

ΕΙΣΑΓΩΓΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Γιώργος Μικρός
ΕΚΠΑ – University of Massachusetts, Boston

Η Μηχανική Μάθηση είναι...

2

Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data.



Η Μηχανική Μάθηση είναι...

3

Machine learning is programming computers to optimize a performance criterion using example data or past experience.

-- Ethem Alpaydin

The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest.

-- Kevin P. Murphy

The field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions.

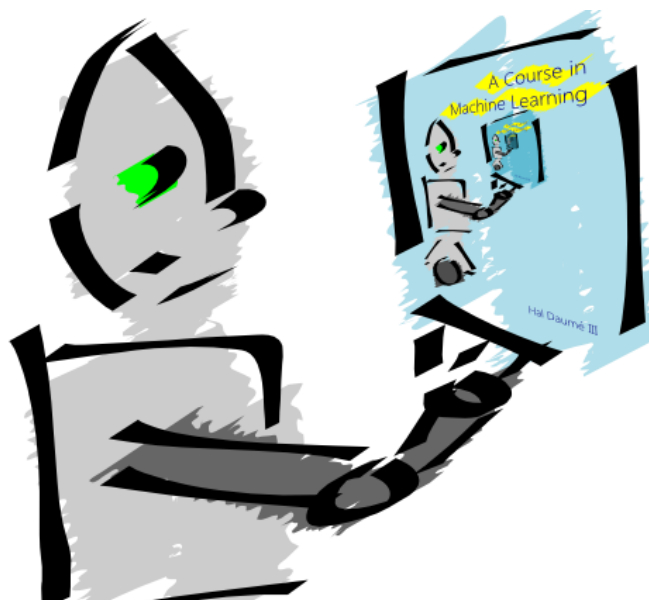
-- Christopher M. Bishop

Η Μηχανική Μάθηση είναι ...

4

Machine learning is about predicting the future based on the past.

-- Hal Daume III

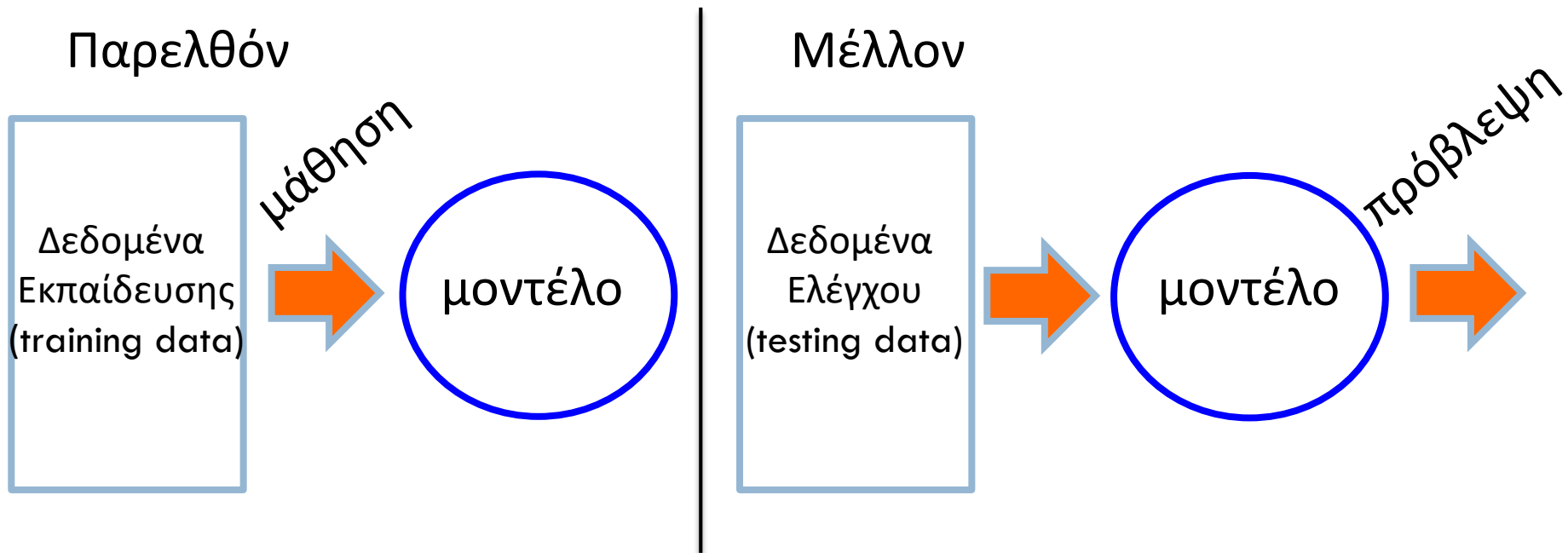


Η Μηχανική Μάθηση είναι ...

5

Machine learning is about predicting the future based on the past.

-- Hal Daume III



Μηχανική Μάθηση aka

6

Εξόρυξη Δεδομένων (data mining): Μηχανική Μάθηση εφαρμοσμένη σε βάσεις δεδομένων

Συμπερασμός (inference) και/ή εκτίμηση (estimation) στην στατιστική

Αναγνώριση προτύπων (pattern recognition) στην πληροφορική

Επεξεργασία σήματος (signal processing) στους μηχανικούς ηλεκτρολογίας (electrical engineering)

Βελτιστοποίηση (optimization)

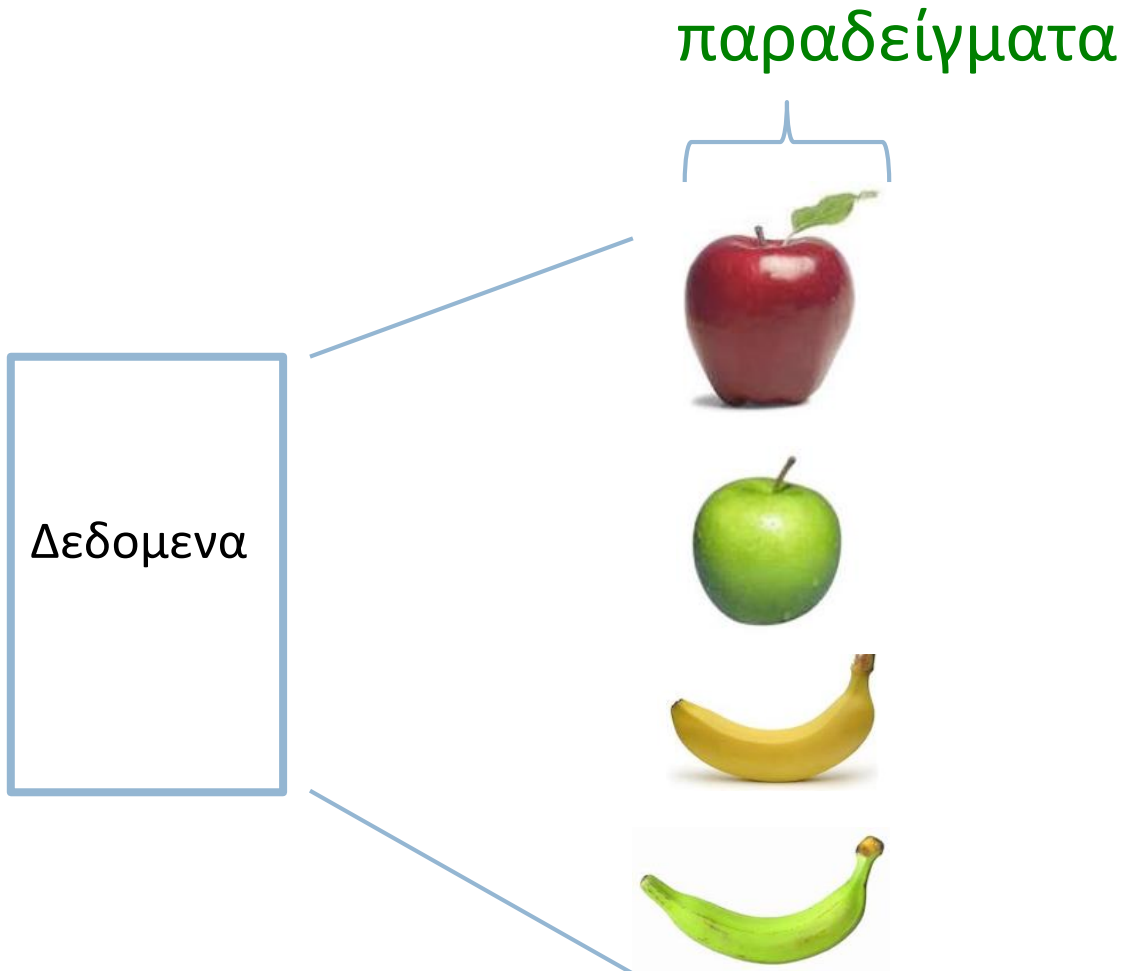
Προβλήματα Μηχανικής Μάθησης

7

Τι προβλήματα MM έχετε δει ή ακούσει μέχρι τώρα;

