητας τη συνολική ακρίβεια του μοντέλου που προκύπτει από τη LibSVM.

- Ένα βασικό GNU Octave script, τέσσερις συναρτήσεις και ένα script επανεκκινήσεων.
- Δέχεται ως εισόδους τον πίνακα δεδομένων, τις παραμέτρους της LibSVM και τις παραμέτρους του γενετικού αλγορίθμου.
- Ο βασικός αλγόριθμος εκτελείται πολλές φορές και επιστρέφει ένα χρωμόσωμα με τις συχνότερα εμφανιζόμενες τιμές των γονιδίων.

ΠΡΟΒΛΗΜΑ 1

ΔΕΔΟΜΕΝΑ AKI

- Έπειτα από χειρουργικές επεμβάσεις στην καρδιά, ορισμένοι ασθενείς εμφανίζουν **οξεία νεφρική βλάβη** (Acute Kidney Injury).
- Σε σύνολο **106 ανθρώπων**, 72 παρέμειναν υγιείς, ενώ 34 εμφάνισαν AKI διαφόρων επιπέδων.
- Για κάθε άτομο είναι διαθέσιμες **701 περιοχές φάσματος NMR** ούρων 24 ώρες μετά την επέμβαση.
- Τα δεδομένα είναι διαθέσιμα στη βάση δεδομένων MetaboLights.

ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΚΙ: ΜΕΜΟΝΩΜΕΝΟΙ ΑΛΓΟΡΙΘΜΟΙ

- Εκπαιδεύτηκαν 26 διαφορετικά μοντέλα, με όλους τους διαθέσιμους στη WEKA αλγορίθμους.
- Υψηλότερη **συνολική** επιτυχία: `Jrip` με 85.5%.
- Υψηλότερη επιτυχία στους **υγιείς**: `Jrip` με 90%.
- Υψηλότερη επιτυχία στους **ασθενείς**: `DecisionStump` με 91%.

ΑΠΟΤΕΛΕΣΜΑΤΑ AKI: ΜΕΜΟΝΩΜΕΝΟΙ ΑΛΓΟΡΙΘΜΟΙ

Jrip

Κλάση	**a**	**b**	**TP rate**
a: Normal	65	7	0.903
b: AKI	8	26	0.765
Επιτυχία			85.8%

LibSVM

Κλάση	**a**	**b**	**TP rate**
a: Normal	63	9	0.875
b: AKI	12	22	0.647
Επιτυχία			80.2%

DecisionStump

Κλάση	**a**	**b**	**TP rate**
a: Normal	53	19	0.736
b: AKI	3	31	0.912
Επιτυχία			79.2%

REPTree

Κλάση	**a**	**b**	**TP rate**
a: Normal	58	14	0.806
b: AKI	7	27	0.794
Επιτυχία			80.2%

ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΚΙ: ΣΥΝΑΙΝΕΤΙΚΑ ΜΟΝΤΕΛΑ

Συνδυάσαμε μερικά από τα καλύτερα μοντέλα που προέκυψαν για τη δημιουργία συναινετικών μοντέλων. Μερικά ενδιαφέροντα αποτελέσματα:

- DecisionStump + REPTree + SMO → 84.0%
- REPTree + LibSVM → 84.9%
- REPTree + SMO → 77.4%
- DecisionStump + Jrip + REPTree + LibSVM → 86.8%
- DesicionStump + Jrip + REPTree + SMO → 87.7%

