

Πανεπιστήμιο Πειραιώς Τμήμα Πληροφορικής



Αποθήκες Δεδομένων και Εξόρυξη Γνώσης (Data Warehousing & Data Mining)

Κατηγοριοποίηση (classification)

Γιάννης Θεοδωρίδης, Νίκος Πελέκης

Ομάδα Διαχείρισης Δεδομένων
Εργαστήριο Πληροφοριακών Συστημάτων

<http://isl.cs.unipi.gr/db>

Περιεχόμενα

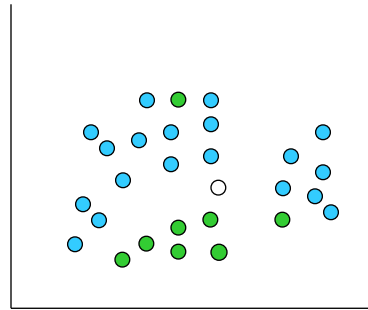


- Το πρόβλημα της κατηγοριοποίησης
- Τεχνικές κατηγοριοποίησης
 - Στατιστικές τεχνικές
 - Τεχνικές βασισμένες στην απόσταση
 - Δένδρα αποφάσεων
 - Νευρωνικά δίκτυα
 - Κανόνες κατηγοριοποίησης

Κατηγοριοποίηση (Classification)



Εκμάθηση μιας τεχνικής να προβλέπει την κλάση ενός στοιχείου επιλέγοντας από προκαθορισμένες τιμές



3

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Εποπτευόμενη vs. Μη εποπτευόμενη μάθηση



■ Εποπτευόμενη μάθηση (κατηγοριοποίηση)

- Επόπτευση: Τα δεδομένα εκπαίδευσης συνοδεύονται από ετικέτες για την κλάση με την οποία ανήκει το καθένα
- Τα νέα δεδομένα κατηγοριοποιούνται με βάση τη γνώση που μας παρέχουν τα δεδομένα εκπαίδευσης

■ Μη εποπτευόμενη μάθηση (συσταδοποίηση)

- Δε γνωρίζουμε την κλάση στην οποία ανήκουν τα δεδομένα εκπαίδευσης
- Μας δίνεται ένα σύνολο μετρήσεων, παρατηρήσεων κλπ. με στόχο να ανακαλύψουμε κλάσεις ή ομάδες μέσα στα δεδομένα

4

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Το πρόβλημα της κατηγοριοποίησης



- Αν μας δοθεί μια βάση δεδομένων $D=\{t_1, t_2, \dots, t_n\}$ και ένα σύνολο κατηγοριών - "κλάσεων" $C=\{C_1, \dots, C_m\}$, το **Πρόβλημα Κατηγοριοποίησης** έγκειται στον ορισμό μιας απεικόνισης $f: D \rightarrow C$ όπου κάθε εγγραφή t_i ανατίθεται σε μία κλάση C_j .
- Ουσιαστικά, η κατηγοριοποίηση διαμερίζει τη D σε **κλάσεις ισοδυναμίας**.
- Η **Πρόβλεψη** είναι παρόμοιο πρόβλημα, αλλά μπορεί να θεωρηθεί ότι έχει άπειρο αριθμό κλάσεων.

Παραδείγματα κατηγοριοποίησης



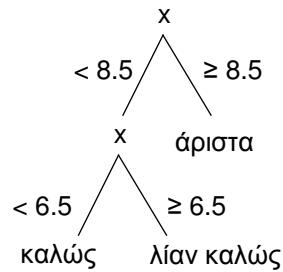
- Οι δάσκαλοι αναθέτουν βαθμούς μέσα από τις κατηγορίες A, B, C, D, F.
- Τα μανιτάρια ταξινομούνται σε δηλητηριώδη και φαγώσιμα.
- Μπορεί να γίνει πρόβλεψη εάν θα πλημμυρίσει ένας ποταμός.
- Οι πελάτες της τράπεζας μπορούν να κατηγοριοποιηθούν ως προς την πιστωτική τους ικανότητα.

Παράδειγμα κατηγοριοποίησης



Βαθμολογία πτυχίου

- If $x \geq 8.5$ then
grade = «άριστα».
- If $6.5 \leq x < 8.5$ then
grade = «λίαν καλώς».
- If $x < 6.5$ then
grade = «καλώς».