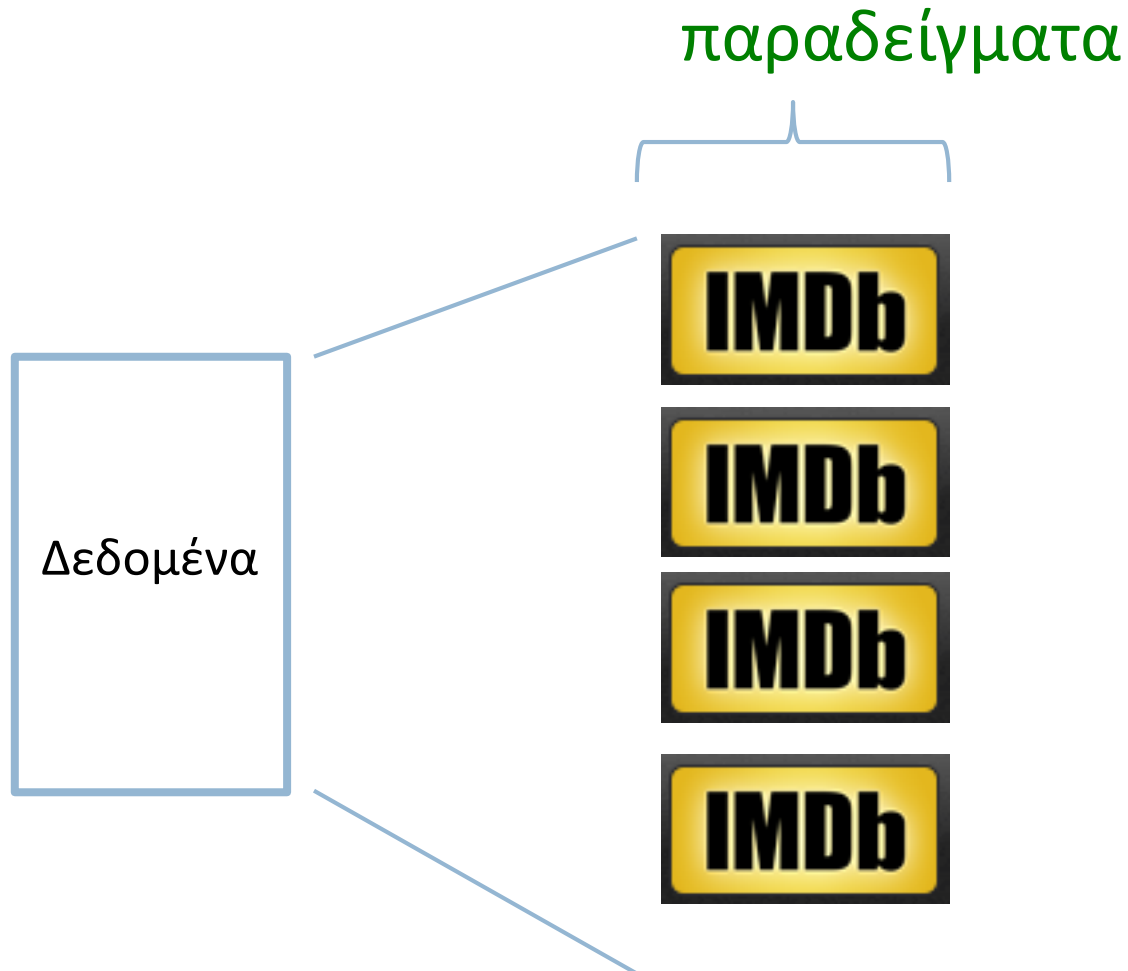
Δεδομένα

8



Δεδομένα

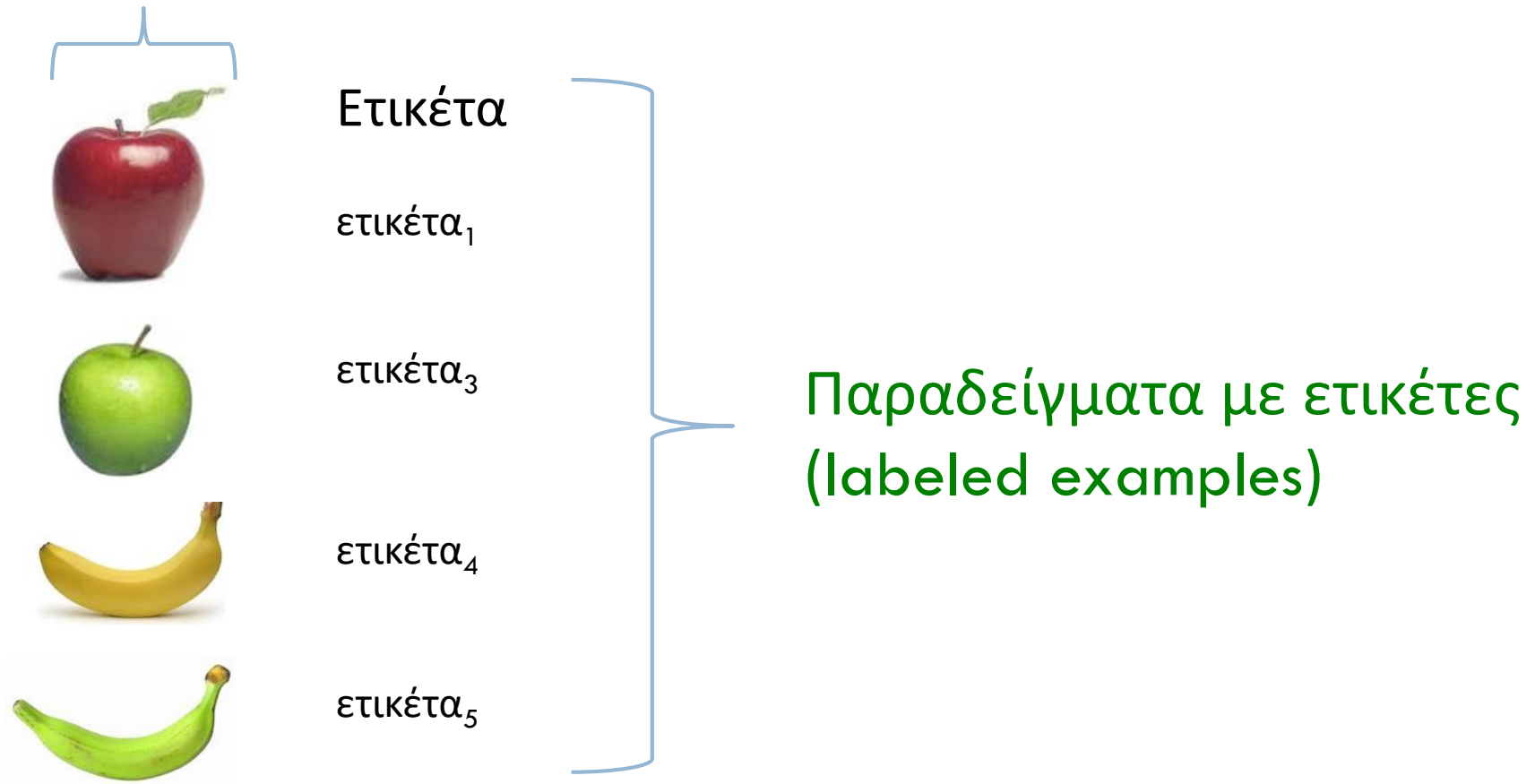
9



Εποπτευόμενη Μάθηση (Supervised learning)

10

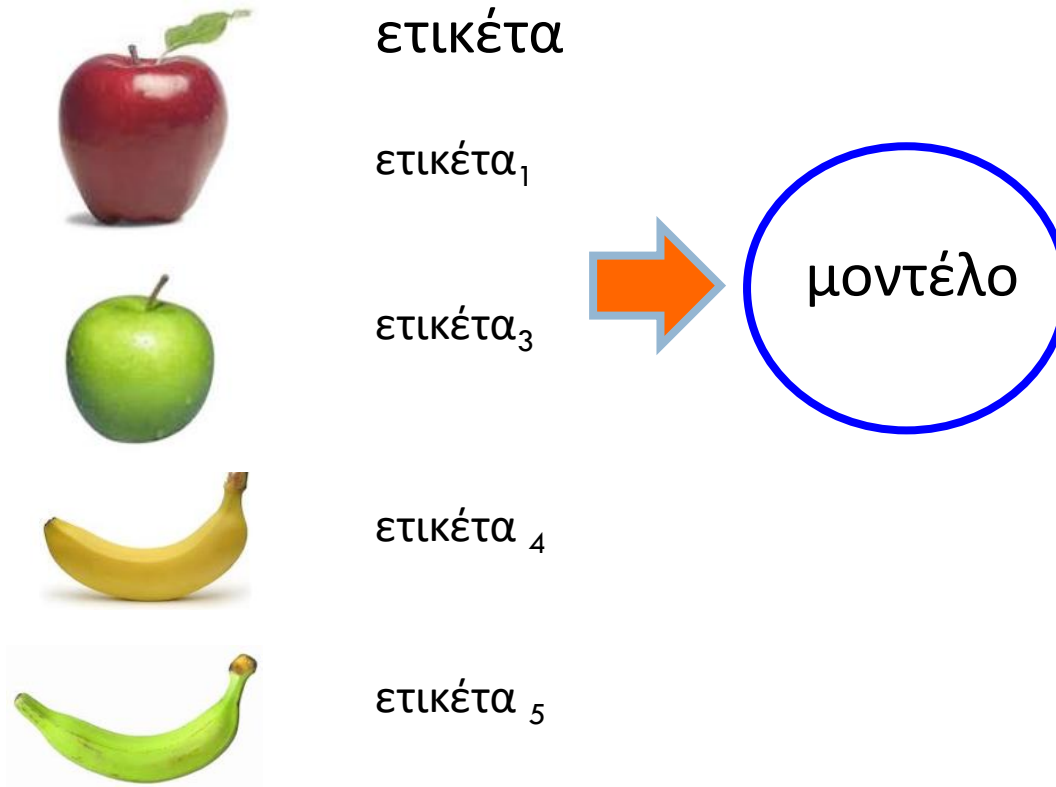
Παραδείγματα



Εποπτευόμενη Μάθηση: να κάνουμε γνωστά στη μηχανή παραδείγματα με ετικέτες

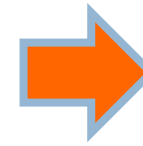
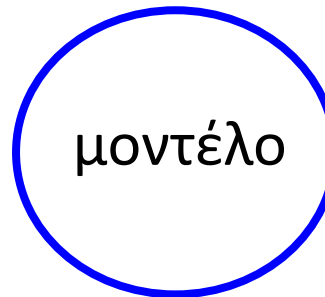
Εποπτευόμενη Μάθηση

11



Εποπτευόμενη Μάθηση

12



Ετικέτα που
προβλέφθηκε

Εποπτευόμενη Μάθηση: μαθαίνει η μηχανή να προβλέπει την ετικέτα ενός νέου παραδείγματος

Εποπτευόμενη Μάθηση:

Κατηγοριοποίηση (classification)

13



Ετικέτα

μήλο



μήλο



μπανάνα



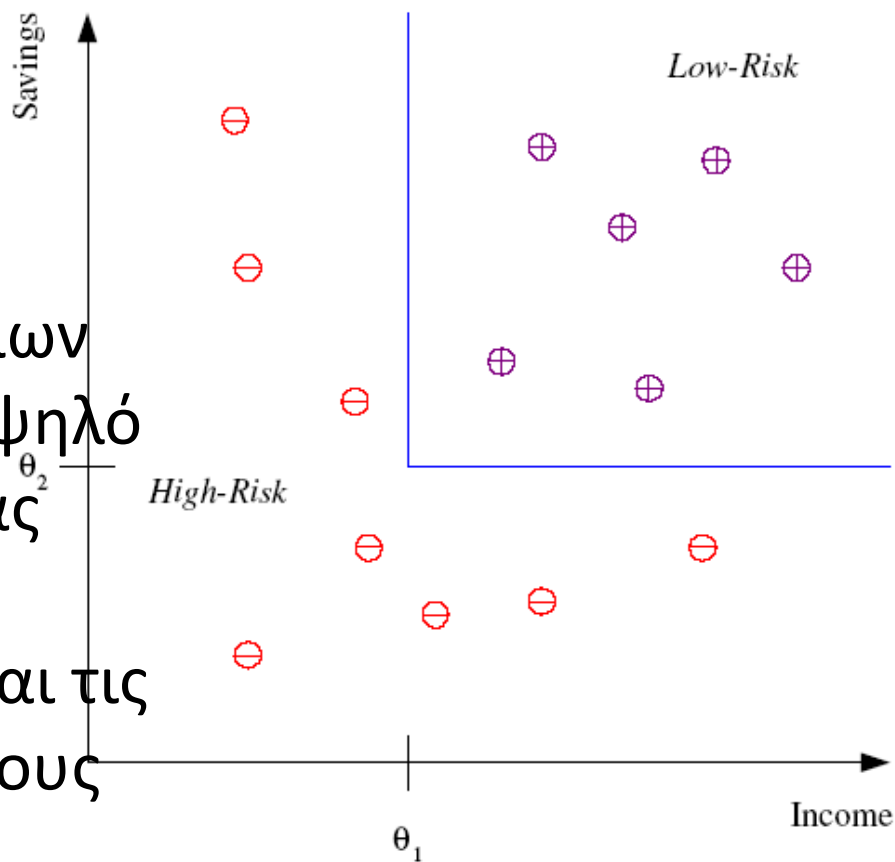
μπανάνα

Κατηγοριοποίηση: πεπερασμένο
σύνολο ετικετών

Παράδειγμα κατηγοριοποίησης

14

Διαφοροποίηση
μεταξύ των ατόμων
με χαμηλό και υψηλό
ρίσκο χρεωκοπίας
βασιζόμενοι στο
εισόδημά τους και τις
αποταμιεύσεις τους



Μη Εποπτευόμενη Μάθηση (Unsupervised learning)

15



Μη Εποπτευόμενη Μάθηση : έχουμε δεδομένα αλλά δεν έχουμε ετικέτες

Αξιολόγηση Εποπτευόμενης Μάθησης








16

Δεδομένα Ετικέτα

Δεδομένα εκπαίδευσης

Δεδομένα ελέγχου

Δεδομένα με ετικέτες








	0
	0
	1
	1
	0
	1
	0

Αξιολόγηση Εποπτευόμενης Μάθησης

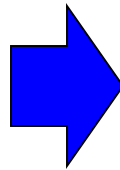
17

Δεδομένα Ετικέτα

Δεδομένα με ετικέτες

	0
	0
	1
	1
	0
	1
	0

Δεδομένα εκπαίδευσης



Μοντέλο

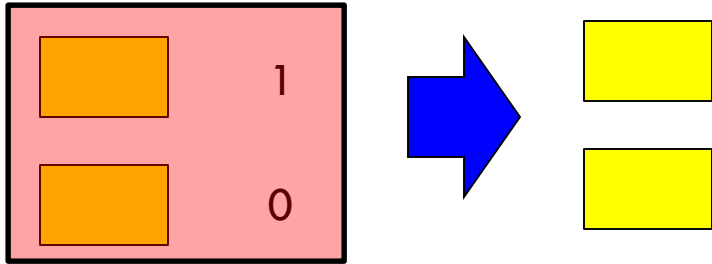
Εκπαιδεύουμε έναν
αλγόριθμο
κατηγοριοποίησης

Δεδομένα ελέγχου

Αξιολόγηση Εποπτευόμενης Μάθησης

18

Δεδομένα Ετικέτα

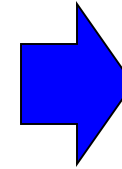
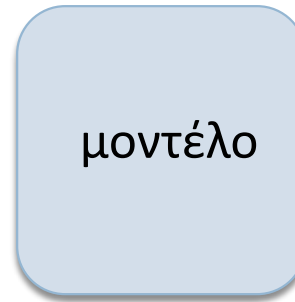
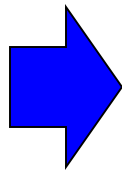
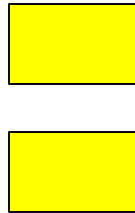
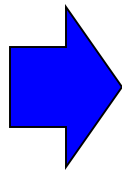
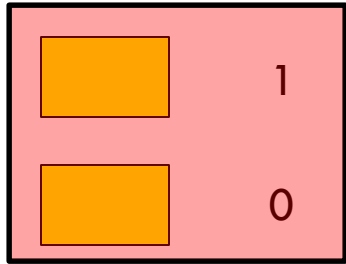


Προσποιούμαστε ότι
δεν γνωρίζουμε τις
ετικέτες

Αξιολόγηση Εποπτευόμενης Μάθησης

19

Δεδομένα Ετικέτα



1
1

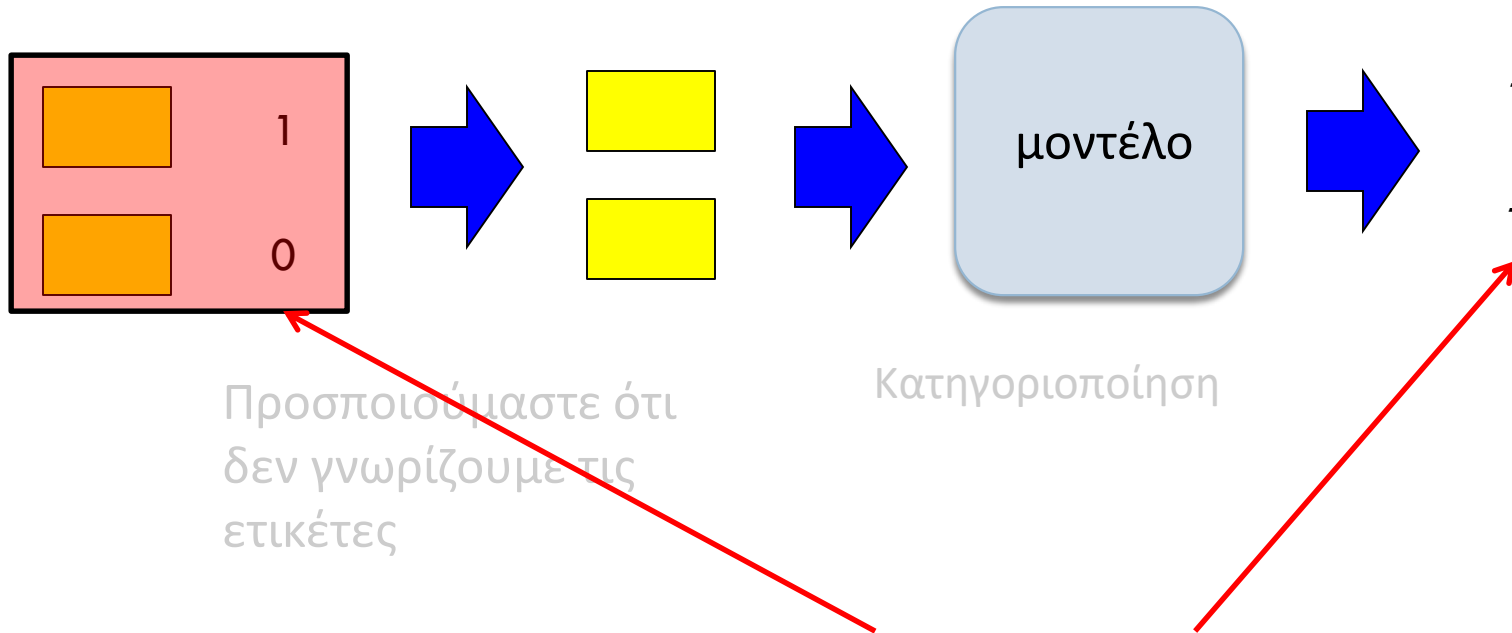
Προσποιούμαστε ότι
δεν γνωρίζουμε τις
ετικέτες

Κατηγοριοποίηση

Αξιολόγηση Εποπτευόμενης Μάθησης

20

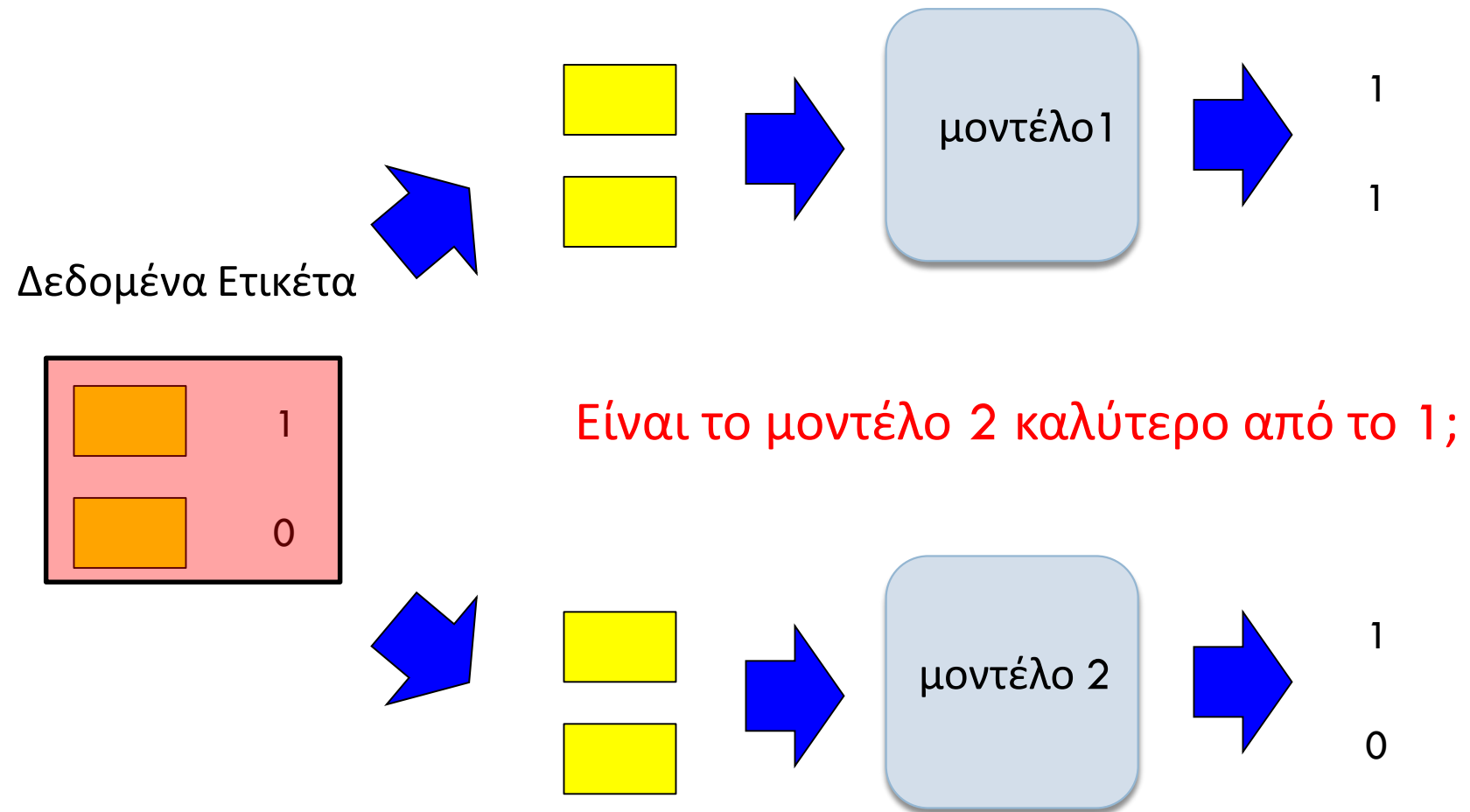
Δεδομένα Ετικέτα



Συγκρίνουμε τις ετικέτες που έχουμε
με αυτές που προβλέφθηκαν από το
μοντέλο

Συγκρίνοντας αλγόριθμους

21



Ιδέα 1

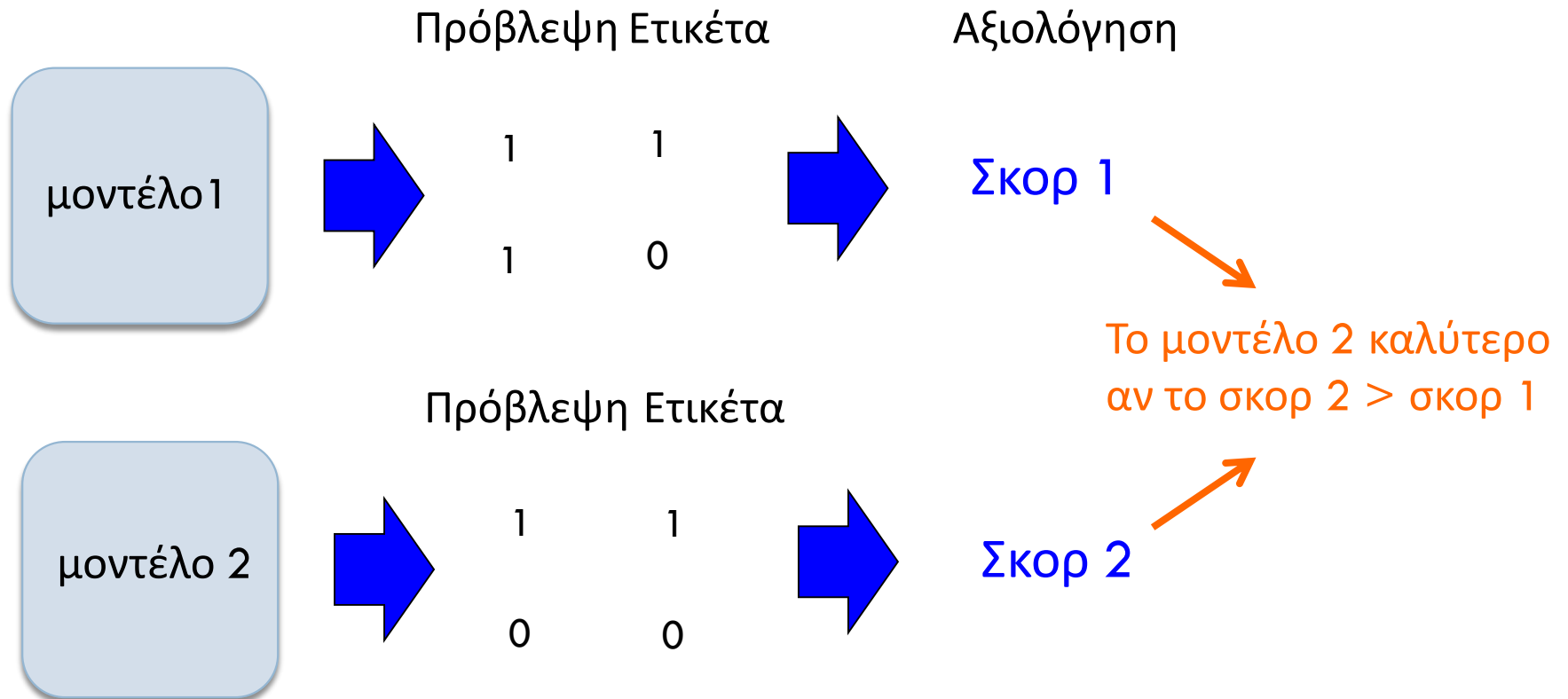
22



Υπάρχουν δεύτερες σκέψεις;

Ιδέα 1

23



Είναι το μοντέλο 2 καλύτερο;

24

Μοντέλο 1: 85% ακρίβεια

Μοντέλο 2: 80% ακρίβεια

Μοντέλο 1: 85.5% ακρίβεια

Μοντέλο 2: 85.0% ακρίβεια

Μοντέλο 1: 0% ακρίβεια

Μοντέλο 2: 100% ακρίβεια

Συγκρίνοντας σκορ: σημαντικότητα

25

- Η απλή σύγκριση των σκορ σε ένα σετ δεδομένων δεν φτάνει!
- Δεν θέλουμε να γνωρίζουμε ποιο σύστημα είναι καλύτερο **σε αυτά τα συγκεκριμένα δεδομένα**, θέλουμε να ξέρουμε αν το μοντέλο 1 είναι καλύτερο από το μοντέλο 2 **γενικά**.
- Ή αλλιώς, θέλουμε να είμαστε σίγουροι ότι η διαφορά είναι πραγματική και όχι λόγω τυχαίας διακύμανσης.

Επαναλαμβανόμενος πειραματισμός

26

Δεδομένα Ετικέτα

Δεδομένα με ετικέτες

0	0
0	0
1	1
1	1
0	0
1	1
0	0

Δεδομένα εκπαίδευσης

Αντί να κάνουμε μία τομή
στα δεδομένα, θα
κάνουμε πολλές

Δεδομένα ελέγχου

Επαναλαμβανόμενος πειραματισμός







27







Δεδομένα εκπαίδευσης







Δεδομένα Ετικέτα

Δεδομένα Ετικέτα


Δεδομένα Ετικέτα


	0
	0
	1
	1
	0
	1

	0
	0
	1
	1
	0
	1

	0
	0
	1
	1
	0
	1

...

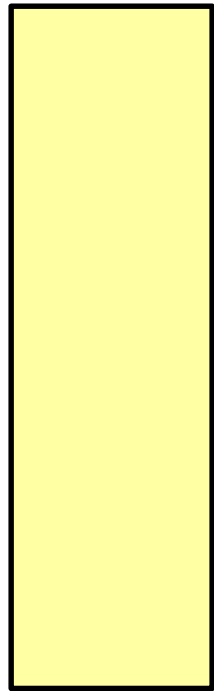
 = εκπαίδευση

 = ανάπτυξη

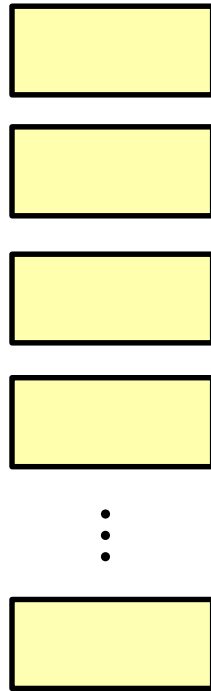
n-fold cross validation

28

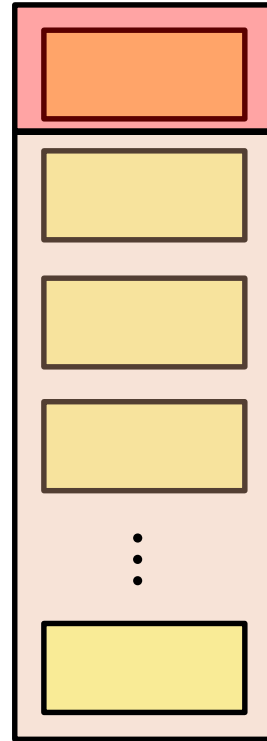
Δεδομένα εκπαίδευσης



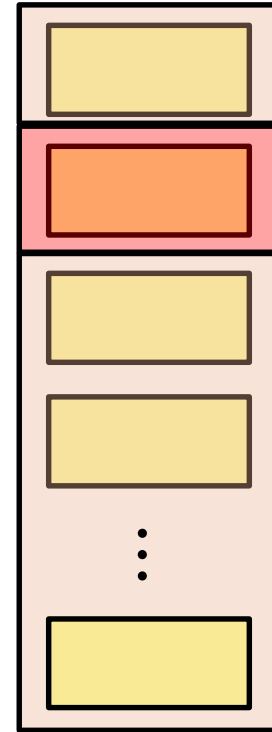
Χωρίζουμε
σε ισομεγέθη
τμήματα



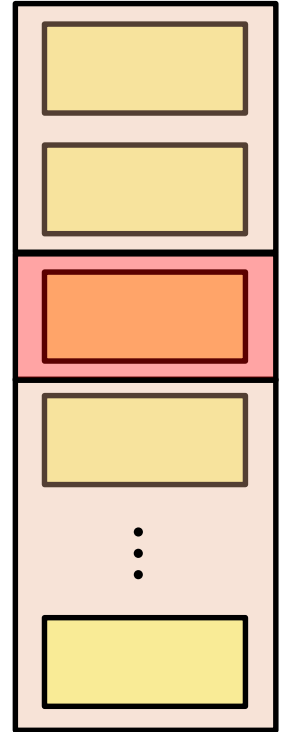
Επαναλαμβάνουμε για όλα τα τμήματα:
Εκπαιδούμε σε $n-1$ τμήματα και ελέγχουμε στα υπόλοιπα



Τμήμα 1



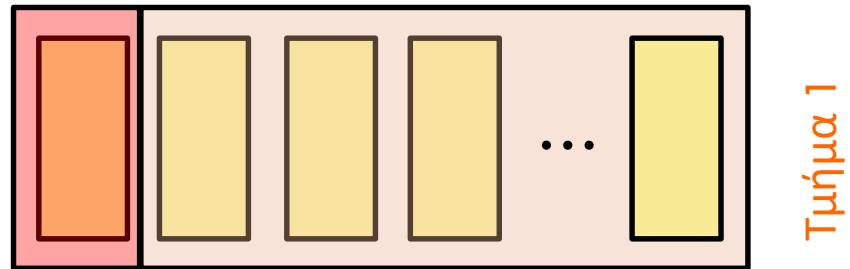
Τμήμα 2



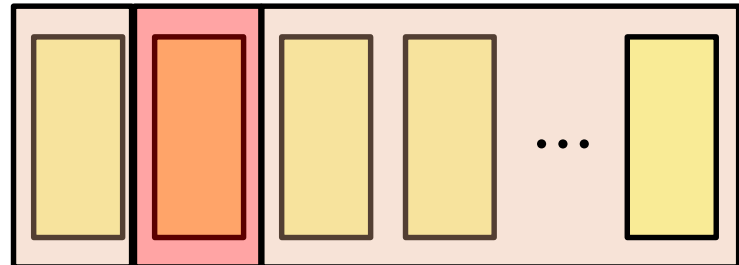
Τμήμα 3

n-fold cross validation

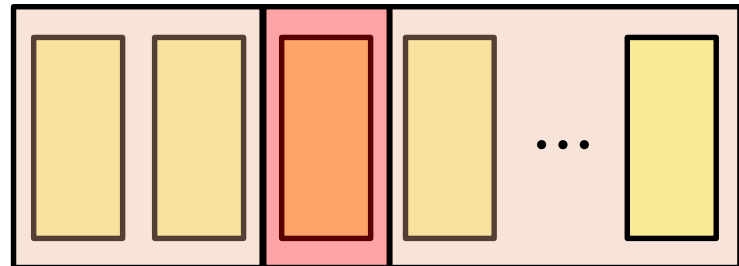
29



Τμήμα 1



Τμήμα 2



Τμήμα 3

...



αξιολόγηση

Σκορ 1

Σκορ 2

Σκορ 3

...

n-fold cross validation

30

Καλύτερη αξιοποίηση των δεδομένων με ετικέτες
Περισσότερο εύρωστη (robust) τεχνική: Δεν υπάρχει
εξάρτηση από ένα μόνο κομμάτι δεδομένων ελέγχου
για να αξιολογήσουμε ένα μοντέλο

Πολλαπλασιάζει το υπολογιστικό φορτίο n φορές
(έχουμε να εκπαιδεύσουμε n μοντέλα και όχι ένα)

10 είναι η πιο συχνή επιλογή για n

Leave-one-out cross validation

31

n-fold cross validation όπου $n =$ ο αριθμός των παραδειγμάτων

Γνωστό και ως “jackknifing”

Πότε το χρησιμοποιούμε;

Leave-one-out cross validation

32

Μπορεί να είναι πολύ «βαριά» υπολογιστικά όταν υπάρχει μεγάλος αριθμός παραδειγμάτων

Χρήσιμη σε περιπτώσεις με περιορισμένο αριθμό δεδομένων: μεγιστοποιεί τα δεδομένα που μπορούμε να χρησιμοποιήσουμε για εκπαίδευση.

Πίνακας Κατηγοριοποίησης (Confusion matrix)

33

Ο πίνακας διαβάζεται από δεξιά προς αριστερά. Το μοντέλο μας κατηγοριοποίησε σωστά 86 κείμενα που είχε γράψει ο Α στον Α. Ωστόσο, 2 κείμενα του Α τα απέδωσε εσφαλμένα στον Β, 4 στον Γ, 18 στον Ε και 1 στον Ζ.

Προβλέψεις

Πραγματικές τιμές

	A	B	Γ	Δ	E	Z
A	86	2	0	4	18	1
B	1	57	5	1	12	13
Γ	0	6	55	4	0	5
Δ	0	15	28	90	4	18
E	7	1	0	0	37	12
Z	6	19	11	0	27	48

Βασικές μετρικές στην αξιολόγηση μοντέλων κατηγοριοποίησης

34

Πραγματικές τιμές	Προβλέψεις	
	Θετικό	Αρνητικό
	Θετικό Αληθώς θετικό ΑΘ	Ψευδώς αρνητικό ΨΑ
Αρνητικό Ψευδώς θετικό ΨΘ	Αληθώς αρνητικό ΑΑ	

Πραγματικές τιμές	Προβλέψεις	
	A	B
	A 40	10
B 5	45	

Ακρίβεια (accuracy)= Πόσο συχνά το μοντέλο είναι σωστό

- $(ΑΘ + ΑΑ)/\text{Σύνολο} = (40+45)/100 = 0,95$

Λάθος κατηγοριοποίησης (misclassification rate): Πόσο συχνά το μοντέλο είναι λάθος

- $1 - \text{Ακρίβεια} = 1 - 0,95 = 0,05$

Ορθότητα (Precision)= Ο αριθμός των σωστών προβλέψεων του Α επί των συνολικών προβλέψεων Α

- $ΑΘ/ΑΘ+ΨΘ = 40/40 + 5 = 40/45 = 0.89$

Ανάκληση (Recall)= Όταν είναι Α στην πραγματικότητα, πόσες φορές προβλέπει Α

- $ΑΘ/ΑΘ+ΨΑ = 40/40+10 = 40/50 = 0,8$

Τυχαία Δάση (Random Forests)

35

- Το Τυχαίο Δάσος είναι μία συλλογή (ensemble) από δέντρα αποφάσεων (Breiman 2001).
- Τα Τυχαία Δάση χρησιμοποιούνται συχνά όταν έχουμε μεγάλα δεδομένα εκπαίδευσης και μεγάλο αριθμό μεταβλητών. Ένα Τυχαίο Δάσος αποτελείται τις περισσότερες φορές από δεκάδες ή και εκατοντάδες δέντρα αποφάσεων.
- Μπορούν να χρησιμοποιηθούν τόσο για κατηγοριοποίηση όσο και για παλινδρόμηση.
- Η ακρίβεια και η σημαντικότητα κάθε μεταβλητής συμπεριλαμβάνεται στα αποτελέσματα
- Για μια πολύ απλή εξήγηση τους δείτε την απάντηση του Edwin Chen's στην Quora: <http://www.quora.com/Machine-Learning/How-do-random-forests-work-in-laymans-terms>



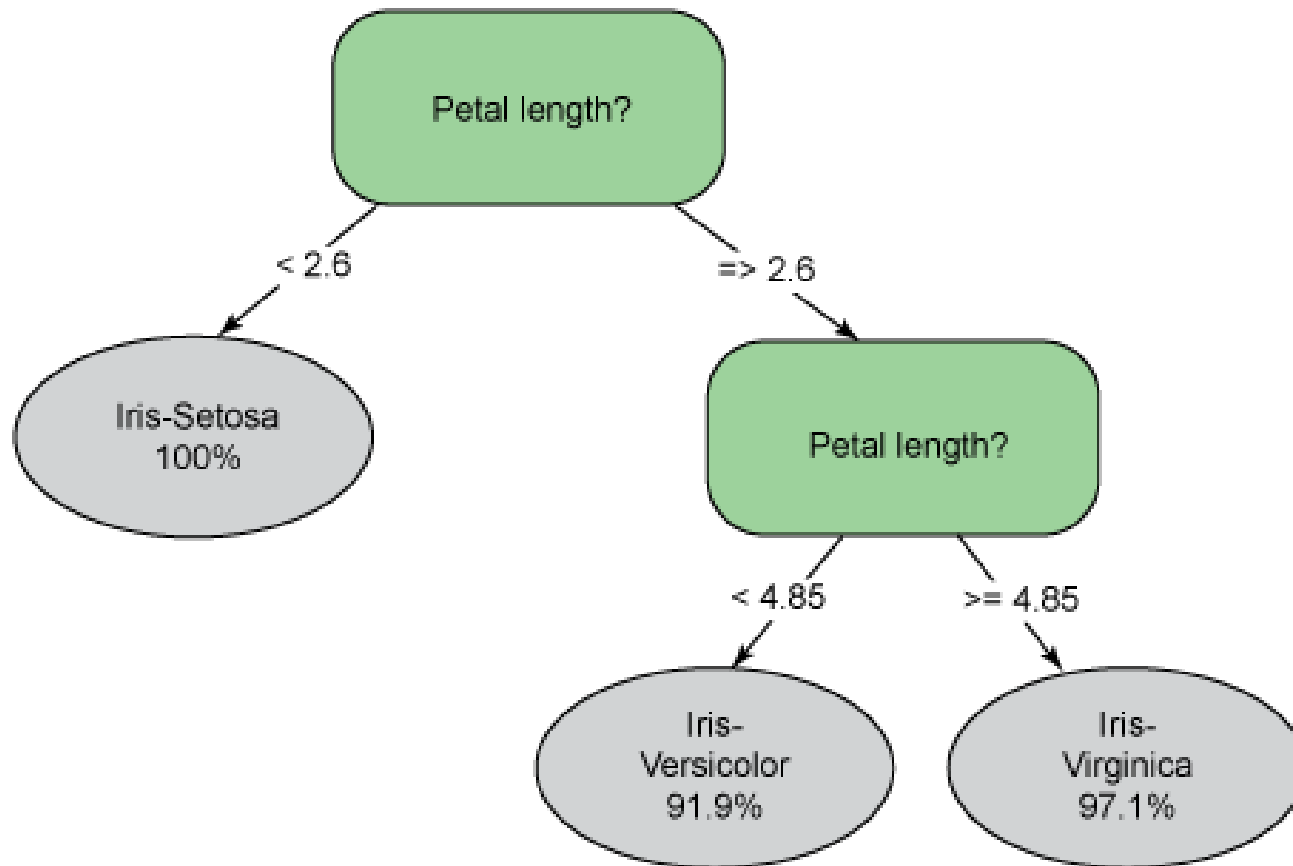
Πώς δουλεύουν;

36

- Κάθε δέντρο απόφασης κατασκευάζεται από ένα τυχαίο υποσύνολο των δεδομένων εκπαίδευσης χρησιμοποιώντας την τεχνική της δειγματοληψίας με αντικατάσταση (replacement sampling) γνωστή και ως bagging. Αυτό σημαίνει ότι κάποια δεδομένα θα περιληφθούν περισσότερες από μία φορές στο δείγμα ενώ άλλα δεν θα εμφανιστούν καθόλου. Σε γενικές γραμμές περίπου τα $2/3$ των δεδομένων θα περιληφθούν στο υποσύνολο των δεδομένων εκπαίδευσης και το $1/3$ δεν θα συμμετάσχει.
- Σε κάθε δείγμα των δεδομένων εκπαίδευσης αναπτύσσεται ένα δέντρο απόφασης χρησιμοποιώντας ένα διαφορετικό τυχαίο υποσύνολο μεταβλητών.
- Τα δέντρα αποφάσεων που έχουν αναπτυχθεί σε διαφορετικά δείγματα και με διαφορετικές συνθέσεις μεταβλητών αντιπροσωπεύουν το τελικό πολυσυλλεκτικό μοντέλο (ensemble model) στο οποίο το κάθε δέντρο απόφασης ψηφίζει για το αποτέλεσμα και η πλειοψηφία κερδίζει.

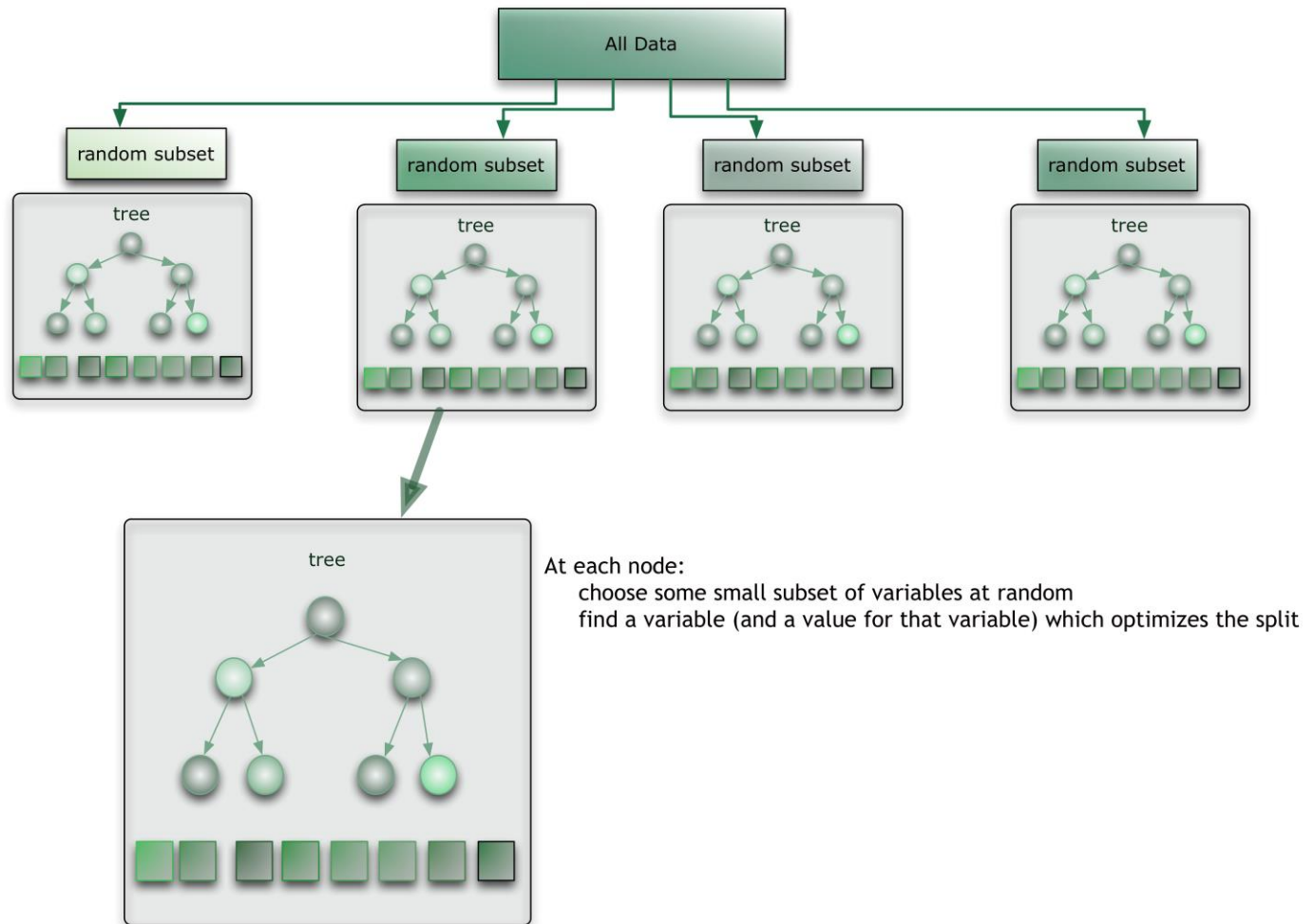
Από τα δέντρα αποφάσεων ...

37



... στα Τυχαία Δάση

38



Πλεονεκτήματα

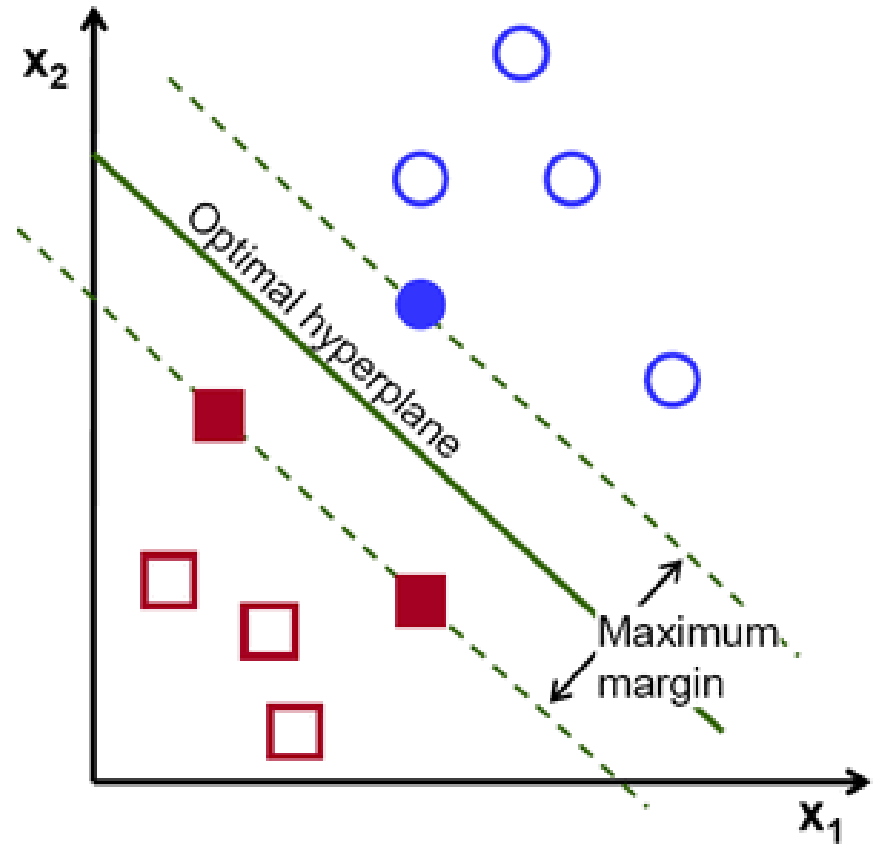
39

- Παράγει μοντέλα κατηγοριοποίησης που είναι ακριβή και γρήγορα.
- Τρέχει αποδοτικά σε μεγάλες βάσεις δεδομένων
- Δεν χρειάζεται προεπεξεργασία δεδομένων (κανονικοποίηση, αντικατάσταση τιμών που λείπουν κ.ά.) και είναι εξαιρετικά ανθεκτικά σε ακραίες τιμές (outliers).
- Μπορεί να χειριστεί χιλιάδες μεταβλητές χωρίς να χρειαστεί να τις μειώσουμε με κάποια διαδικασία επιλογής (variable selection).
- Επειδή αναπτύσσονται πολλά δέντρα αποφάσεων και υπάρχουν δύο επίπεδα τυχαιότητας και στην ουσία κάθε δέντρο είναι ένα ανεξάρτητο μοντέλο, τα Τυχαία Δάση δεν κάνουν υπερπροσαρμογή (overfit) στα δεδομένα εκπαίδευσης.

Μηχανές Διανυσμάτων Υποστήριξης – Support Vector Machines - SVM

40

- Οι Μηχανές Διανυσμάτων Υποστήριξης είναι ένας αλγόριθμος εποπτευόμενης Μηχανικής Μάθησης που μπορεί να χρησιμοποιηθεί τόσο σε δεδομένα κατηγοριοποίησης όσο και σε δεδομένα παλινδρόμησης (Varnik 1995).
- Ο αλγόριθμος βρίσκει το υπερπλάνο (hyperplane) (γραμμή στις 2 διαστάσεις, πλάνο (plane) στις 3 διαστάσεις και υπερπλάνο (hyperplane) σε μεγαλύτερες διαστάσεις).
- Πιο τυπικά, ένα υπερπλάνο είναι ένας $n-1$ διάστατος υποχώρος ενός n -διάστατος χώρος) που διακρίνει με τον βέλτιστο τρόπο δύο κατηγορίες σημείων με το μέγιστο περιθώριο μεταξύ τους (maximum margin).
- Τα σημεία που «υποστηρίζουν» το υπερπλάνο στις δύο μεριές ονομάζονται «διανύσματα υποστήριξης» "support vectors".
- Στις περιπτώσεις που οι δύο κατηγορίες δεν μπορούν να διαχωριστούν με γραμμικό τρόπο, τα σημεία τους προβάλλονται σε έναν χώρο υψηλότερης διάστασης στον οποίον μπορεί να είναι εφικτή η γραμμική διαφοροποίηση

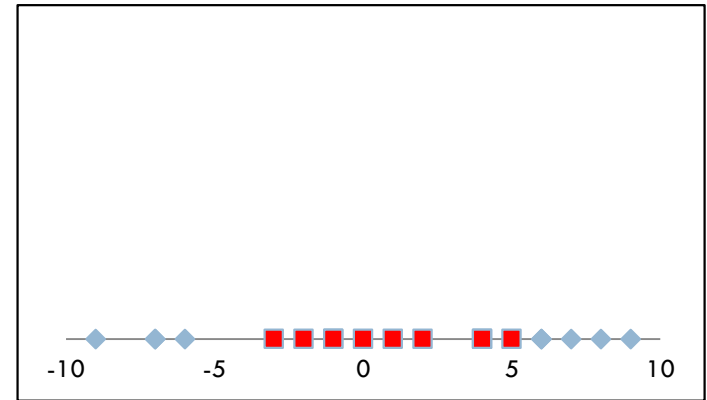


Η συνάρτηση πυρήνα (kernel function)

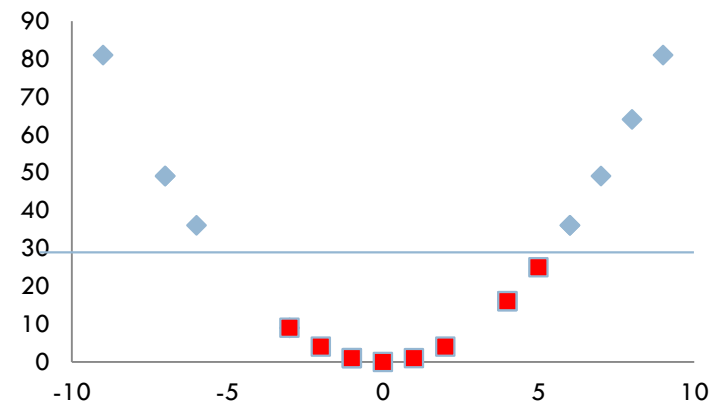
41

- Όταν εξετάζουμε δεδομένα σε μία διάσταση (μία μεταβλητή) μπορούμε να τα διατάξουμε κατά μήκος μίας γραμμής.
- Στην Εικ. 1 δεν μπορούμε να διαχωρίσουμε γραμμικά τις κόκκινες από τις μπλε τελείες αφού οι κόκκινες είναι στο μέσο μεταξύ των μπλε τελειών.
- Μπορούμε να λύσουμε το πρόβλημα προσθέτοντας μια υψηλότερη διάσταση στα δεδομένα μας υψώνοντάς τα στο τετράγωνο.
- Στην Εικ. 2 έχουμε ένα δισδιάστατο γράφημα (x vs. x^2) και τα δεδομένα μας τώρα μπορούν να διαχωριστούν γραμμικά.
- Η συνάρτηση πυρήνα είναι μία μέθοδος που επιτρέπει στις Μηχανές Διανυσμάτων Υποστήριξης να προβάλλουν δεδομένα σε υψηλότερο χώρο διαστάσεων. Αποδεικνύεται ότι για κάθε σύνολο δεδομένων υπάρχει μία συνάρτηση πυρήνα που μπορεί να το διακρίνει γραμμικά.

Εικ. 1



Εικ. 2



Πηγές μάθησης

42

□ R + Machine Learning

- Introduction to R language. YouTube videos by Google:
<https://www.youtube.com/playlist?list=PLOU2XLYxmslK9qQfztXeybpHvru-TrqAP>
- R Learning Path: From beginner to expert in R in 7 steps:
<https://www.kdnuggets.com/2016/03/datacamp-r-learning-path-7-steps.html>
- R Studio Learning Links:
<https://www.rstudio.com/online-learning/>