Θυμηθείτε: Jrip → 85.8%

ΠΡΟΒΛΗΜΑ 2

ΔΕΔΟΜΕΝΑ DXR

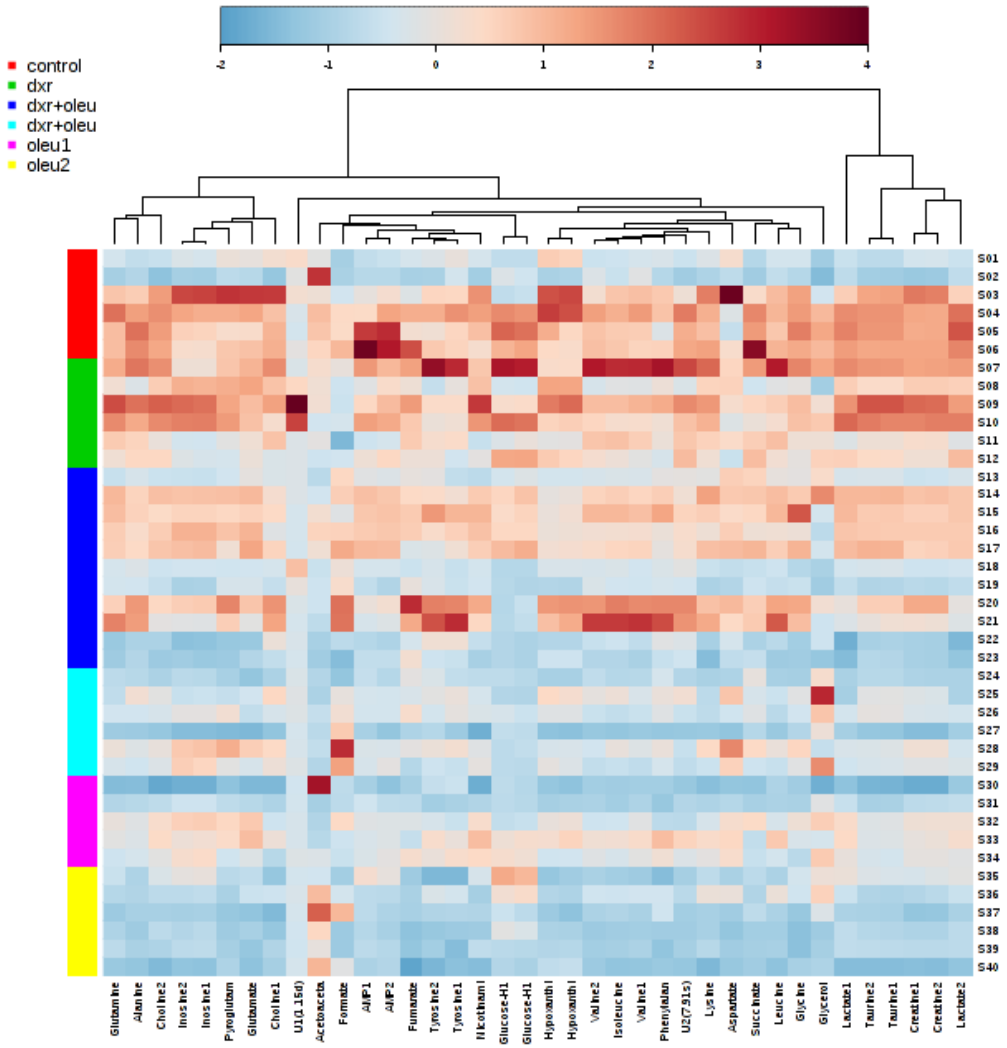
Η **αδριαμυκίνη** (DXR) χορηγείται ως αντινεοπλασματικός παράγοντας για την αντιμετώπιση πολλών μορφών καρκίνου, έχει όμως ανεπιθύμητες παρενέργειες, όπως **καρδιοτοξικότητα**.

Εξετάζεται αν η συγχορήγηση **ολευρωπαΐνης** μπορεί να αναστείλει την καρδιοτοξική επίπτωση της DXR.

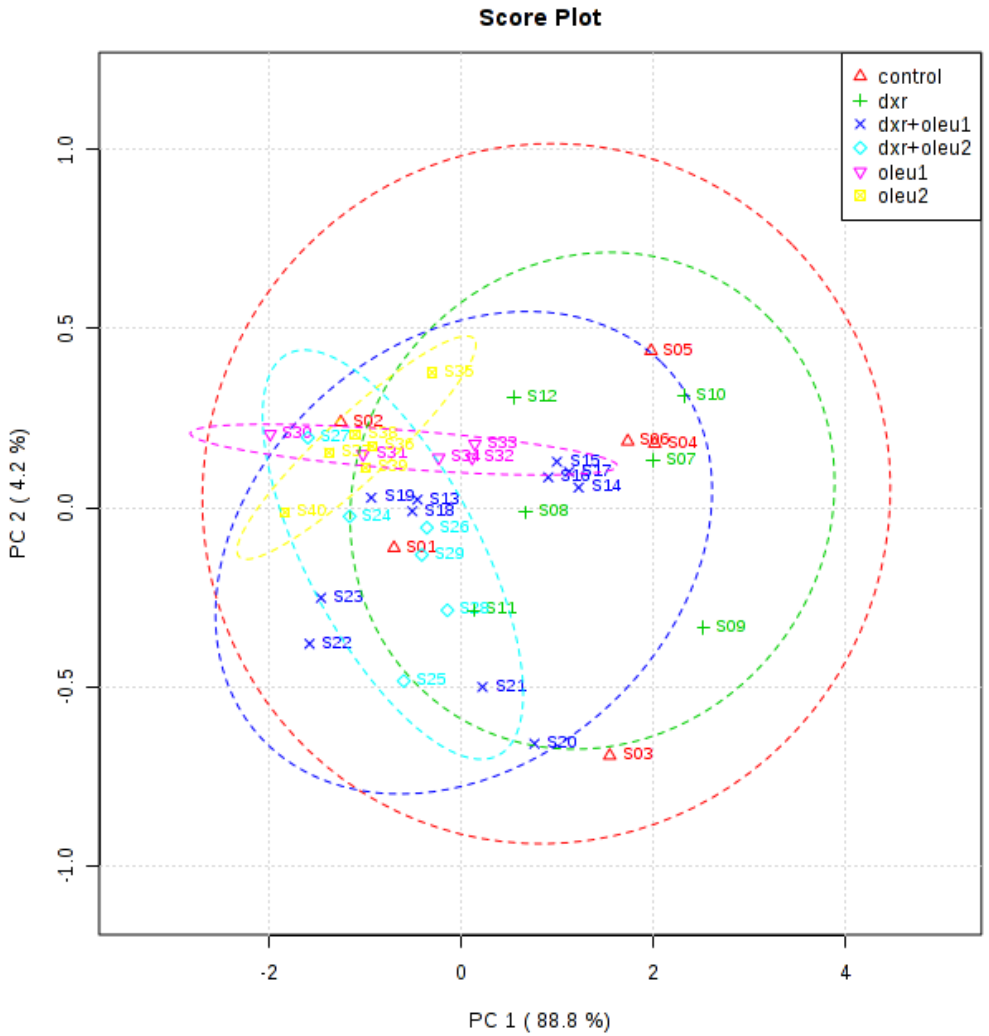
38 περιοχές φασμάτων NMR από εκχυλίσματα ιστών **40 επιμύων**, χωρισμένων σε **6 κλάσεις**: control, dxr, dxr+oleu1 (μικρή δόση), dxr+oleu2 (διπλάσια δόση), oleu1, oleu2.

Τα δεδομένα προέρχονται από την ερευνητική ομάδα του καθηγητή Εμμ. Μικρού (Τμήμα Φαρμακευτικής ΕΚΠΑ).

DXR: ΑΝΑΛΥΣΗ ΜΕ ΤΟ METABOANALYST (HEATMAP)



DXR: ΑΝΑΛΥΣΗ ΜΕ ΤΟ METABOANALYST (PCA)



DXR: ΑΝΑΛΥΣΗ ΜΕ ΤΟ METABOANALYST

Ταξινόμηση με τον παρεχόμενο αλγόριθμο Random Forest:

	a	b	c	d	e	f	TP rate
a: control	**1**	3	0	1	0	1	0.270
b: dxr	1	**2**	3	0	0	0	0.330
c: dxr+oleu1	0	0	**10**	1	0	0	0.910
d: dxr+oleu2	0	0	3	**1**	1	1	0.270
e: oleu1	0	0	2	1	**0**	2	0.000
f: oleu2	0	0	0	0	1	**5**	0.830
Επιτυχία							47.5%

DXR: ΑΝΑΛΥΣΗ ΜΕ ΤΟ METABOANALYST

Τα αποτελέσματα δεν δείχνουν πολύ ενθαρρυντικά!

Μπορούμε να κάνουμε κάτι καλύτερο;

DXR: ΑΝΑΛΥΣΗ ΜΕ ΤΗ WEKA

Εκπαιδεύτηκαν 25 μοντέλα με όλους τους διαθέσιμους στη WEKA αλγορίθμους που εφαρμόζονται σε προβλήματα πολλών κλάσεων.

Υψηλότερη συνολική επιτυχία: 60.0%.