7

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Τεχνικές κατηγοριοποίησης

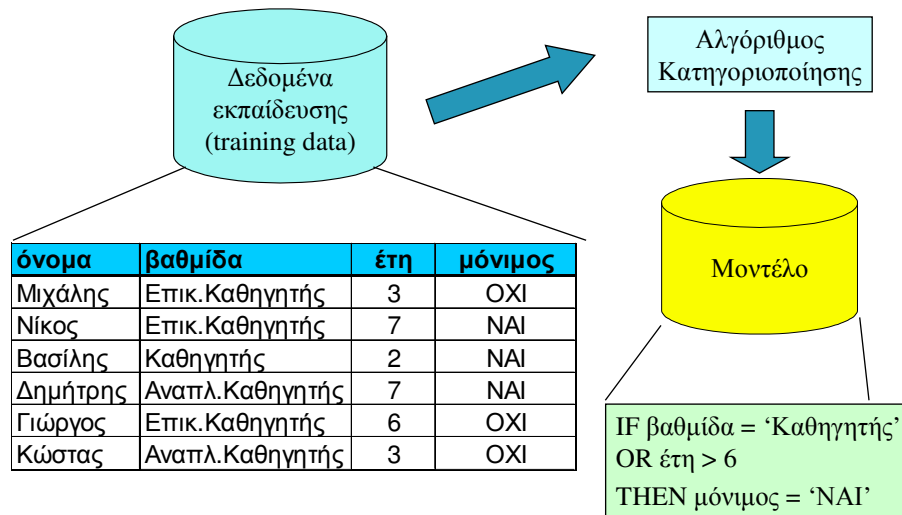


- Τυπική προσέγγιση:
 1. Δημιουργία ενός μοντέλου μέσω της αξιολόγησης ενός συνόλου δεδομένων εκπαίδευσης (training data) (ή μέσω της γνώσης ειδικών του πεδίου).
 2. Εφαρμογή του μοντέλου σε νέα δεδομένα.
- Οι κλάσεις πρέπει να είναι προκαθορισμένες
- Οι πιο κοινές τεχνικές είναι τα δένδρα αποφάσεων, τα νευρωνικά δίκτυα και τεχνικές βασισμένες σε απόσταση ή σε στατιστικές μεθόδους.

8

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

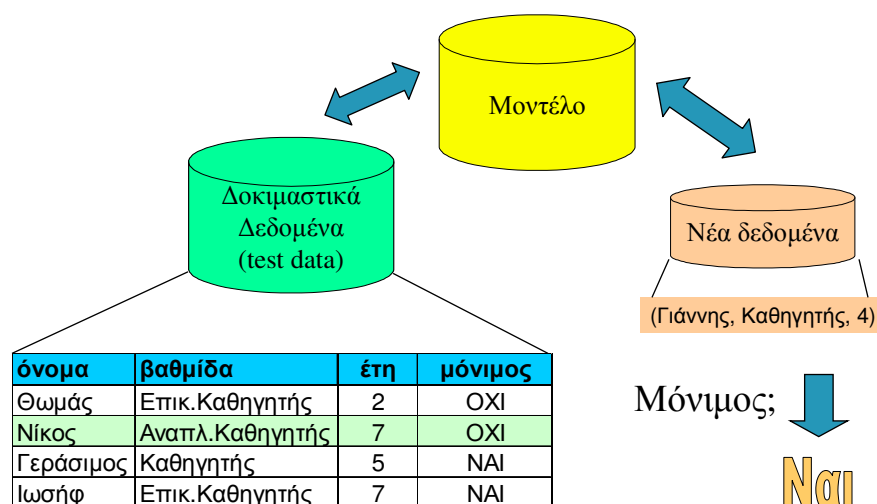
1ο βήμα: Δημιουργία μοντέλου



9

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

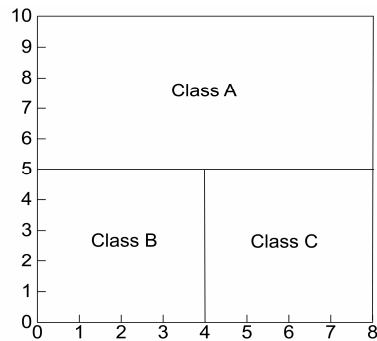
2ο βήμα: Εφαρμογή μοντέλου



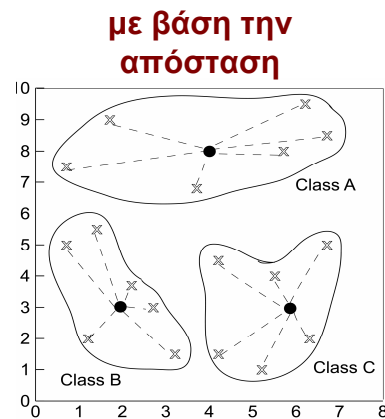
10

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Προσδιορισμός κλάσεων



**με βάση τη
διαμέριση**



11

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Ζητήματα που προκύπτουν



- Ελλιπή δεδομένα (missing data)
 - τα αγνοούμε
 - τα αντικαθιστούμε με ειδικές τιμές
- Μέτρηση απόδοσης
 - Μέτρηση ακρίβειας με χρήση συνόλου δοκιμαστικών δεδομένων (test data)
 - Μήτρα σύγχυσης (confusion matrix)
 - Καμπύλη OC (operating characteristic)

12

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Παράδειγμα με δοκιμαστικά δεδομένα



Name	Gender	Height	Output1	Output2
Kristina	F	1.6m	Short	Medium
Jim	M	2m	Tall	Medium
Maggie	F	1.9m	Medium	Tall
Martha	F	1.88m	Medium	Tall
Stephanie	F	1.7m	Short	Medium
Bob	M	1.85m	Medium	Medium
Kathy	F	1.6m	Short	Medium
Dave	M	1.7m	Short	Medium
Worth	M	2.2m	Tall	Tall
Steven	M	2.1m	Tall	Tall
Debbie	F	1.8m	Medium	Medium
Todd	M	1.95m	Medium	Medium
Kim	F	1.9m	Medium	Tall
Amy	F	1.8m	Medium	Medium
Wynette	F	1.75m	Medium	Medium

13

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Ακρίβεια κατηγοριοποίησης

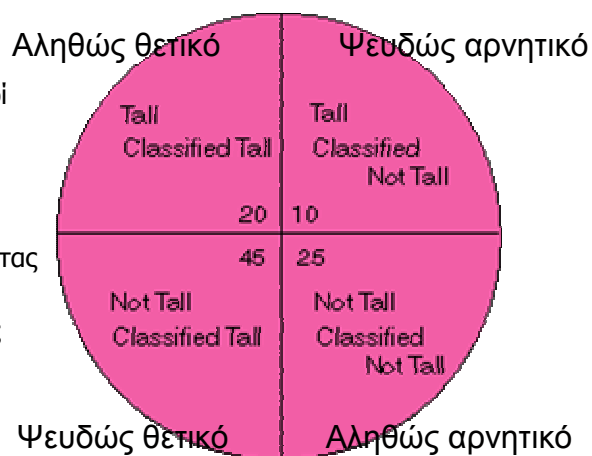


Για 2 κλάσεις
(π.χ. Tall/Medium)
υπάρχουν 4
πιθανοί συνδυασμοί

(n κλάσεις →
 n^2 συνδυασμοί)

Οπτικοποίηση ποιότητας
κατηγοριοποίησης:

- Μήτρα σύγχυσης
- Καμπύλη OC



14

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Μήτρα σύγχυσης



Έστω **Output1** η ορθή κατηγοριοποίηση και **Output2** η ανάθεση σε κλάσεις που προέκυψε από την (όποια) τεχνική κατηγοριοποίησης

Name	Gender	Height	Output1	Output2
Kristina	F	1.6m	Short	Medium
Jim	M	2m	Tall	Medium
Maggie	F	1.9m	Medium	Tall
Martha	F	1.88m	Medium	Tall
Stephanie	F	1.7m	Short	Medium
Bob	M	1.85m	Medium	Medium
Kathy	F	1.6m	Short	Medium
Dave	M	1.7m	Short	Medium
Worth	M	2.2m	Tall	Tall
Steven	M	2.1m	Tall	Tall
Debbie	F	1.8m	Medium	Medium
Todd	M	1.95m	Medium	Medium
Kim	F	1.9m	Medium	Tall
Amy	F	1.8m	Medium	Medium
Wynette	F	1.75m	Medium	Medium

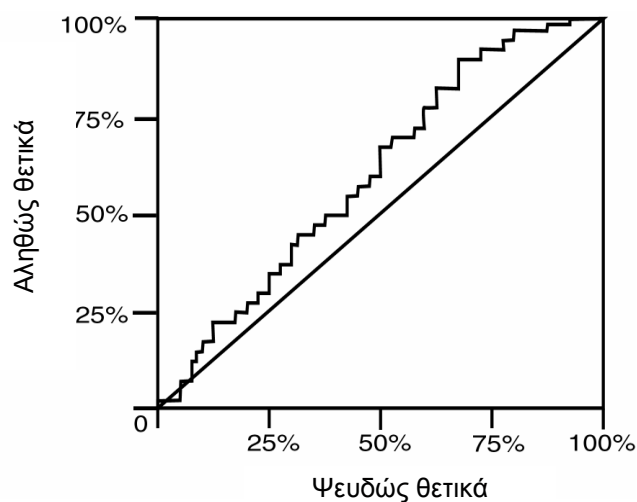
Μήτρα σύγχυσης
(confusion matrix):

Πραγματική κλάση	Ανάθεση		
	Short	Medium	Tall
Short	0	4	0
Medium	0	5	3
Tall	0	1	2

15

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Καμπύλη OC (Operating Characteristic)



16

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Παλινδρόμηση



- Κάνουμε την παραδοχή ότι τα δεδομένα ταιριάζουν σε μία συνάρτηση: $y = c_0 + c_1x_1 + \dots + c_nx_n$
- Το πρόβλημα είναι ο προσδιορισμός των **συντελεστών παλινδρόμησης** $c_0, c_1, \dots, c_n$
- Παραδοχή σφάλματος: $y = c_0 + c_1x_1 + \dots + c_nx_n + \varepsilon$
- Εκτίμηση σφάλματος με χρήση σφάλματος τετραγωνικού μέσου πάνω στο σύνολο δοκιμαστικών δεδομένων:

$$y_i = c_0 + c_1x_{1i} + \varepsilon_i, i = 1, \dots, k$$

$$L = \sum_{i=1}^k \varepsilon_i^2 = \sum_{i=1}^k (y_i - c_0 - c_1x_{1i})^2$$

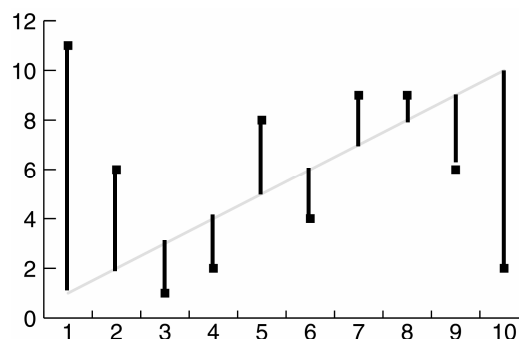
17

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Γραμμική παλινδρόμηση



- Φτωχή απόδοση (μέτριο ταίριασμα)



18

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Κατηγοριοποίηση με χρήση παλινδρόμησης



- **Διαίρεση:** χρησιμοποιούμε τη συνάρτηση παλινδρόμησης για να διαιρέσουμε το χώρο σε περιοχές.
- **Πρόβλεψη:** χρησιμοποιούμε τη συνάρτηση παλινδρόμησης για να προβλέψουμε τη συνάρτηση μέλους για μια κλάση. Η επιθυμητή κλάση δίνεται ως είσοδος στο πρόβλημα.

19

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Διαίρεση

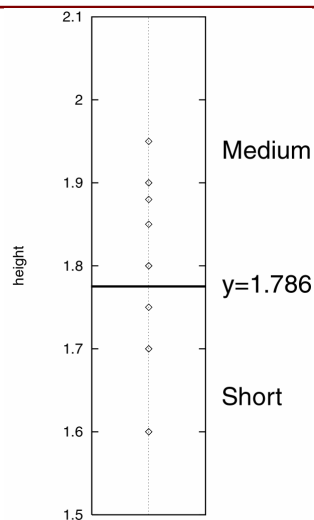


$$y = c_0 + \varepsilon$$

Θέλουμε να
ελαχιστοποιήσουμε
το L ως προς $c_0 \rightarrow$

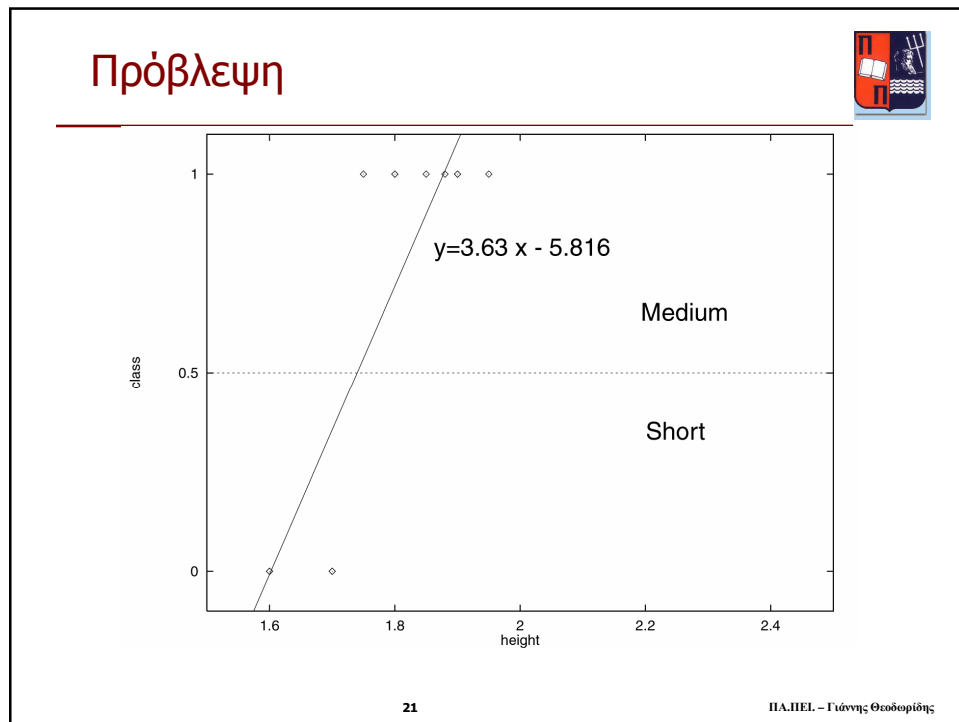
$$c_0 = \dots = 1.786 \rightarrow$$

$$y = 1.786$$



20

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης



Bayesian κατηγοριοποίηση

- Δύο παραδοχές για τα γνωρίσματα
 - Εξίσου σημαντικά
 - Στατιστικώς ανεξάρτητα (δοθείσης της τιμής μιας κλάσης)
 - Δηλαδή, αν γνωρίζουμε την τιμή ενός γνωρίσματος δεν μπορούμε να πούμε τίποτα για την τιμή ενός άλλου γνωρίσματος (με δεδομένο ότι γνωρίζουμε την κλάση)
- Η παραδοχή για την ανεξαρτησία των γνωρισμάτων σχεδόν ποτέ δεν ισχύει!
 - αλλά ... αυτό το σχήμα δείχνει να δουλεύει καλά στην πράξη

22

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Πρόγνωση καιρού



Outlook			Temperature			Humidity			Windy			Play	
Yes No			Yes No			Yes No			Yes No			Yes	No
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

23

Πρόγνωση καιρού (συν.)



Outlook			Temperature			Humidity			Windy			Play	
Yes No			Yes No			Yes No			Yes No			Yes	No
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

- Μια νέα ημέρα:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Πιθανοφάνειες για τις δύο κλάσεις

$$\text{"yes"} : 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$$

$$\text{"no"} : 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0206$$

Πιθανότητες (μετά την κανονικοποίηση):

$$P(\text{"yes"}) = 0.0053 / (0.0053 + 0.0206) = 0.205$$

$$P(\text{"no"}) = 0.0206 / (0.0053 + 0.0206) = 0.795$$

24

ΠΑ.ΙΕΛ - Γιάννης Θεοδωρίδης

Ο κανόνας του Bayes



- Η πιθανότητα να συμβεί ένα γεγονός H δοθείσης μιας μαρτυρίας E :

$$\Pr[H | E] = \frac{\Pr[E | H] \Pr[H]}{\Pr[E]}$$

- $A priori$ πιθανότητα του H : $\Pr[H]$
 - Η πιθανότητα του γεγονότος *χωρίς* την επίκληση της μαρτυρίας
- $A posteriori$ πιθανότητα του H : $\Pr[H | E]$
 - Η πιθανότητα του γεγονότος *με* την επίκληση της μαρτυρίας

Κατηγοριοποίηση Naïve Bayes



- Εκμάθηση κατηγοριοποίησης: ποια η πιθανότητα μιας κλάσης δοθείσης μιας μαρτυρίας;
 - Η μαρτυρία E είναι η εγγραφή στη ΒΔ
 - Το γεγονός H είναι η κλάση της εγγραφής
- Απλοϊκή (naïve) παραδοχή: η μαρτυρία διαιρείται σε μέρη (όσο και τα γνωρίσματα) που είναι ανεξάρτητα μεταξύ τους

$$\Pr[H | E] = \frac{\Pr[E_1 | H] \Pr[E_2 | H] \dots \Pr[E_n | H] \Pr[H]}{\Pr[E]}$$

Παράδειγμα πρόγνωσης καιρού



Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← μαρτυρία E

Πιθανότητα
κλάσης "yes"

$$\begin{aligned}
 \Pr[\text{yes} | E] &= \Pr[\text{Outlook} = \text{Sunny} | \text{yes}] \\
 &\quad \times \Pr[\text{Temperature} = \text{Cool} | \text{yes}] \\
 &\quad \times \Pr[\text{Humidity} = \text{High} | \text{yes}] \\
 &\quad \times \Pr[\text{Windy} = \text{True} | \text{yes}] \\
 &\quad \times \frac{\Pr[\text{yes}]}{\Pr[E]} \\
 &= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{\Pr[E]}
 \end{aligned}$$

27

ΠΑ.ΠΕΛ – Γιάννης Θεοδωρίδης

Το πρόβλημα της "μηδενικής συχνότητας"



- Τι θα συμβεί εάν δεν εμφανίζεται μια τιμή γνωρίσματος σε κάθε κλάση;
(π.χ. "Humidity = high" για την κλάση "yes")
 - Η πιθανότητα θα είναι μηδέν!

$$\Pr[\text{Humidity} = \text{High} | \text{yes}] = 0$$
 - Η *a posteriori* πιθανότητα θα είναι επίσης μηδέν!
(άσχετα με το ποιες είναι οι υπόλοιπες τιμές!)

$$\Pr[\text{yes} | E] = 0$$
- Τέχνασμα: προσθέτουμε 1 στο μετρητή κάθε ζευγαριού τιμής γνωρίσματος – κλάσης (εκτιμήτρια *Laplace*)
- Αποτέλεσμα: οι πιθανότητες δεν είναι ποτέ μηδέν!

28

ΠΑ.ΠΕΛ – Γιάννης Θεοδωρίδης

Σχολιασμός Naïve Bayes



- Η κατηγοριοποίηση Naïve Bayes περιέργως δουλεύει καλά!
 - ακόμη και αν καταστρατηγείται φανερά η παραδοχή περί ανεξαρτησίας γνωρισμάτων
- **Γιατί;** Επειδή η κατηγοριοποίηση δεν απαιτεί ακριβείς εκτιμήσεις πιθανοτήτων
 - αρκεί η μέγιστη πιθανότητα να αντιστοιχεί στη σωστή κλάση
- **Όμως:** η προσθήκη επιπλέον γνωρισμάτων μπορεί να δημιουργήσει προβλήματα
 - π.χ. ταυτόσημα γνωρίσματα

Κατηγοριοποίηση με χρήση απόστασης



- Τοποθετούμε τα δεδομένα στην «πλησιέστερη" (με όρους απόστασης) κλάση.
- Πρέπει να προσδιορίσουμε την απόσταση μεταξύ ενός στοιχείου και μιας κλάσης.
- Κάθε κλάση μπορεί να αναπαρασταθεί με
 - **Κέντρο βάρους (Centroid):** η κεντρική τιμή της κλάσης
 - **Κεντρικό στοιχείο (Medoid):** ένα αντιπροσωπευτικό σημείο – μέλος της.
 - Σύνολο από ενδεικτικά σημεία
- Αλγόριθμος: k- nearest neighbors (KNN)

Η προσέγγιση KNN



- Το σύνολο δεδομένων εκπαίδευσης περιλαμβάνει τις κλάσεις.
- Για να αναθέσουμε ένα νέο στοιχείο σε μια κλάση εξετάζουμε τα K πλησιέστερα σ' αυτό σημεία.
- Τοποθετούμε το νέο στοιχείο στην κλάση που έχει την πλειοψηφία μέσα στα κοντινά στοιχεία.
- Πολυπλοκότητα $O(q)$ για κάθε νέο στοιχείο (q είναι το μέγεθος του συνόλου δεδομένων εκπαίδευσης).

31

ΠΑ.ΠΕΛ - Γιάννης Θεοδωρίδης

Αλγόριθμος KNN



```

Input:
T           //training data
K           //Number of neighbors
t           //Input tuple to classify

Output:
c           //Class to which t is assigned

KNN algorithm: //Algorithm to classify tuple using KNN
begin
N = ∅;      //Find set of neighbors, N, for t
for each d ∈ T do
  if |N| ≤ K, then
    N = N ∪ {d};
  else
    if ∃ u ∈ N such that sim(t,u) ≤ sim(t,d),
    then
      begin N = N - {u}; N = N ∪ {d}; end
  //Find class for classification
c = class to which the most u ∈ N are classified
end

```

Υποθέτει ότι N είναι μια ειδική δομή, οργανωμένη με βάση την ομοιότητα $\text{sim}(t,u)$ – π.χ. σωρός ελαχίστων

32

ΠΑ.ΠΕΛ - Γιάννης Θεοδωρίδης

Παράδειγμα KNN



Name	Gender	Height	Output1	
Kristina	F	1.6m	Short	1
Jim	M	2m	Tall	
Maggie	F	1.9m	Medium	
Martha	F	1.88m	Medium	
Stephanie	F	1.7m	Short	3
Bob	M	1.85m	Medium	
Kathy	F	1.6m	Short	2
Dave	M	1.7m	Short	4
Worth	M	2.2m	Tall	
Steven	M	2.1m	Tall	
Debbie	F	1.8m	Medium	
Todd	M	1.95m	Medium	
Kim	F	1.9m	Medium	
Amy	F	1.8m	Medium	
Wynette	F	1.75m	Medium	5
Pat	F	1.6m	?	Short

33

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

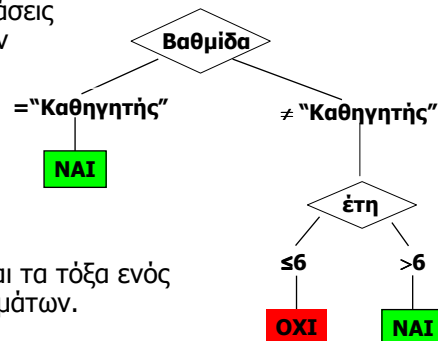
Κατηγοριοποίηση με δένδρα αποφάσεων (decision trees)



Κατηγοριοποίηση βασισμένη στη διαμέριση:

διαίρεση του χώρου σε ορθογώνιες περιοχές

- Οι εγγραφές ανατίθενται σε κλάσεις με βάση την περιοχή μέσα στην οποία πέφτουν.
- Οι τεχνικές ΔΑ διαφέρουν μεταξύ τους στον τρόπο κατασκευής του δένδρου (επαγωγή ΔΑ)
- Οι εσωτερικοί κόμβοι ενός ΔΑ αντιστοιχούν σε γνωρίσματα και τα τόξα ενός ΔΑ σε τιμές αυτών των γνωρισμάτων.
- Αλγόριθμοι: ID3, C4.5, CART



34

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Παράδειγμα: πρόγνωση καιρού

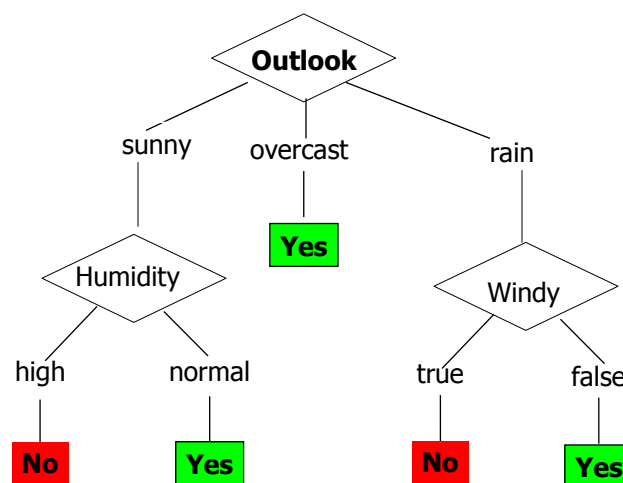


Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

35

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Παράδειγμα ΔΑ για το γνώρισμα "Play?"



36

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Δένδρο Απόφασης



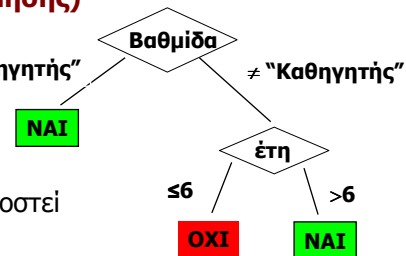
Δοθέντων:

- μιας βάσης δεδομένων $D = \{t_1, \dots, t_n\}$ όπου $t_i = \langle t_{i1}, \dots, t_{ih} \rangle$
- του σχήματος της ΒΔ $\{A_1, A_2, \dots, A_h\}$
- ενός συνόλου κλάσεων $C = \{C_1, \dots, C_m\}$

Δένδρο απόφασης (ή κατηγοριοποίησης)

είναι ένα δένδρο συσχετισμένο με τη D έτσι ώστε

- Κάθε εσωτερικός κόμβος έχει ως ετικέτα ένα γνώρισμα, A_i
- Κάθε τόξο έχει ως ετικέτα ένα κατηγορημα που μπορεί να εφαρμοστεί στο γνώρισμα του κόμβου-γονέα
- Κάθε φύλλο (τερματικός κόμβος) έχει ως ετικέτα μια κλάση, C_j



37

ΠΑ.ΠΕΛ - Γιάννης Θεοδωρίδης

Επαγωγή ΔΑ



```

Input:
D //Training data
Output:
T //Decision tree
DTBuild algorithm: //Simplistic algorithm to illustrate
naïve approach to building DT
begin
  T = ∅;
  Determine splitting criterion;
  T = Create root node and label with splitting attribute;
  T = Add arc to root node for each split predicate
    and label;
  for each arc do
    begin
      D = Database created by applying splitting predicate
        to D;
      if stopping point reached for this path, then
        T' = Create leaf node and label with appropriate
          class;
      else
        T' = DTBuild(D);
      T = Add T' to arc;
    end
  end
end

```

38

ΠΑ.ΠΕΛ - Γιάννης Θεοδωρίδης

Ζητήματα στα ΔΑ



- Αρχική επιλογή των γνωρισμάτων διάσπασης
 - Κάποια από τα γνωρίσματα της ΒΔ πρέπει να παραλειφθούν (δεν εξυπηρετούν την κατηγοριοποίηση)
- Κριτήριο διάσπασης
 - Επιλογή του γνωρίσματος διάσπασης
 - Επιλογή των κατηγορημάτων διάσπασης (πάνω στο γνώρισμα διάσπασης)
- Δενδρική δομή
 - επιθυμητό: ισοζυγισμένο δένδρο με λίγα επίπεδα
 - κάποιες τεχνικές παράγουν μόνο δυαδικά δένδρα
- Κριτήρια τερματισμού
 - ακρίβεια κατηγοριοποίησης vs. απόδοση vs. υπερπροσαρμογή
- Κλάδεμα (pruning)
 - εκ των υστέρων «τακτοποίηση» του ΔΑ για καλύτερη απόδοση

39

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Ένα κριτήριο για την επιλογή του κατάλληλου γνωρίσματος διάσπασης



- Ποιο είναι το καλύτερο;
 - Αυτό που θα οδηγήσει στο μικρότερο δένδρο
 - Ένας ευριστικός κανόνας (heuristic): επιλέγουμε το γνώρισμα που παράγει τους πιο "αγνούς" κόμβους. Για το σκοπό αυτό, χρησιμοποιείται μια **συνάρτηση καταλληλότητας** (fitness function).
- Στρατηγική: επιλέγουμε το γνώρισμα που μεγιστοποιεί τη συνάρτηση καταλληλότητας
 - Χαρακτηριστικές συναρτήσεις καταλληλότητας:
 - Κέρδος πληροφορίας – Gain (ID3)
 - Λόγος κέρδους πληροφορίας – GainRatio (C4.5)
 - gini index (SPRINT)

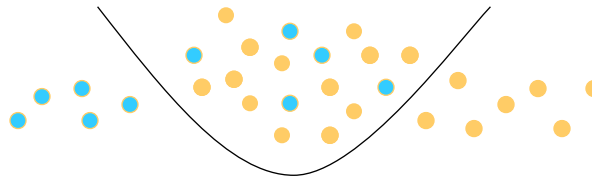
40

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Θεωρία Πληροφορίας



Η επαγωγή ΔΑ βασίζεται συχνά στη Θεωρία Πληροφορίας



41

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Πληροφορία / Εντροπία

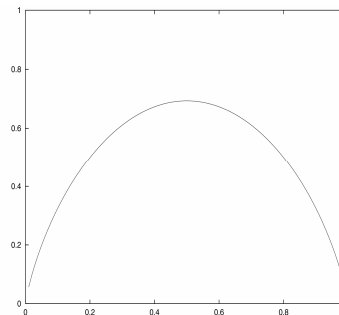


- Εστω πιθανότητες $p_1, p_2, \dots, p_s$ των οποίων το άθροισμα είναι 1.
Η **Εντροπία** ορίζεται ως εξής:

$$H(p_1, p_2, \dots, p_s) = \sum_{i=1}^s \left(p_i \log \left(\frac{1}{p_i} \right) \right)$$

η βάση του λογάριθμου δεν προσδιορίζεται
(συνήθως, 10 ή 2)

- Η εντροπία είναι ποσοτικοποίηση της τυχαιότητας (έκπληξης, αβεβαιότητας)
- Ο στόχος της κατηγοριοποίησης
 - καθόλου έκπληξη
 - εντροπία = 0



$H(p, 1-p)$

42

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Αλγόριθμος ID3



- Δημιουργεί ΔΑ με χρήση στοιχείων από τη θεωρία πληροφορίας (εντροπία)
- Επιλέγει για διάσπαση το γνώρισμα με το μεγαλύτερο κέρδος πληροφορίας (information gain):

$$Gain(D, S) = H(D) - \sum_{i=1}^s P(D_i) H(D_i)$$

- $H(D)$ η εντροπία του D (πριν το διαχωρισμό)
- $H(D_i)$ η εντροπία των επιμέρους D_i (μετά το διαχωρισμό)
- Όσο μεγαλύτερη είναι η μείωση (το «άλμα» προς το 0), τόσο μεγαλύτερο είναι το κέρδος $Gain(D, S)$

43

ΠΑ.ΠΕΛ – Γιάννης Θεοδωρίδης

Παράδειγμα ID3



- Αρχική κατάσταση εντροπίας:
 $H(D) = 4/15 \log(15/4) + 8/15 \log(15/8) + 3/15 \log(15/3) = 0.4384$

- Κέρδος αν γίνει διάσπαση στο **gender**:

- Gender='F':
 $3/9 \log(9/3) + 6/9 \log(9/6) = 0.2764$

- Gender='M':
 $1/6 \log(6/1) + 2/6 \log(6/2) + 3/6 \log(6/3) = 0.4392$

- Weighted sum:
 $(9/15)(0.2764) + (6/15)(0.4392) = 0.3415$

- Gain: $0.4384 - 0.3415 = \mathbf{0.0969}$

- Κέρδος αν γίνει διάσπαση στο **height**:

- Weighted sum: $0 + 0 + \dots + (2/15)(0.301) = 0.0401$

- Gain: $0.4384 - 0.0401 = \mathbf{0.3983}$

- Κατηγορήματα διάσπ.: $(0, 1.6], (1.6, 1.7], (1.7, 1.8], (1.8, 1.9], (1.9, 2.0], (2.0, \infty)$

- Επιλέγουμε **height**

Name	Gender	Height	Output1
Kristina	F	1.6m	Short
Jim	M	2m	Tall
Maggie	F	1.9m	Medium
Martha	F	1.88m	Medium
Stephanie	F	1.7m	Short
Bob	M	1.85m	Medium
Kathy	F	1.6m	Short
Dave	M	1.7m	Short
Worth	M	2.2m	Tall
Steven	M	2.1m	Tall
Debbie	F	1.8m	Medium
Todd	M	1.95m	Medium
Kim	F	1.9m	Medium
Amy	F	1.8m	Medium
Wynette	F	1.75m	Medium

44

ΠΑ.ΠΕΛ – Γιάννης Θεοδωρίδης

Αλγόριθμος C4.5



- Ο αλγόριθμος ID3 μεροληπτεί υπέρ των γνωρισμάτων με μεγάλο αριθμό διαιρέσεων
- Ο αλγόριθμος C4.5 αποτελεί βελτιωμένη έκδοχή του ID3:
 - Καλύτερη διαχείριση ελλিপών / συνεχών δεδομένων
 - Κλάδεμα – 2 τεχνικές:
 - αντικατάσταση υποδένδρου / ανύψωση υποδένδρου
 - Κανόνες αποφάσεων (που παράγονται από τα ΔΑ)
 - Βελτιωμένη συνάρτηση καταλληλότητας (για αποφυγή υπερπροσαρμογής):

$$GainRatio(D, S) = \frac{Gain(D, S)}{H\left(\frac{|D_1|}{|D|}, \dots, \frac{|D_s|}{|D|}\right)}$$

45

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Αλγόριθμος CART



- Δημιουργεί δυαδικό δένδρο
- Χρησιμοποιεί εντροπία
- Μαθηματικός τύπος για την επιλογή του σημείου διάσπασης, s , για τον κόμβο t :

$$\Phi(s/t) = 2P_L P_R \sum_{j=1}^m |P(C_j | t_L) - P(C_j | t_R)|$$

- Οι πιθανότητες P_L, P_R αντιστοιχούν στην πιθανότητα μια εγγραφή να βρεθεί στην αριστερή ή τη δεξιά πλευρά, αντίστοιχα, του δένδρου.

46

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Παράδειγμα CART



- Στο ξεκίνημα, υπάρχουν έξι επιλογές για σημείο διάσπασης:
 - Gender='M', height=1.6, height=1.7, height=1.8, height=1.9, height=2.0 (παραδοχή: η ισότητα οδηγεί στο δεξί κλαδί):
 - $\Phi(\text{Gender}='M') = \frac{2 \cdot (6/15) \cdot (9/15)}{(2/15 + 4/15 + 3/15)} = 0.224$
 - $\Phi(\text{height}=1.6) = 0$
 - $\Phi(\text{height}=1.7) = \frac{2 \cdot (2/15) \cdot (13/15)}{(0 + 8/15 + 3/15)} = 0.169$
 - $\Phi(\text{height}=1.8) = \frac{2 \cdot (5/15) \cdot (10/15)}{(4/15 + 6/15 + 3/15)} = 0.385$
 - $\Phi(\text{height}=1.9) = \frac{2 \cdot (9/15) \cdot (6/15)}{(4/15 + 2/15 + 3/15)} = 0.256$
 - $\Phi(\text{height}=2.0) = \frac{2 \cdot (12/15) \cdot (3/15)}{(4/15 + 8/15 + 3/15)} = 0.32$
- Αποφασίζεται διάσπαση στο height=1.8
- ΚΟΚ.

Name	Gender	Height	Output1
Kristina	F	1.6m	Short
Jim	M	2m	Tall
Maggie	F	1.9m	Medium
Martha	F	1.88m	Medium
Stephanie	F	1.7m	Short
Bob	M	1.85m	Medium
Kathy	F	1.6m	Short
Dave	M	1.7m	Short
Worth	M	2.2m	Tall
Steven	M	2.1m	Tall
Debbie	F	1.8m	Medium
Todd	M	1.95m	Medium
Kim	F	1.9m	Medium
Amy	F	1.8m	Medium
Wynette	F	1.75m	Medium

47

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Σύνοψη



- Κατηγοριοποίηση (ή ταξινόμηση): η ανάθεση ετικετών στις εγγραφές της βάσης δεδομένων σχετικά με την κλάση στην οποία ανήκει η καθεμία
 - Αλλιώς, διαμέριση της βάσης δεδομένων σε (προκαθορισμένες) κατηγορίες
- Τεχνικές:
 - στατιστικές (παλινδρόμηση, Bayesian, ...)
 - βασισμένες σε απόσταση (k-NN, ...)
 - δένδρα αποφάσεων (ID3, C4.5, CART, ...)

48

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης