Τοπολογική Ανάλυση Δεδομένων

Εαρινό Εξάμηνο 2026

Φροντιστηριακή Διάλεξη 3: Εισαγωγή Στη Μηχανική Μάθηση

Διδάσκοντες:

Αθανάσιος Ανδρικόπουλος
Ιωάννης Γουναρίδης

Επικουρικό Έργο: Αλέξανδρος - Μάριος Αφράτης

Πάτρα, 19 Μαρτίου 2026

Πίνακας Περιεχομένων

Επιβλεπόμενη Μάθηση (Supervised Learning)

Βασικές Έννοιες (Basic Concepts)

- Ένα σύνολο **αντικειμένων** (items)
- Κάθε αντικείμενο χαρακτηρίζεται από **ιδιότητες** (attributes) $(a_1, a_2, \ldots, a_k)$
- Κάθε αντικείμενο ανήκει σε μία **κλάση** (class) ή **κατηγορία** (category) c
- Δοθέντος συνόλου παραδειγμάτων, πρόβλεψη του c για νέο αντικείμενο με ιδιότητες $(a'_1, a'_2, \ldots, a'_k)$
- Τα παραδείγματα αποκαλούνται **δεδομένα εκπαίδευσης** (training data)
- Στόχος: εκμάθηση **μαθηματικού μοντέλου** (mathematical model) που *γενικεύει* τα δεδομένα εκπαίδευσης
- Το μοντέλο πρέπει να επεκτείνεται σε *προηγουμένως άγνωστες* εισόδους (previously unseen inputs)

Πρόβλημα Ταξινόμησης (Classification Problem)

Συνήθως δυαδικό — δύο κλάσεις.

Παράδειγμα: Δεδομένα Αίτησης Δανείου

ID	Age	Has_job	Own_house	Credit_rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**

Σχ. 3.1: Δεδομένα Αίτησης Δανείου

Πίνακας Περιεχομένων

Βασικές Παραδοχές Μηχανικής Μάθησης

Θεμελιώδης Παραδοχή (Fundamental Assumption)

Η κατανομή των παραδειγμάτων εκπαίδευσης είναι *πανομοιότυπη* με την κατανομή των αόρατων δεδομένων.

Τι σημαίνει να μαθαίνουμε από δεδομένα; (What does it mean to learn?)

- Κατασκευή μοντέλου που αποδίδει **καλύτερα από τυχαία εικασία** (better than random guessing)
- Στα δεδομένα δανείου, πάντοτε «Yes/Ναι» θα ήταν σωστό $\approx 9/15$ φορές
- Η απόδοση πρέπει να *βελτιώνεται* με περισσότερα δεδομένα εκπαίδευσης

Αξιολόγηση Μοντέλου (Model Evaluation)

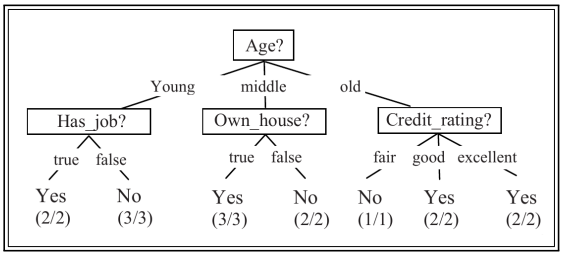
- Το μοντέλο βελτιστοποιείται για τα δεδομένα εκπαίδευσης — πόσο καλά δουλεύει σε προηγουμένως άγνωστα δεδομένα;
- Δεν γνωρίζουμε τις σωστές απαντήσεις εκ των προτέρων — διαφορετικό από τον συνηθισμένο έλεγχο λογισμικού

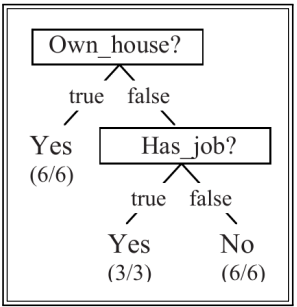
Πίνακας Περιεχομένων

Αλγόριθμος Κατασκευής

- A: τρέχον σύνολο ιδιοτήτων
- Επιλογή $a \in A$, δημιουργία παιδιών αντίστοιχα με τον διαμερισμό, με ιδιότητες $A \setminus \{a\}$
- **Κριτήριο τερματισμού** (Stopping criterion):
  - Ο τρέχων κόμβος έχει ομοιόμορφη ετικέτα κλάσης
  - $A = \emptyset$ — δεν υπάρχουν άλλες ιδιότητες
- Αν ένας φυλλοκόμβος δεν είναι ομοιόμορφος, χρήση **πλειοψηφικής κλάσης** (majority class) ως πρόβλεψη
  - Πανομοιότυπος συνδυασμός χαρακτηριστικών, αλλά διαφορετικές κλάσεις
  - Τα χαρακτηριστικά δεν καταγράφουν όλα τα κριτήρια που χρησιμοποιούνται για την ταξινόμηση

Παράδειγμα Δέντρου Απόφασης (Δεδομένα Αίτησης Δανείου)





Σχ. 3.2: Παράδειγμα Δέντρου Απόφασης

Ευρετική Για Μικρά Δέντρα (Heuristic For Small Trees)

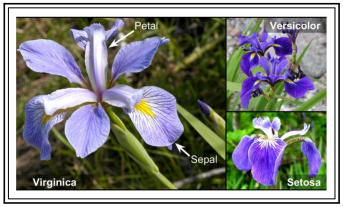
Στόχος: Μικρά Δέντρα

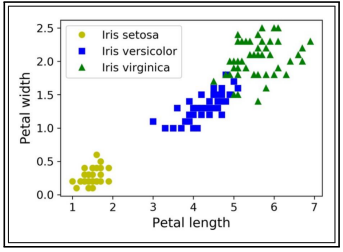
- Προτίμηση **μικρών δέντρων** (small trees) — εύρεση του μικρότερου είναι *αδύνατο υπολογιστικά* (intractable)
- Στόχος: διαμερισμός με ομοιόμορφη κατηγορία — **ανόθευτο φύλλο** (pure leaf)
- **Νοθευμένος κόμβος** (Impure node) — καλύτερη πρόβλεψη είναι η πλειοψηφική τιμή
- **Αναλογία μειοψηφίας** (Minority ratio) = **νοθεία** (impurity)

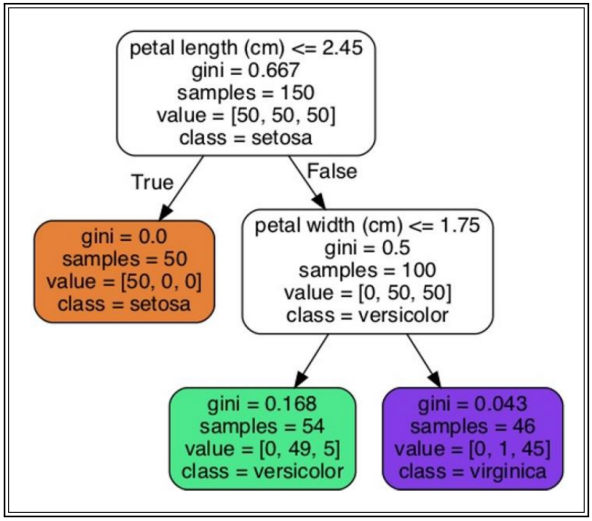
Ευρετική Στρατηγική (Heuristic Strategy)

- Μείωση νοθείας όσο το δυνατόν περισσότερο
- Για κάθε ιδιότητα, υπολογισμός *σταθμισμένης μέσης νοθείας* (weighted average impurity) των παιδιών
- Επιλογή εκείνης με *ελάχιστη* σταθμισμένη νοθεία

Οπτικοποίηση & Διαχωρισμός (Σύνολο Δεδομένων Iris)







Σχ. 3.3: Σύνολο Δεδομένων Iris

Περιγραφή & Παρατηρήσεις (Σύνολο Δεδομένων Iris)

Περιγραφή (Description)

- Τρία είδη: *iris setosa*, *iris versicolor*, *iris virginica*
- **150 λουλούδια**: μήκος/πλάτος σεπάλου και πετάλου
- Γράφημα διασποράς για **μήκος** και **πλάτος πετάλου**

Δέντρο Απόφασης με Gini

- Ρίζα: **μήκος πετάλου** ≤ 2.45, gini $= 0.667$, $n = 150$
- Αληθής κλάδος: gini $= 0$, $n = 50$ (*setosa*)
- Ψευδής κλάδος: **πλάτος πετάλου** ≤ 1.75, gini $= 0.5$, $n = 100$

Αποτέλεσμα Διαχωρισμού

- *Setosa*: gini $= 0$, $n = 50$, $[50, 0, 0]$
- *Versicolor*: gini $= 0.168$, $n = 54$, $[0, 49, 5]$
- *Virginica*: gini $= 0.043$, $n = 46$, $[0, 1, 45]$

Παρατήρηση

Χρήση της **νοθείας Gini** (Gini impurity) ως κριτήριο διαχωρισμού.