DXR: ΑΝΑΛΥΣΗ ΜΕ ΤΗ WEKA

Τα καλύτερα αποτελέσματα έδωσε η **LibSVM**:

	a	**b**	**c**	**d**	**e**	**f**	**TP rate**
a: control	**1**	2	1	0	2	0	0.167
b: dxr	3	**2**	1	0	0	0	0.333
c: dxr+oleu1	0	0	**11**	0	0	0	1.000
d: dxr+oleu2	0	0	1	**4**	1	0	0.667
e: oleu1	1	1	0	2	**1**	0	0.200
f: oleu2	0	0	0	0	1	**5**	0.833
Επιτυχία							60.0%

DXR: ΑΝΑΛΥΣΗ ΜΕ ΤΗ WEKA

...ακολουθούμενη από τον **IBk**:

	a	**b**	**c**	**d**	**e**	**f**	**TP rate**
a: control	**3**	1	1	0	0	1	0.500
b: dxr	2	**1**	1	0	2	0	0.167
c: dxr+oleu1	0	0	**10**	1	0	0	0.909
d: dxr+oleu2	0	0	2	**3**	0	1	0.500
e: oleu1	0	0	0	3	**1**	1	0.200
f: oleu2	0	0	0	0	1	**5**	0.833
Επιτυχία							57.5%

Προσέξτε ότι μεταβολή στην πρόβλεψη ενός δείγματος επιφέρει ποσοστιαία μεταβολή 2.5%.

DXR: ΑΝΑΛΥΣΗ ΜΕ ΤΗ WEKA

Μπορούμε και εδώ να παράγουμε **συναινετικά μοντέλα** με υψηλότερη ακρίβεια;

Συνδυάζοντας τα μοντέλα **Logistic** (50.0%), **LWL-BayesNet** (47.5%) και **J48** (50.0%) παράγουμε μοντέλο με συνολική ακρίβεια 52.5%, χρησιμοποιώντας τον `vote` meta-classifier.

	a	**b**	**c**	**d**	**e**	**f**	**TP rate**
a: control	**2**	2	0	0	2	0	0.333
b: dxr	2	**3**	1	0	0	0	0.500
c: dxr+oleu1	1	3	**7**	0	0	0	0.636
d: dxr+oleu2	1	0	1	**4**	1	0	0.667
e: oleu1	0	1	1	2	**0**	1	0.000
f: oleu2	1	0	0	0	0	**5**	0.833
Επιτυχία							52.5%

DXR: ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ

α/α	ID	Όνομα χαρακτηριστικού	Συχνότητα επιλογής
1	21	Glycerol	0.83
2	33	U2(7.91s)	0.73
3	38	Formate	0.73
4	7	Alanine	0.70
5	25	Glucose-H1a	0.70
6	15	Creatine1	0.67
7	23	Lactate2	0.67
8	13	Aspartate	0.63
9	3	Valine1	0.60
10	4	Valine2	0.60
11	6	Lactate1	0.60
12	8	Acetoacetate	0.60

DXR: ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ

LibSVM μέσω GNU Octave με τις 12 πρώτες επιλεγμένες μεταβλητές:

	a	**b**	**c**	**d**	**e**	**f**	**TP rate**
a: control	**3**	0	2	0	1	0	0.500
b: dxr	1	**2**	2	0	1	0	0.333
c: dxr+oleu1	1	0	**10**	0	0	0	0.909
d: dxr+oleu2	0	0	1	**3**	0	2	0.500
e: oleu1	0	0	2	1	**1**	1	0.200
f: oleu2	0	0	0	0	1	**5**	0.833
Επιτυχία							60.0%

DXR: ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ

LibSVM μέσω GNU Octave με τις 2 πρώτες επιλεγμένες μεταβλητές:

	a	**b**	**c**	**d**	**e**	**f**	**TP rate**
a: control	**2**	1	2	0	0	1	0.333
b: dxr	2	**0**	4	0	0	0	0.000
c: dxr+oleu1	0	2	**8**	1	0	0	0.727
d: dxr+oleu2	0	0	2	**2**	1	1	0.333
e: oleu1	0	0	1	2	**0**	2	0.000
f: oleu2	0	0	0	1	0	**5**	0.833
Επιτυχία							42.5%

ΣΥΜΠΕΡΑΣΜΑΤΑ

ΣΥΜΠΕΡΑΣΜΑΤΑ

- Η **επιτυχία ενός αλγορίθμου** μηχανικής μάθησης εξαρτάται σε μεγάλο βαθμό από τη **φύση του προβλήματος** στο οποίο εφαρμόζεται. Σε απλούστερα προβλήματα ενδέχεται να αποδίδουν καλύτερα απλούστεροι αλγόριθμοι, ενώ σε συνθετότερα φαίνεται η δύναμη πιο εξελιγμένων αλγορίθμων.
- Χρειάζεται προσοχή για **υπερπροσαρμογή** μοντέλων στα δεδομένα. Χρησιμοποιήστε Cross-Validation και ελέγχετε τους confusion matrices.
- **Συναινετικά μοντέλα** μπορούν να επιτύχουν υψηλότερη ακρίβεια σε σχέση με μεμονωμένα.

ΣΥΜΠΕΡΑΣΜΑΤΑ

- Η **μείωση των μεταβλητών** μπορεί να οδηγήσει σε εξίσου ακριβή αλλά και ταυτόχρονα απλούστερα μοντέλα. Στη Μεταβολομική, η απλότητα των μοντέλων και η επιλογή μεταβλητών είναι πολύ σημαντικά.
- Η μεγάλη ποικιλία εργαλείων data mining και οι προοπτικές βελτίωσής τους **μπορούν να προσφέρουν στη Μεταβολομική** αποτελεσματικότερα μοντέλα και περισσότερες δυνατότητες παραμετροποίησης.

ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

- Εστιάστε σε **συγκεκριμένους αλγορίθμους μάθησης**.
- Μελετήστε διαφορετικές **μεθόδους προεπεξεργασίας** σε συνδυασμό με αλγορίθμους μηχανικής μάθησης.
- Μελετήστε την πλειάδα διαθέσιμων **meta-classifiers**.
- Εστιάστε στην **επιλογή μεταβλητών**, ενδεχομένως με κριτήρια που να συνδυάζουν την ευκολία προσδιορισμού των διαφόρων μεταβολιτών. Δοκιμάστε άλλες μεθόδους, όπως Προσομοιωμένη Ανόπτηση.
- **Υλοποιήστε παράλληλα** κομμάτια της εξόρυξης δεδομένων (πχ Cross-Validation ή γενετική έρευνα) για αύξηση της ταχύτητάς τους.
- Αναπτύξτε έναν **αλγόριθμο ο οποίος να προσαρμόζεται** στα χαρακτηριστικά του προβλήματος και να **βελτιστοποιεί τις παραμέτρους** του.

ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

- Εστιάστε σε **συγκεκριμένους αλγορίθμους μάθησης**.
- Μελετήστε διαφορετικές **μεθόδους προεπεξεργασίας** σε συνδυασμό με αλγορίθμους μηχανικής μάθησης.
- Μελετήστε την πλειάδα διαθέσιμων **meta-classifiers**.
- Εστιάστε στην **επιλογή μεταβλητών**, ενδεχομένως με κριτήρια που να συνδυάζουν την ευκολία προσδιορισμού των διαφόρων μεταβολιτών. Δοκιμάστε άλλες μεθόδους, όπως Προσομοιωμένη Ανόπτηση.
- **Υλοποιήστε παράλληλα** κομμάτια της εξόρυξης δεδομένων (πχ Cross-Validation ή γενετική έρευνα) για αύξηση της ταχύτητάς τους.
- Αναπτύξτε έναν **αλγόριθμο ο οποίος να προσαρμόζεται** στα χαρακτηριστικά του προβλήματος και να **βελτιστοποιεί τις παραμέτρους** του.

ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

- Εστιάστε σε **συγκεκριμένους αλγορίθμους μάθησης**.
- Μελετήστε διαφορετικές **μεθόδους προεπεξεργασίας** σε συνδυασμό με αλγορίθμους μηχανικής μάθησης.
- Μελετήστε την πλειάδα διαθέσιμων **meta-classifiers**.
- Εστιάστε στην **επιλογή μεταβλητών**, ενδεχομένως με κριτήρια που να συνδυάζουν την ευκολία προσδιορισμού των διαφόρων μεταβολιτών. Δοκιμάστε άλλες μεθόδους, όπως Προσομοιωμένη Ανόπτηση.
- **Υλοποιήστε παράλληλα** κομμάτια της εξόρυξης δεδομένων (πχ Cross-Validation ή γενετική έρευνα) για αύξηση της ταχύτητάς τους.
- Αναπτύξτε έναν **αλγόριθμο ο οποίος να προσαρμόζεται** στα χαρακτηριστικά του προβλήματος και να **βελτιστοποιεί τις παραμέτρους** του.

ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

- Εστιάστε σε **συγκεκριμένους αλγορίθμους μάθησης**.
- Μελετήστε διαφορετικές **μεθόδους προεπεξεργασίας** σε συνδυασμό με αλγορίθμους μηχανικής μάθησης.
- Μελετήστε την πλειάδα διαθέσιμων **meta-classifiers**.
- Εστιάστε στην **επιλογή μεταβλητών**, ενδεχομένως με κριτήρια που να συνδυάζουν την ευκολία προσδιορισμού των διαφόρων μεταβολιτών. Δοκιμάστε άλλες μεθόδους, όπως Προσομοιωμένη Ανόπτηση.
- **Υλοποιήστε παράλληλα** κομμάτια της εξόρυξης δεδομένων (πχ Cross-Validation ή γενετική έρευνα) για αύξηση της ταχύτητάς τους.
- Αναπτύξτε έναν **αλγόριθμο ο οποίος να προσαρμόζεται** στα χαρακτηριστικά του προβλήματος και να **βελτιστοποιεί τις παραμέτρους** του.

ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

- Εστιάστε σε **συγκεκριμένους αλγορίθμους μάθησης**.
- Μελετήστε διαφορετικές **μεθόδους προεπεξεργασίας** σε συνδυασμό με αλγορίθμους μηχανικής μάθησης.
- Μελετήστε την πλειάδα διαθέσιμων **meta-classifiers**.
- Εστιάστε στην **επιλογή μεταβλητών**, ενδεχομένως με κριτήρια που να συνδυάζουν την ευκολία προσδιορισμού των διαφόρων μεταβολιτών. Δοκιμάστε άλλες μεθόδους, όπως Προσομοιωμένη Ανόπτηση.
- **Υλοποιήστε παράλληλα** κομμάτια της εξόρυξης δεδομένων (πχ Cross-Validation ή γενετική έρευνα) για αύξηση της ταχύτητάς τους.
- Αναπτύξτε έναν **αλγόριθμο ο οποίος να προσαρμόζεται** στα χαρακτηριστικά του προβλήματος και να **βελτιστοποιεί τις παραμέτρους** του.

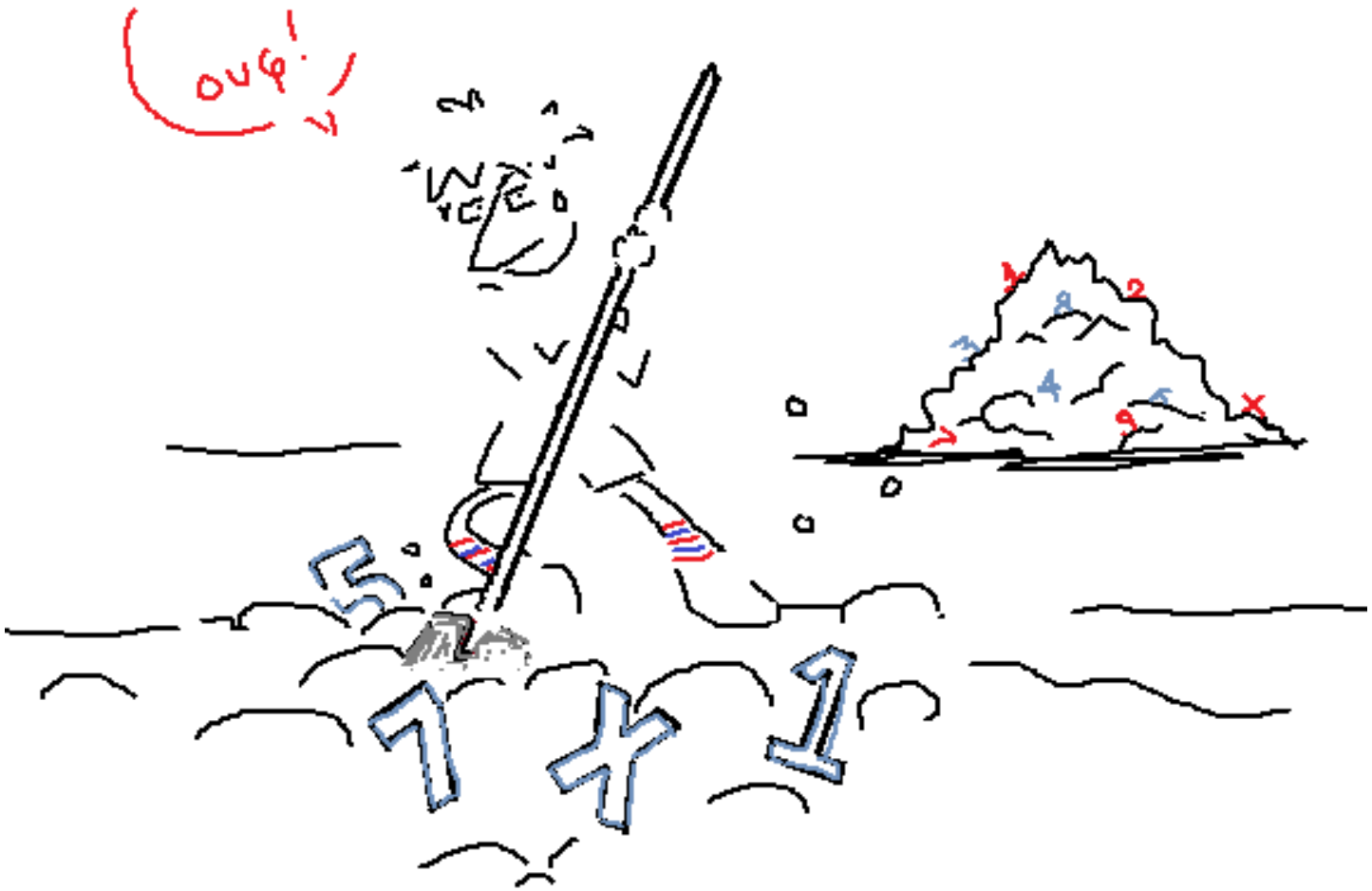
ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

- Εστιάστε σε **συγκεκριμένους αλγορίθμους μάθησης**.
- Μελετήστε διαφορετικές **μεθόδους προεπεξεργασίας** σε συνδυασμό με αλγορίθμους μηχανικής μάθησης.
- Μελετήστε την πλειάδα διαθέσιμων **meta-classifiers**.
- Εστιάστε στην **επιλογή μεταβλητών**, ενδεχομένως με κριτήρια που να συνδυάζουν την ευκολία προσδιορισμού των διαφόρων μεταβολιτών. Δοκιμάστε άλλες μεθόδους, όπως Προσομοιωμένη Ανόπτηση.
- **Υλοποιήστε παράλληλα** κομμάτια της εξόρυξης δεδομένων (πχ Cross-Validation ή γενετική έρευνα) για αύξηση της ταχύτητάς τους.
- Αναπτύξτε έναν **αλγόριθμο ο οποίος να προσαρμόζεται** στα χαρακτηριστικά του προβλήματος και να **βελτιστοποιεί τις παραμέτρους** του.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Εργασία σχετική με το **Πρόβλημα 1**:
  H.U. Zacharias et al.: "Analysis of human urine reveals metabolic changes related to the development of acute kidney injury following cardiac surgery". *Metabolomics* **9**, 697–707 (2013).
- Εργασία σχετική με το **Πρόβλημα 2**:
  Κ.Ε. Ιωαννίδης: *«Επίδραση της ολευρωπαΐνης στην χρόνια καρδιακή ανεπάρκεια που προκαλεί η χορήγηση της αδριαμυκίνης. Μελέτη του μηχανισμού δράσης.»*
  Διδακτορική διατριβή, Τμήμα Φαρμακευτικής ΕΚΠΑ. (2012)

ΕΡΩΤΗΣΕΙΣ;



Οι εικόνες στις διαφάνειες προέρχονται από:

- 7, 12, 15, 30–38: own work
- 6, 14, 41–43, 53, 54: own work (MetaboAnalyst, WEKA, LibSVM)
- 3: δημιουργήθηκε με το Wordle.net
- 8: παρουσίαση διδακτορικού του Κ. Ιωαννίδη
- 9, 19, 22, 28: Wikimedia Commons (CC BY-SA ή public domain)
- 69: Μυρτώ Παπαδημητράκη (CC BY-SA)

Επικοινωνία: **makishourdakis@gmail.com**