Επερώτηση Αριθμητικών Ιδιοτήτων (Querying Numerical Attributes)

Πρόβλημα

- Επιλογή τιμής v στο εύρος και έλεγχος $A \leq v$ — *άπειρες επιλογές* για v
- Μόνο n τιμές για A στα δεδομένα εκπαίδευσης

Λύση: Ωφέλιμα Διαστήματα (Useful Intervals)

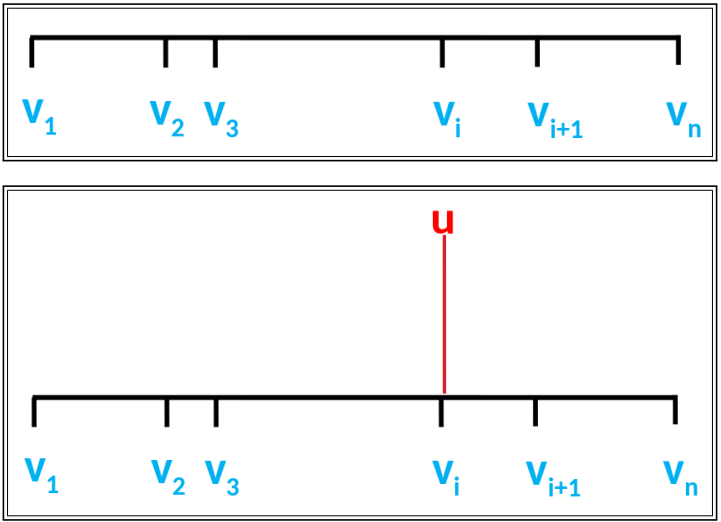
- $[v_i, v_{i+1}]$: για κάθε $v_i \leq u < v_{i+1}$, το ερώτημα $A \leq u$ δίνει την *ίδια απάντηση*
- Μόνο $\mathbf{n - 1}$ **ωφέλιμα διαστήματα** (useful intervals) να ελεγχθούν
- Επιλογή μέσης τιμής $u_i = (v_i + v_{i+1})/2$ ως τιμή ερωτήματος για κάθε διάστημα

Κανόνας Επιλογής

$$u_i = \frac{v_i + v_{i+1}}{2} \quad \text{(σημείο τομής – cut point)}$$

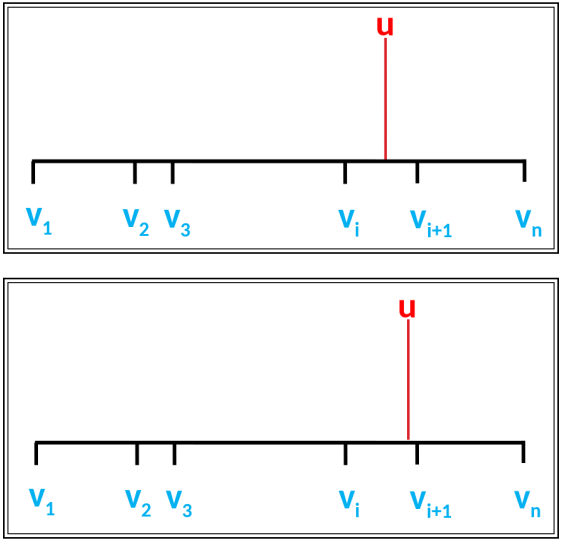
Επιλογή του ερωτήματος $A \leq u_i$ με τη **μέγιστη μείωση νοθείας** (maximum impurity reduction).

Οπτικοποίηση (1/3)



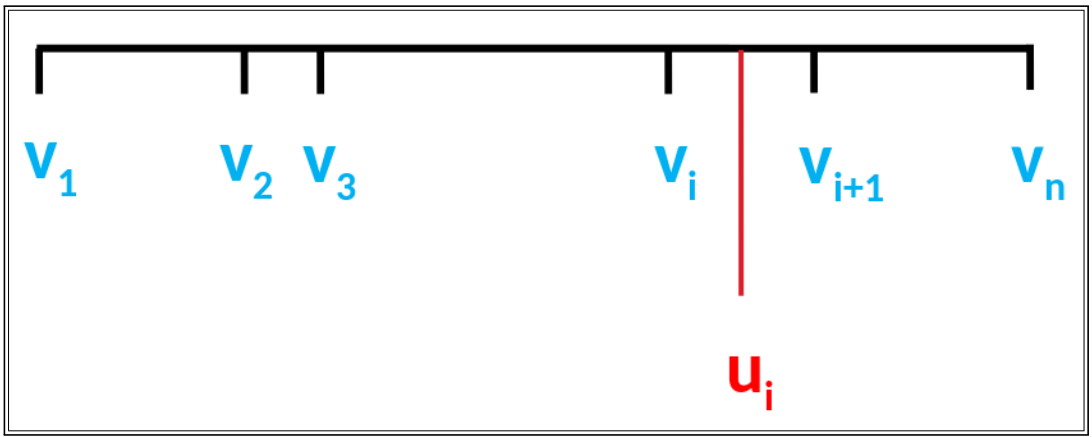
Σχ. 3.4: Οπτικοποίηση Διαδικασίας (Βήματα 1-2)

Οπτικοποίηση (2/3)



Σχ. 3.5: Οπτικοποίηση Διαδικασίας (Βήματα 3-4)

Οπτικοποίηση (3/3)



Σχ. 3.6: Οπτικοποίηση Διαδικασίας (Βήμα 5)

Πίνακας Περιεχομένων

Έλεγχος Μοντέλου Επιβλεπόμενης Μάθησης

Πώς επικυρώνουμε λογισμικό;

- Σύνολο δοκιμών από επιλεγμένες εισόδους
- Σύγκριση εξόδου με αναμενόμενες απαντήσεις

Τι ισχύει για μοντέλα ταξινόμησης;

- Εξ ορισμού, εφαρμόζονται σε δεδομένα όπου το αποτέλεσμα είναι *άγνωστο*
- Αν η αναμενόμενη απάντηση ήταν γνωστή, θα είχαμε ντετερμινιστική λύση — το μοντέλο δεν θα χρειαζόταν!

Ποιό είναι το κριτήριο αξιολόγησης;

Με τι κριτήριο μπορούμε να αξιολογήσουμε ένα μοντέλο εποπτευόμενης μάθησης;

Δημιουργία Συνόλου Δοκιμών (Creating A Test Set)

- Τα δεδομένα εκπαίδευσης είναι *επισημειωμένα* (labelled)
- **Διαχωρισμός** μέρους δεδομένων εκπαίδευσης για δοκιμή
- Ορολογία: **σύνολο εκπαίδευσης** (training set) και **σύνολο δοκιμών** (test set)
- Χρήση **στρωματοποιημένης δειγματοληψίας** (stratified sampling) για διατήρηση σχετικών αναλογιών

Πίνακας Περιεχομένων

Παράδειγμα Συνόλου Δεδομένων Προς Ταξινόμηση

Σχ. 3.7: Σύνολο Δεδομένων Προς Ταξινόμηση

Τι Μετράμε; (What Are We Measuring?)

Ακρίβεια (Accuracy)

- Κλάσμα εισόδων όπου η ταξινόμηση είναι σωστή

Ασύμμετρες Καταστάσεις (Asymmetric Situations)

- Λιγότερο από 1% των συναλλαγών πιστωτικών καρτών είναι απάτη
- Τετριμμένος ταξινομητής — πάντοτε απάντηση «Όχι»
- Ακρίβεια $> 99\%$, αλλά **πρακτικά μη χρήσιμος**!

Επιθετικός Ταξινομητής (Aggressive Classifier)

- Χαρακτηρίζει οριακά «Όχι» ως «Ναι»
- **Ψευδώς θετικά** (False Positives)

Συντηρητικός Ταξινομητής (Cautious Classifier)

- Χαρακτηρίζει οριακά «Ναι» ως «Όχι»
- **Ψευδώς αρνητικά** (False Negatives)

Μητρώο Σύγχυσης & Μετρικές Απόδοσης (Confusion Matrix & Performance Measures)

Μητρώο Σύγχυσης (Confusion Matrix)

	Ταξινομήθηκε Θετικό	**Ταξινομήθηκε Αρνητικό**
Πραγματικά Θετικό	Αληθώς Θετικό (TP)	Ψευδώς Αρνητικό (FN)
Πραγματικά Αρνητικό	Ψευδώς Θετικό (FP)	Αληθώς Αρνητικό (TN)

Ακρίβεια (Precision)

Τι ποσοστό θετικών προβλέψεων είναι σωστό;

$$\text{Ακρίβεια} = \frac{TP}{TP + FP}$$

Ανάκληση (Recall)

Τι ποσοστό πραγματικών θετικών περιπτώσεων ανακαλύπτεται;

$$\text{Ανάκληση} = \frac{TP}{TP + FN}$$

Αντιστάθμιση Ακρίβειας–Ανάκλησης (Precision–Recall Tradeoff)

Παραδείγματα

Σενάριο	TP	FN	FP	TN
Ακρίβεια 1, Ανάκληση 0.01	1	99	0	900
Ανάκληση 0.4, Ακρίβεια 0.29	40	60	100	800
Ανάκληση 0.99, Ακρίβεια 0.165	99	1	500	400

Κύρια Παρατήρηση

- **Αυστηροί ταξινομητές**: λιγότερα ψευδώς θετικά (υψηλή ακρίβεια), χάνουν περισσότερα πραγματικά θετικά (χαμηλή ανάκληση)
- **Επιτρεπτικοί ταξινομητές**: εντοπίζουν περισσότερα θετικά (υψηλή ανάκληση), αλλά με περισσότερα ψευδώς θετικά (χαμηλή ακρίβεια)

Πίνακας Περιεχομένων

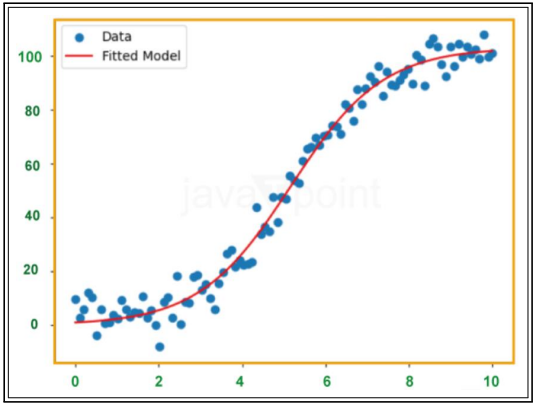
Παλινδρόμηση (Regression)

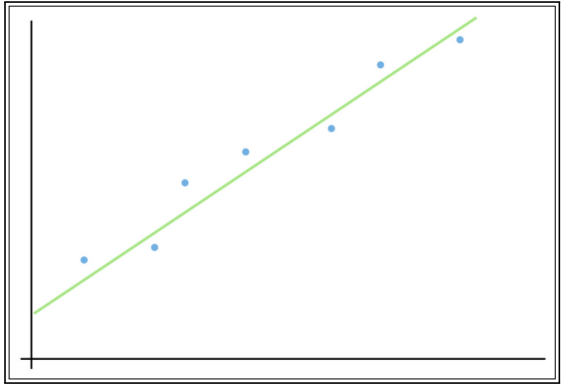
- Προσαρμογή καμπύλης σε σύνολο παρατηρήσεων — **παλινδρόμηση** (regression)
- Απλούστερη περίπτωση: γραμμή — **γραμμική παλινδρόμηση** (linear regression):

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_k x_k$$

- Για ταξινόμηση: ορισμός **κατωφλίου** (threshold):
  - Έξοδος κάτω από κατώφλι: 0 (Όχι)
  - Έξοδος πάνω από κατώφλι: 1 (Ναι)

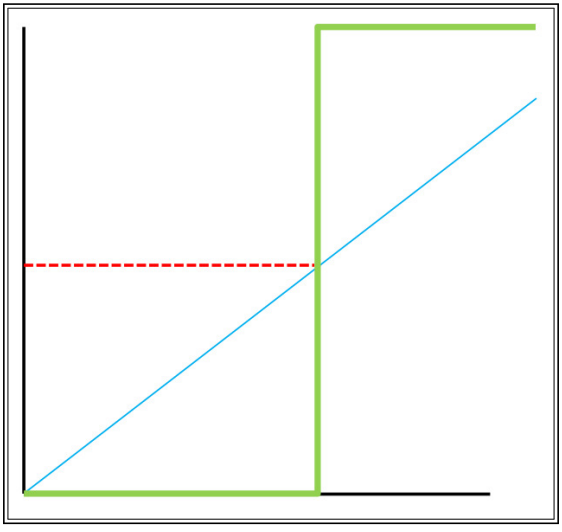
Οπτικοποίηση Παλινδρόμησης (1/3)





Σχ. 3.8: Παράδειγμα Παλινδρόμησης (Αριστερά) & Γραμμικής Παλινδρόμησης (Δεξιά)

Οπτικοποίηση Παλινδρόμησης (2/3)



Σχ. 3.9: Ορισμός Κατωφλίου Για Ταξινόμηση

Συνάρτηση Sigmoid (Sigmoid Function)

Εξομάλυνση του βήματος — **σιγμοειδής συνάρτηση** (sigmoid function):

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \qquad z = \theta_0 + \theta_1 x_1 + \cdots + \theta_k x_k$$

$$\sigma(z) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \cdots + \theta_k x_k)}}$$

Η sigmoid απεικονίζει κάθε πραγματική τιμή ομαλά στο $(0, 1)$ — φυσικά ερμηνεύεται ως πιθανότητα.

Δεδομένα Εκπαίδευσης

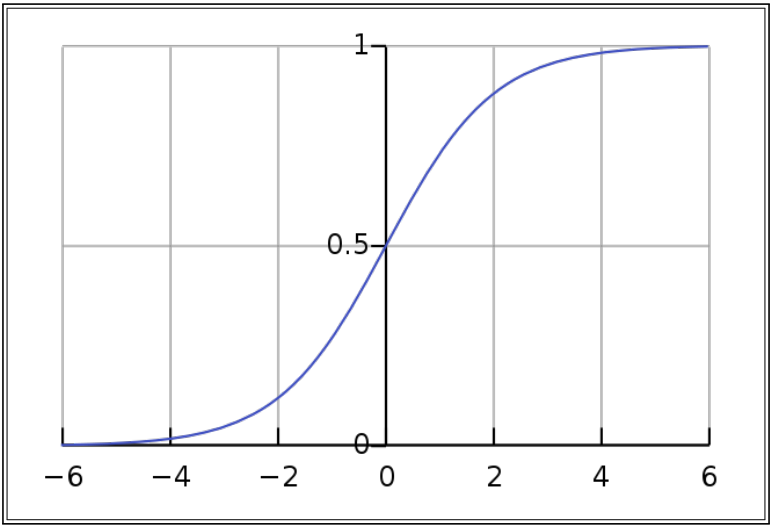
- Είσοδος εκπαίδευσης: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
  - Κάθε x_i είναι διάνυσμα $(x_i^1, \ldots, x_i^k)$
  - $y_i \in \{0, 1\}$ είναι η πραγματική έξοδος
- Στόχος: εύρεση $\theta = (\theta_0, \theta_1, \ldots, \theta_k)$ ώστε, με

$$z = \theta_0 + \theta_1 x_1 + \cdots + \theta_k x_k$$

να χρησιμοποιούμε $\sigma(z) = \dfrac{1}{1 + e^{-z}}$ για ταξινόμηση εισόδων

- Αν $\sigma(z) > 0.5$, έξοδος 1, αλλιώς 0

Οπτικοποίηση Παλινδρόμησης (3/3)



Σχ. 3.10: Εξομάλυνση Βήματος Μέσω Sigmoid (Κατασκευή Ταξινομητή)

Συνάρτηση Κόστους (Cost / Loss Function)

- Πόσο καλή είναι η εκτίμησή μας για θ;
- Ορισμός **συνάρτησης κόστους** (cost function) $J(\theta)$ που μετρά απόκλιση μεταξύ προβλέψεων και πραγματικών απαντήσεων

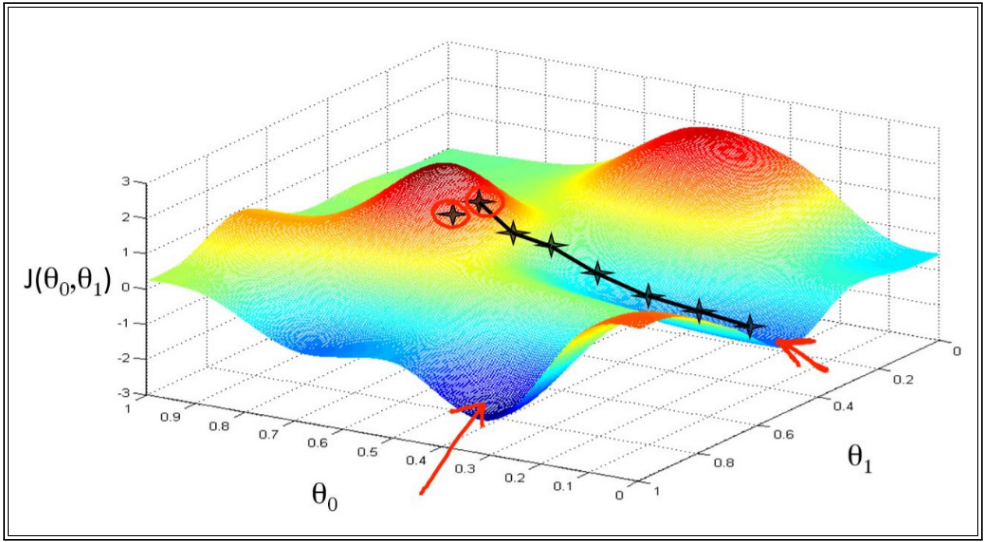
Αλγόριθμος Κατηφορικής Κλίσης (Gradient Descent)

- Πώς μεταβάλλεται το κόστος ως προς $\theta = (\theta_0, \theta_1, \ldots, \theta_k)$;
  - Κλίσεις (Gradients): $\frac{\partial}{\partial \theta_i} J(\theta)$
- Προσαρμογή κάθε παραμέτρου *αντίθετα* προς την κλίση:

$$\theta_i = \theta_i - \alpha \frac{\partial}{\partial \theta_i} J(\theta)$$

- Επανάληψη μέχρι **σύγκλιση** (convergence)

Οπτικοποίηση Κατηφορικής Κλίσης (Gradient Descent Visualization)



Σχ. 3.11: Αλγόριθμος Κατηφορικής Κλίσης

Πιθανοφάνεια (Likelihood)

Εκτιμητής Μέγιστης Πιθανοφάνειας (MLE)

- Εύρεση θ που μεγιστοποιεί:

$$L(\theta) = \prod_{i=1}^{n} \Pr(y_i \mid x_i; \theta)$$

- Ισοδύναμα, μεγιστοποίηση **λογαριθμικής πιθανοφάνειας** (log-likelihood):

$$\ell(\theta) = \sum_{i=1}^{n} \log \Pr(y_i \mid x_i; \theta)$$

- Ευκολότερο να εργαζόμαστε με άθροισμα παρά με γινόμενο

Συνάρτηση Απώλειας Λογιστικής Παλινδρόμησης (Loss Function)

Έστω $h_\theta(x_i) = \sigma(z_i)$. Τότε:

$$\Pr(y_i = 1 \mid x_i;\, \theta) = h_\theta(x_i), \qquad \Pr(y_i = 0 \mid x_i;\, \theta) = 1 - h_\theta(x_i)$$

Πιθανοφάνεια:

$$L(\theta) = \prod_{i=1}^{n} h_\theta(x_i)^{y_i} \cdot (1 - h_\theta(x_i))^{1-y_i}$$

Μεγιστοποίηση Λογιστικής Πιθανοφάνειας:

$$\ell(\theta) = \sum_{i=1}^{n} \left[y_i \log h_\theta(x_i) + (1 - y_i) \log(1 - h_\theta(x_i)) \right]$$

Ελαχιστοποίηση Σταυρωτής Εντροπίας (Cross-Entropy)

$$\mathcal{L}(\theta) = -\sum_{i=1}^{n} \left[y_i \log h_\theta(x_i) + (1 - y_i) \log(1 - h_\theta(x_i)) \right]$$

Σταυρωτή Εντροπία & Κατηφορική Κλίση (Cross Entropy & Gradient Descent)

Με $C = -[y \ln(\sigma(z)) + (1-y)\ln(1-\sigma(z))]$, η κλίση ως προς θ_j:

$$\frac{\partial C}{\partial \theta_j} = -\left(\frac{y}{\sigma(z)} - \frac{1-y}{1-\sigma(z)}\right)\sigma'(z)\, x_j$$

$$= -\left(\frac{y(1-\sigma(z)) - (1-y)\sigma(z)}{\sigma(z)(1-\sigma(z))}\right)\sigma'(z)\, x_j$$

Χρησιμοποιώντας $\sigma'(z) = \sigma(z)(1-\sigma(z))$:

$$\frac{\partial C}{\partial \theta_j} = (\sigma(z) - y)\, x_j, \qquad \frac{\partial C}{\partial \theta_0} = \sigma(z) - y$$

Ερμηνεία

- Οι κλίσεις είναι ανάλογες του $\sigma(z) - y$
- Όσο μεγαλύτερο το σφάλμα, τόσο ταχύτερη η εκμάθηση

Πίνακας Περιεχομένων

Γεωμετρική Θεώρηση Επιβλεπόμενης Μάθησης (Geometric View)

Δεδομένα ως Σημεία στον Χώρο

- Τα δεδομένα νοούνται ως **σημεία στον χώρο** (points in space)
- Εύρεση **διαχωριστικής καμπύλης** (separating curve / surface)

Διαχωρίσιμη Περίπτωση (Separable Case)

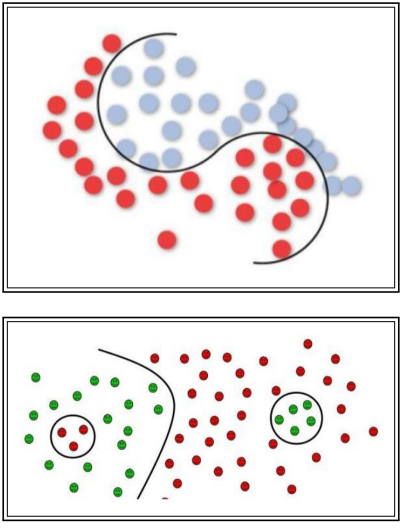
- Κάθε κλάση αποτελεί συνεκτική περιοχή
- Μία μόνο καμπύλη μπορεί να τις διαχωρίσει

Πιο Σύνθετο Σενάριο (More Complex Scenario)

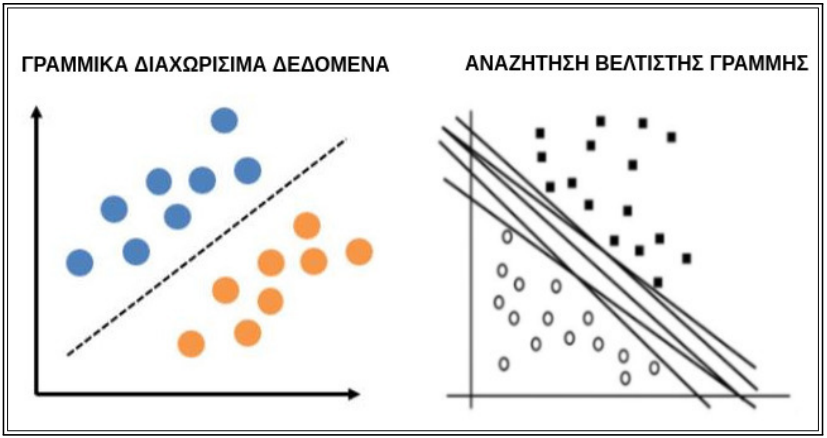
- Οι κλάσεις σχηματίζουν πολλαπλές συνεκτικές περιοχές
- Απαιτούνται πολλαπλοί διαχωριστές

Γραμμικοί Διαχωριστές (Linear Separators)

Απλούστερη περίπτωση: **γραμμικά διαχωρίσιμα δεδομένα** (linearly separable data). Πολλές γραμμές είναι υποψήφιες — ποια είναι η βέλτιστη;



Σχ. 3.12: Οπτικοποίηση Επιβλεπόμενης Μάθησης

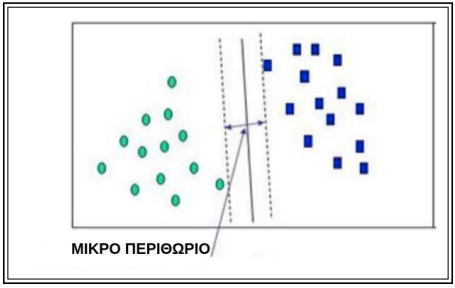


Σχ. 3.13: Διαχωρισιμότητα & Βέλτιστη Γραμμή

Περιθώριο & Ταξινομητής Μέγιστου Περιθωρίου (Margin)

Ορισμός Περιθωρίου (Margin Definition)

- Κάθε διαχωριστής ορίζει ένα **περιθώριο** (margin):
  - Κενός διάδρομος που διαχωρίζει τα σημεία
  - Ο διαχωριστής είναι η κεντρική γραμμή του περιθωρίου
- Ευρύτερο περιθώριο $\Rightarrow$ πιο **ισχυρός ταξινομητής** (robust classifier)
- Βέλτιστος ταξινομητής: **μεγιστοποιεί το πλάτος** του περιθωρίου
- Το περιθώριο ορίζεται από σημεία εκπαίδευσης στο όριο
  - **Διανύσματα Υποστήριξης** (Support Vectors)



Σχ. 3.14: Περιθώρια & Διανύσματα Υποστήριξης

Εύρεση Ταξινομητή Μέγιστου Περιθωρίου (Finding A Max. Mar. Cl.)

Γραμμικός Ταξινομητής

- $w_1x_1 + \cdots + w_nx_n + b > 0$: ταξινόμηση ναι, $+1$
- $w_1x_1 + \cdots + w_nx_n + b < 0$: ταξινόμηση όχι, -1

Κλιμάκωση Περιθωρίου

- $w_1x_1 + \cdots + w_nx_n + b > 1$: ταξινόμηση ναι, $+1$
- $w_1x_1 + \cdots + w_nx_n + b < -1$: ταξινόμηση όχι, -1

Κάθετη απόσταση στο πλησιέστερο διάνυσμα υποστήριξης (Πυθαγόρειο Θεώρημα): $\frac{1}{|w|}$, όπου $|w| = \sqrt{w_1^2 + \cdots + w_n^2}$.

Πρόβλημα Βελτιστοποίησης (Optimization Problem)

Μεγιστοποίηση $\frac{2}{|w|} \Leftrightarrow$ **Ελαχιστοποίηση** $\frac{|w|}{2}$ υπό γραμμικούς περιορισμούς. Αυτό είναι πρόβλημα **τετραγωνικής βελτιστοποίησης** (quadratic optimization).

Λύση Προβλήματος Βελτιστοποίησης

- **Κυρτή βελτιστοποίηση** (Convex optimization) — επιλύσιμο με υπολογιστικές τεχνικές
- Λύση εκφρασμένη με **πολλαπλασιαστές Lagrange** (Lagrange multipliers) $\alpha_1, \ldots, \alpha_N$
- $\alpha_i \neq 0$ αν και μόνο αν x_i είναι διάνυσμα υποστήριξης
- Ταξινομητής για νέα είσοδο z:

$$\text{sign}\left[\sum_{i \in sv} y_i \alpha_i (x_i \cdot z) + b\right]$$

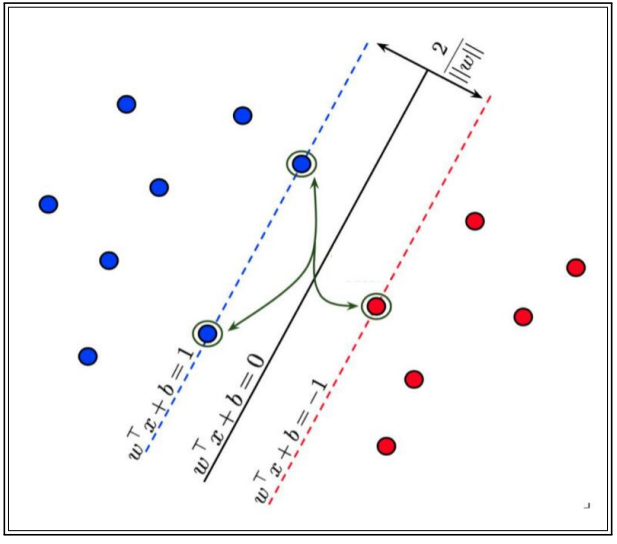
Βασική Ιδιότητα (Key Property)

- Η λύση εξαρτάται *μόνο* από τα διανύσματα υποστήριξης
- Αν προσθέσουμε δεδομένα μακριά από τα διανύσματα υποστήριξης, ο διαχωριστής *δεν αλλάζει*

Μη Γραμμικά Διαχωρίσιμα Δεδομένα

Γεωμετρικός μετασχηματισμός σε υψηλότερη διάσταση — **Μέθοδοι Πυρήνα** (Kernel Methods)

Οπτικοποίηση Μηχανής Διανυσμάτων Υποστήριξης (SVM)



Σχ. 3.15: Οπτικοποίηση Μηχανής Διανυσμάτων Υποστήριξης (SVM)

SVM Μετά Μετασχηματισμό

Μετά την εφαρμογή μετασχηματισμού φ, ο ταξινομητής γίνεται:

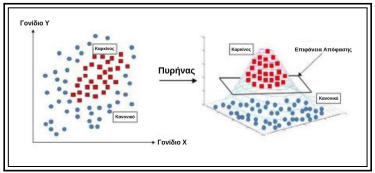
$$\text{sign}\left[\sum_{i \in sv} y_i \alpha_i \left(\varphi(x_i) \cdot \varphi(z)\right) + b\right]$$

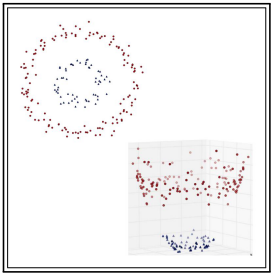
Χρειαζόμαστε μόνο *εσωτερικά γινόμενα* (dot products) στον μετασχηματισμένο χώρο.

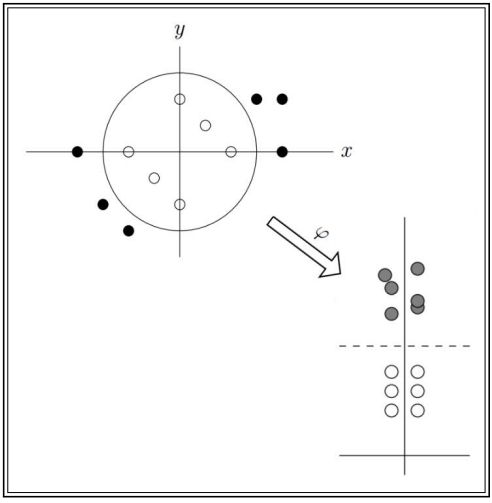
Ορισμός Πυρήνα (Kernel Definition)

K είναι **πυρήνας** (kernel) για μετασχηματισμό φ αν $K(x, z) = \varphi(x) \cdot \varphi(z)$.

Οπτικοποίηση Πυρήνων







Σχ. 3.16: Οπτικοποίηση Πυρήνων

Γνωστοί Πυρήνες (Known Kernels)

- **Πολυωνυμικοί πυρήνες** (Polynomial kernels): $K(x, z) = (1 + x \cdot z)^k$
- **Γκαουσιανή ακτινική βάση** (Gaussian RBF): $K(x, z) = e^{-c|x-z|^2}$
- Κάθε $K(x, z)$ που αντιπροσωπεύει *μέτρο ομοιότητας*

Εγκυρότητα Πυρήνων: Θεώρημα Mercer

Αν γνωρίζουμε ότι K είναι πυρήνας για κάποιο φ, μπορούμε να χρησιμοποιήσουμε K χωρίς να γνωρίζουμε τον φ.

Πίνακας Περιεχομένων

Περιορισμοί Μοντέλων Ταξινόμησης

Βασικοί Περιορισμοί

- **Μεροληψία (Bias)**: Η εκφραστικότητα του μοντέλου περιορίζει την ταξινόμηση
  - Π.χ., γραμμική λογιστική παλινδρόμηση
- **Διακύμανση (Variance)**: Μεταβολή μοντέλου βάσει δείγματος δεδομένων εκπαίδευσης
  - Η μορφή δέντρου απόφασης μεταβάλλεται με την κατανομή εισόδων
  - Μοντέλα υψηλής διακύμανσης είναι εκφραστικά αλλά *ασταθή*
  - Αρχικά, ένα δέντρο απόφασης μπορεί να αποτυπώσει αυθαίρετα σύνθετα κριτήρια

Κίνδυνος Υπερπροσαρμογής (Overfitting)

Μοντέλο συνδεδεμένο υπερβολικά στενά με το σύνολο εκπαίδευσης — κακή γενίκευση σε προηγουμένως άγνωστα δεδομένα.

Πίνακας Περιεχομένων

Μοντέλα Συνόλου (Ensemble Models)

Βασική Ιδέα

- Ακολουθία ανεξάρτητων συνόλων εκπαίδευσης $D_1, D_2, \ldots, D_k$
- Δημιουργία μοντέλων $M_1, M_2, \ldots, M_k$
- «Μέσος όρος» αυτού του **συνόλου μοντέλων** (ensemble of models):
  - Για *παλινδρόμηση*: μέση τιμή προβλέψεων
  - Για *ταξινόμηση*: ψηφοφορία και επιλογή πιο δημοφιλούς αποτελέσματος

Πρόκληση

Ανέφικτη λήψη μεγάλου αριθμού ανεξάρτητων συνόλων εκπαίδευσης. Μπορούμε να κατασκευάσουμε ανεξάρτητα μοντέλα από ένα *μόνο* σύνολο εκπαίδευσης;

Συλλογική Ενοποίηση Bootstrap = Bagging (Bootstrap Aggregating)

Bootstrap Δειγματοληψία (Bootstrap Sampling)

Δεδομένα εκπαίδευσης N αντικειμένων: $TD = \{d_1, \ldots, d_N\}$

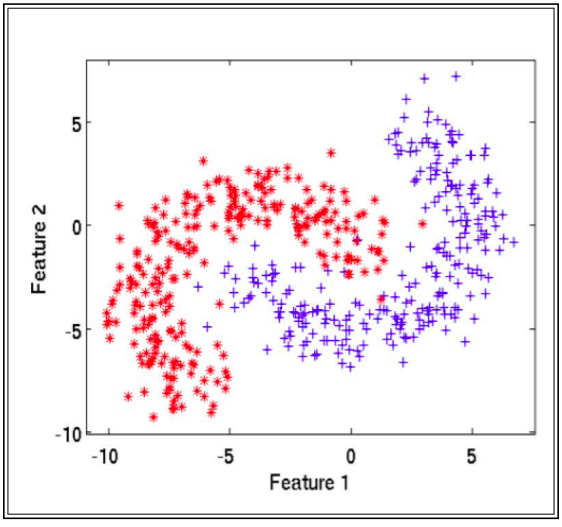
Τυχαίο δείγμα με επανάθεση:

- Επιλογή αντικειμένου τυχαία (πιθ. $\frac{1}{N}$), επιστροφή του, επανάληψη K φορές
- Ορισμένα αντικείμενα επαναλαμβάνονται
- Αν μέγεθος δείγματος $= N$: αναμενόμενα διακριτά αντικείμενα $\approx 63.2\%$

Διαδικασία Bagging

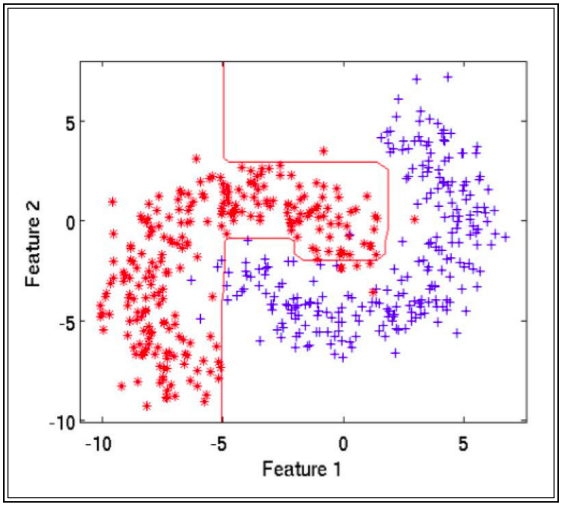
- **Bootstrap δείγμα** μεγέθους N: $\approx 2/3$ πλήρους συνόλου εκπαίδευσης
- k τέτοια δείγματα; κατασκευή μοντέλου για κάθε ένα
- Τελικός ταξινομητής: **πλειοψηφική απάντηση** (majority answer)
- Αποδεδειγμένα *μειώνει τη διακύμανση*

Bagging με Δέντρα Απόφασης — Παράδειγμα (1/6)



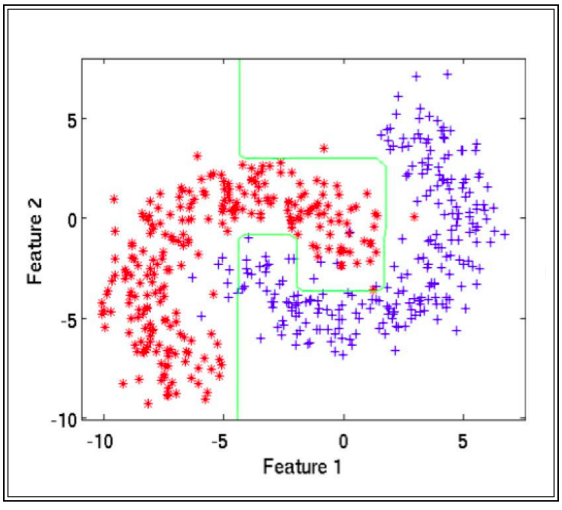
Σχ. 3.17: Δεδομένα Εκπαίδευσης: δύο διαπλεκόμενες κλάσεις σε 2-D χώρο

Bagging με Δέντρα Απόφασης — Παράδειγμα (2/6)



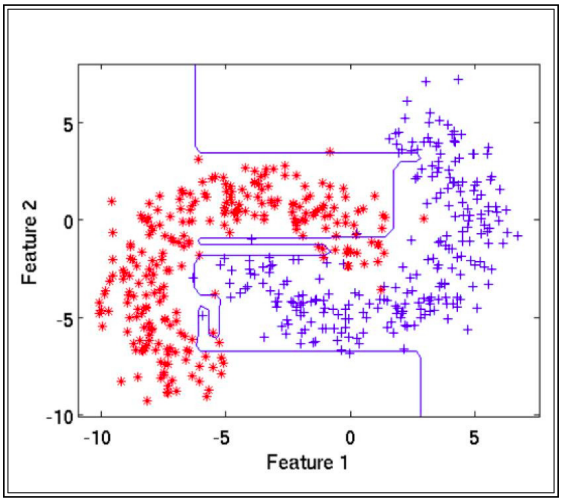
Σχ. 3.18: Δέντρο 1: ορθογώνιες περιοχές απόφασης από πρώτο bootstrap δείγμα

Bagging με Δέντρα Απόφασης — Παράδειγμα (3/6)



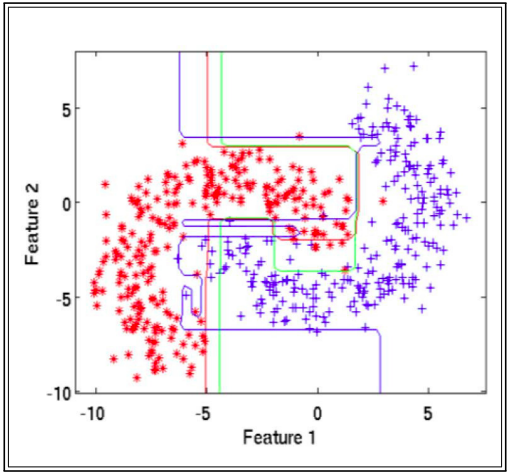
Σχ. 3.19: Δέντρο 2: ελαφρώς διαφορετικό όριο από δεύτερο bootstrap δείγμα

Bagging με Δέντρα Απόφασης — Παράδειγμα (4/6)



Σχ. 3.20: Δέντρο 3: διαφορετικό όριο από το τρίτο δείγμα

Bagging με Δέντρα Απόφασης — Παράδειγμα (5/6)



Σχ. 3.21: Τρία Επικαλυπτόμενα Δέντρα: τα όρια διαφέρουν εμφανώς, δείχνοντας τη διακύμανση

Bagging με Δέντρα Απόφασης — Παράδειγμα (6/6)

Συγκεντρωτικές Οπτικές Παρατηρήσεις

- **Δεδομένα Εκπαίδευσης**: δύο διαπλεκόμενες κλάσεις σε 2-D χώρο
- **Δέντρο 1**: ορθογώνιες περιοχές απόφασης από πρώτο bootstrap δείγμα
- **Δέντρο 2**: ελαφρώς διαφορετικό όριο από δεύτερο bootstrap δείγμα
- **Δέντρο 3**: διαφορετικό όριο από το τρίτο δείγμα
- **Τρία Επικαλυπτόμενα Δέντρα**: τα όρια διαφέρουν εμφανώς, δείχνοντας τη διακύμανση

Αποτέλεσμα Bagging

Το όριο πλειοψηφίας (ψηφοφορίας) είναι **ομαλότερο** και **ακριβέστερο** από οποιοδήποτε μεμονωμένο δέντρο.

Πότε Χρησιμοποιούμε Bagging

Συνθήκες Αποτελεσματικότητας

- Το Bagging βελτιώνει την απόδοση όταν υπάρχει **υψηλή διακύμανση** (high variance)
  - Ανεξάρτητα δείγματα παράγουν αρκετά διαφορετικά μοντέλα
- Μοντέλο με *χαμηλή διακύμανση* δεν θα δείξει βελτίωση

Παράδειγμα: *k*-Πλησιέστεροι Γείτονες (*k*-NN)

- Δοθείσης άγνωστης εισόδου, εύρεση *k* πλησιέστερων γειτόνων και επιλογή πλειοψηφίας
- Σε διαφορετικά υποσύνολα δεδομένων, η διακύμανση στους *k* πλησιέστερους γείτονες είναι σχετικά μικρή
- Τα bootstrap δείγματα παράγουν παρόμοια μοντέλα — **το Bagging δεν βοηθάει ιδιαίτερα**

Πίνακας Περιεχομένων

Τυχαίο Δάσος (Random Forest)

Bagging Δέντρων Απόφασης με Επιπλέον Τεχνική

- k bootstrap δείγματα $D_1, D_2, \ldots, D_k$
- Για κάθε D_i, κατασκευή δέντρου T_i ως εξής:
  - Κάθε αντικείμενο δεδομένων έχει M ιδιότητες
  - Συνήθως: επιλογή μέγιστου κέρδους ακαθαρσίας μεταξύ M ιδιοτήτων
  - Αντί αυτού: ορισμός μικρού ορίου $m < M$ — π.χ. $m = \lfloor \log_2 M \rfloor + 1$
  - Σε κάθε επίπεδο, επιλογή *τυχαίου υποσυνόλου* ιδιοτήτων μεγέθους m
  - Αξιολόγηση μόνο αυτών των m ιδιοτήτων για το επόμενο ερώτημα
- Τελικός ταξινομητής: ψηφοφορία επί αποτελεσμάτων $T_1, T_2, \ldots, T_k$

Ανάλυση Ποσοστού Σφάλματος

- **Συσχέτιση** (Correlation) μεταξύ ζευγών δέντρων: υψηλή $\Rightarrow$ υψηλότερο σφάλμα
- **Ισχύς** (Strength) κάθε δέντρου: υψηλή $\Rightarrow$ χαμηλότερο σφάλμα
- Σχέση m (με συσχέτιση & ισχύ): Ευθέως Ανάλογα Μεγέθη
- Αναζητούμε τιμή του m που ελαχιστοποιεί το συνολικό ποσοστό σφάλματος

Πίνακας Περιεχομένων

Αντιμετώπιση Μεροληψίας — Ενίσχυση (Boosting)

Αντιμετώπιση Υψηλής Μεροληψίας (Dealing with Bias)

- Ένα μεροληπτικό μοντέλο *πάντοτε* κάνει λάθη
- Κατασκευή συνόλου μοντέλων για αντιστάθμιση λαθών
  - Τα λάθη να *αντισταθμίζονται* μεταξύ μοντέλων στο σύνολο
- Πώς να κατασκευαστεί ακολουθία μοντέλων, καθένα *διαφορετικά* μεροληπτικό;

Ενίσχυση (Boosting)

Κατασκευή ακολουθίας **αδύναμων ταξινομητών** (weak classifiers) $M_1, M_2, \ldots, M_n$ σε εισόδους $D_1, D_2, \ldots, D_n$:

- **Αδύναμος ταξινομητής**: ποσοστό σφάλματος αυστηρά κάτω από 50%
- Κάθε D_i είναι *σταθμισμένη* παραλλαγή αρχικών δεδομένων D
- Αρχικά: ίσα βάρη (D_1)
- Μετάβαση $D_i \to D_{i+1}$: **αύξηση βαρών** εκεί όπου ο M_i κάνει λάθη

Ενίσχυση: Έξοδος Συνόλου & Βάρη Μοντέλου

Βάρος Κάθε Μοντέλου

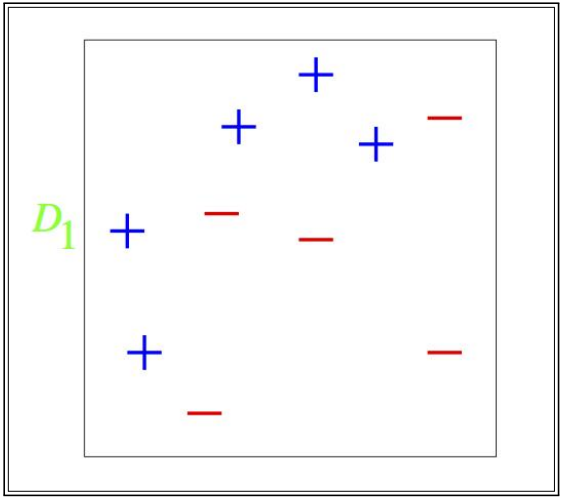
- Κάθε μοντέλο M_i λαμβάνει βάρος α_i βάσει της *ακρίβειάς* του στο D_i

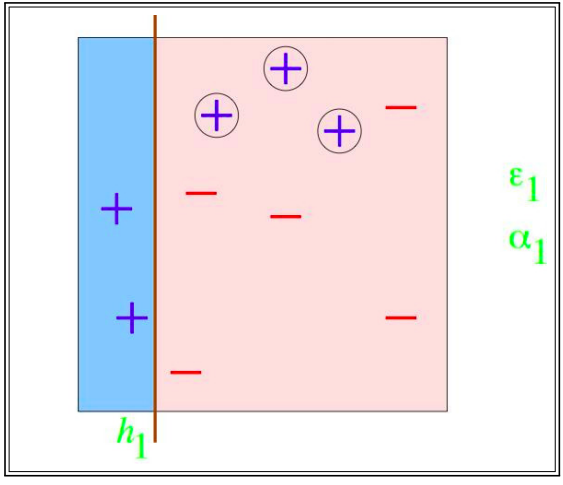
Έξοδος Συνόλου (Ensemble Output)

Μεμονωμένα αποτελέσματα ταξινόμησης: $\{-1, +1\}$. Για άγνωστη είσοδο x:

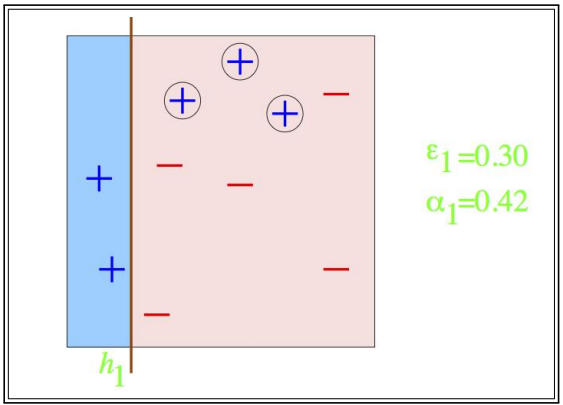
$$\text{Αποτέλεσμα Συνόλου} = \sum_{i=1}^{n} \alpha_i M_i(x)$$

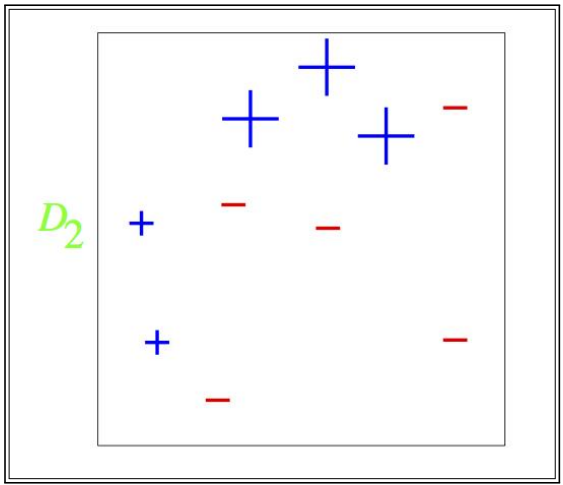
Χρήζει ελέγχου η κλάση του σταθμισμένου αθροίσματος, δηλαδή εάν το σταθμισμένο άθροισμα είναι αρνητικό (κλάση -1) ή θετικό (κλάση $+1$).



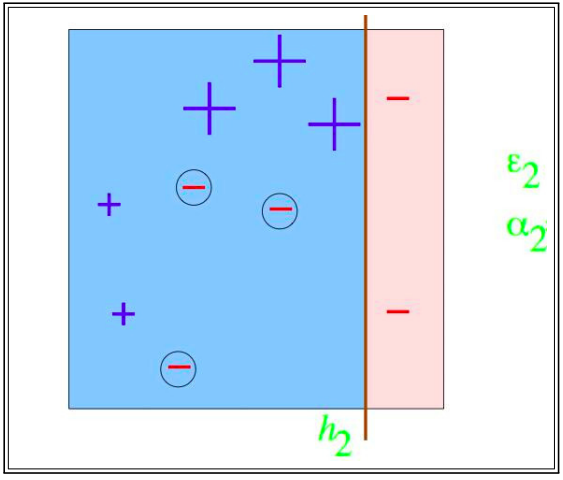


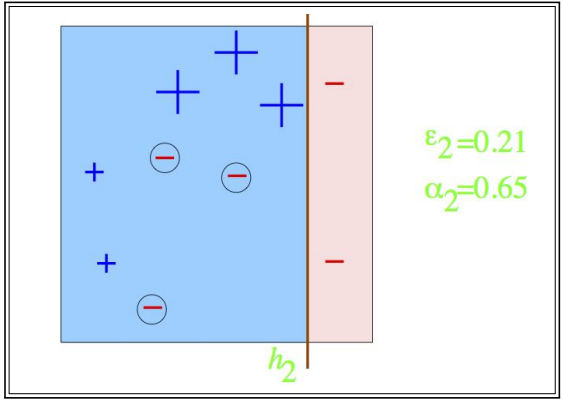
Σχ. 3.22: Οπτικοποίηση Boosting - Ενίσχυσης: Παράδειγμα (1/7)



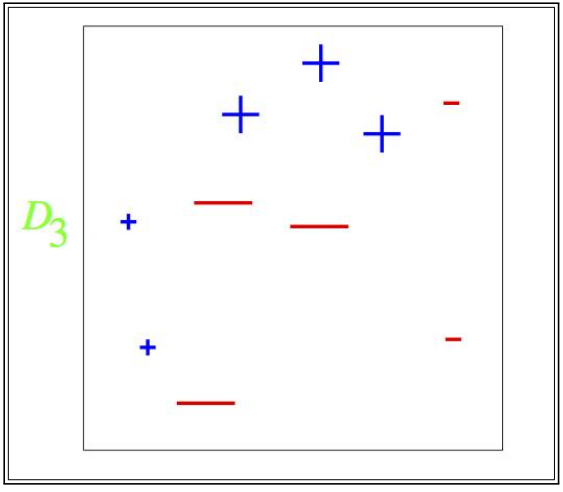


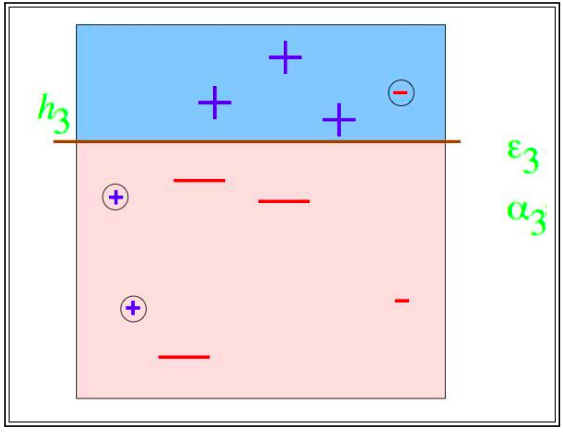
Σχ. 3.23: Οπτικοποίηση Boosting - Ενίσχυσης: Παράδειγμα (2/7)



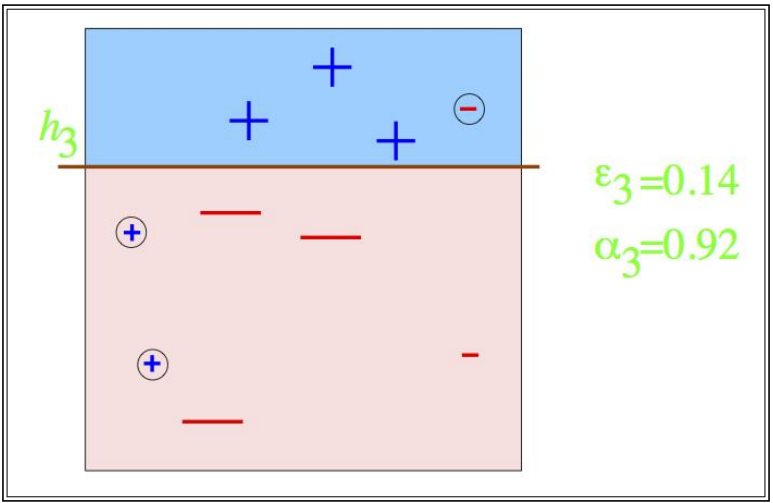


Σχ. 3.24: Οπτικοποίηση Boosting - Ενίσχυσης: Παράδειγμα (3/7)



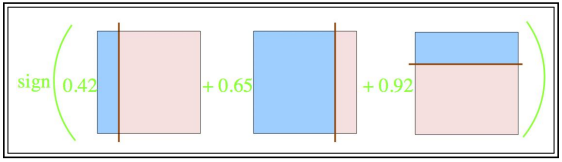


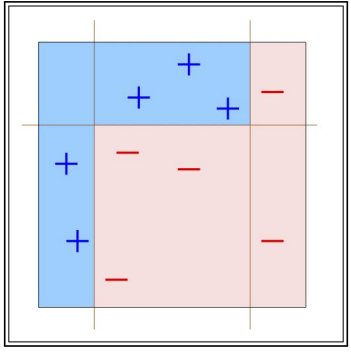
Σχ. 3.25: Οπτικοποίηση Boosting - Ενίσχυσης: Παράδειγμα (4/7)



Σχ. 3.26: Οπτικοποίηση Boosting - Ενίσχυσης: Παράδειγμα (5/7)

Boosting - Ενίσχυση: Παράδειγμα (6/7)





Σχ. 3.27: Οπτικοποίηση Boosting - Ενίσχυσης: Παράδειγμα (6/7)

Boosting - Ενίσχυση: Παράδειγμα (7/7)

Ρύθμιση: Αδύναμοι Ταξινομητές = Οριζόντιες/Κατακόρυφες Γραμμές

- **1ος Γύρος**: Διαχωριστής h_1 (κατακόρυφη γραμμή), σφάλμα $\varepsilon_1 = 0.30$, βάρος $\alpha_1 = 0.42$
  Λανθασμένα σημεία: αύξηση βάρους $\Rightarrow D_2$
- **2ος Γύρος**: Διαχωριστής h_2 (κατακόρυφη δεξιά), σφάλμα $\varepsilon_2 = 0.21$, βάρος $\alpha_2 = 0.65$
  Λανθασμένα σημεία: αύξηση βάρους $\Rightarrow D_3$
- **3ος Γύρος**: Διαχωριστής h_3 (οριζόντια γραμμή), σφάλμα $\varepsilon_3 = 0.14$, βάρος $\alpha_3 = 0.92$

Τελικός Ταξινομητής

$$\text{sign}\big(0.42 \cdot h_1(x) + 0.65 \cdot h_2(x) + 0.92 \cdot h_3(x)\big)$$

Ο συνδυασμός τριών αδύναμων ταξινομητών παράγει **μη-γραμμική περιοχή απόφασης** που ταξινομεί σωστά όλα τα σημεία εκπαίδευσης.

Σύνοψη Εννοιών (Summary Of Concepts)

Επιβλεπόμενη Μάθηση & Δέντρα Απόφασης

- **Ταξινόμηση**: Δεδομένα εκπαίδευσης, γενίκευση, αξιολόγηση με test set
- **Δέντρα Απόφασης**: Αναδρομικός διαμερισμός, κριτήριο νοθείας (Gini)
- **Μετρικές**: Πίνακας σύγχυσης, Ακρίβεια (Precision), Ανάκληση (Recall)

Λογιστική Παλινδρόμηση & SVM

- **Λογιστική Παλινδρόμηση**: Sigmoid, cross-entropy, κατηφορική κλίση
- **SVM**: Μέγιστο περιθώριο, διανύσματα υποστήριξης, πυρήνες (kernels)

Μέθοδοι Συνόλου

- **Bagging**: Bootstrap δειγματοληψία, μείωση διακύμανσης
- **Τυχαίο Δάσος**: Τυχαία υποσύνολα ιδιοτήτων, ψηφοφορία
- **Boosting**: Αδύναμοι ταξινομητές, σταθμισμένα βάρη, αντιμετώπιση μεροληψίας

Σας ευχαριστώ θερμά για την προσοχή σας!

Ερωτήσεις;

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS