Κατηγοριοποίηση (classification)

Κατηγοριοποίηση (classification)

Τί είναι η κατηγοριοποίηση;

Κατηγοριοποίηση

Κατηγοριοποίηση είναι η τοποθέτηση ενός αντικειμένου σε μια ή περισσότερες **προκαθορισμένες κατηγορίες** (ομάδες) με βάση κάποια χαρακτηριστικά του.

Παραδείγματα

- Εντοπισμός spam email, με βάση τον header ή το περιεχόμενό τους.
- Αξιολόγηση καρκινικών κυττάρων ως καλοήθη ή κακοήθη.
- Χαρακτηρισμός συναλλαγών με πιστωτικές κάρτες ως νόμιμες ή προϊόν απάτης.
- Χαρακτηρισμός ειδήσεων ως οικονομικές, αθλητικές, πολιτιστικές, πρόβλεψης καιρού.

Κατηγοριοποίηση (classification)

Τί είναι η κατηγοριοποίηση;

Έστω μια οικογένεια αντικειμένων της οποίας κάθε αντικείμενο κωδικοποιείται από ορισμένα χαρακτηριστικά $x_1, x_2, \ldots, x_n, y$, όπου

- $x_1, x_2, \ldots, x_n$ είναι τα γνωρίσματα/ανεξάρτητες μεταβλητές/είσοδος
- y είναι η κατηγορία/εξαρτημένη μεταβλητή/έξοδος

Ο στόχος της κατηγοριοποίησης είναι

- με βάση ένα σύνολο αντικειμένων για το οποίο γνωρίζουμε όλα τα χαρακτηριστικά $x_1, x_2, \ldots, x_n, y$ (**σύνολο εκμάθησης/εκπαίδευσης**),
- να βρεθεί μια συνάρτηση/ένας κανόνας (**μοντέλο κατηγοριοποίησης**) ο οποίος συσχετίζει τα γνωρίσματα $x_1, x_2, \ldots, x_n$ με την κατηγορία y,
- και όταν δοθεί ένα αντικείμενο για το οποίο γνωρίζουμε μόνο τα $x_1, x_2, \ldots, x_n$ να μπορούμε με βάση το μοντέλο να προσδιορίσουμε την κατηγορία y.

Κατηγοριοποίηση (classification)

Τί είναι η κατηγοριοποίηση;

Παράδειγμα

Tid	Επιστροφή (κατηγορικό)	Οικογενειακή Κατάσταση (κατηγορικό)	Φορολογητέο Εισόδημα (συνεχές)	Απάτη (κλάση)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Κατηγοριοποίηση (classification)

Τί είναι η κατηγοριοποίηση;

Το μοντέλο κατηγοροποίησης μπορεί να είναι

- δένδρο απόφασης
- κανόνες
- μαθηματικός τύπος, κ.ο.κ.

Κατηγοριοποίηση (classification)

Τί είναι η κατηγοριοποίηση;

Τα μοντέλα κατηγοριοποίησης χρησιμοποιούνται είτε

- **ως μοντέλα πρόβλεψης (predictive models)** για την πρόβλεψη της κλάσης άγνωστων αντικειμένων, π.χ. με βάση το header ενός email να χαρακτηριστεί spam ή όχι.
- **ως μοντέλα περιγραφής (descriptive models)** για την εύρεση ποιων χαρακτηριστικών είναι σημαντικά για την κατηγοριοποίηση ενός αντικειμένου σε μια ομάδα, π.χ. για να χαρακτηριστεί ένας δανειολήπτης ως αξιόπιστος ή όχι θα εξετάσουμε το περσινό εισόδημά του και όχι π.χ. την ποδοσφαιρική ομάδα που προτιμά.

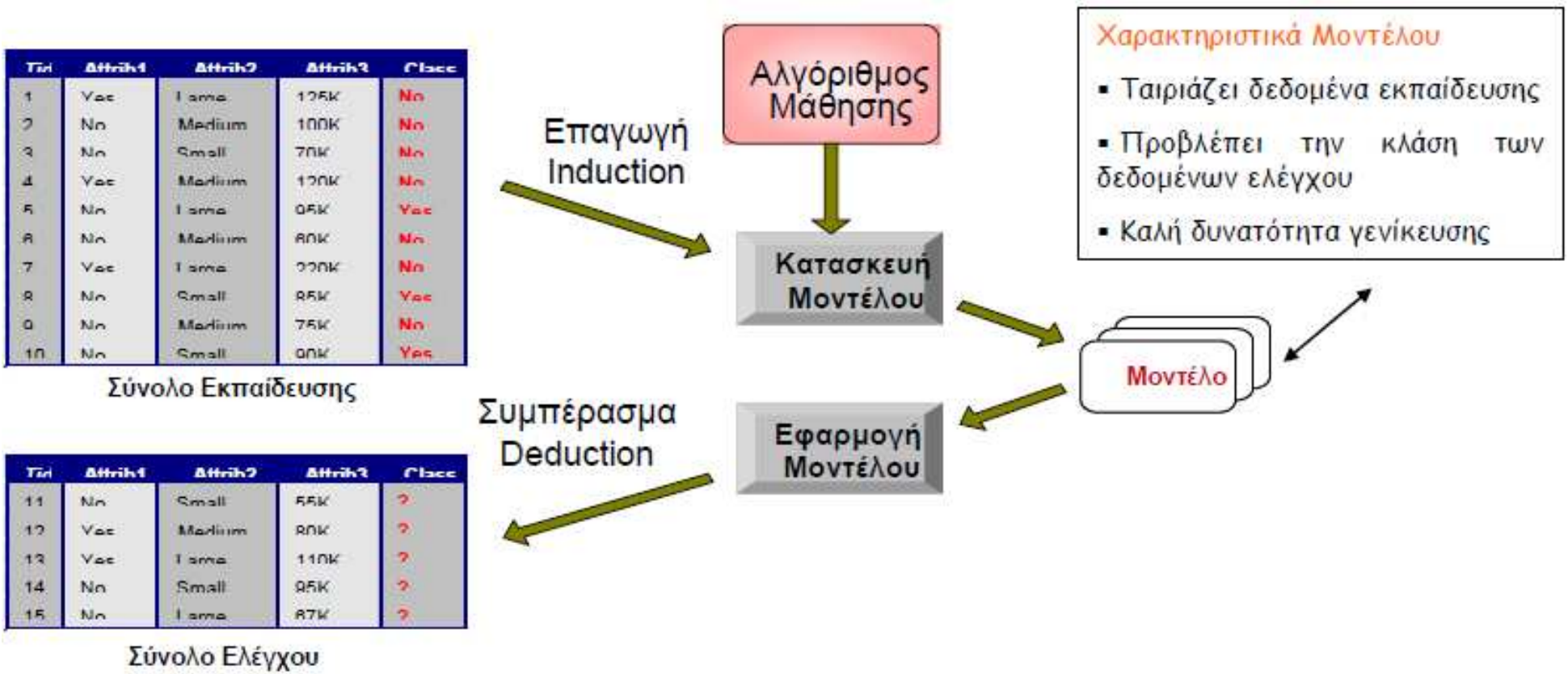
Κατηγοριοποίηση (classification)

Τί είναι η κατηγοριοποίηση;

Συνήθως εκτός από το **σύνολο εκπαίδευσης** (training set) έχουμε και ένα **σύνολο ελέγχου** (test set).
Το σύνολο εκπαίδευσης χρησιμοποιείται για την **κατασκευή** του μοντέλου, ενώ το σύνολο ελέγχου για την **αξιολόγηση** του μοντέλου πριν εφαρμοσθεί.
Για παράδειγμα, μετράμε τον **ρυθμό ακρίβειας** του μοντέλου: το ποσοστό των εγγραφών του συνόλου ελέγχου που ταξινομούνται σωστά από το μοντέλο.

Κατηγοριοποίηση (classification)

Διαδικασία κατηγοροποίησης

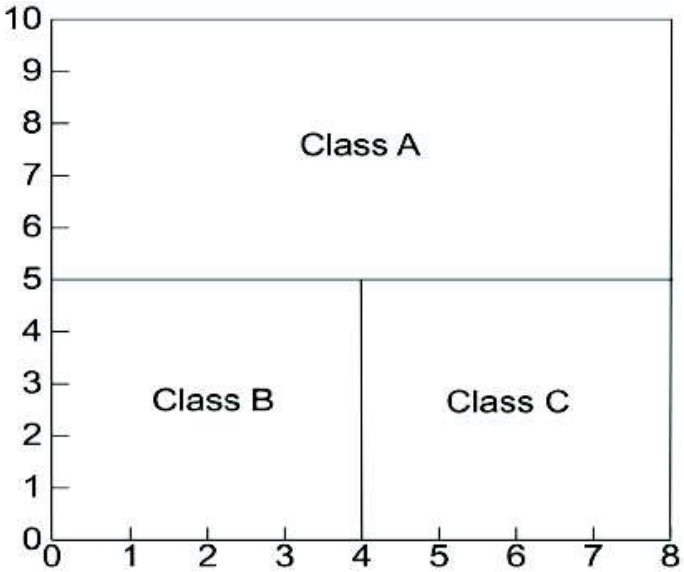


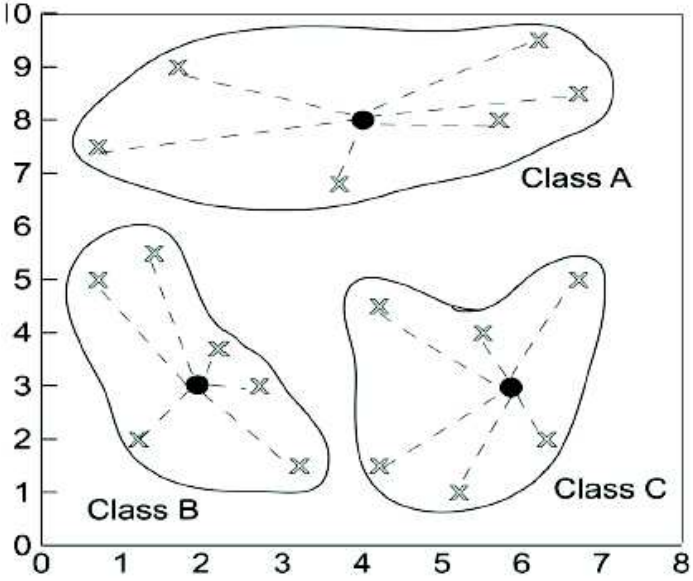
Κατηγοριοποίηση (classification)

Μέθοδοι κατηγοριοποίησης

Οι μέθοδοι κατηγοριοποίησης μπορούν να χωρισθούν σε 3 μεγάλες κατηγορίες:

- Με βάση τη διαμέριση του χώρου (δένδρα απόφασης, κανόνες, κ.α.)





- Με βάση την ομοιότητα των αντικειμένων (k-nearest neighbors)
- Στατιστικές τεχνικές (Παλινδρόμηση, Bayesian κατηγοριοποίηση, κ.α.)

Κατηγοριοποίηση (classification)

Μέθοδοι κατηγοριοποίησης

Πιο αναλυτικά υπάρχουν μέθοδοι κατηγοριοποίησης που βασίζονται σε

1. δένδρα απόφασης (decision trees)
2. κανόνες (rule-based methods)
3. αλγόριθμους κοντινότερου γείτονα (nearest neighbor methods)
4. νευρωνικά δίκτυα (neural networks)
5. naive Bayes και Bayes networks
6. support vector machines
7. κ.α.

Δένδρα απόφασης

Εδώ το μοντέλο είναι ένα δένδρο.

- **Εσωτερικοί κόμβοι**: Γνωρίσματα του αντικειμένου
- **Φύλλα (εξωτερικοί κόμβοι)**: Κατηγορία του αντικειμένου
- **Ακμές**: Συνθήκες διαχωρισμού (με βάση τις οποίες επιλέγουμε κάποιο παιδί ενός κόμβου.

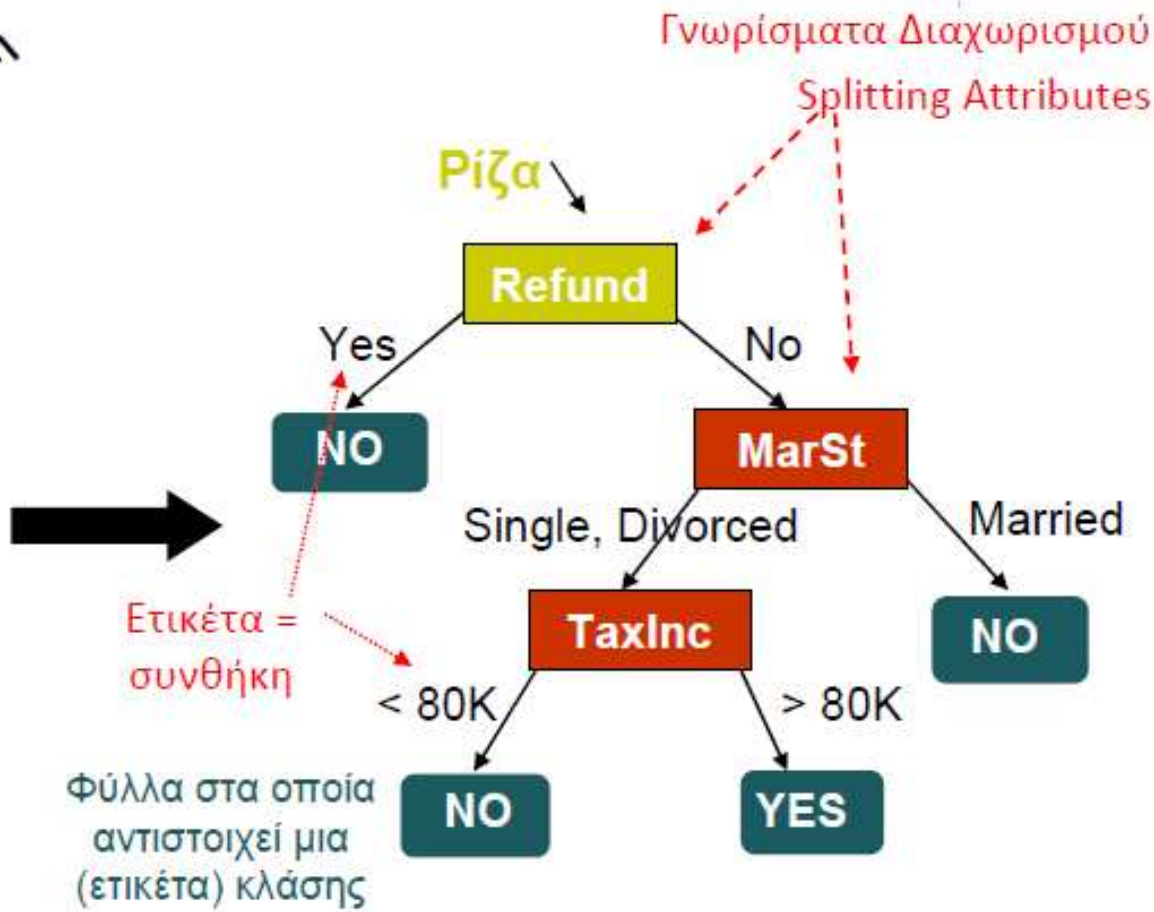
Η εφαρμογή του μοντέλου συνίσταται στην **διάσχιση του δένδρου από την ρίζα προς κάποιο φύλλο** με βάση τα χαρακτηριστικά ενός αντικείμενου και τις συνθήκες διαχωρισμού των ακμών.

Κατηγοριοποίηση (classification)

Παράδειγμα δένδρου απόφασης

Δεδομένα Εκπαίδευσης

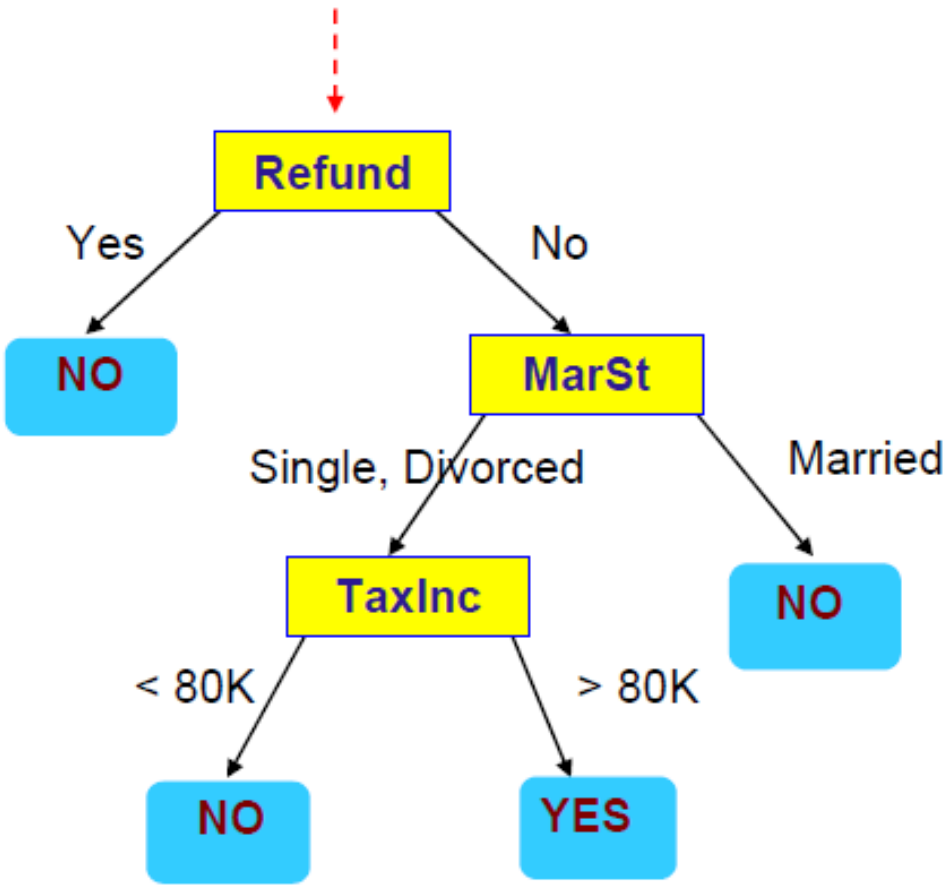
Tid	Refund (κατηγορικό)	Marital Status (κατηγορικό)	Taxable Income (συνεχές)	Cheat (κλάση)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Κατηγοριοποίηση (classification)

Παράδειγμα εφαρμογής δένδρου απόφασης

Ξεκίνα από τη ρίζα του δέντρου.



Δεδομένα Ελέγχου

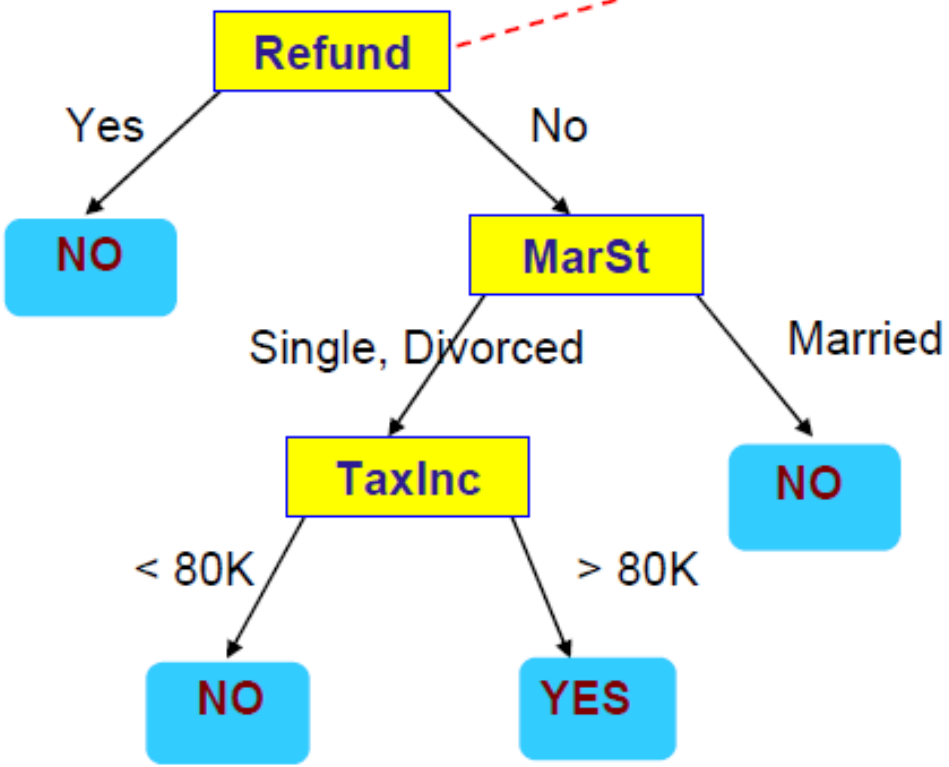
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Κατηγοριοποίηση (classification)

Παράδειγμα εφαρμογής δένδρου απόφασης

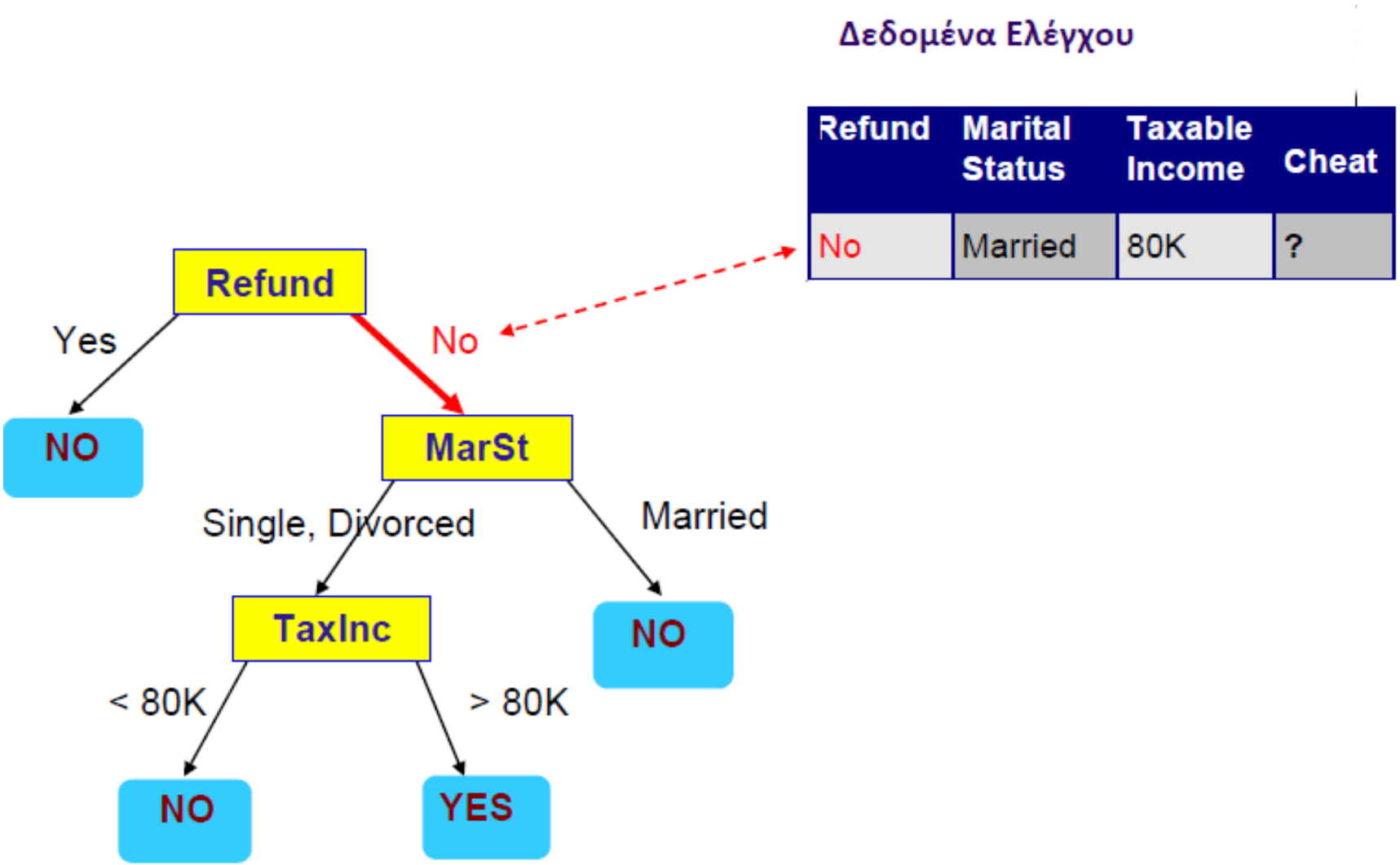
Δεδομένα Ελέγχου

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



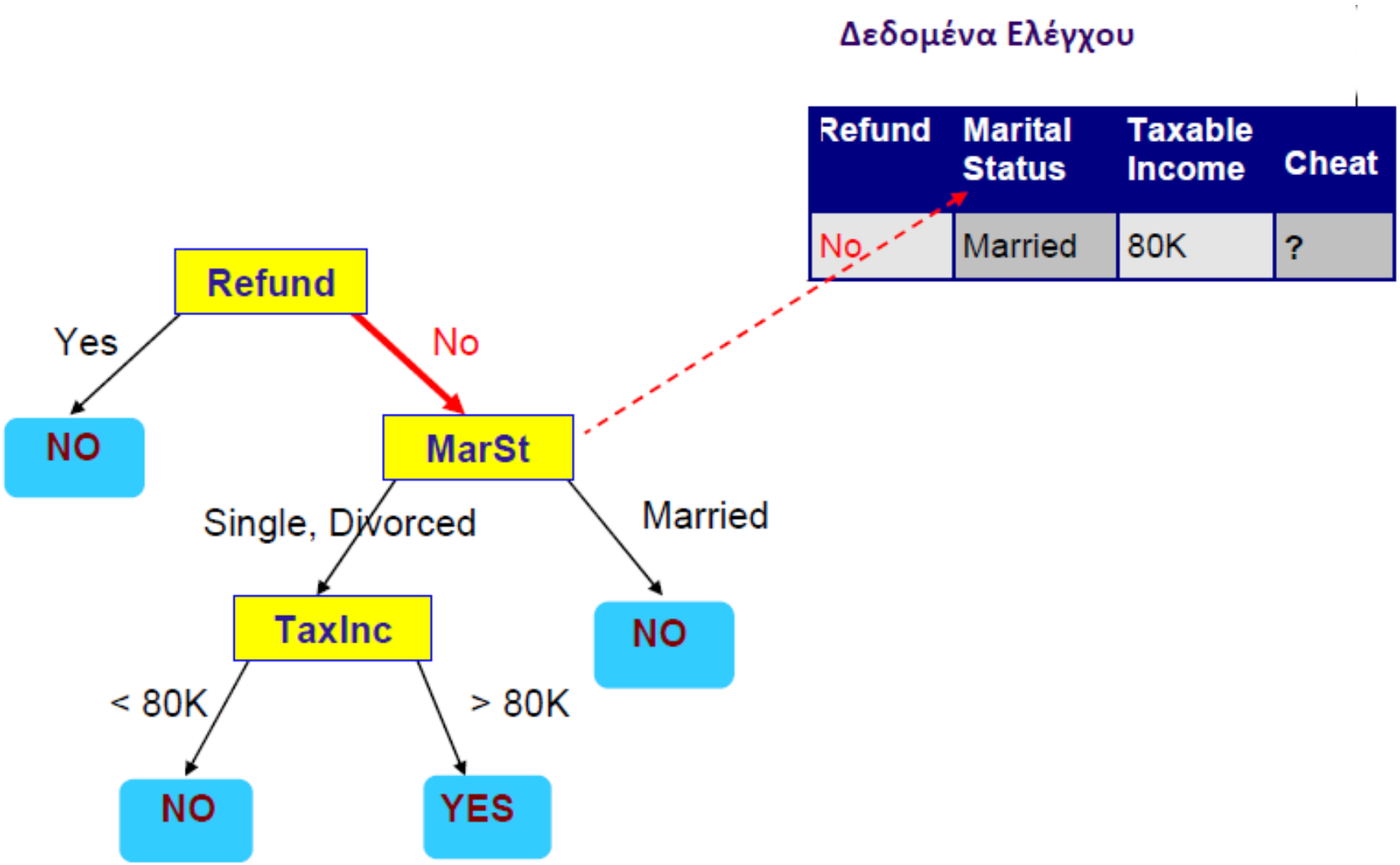
Κατηγοριοποίηση (classification)

Παράδειγμα εφαρμογής δένδρου απόφασης



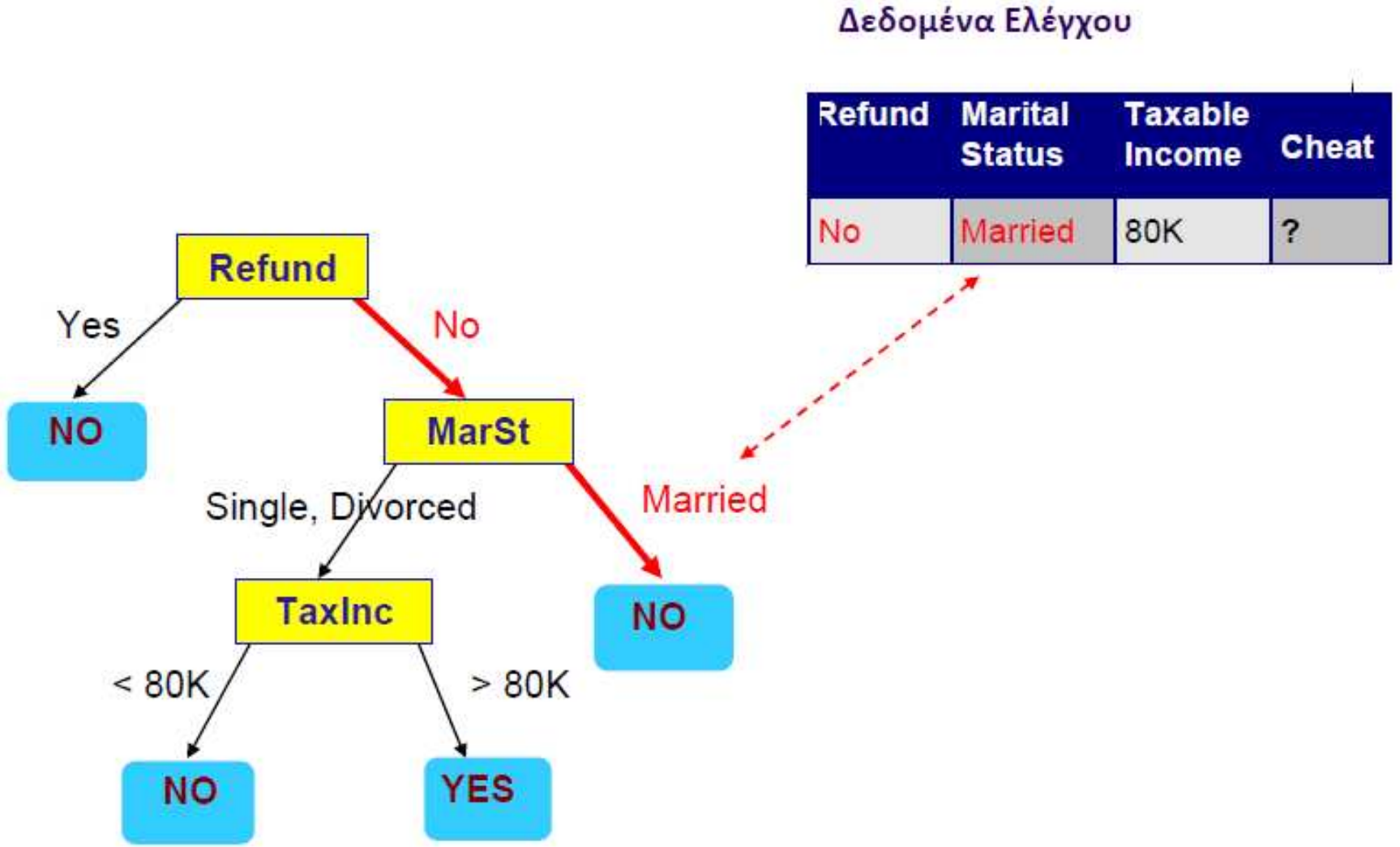
Κατηγοριοποίηση (classification)

Παράδειγμα εφαρμογής δένδρου απόφασης



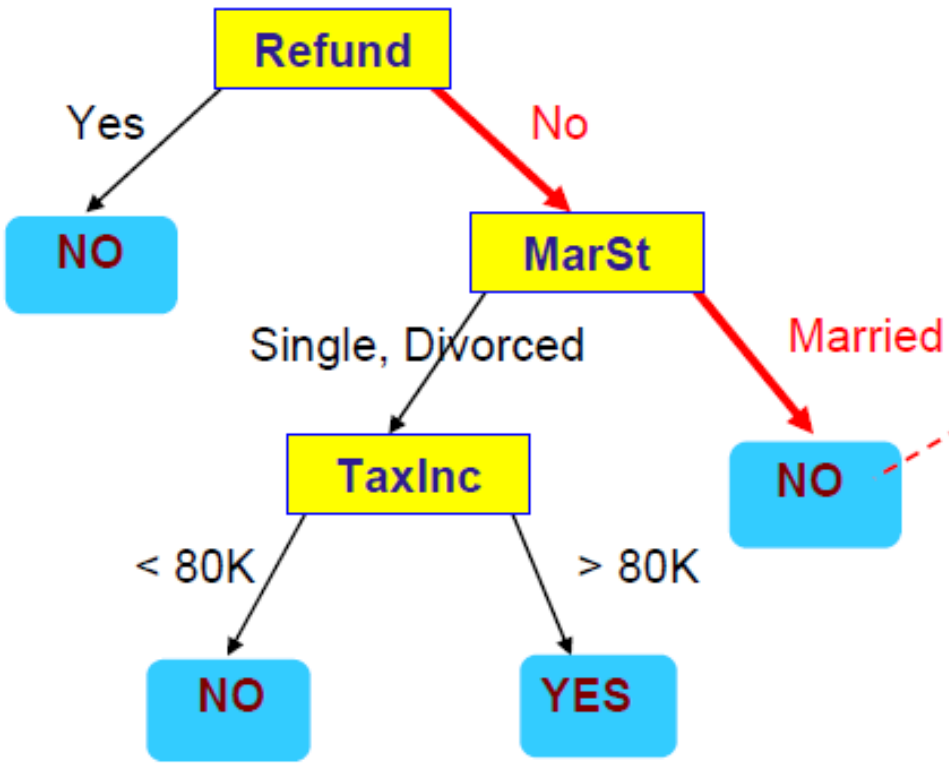
Κατηγοριοποίηση (classification)

Παράδειγμα εφαρμογής δένδρου απόφασης



Κατηγοριοποίηση (classification)

Παράδειγμα εφαρμογής δένδρου απόφασης



Δεδομένα Ελέγχου

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Ανάθεση στο Cheat “No”

Κατηγοριοποίηση (classification)

Κατασκευή δένδρων απόφασης

Η βασική τεχνική για την κατασκευή ενός δένδρου απόφασης είναι η εξής:

1. Ξεκινάμε με έναν κόμβο που περιέχει όλες τα αντικείμενα του συνόλου εκπαίδευσης
2. Διασπάμε τον κόμβο (μοίρασμα των αντικειμένων) με βάση κάποια **συνθήκη διαχωρισμού** για κάποιο/κάποια από τα γνωρίσματα των αντικειμένων.
3. Αναδρομική κλήση των δύο προηγούμενων βημάτων στα διασπασμένα σύνολα.
4. Αφού ολοκληρωθεί η αναδρομική κατασκευή γίνονται κάποιες βελτιώσεις (π.χ. κλαδεύουμε αν όλα τα φύλλα που βρίσκονται σε κάποιο υποδένδρο έχουν την ίδια τιμή).

Ερώτημα

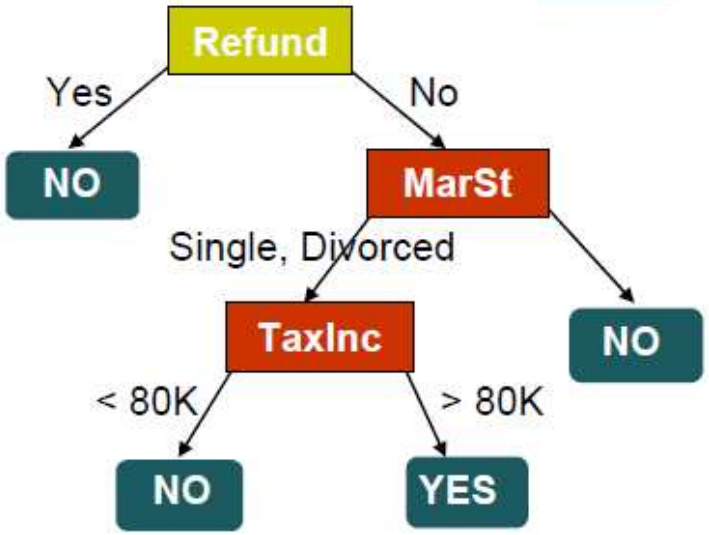
Ποια συνθήκη διαχωρισμού και με βάση ποιο γνώρισμα πρέπει να χρησιμοποιήσουμε για την διάσπαση κάθε κόμβου;

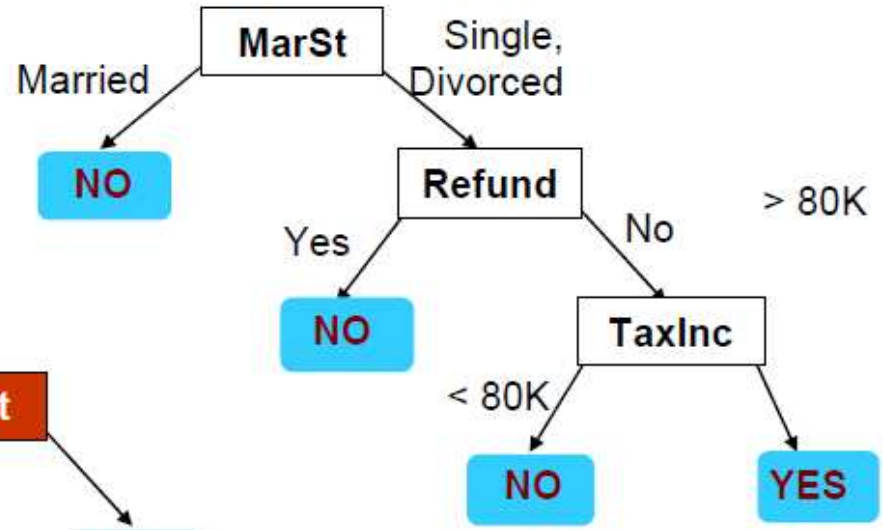
Κατηγοριοποίηση (classification)

Κατασκευή δένδρων απόφασης

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Για το ίδιο σύνολο εκπαίδευσης υπάρχουν διαφορετικά δέντρα





Κατηγοριοποίηση (classification)

Κατασκευή δένδρων απόφασης

Οι γνωστότεροι αλγόριθμοι κατασκευής δένδρων απόφασης είναι οι εξής:

1. αλγόριθμος του Hunt (από τους παλαιότερους αλγόριθμους)
2. CART
3. ID3, C4.5
4. SLIQ, SPRINT

Αλγόριθμος του Hunt

Ο αλγόριθμος του Hunt παράγει το δένδρο αναδρομικά: Αρχικά όλα τα αντικείμενα του συνόλου εκπαίδευσης βρίσκονται σε έναν κόμβο (ρίζα). Έστω D_t το σύνολο των αντικειμένων που βρίσκονται στον κόμβο t.
Σε κάθε κόμβο εφαρμόζονται αναδρομικά τα εξής βήματα

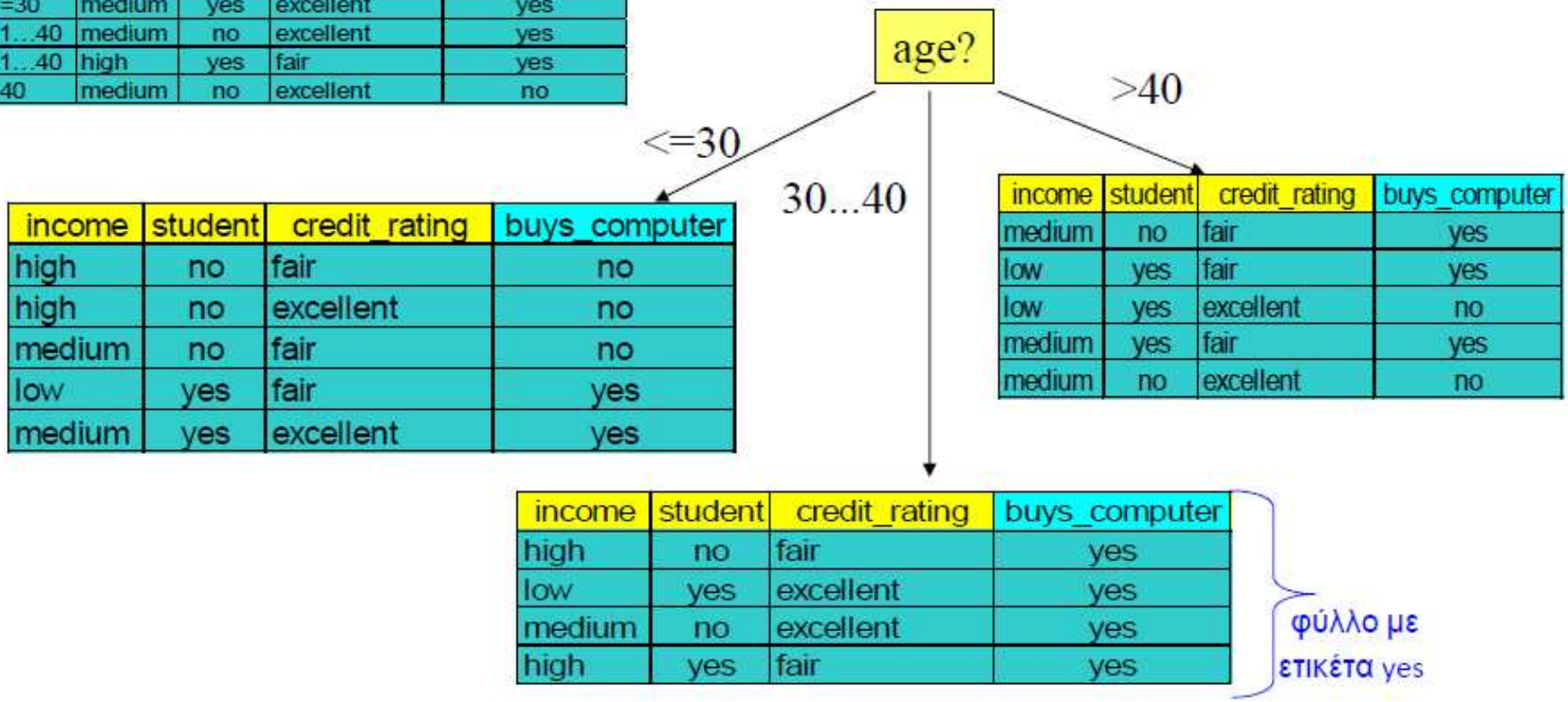
- Αν το D_t περιέχει αντικείμενα που ανήκουν στην ίδια κλάση γ, τότε ο κόμβος t είναι γίνεται φύλλο με ετικέτα γ.
- Αν το D_t είναι το κενό σύνολο (δηλαδή στο σύνολο εκπαίδευσης δεν υπάρχει αντικείμενο με αυτό τον συνδυασμό τιμών), τότε το D_t γίνεται φύλλο με ετικέτα την μεγαλύτερη κλάση του συνόλου εκπαίδευσης ή κάποιας άλλης default κλάσης.
- Αν το D_t περιέχει αντικειμένα που ανήκουν σε περισσότερες από μια κλάσεις τότε χρησιμοποίησε **μια συνθήκη διαχωρισμού** για τον διαχωρισμό των δεδομένων σε μικρότερα υποσύνολα.
  Ενδέχεται να μην είναι δυνατός ο διαχωρισμός: αν ο ίδιος συνδυασμός γνωρισμάτων αντιστοιχεί σε περισσότερες από μια κλάσεις. Τότε, ο κόμβος t γίνεται φύλλο όπως και στην προηγούμενη περίπτωση.

Αλγόριθμος του Hunt - Παράδειγμα 1

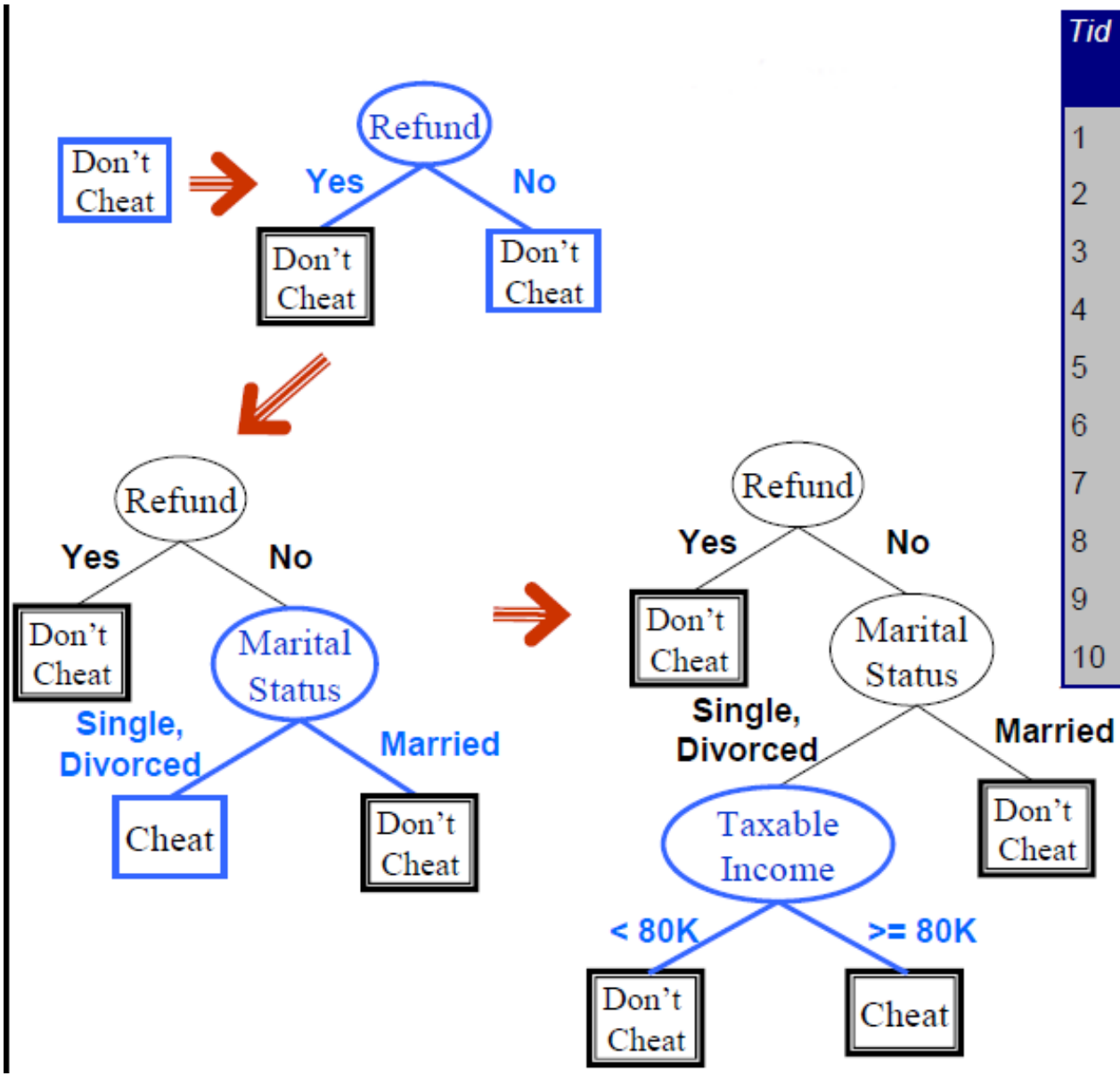
age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

κλάση

- Ποιο γνώρισμα (πχ age)
- Ποια συνθήκη

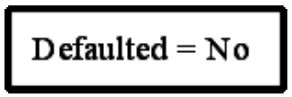


Αλγόριθμος του Hunt - Παράδειγμα 2



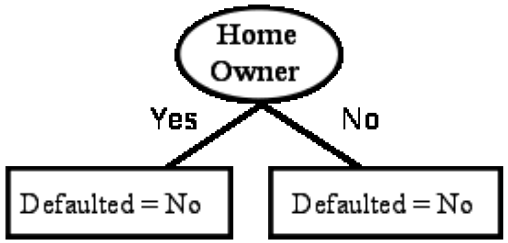
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Αλγόριθμος του Hunt - Παράδειγμα 3



(7,3)

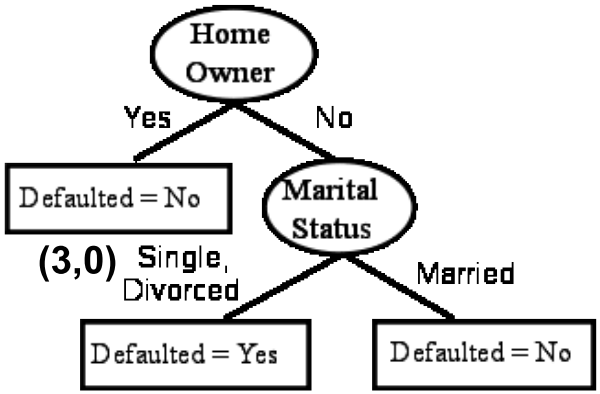
(a)



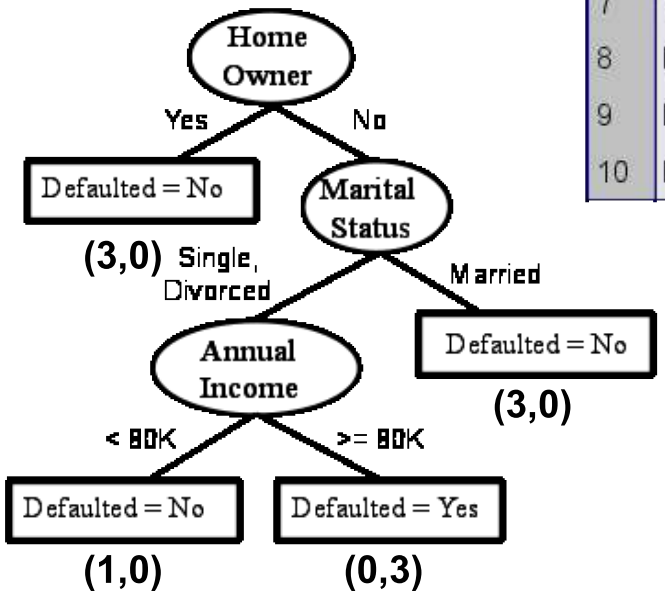
(3,0) (4,3)

(b)

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



(c)



(d)

Συνθήκες διαχωρισμού

Ερώτημα

Ποια συνθήκη διαχωρισμού και με βάση ποιο γνώρισμα πρέπει να χρησιμοποιήσουμε για την διάσπαση κάθε κόμβου;

Πριν δώσουμε απαντήσεις στο ερώτημα αυτό ας δούμε μερικούς **τύπους συνθηκών διαχωρισμού**:

- **δυαδικός διαχωρισμός** (2-way split, binary split)
- **πολλαπλός διαχωρισμός** (multiway split)

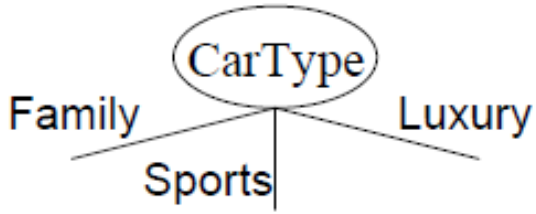
Ο τύπος διαχωρισμού εξαρτάται από τον τύπο των γνωρισμάτων των αντικειμένων

- διακριτές (nominal) (π.χ. φύλο (άνδρας, γυναίκα) ή χρώμα (κόκκινο, κίτρινο, μπλε))
- διατεταγμένες (ordinal) (π.χ. ηλικία σε έτη)
- συνεχείς (continuous) (π.χ. βάρος)

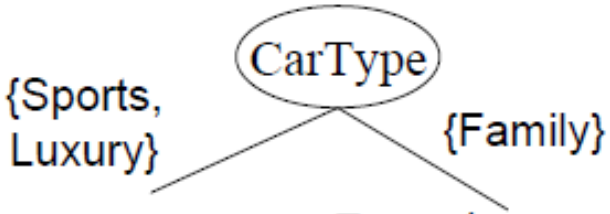
Συνθήκες διαχωρισμού

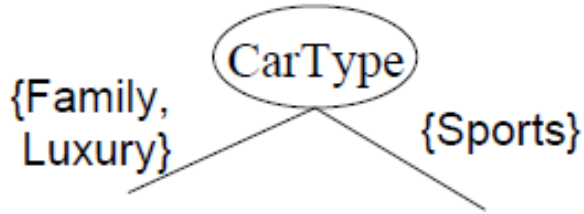
Διαχωρισμός βασισμένος σε διακριτές τιμές

- Πολλαπλός διαχωρισμός: Χρησιμοποίησε τόσες διασπάσεις όσες οι διαφορετικές τιμές



- Δυαδικός Διαχωρισμός: Χωρίζει τις τιμές σε δύο υποσύνολα. Πρέπει να βρει το *βέλτιστο* διαχωρισμό (partitioning).

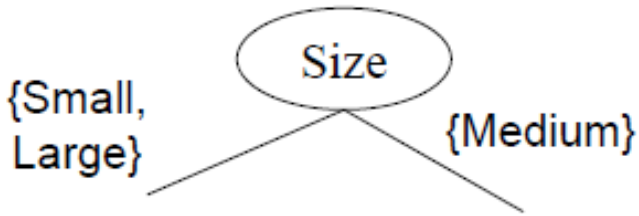




Γενικά, αν κ τιμές, $2^{κ-1} - 1$ τρόποι

Όταν υπάρχει διάταξη, πρέπει οι διασπάσεις να μη την παραβιάζουν

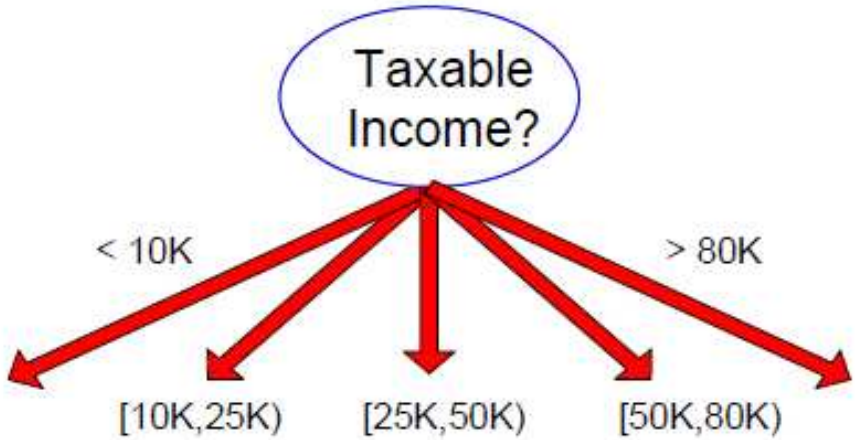
Είναι καλός αυτός το διαχωρισμός;



Συνθήκες διαχωρισμού

Διαχωρισμός βασισμένος σε συνεχείς τιμές

Δυαδικός διαχωρισμός



Πολλαπλός διαχωρισμός

Συνθήκες διαχωρισμού

Διαχωρισμός βασισμένος σε συνεχείς τιμές

Τρόποι χειρισμού

- Discretization (διακριτοποίηση) ώστε να προκύψει ένα διατεταγμένο κατηγορικό γνώρισμα

  Ταξινόμηση των τιμών και χωρισμός τους σε περιοχές καθορίζοντας *n* – 1 σημεία διαχωρισμού, απεικόνιση όλων των τιμών μιας περιοχής στην ίδια κατηγορική τιμή

  Στατικό – μια φορά στην αρχή

  Δυναμικό – εύρεση των περιοχών πχ έτσι ώστε οι περιοχές να έχουν το ίδιο διάστημα ή τις ίδιες συχνότητες εμφάνισης ή με χρήση συσταδοποίησης

- Δυαδική Απόφαση: $(A < v)$ or $(A \geq v)$

  εξετάζει όλους τους δυνατούς διαχωρισμούς (τιμές του v) και επιλέγει τον καλύτερο – υπολογιστικά βαρύ

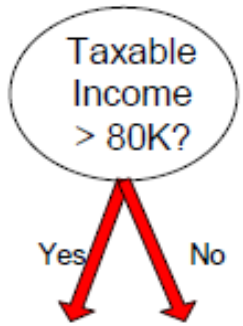
Συνθήκες διαχωρισμού

Συνεχή Γνωρίσματα (δυαδικός διαχωρισμός αναλυτικά)

Πχ, χρήση **δυαδικών αποφάσεων** πάνω σε μία τιμή

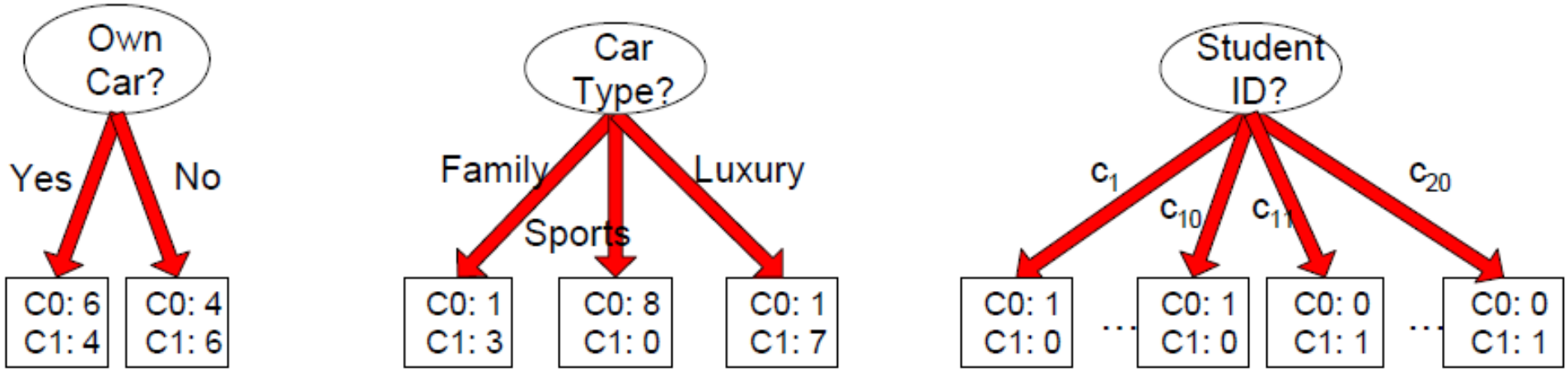
- Πολλές επιλογές για την τιμή διαχωρισμού
  - Αριθμός πιθανών διαχωρισμών = Αριθμός διαφορετικών τιμών – *έστω N*
- Κάθε τιμή διαχωρισμού v συσχετίζεται με έναν πίνακα μετρητών
  - Μετρητές των κλάσεων για κάθε μια από τις δύο διασπάσεις, $A < v$ and $A \geq v$
- Απλή μέθοδος για την επιλογή της καλύτερης τιμής v (βέλτιστη τιμή διαχωρισμού – best split point)
  - Διάταξε τις τιμές του A σε αύξουσα διάταξη
  - Συνήθως επιλέγεται το μεσαίο σημείο ανάμεσα σε γειτονικές τιμές ας υποψήφιο
  - $(a_i+a_{i+1})/2$ **μέσο των τιμών** a_i **και** a_{i+1}
  - Επέλεξε το «βέλτιστο» ανάμεσα στα υποψήφια

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Βέλτιστες συνθήκες διαχωρισμού

Έστω ότι πριν το διαχωρισμό: 10 εγγραφές της κλάσης 0,
10 εγγραφές της κλάσης 1

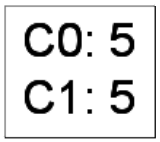


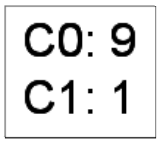
Ποια από τις 3 διασπάσεις να προτιμήσουμε; (Δηλαδή, ποια συνθήκη ελέγχου είναι καλύτερη;)

=> ορισμός κριτηρίου βέλτιστου διαχωρισμού

Βέλτιστες συνθήκες διαχωρισμού

Ευρετικός κανόνας: Διαισθητικά προτιμώνται οι κόμβοι με **ομοιογενείς (pure - homogeneous)** κατανομές κλάσεων, έναντι των κόμβων με **ανομοιογενείς (impure)** κατανομές κλάσεων





High degree of impurity **Low degree of impurity**

Ιδανικά θέλουμε όλα τα αντικείμενα να ανήκουν στην ίδια κλάση. Για το σκοπό αυτό έχουν ορισθεί μέτρα της **μη καθαρότητας (impurity)**, ή ισοδύναμα της **ομοιογένειας (homegenicity)** ενός συνόλου αντικειμένων.

Βέλτιστες συνθήκες διαχωρισμού - Μέτρα μη καθαρότητας

Έστω c **ο συνολικός αριθμός των κλασεων** και $P(j|t)$ **το ποσοστό των αντικείμενων του κόμβου** t **που ανήκουν στην κλάση** j.

1. Δείκτης Gini (Gini index)

$$\text{GINI}(t) = 1 - \sum_{j=1}^{c} P(j|t)^2$$

   Χρησιμοποιείται στους αλγορίθμους CART, SLIQ, SPRINT.

2. Εντροπία (Entropy)

$$I(t) = \text{Info}(t) = \text{Entropy}(t) = -\sum_{j=1}^{c} P(j|t) \log P(j|t)$$

   Χρησιμοποιείται στους αλγορίθμους ID3 (κέρδος πληροφορίας), C4.5 (λόγος κέρδους πληροφορίας)

3. Λάθος ταξινόμησης (Misclassification error)

$$\text{Error}(t) = 1 - \max_j P(j|t)$$

Βέλτιστες συνθήκες διαχωρισμού - Μέτρα μη καθαρότητας

Δεδομένου ότι έχουμε διαλέξει ένα μέτρο μη καθαρότητας, για να βρούμε την βέλτιστη συνθήκη διαχωρισμού ακολουθούμε τα εξής βήματα:

1. Υπολογίζουμε την μη καθαρότητα P ενός κόμβου t πριν τον διαχωρισμό.
2. Διαλέγουμε μια συνθήκη διαχωρισμού και υπολογίζουμε τον σταθμισμένο μέσο M της μη καθαρότητας των παιδιών του κόμβου t μετά το διαχωρισμό.
3. Διαλέγουμε εκείνη την συνθήκη διαχωρισμού που δίνει το μεγαλύτερο **κέρδος (gain)** Δ

$$\Delta = \text{Gain} = P - M,$$

ή, ισοδύναμα την ελάχιστη μη καθαρότητα M μετά τον διαχωρισμό.

Ο δείκτης Gini

Έστω c **ο συνολικός αριθμός των κλασεων** και $P(j|t)$ **το ποσοστό των αντικείμενων του κόμβου** t **που ανήκουν στην κλάση** j.

$$\text{GINI}(t) = 1 - \sum_{j=1}^{c} P(j|t)^2$$

Παραδείγματα

v1	
C1	0
C2	6
Gini=0.000	

v2	
C1	1
C2	5
Gini=0.278	

v3	
C1	2
C2	4
Gini=0.444	

v4	
C1	3
C2	3
Gini=0.500	

1. Η ελάχιστη τιμή του $\text{GINI}(t)$ είναι 0 (όταν όλα τα αντικείμενα ανήκουν στην ίδια κλάση)
2. Η μέγιστη τιμή του $\text{GINI}(t)$ είναι $1 - 1/c$ (όταν τα αντικείμενα είναι ομοιόμορφα κατανεμημένα στις κλάσεις). Εξαρτάται από τον αριθμό των κλάσεων.

Βέλτιστη συνθήκη διαχωρισμού με τον δείκτη Gini

Όταν ένας κόμβος t διασπάται σε k κόμβους (παιδιά) (το οποίο σημαίνει ότι το σύνολο των αντικειμένων του κόμβου t διαμερίζεται σε k υποσύνολα) η μη καθαρότητα του διαχωρισμού ορίζεται ως σταθμισμένος μέσος της μη καθαρότητας των παιδιών:

$$\text{GINI}_{\text{split}} = \text{GINI}(\text{Children}) = \frac{1}{n}\sum_{i=1}^{k} n_i \text{GINI}(i).$$

όπου

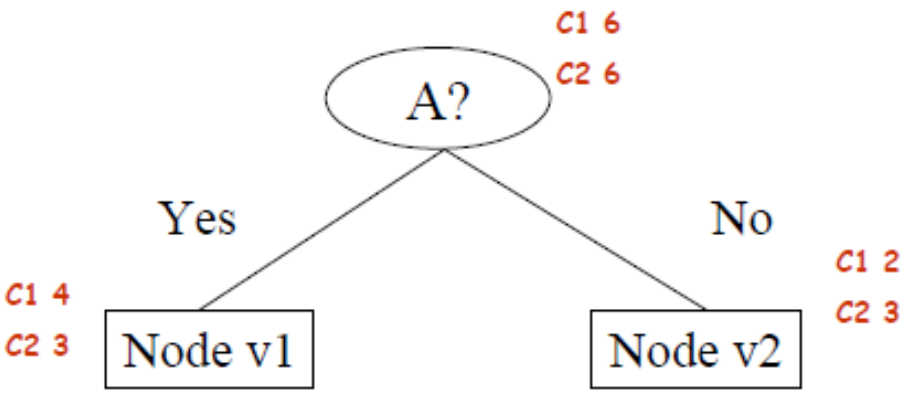
n_i ο αριθμός των αντικειμένων του παιδιού i.

n ο αριθμός των αντικειμένων του κόμβου t.

Βέλτιστη συνθήκη διαχωρισμού με τον δείκτη Gini

Παράδειγμα με δυαδικά γνωρίσματα

Έστω ότι έχουμε να διαλέξουμε μεταξύ δύο συνθηκών διαχωρισμού A και B για τα αντικείμενα ενός (αρχικού) κόμβου.



Αρχικός κόμβος

	Parent
C1	6
C2	6
Gini = 0.500	

	v1	v2
C1	4	2
C2	3	3
Gini=0.486		

Gini(v1)
= 1 – $(4/7)^2$ – $(3/7)^2$ = 0.49

Gini(v2)
= 1 – $(2/5)^2$ – $(3/5)^2$ = 0.48

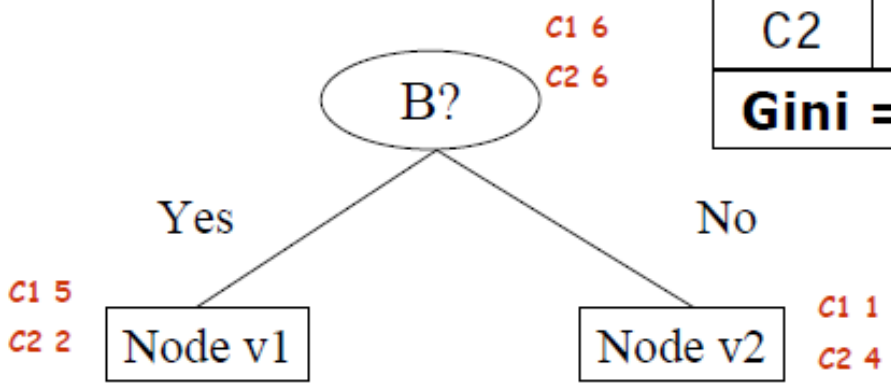
Gini(Children)
= 7/12 * 0.49 +
5/12 * 0.48
= 0.486

Κέρδος Δ = 0.500 – 0.486

Βέλτιστη συνθήκη διαχωρισμού με τον δείκτη Gini

Παράδειγμα με δυαδικά γνωρίσματα

Μεταξύ δύο συνθηκών διαχωρισμού *A* και *B*



	Parent
C1	6
C2	6
Gini = 0.500	

Υπενθύμιση: με βάση το *A*

	v1	v2
C1	4	2
C2	3	3
Gini=0.486		

	v1	v2
C1	5	1
C2	2	4
Gini=0.371		

Gini(v1)
= 1 – (5/7)² – (2/7)²
= 0.408

Gini(v2)
= 1 – (1/5)² – (4/5)²
= 0.32

Gini(Children)
= 7/12 * 0.408 +
5/12 * 0.32
= 0.371

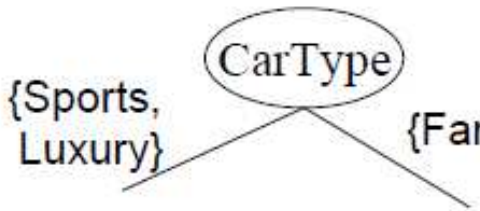
Κέρδος Δ = 0.500 – 0.371

Άρα διαλέγουμε το B

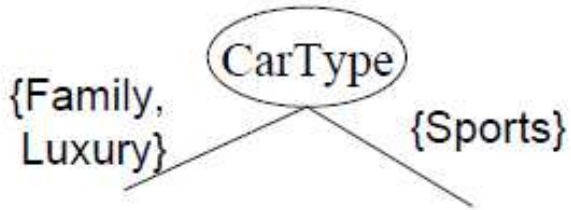
Βέλτιστη συνθήκη διαχωρισμού με τον δείκτη Gini

Παράδειγμα με κατηγορικά γνωρίσματα

Όταν έχουμε κατηγορικά γνωρίσματα ακολουθούμε την εξής μέθοδο:
Έστω ότι υπάρχουν k διαφορετικές τιμές για ένα κατηγορικό γνώρισμα. Δημιουργούμε ένα πίνακα όπου καταγράφουμε πόσα αντικείμενα κάθε κλάσης έχουν κάθε μια από τις k διαφορετικές τιμές.
Για παράδειγμα για το γνώρισμα CarType που λαμβάνει 3 διαφορετικές τιμές Sports, Family, Luxury δύο πιθανά σενάρια δυαδικής διάσπασης θα μπορούσαν να είναι:



	CarType	
	{Sports, Luxury}	{Family}
C1	9	1
C2	7	3
Gini	0.468	



	CarType	
	{Sports}	{Family, Luxury}
C1	8	2
C2	0	10
Gini	0.167	

Βέλτιστη συνθήκη διαχωρισμού με τον δείκτη Gini

Παράδειγμα με κατηγορικά γνωρίσματα

Όταν έχουμε **κατηγορικά** γνωρίσματα ακολουθούμε την εξής μέθοδο:
Έστω ότι υπάρχουν k διαφορετικές τιμές για ένα κατηγορικό γνώρισμα. Δημιουργούμε ένα πίνακα όπου καταγράφουμε πόσα αντικείμενα κάθε κλάσης έχουν κάθε μια από τις k διαφορετικές τιμές.
Για παράδειγμα για το γνώρισμα CarType, που λαμβάνει 3 διαφορετικές τιμές Sports, Family, Luxury, ένα πιθανό σενάριο πολλαπλής διάσπασης θα μπορούσε να είναι:

	CarType		
	Family	Sports	Luxury
C1	1	8	1
C2	3	0	7
Gini	0.163		

Βέλτιστη συνθήκη διαχωρισμού με τον δείκτη Gini

Παράδειγμα με κατηγορικά γνωρίσματα

κλάση

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Αρχικό Gini για αυτό τον κόμβο είναι: $\text{Gini} = 1 - (\frac{9}{14})^2 - (\frac{5}{14})^2 = 0.459$.
Μια επιλογή είναι να το διασπάσουμε με βάση το γνώρισμα income.
Πρέπει να θεωρήσουμε όλες τις δυνατές διασπάσεις. Για το income πρέπει να διαλέξουμε δυαδικές ή μή.
Ομοίως, εξετάζουμε και άλλες πιθανές διασπάσεις με άλλα γνωρίσματα (δηλαδή age, student, credit_rating).

Βέλτιστη συνθήκη διαχωρισμού με τον δείκτη Gini

Παράδειγμα με συνεχή γνωρίσματα

Έστω ένα **συνεχές** γνώρισμα A (π.χ. εισόδημα).
Μπορούμε να χρησιμοποιήσουμε δυαδικό διαχωρισμό πάνω σε μια τιμή του A.
Κάθε τιμή διαχωρισμού v συσχετίζεται με ένα πίνακα μετρήσεων:
Μετράμε πόσα αντικείμενα που έχουν $A < v$ και $A \geq v$ ανήκουν σε κάθε μια από τις δυνατές κλάσεις.
Για να βρούμε την βέλτιστη επιλογή για το v συνήθως

- ταξινομούμε (sort) τα αντικείμενα με βάση τις τιμές του γνωρίσματος (πολυπλοκότητα $O(N \log N)$ για N διαφορετικές τιμές του γνωρίσματος στο σύνολο εκπαίδευσης)
- σαρώνουμε σειριακά τις τιμές, ενημερώνοντας κάθε φορά τους μετρητές και υπολογίζοντας τον δείκτη Gini.
- επιλέγουμε τον (δυαδικό) διαχωρισμό με το μικρότερο δείκτη Gini.

Βέλτιστη συνθήκη διαχωρισμού με τον δείκτη Gini

Παράδειγμα με συνεχή γνωρίσματα

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Διαχωρισμός με βάση το γνώρισμα income:

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
Taxable Income										
Ταξινόμηση Τιμών →	60	70	75	85	90	95	100	120	125	220
Τιμές διαχωρισμού →	55	65	72	80	87	92	97	110	122	172	230
	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >
Yes	0 3	0 3	0 3	0 3	1 2	2 1	3 0	3 0	3 0	3 0	3 0
No	0 7	1 6	2 5	3 4	3 4	3 4	3 4	4 3	5 2	6 1	7 0
Gini	0.420	0.400	0.375	0.343	0.417	0.400	*0.300*	0.343	0.375	0.400	0.420

Βέλτιστη συνθήκη διαχωρισμού με τον δείκτη Gini

Παράδειγμα με συνεχή γνωρίσματα

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Για <55, δεν υπάρχει εγγραφή οπότε 0

Για <65, κοιτάμε το μικρότερο το 60, NO 0->1, 7->6 YES δεν αλλάζει

Για <72, κοιτάμε το μικρότερο το 70, NO 1->2 6->5, YES δεν αλλάζει

κοκ

Καλύτερα; Αγνοούμε τα σημεία στα οποία δεν υπάρχει αλλαγή κλάσης (αυτά δε μπορεί να είναι σημεία διαχωρισμού)

Άρα, στο παράδειγμα, αγνοούνται τα σημεία 55, 65, 72, 87, 92, 122, 172, 230

Από 11 πιθανά σημεία διαχωρισμού μας μένουν μόνο 2

Cheat	No		No		No		Yes		Yes		Yes		No		No		No		No		No	
	Taxable Income																					
Ταξινομημένες Τιμές →	60		70		75		85		90		95		100		120		125		220			
Τιμές Διαχωρισμού →	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
No	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420		0.400		0.375		0.343		0.417		0.400		0.300		0.343		0.375		0.400		0.420	

Εντροπία

Έστω c **ο συνολικός αριθμός των κλασεων** και $P(j|t)$ **το ποσοστό των αντικείμενων του κόμβου** t **που ανήκουν στην κλάση** j.

$$I(t) = \text{Info}(t) = \text{Entropy}(t) = -\sum_{j=1}^{c} P(j|t) \log_2 P(j|t)$$

Παραδείγματα

v1	
C1	0
C2	6
Entropy=0.000	

Gini = 0.000

v2	
C1	1
C2	5
Entropy=0.650	

Gini = 0.278

v3	
C1	2
C2	4
Entropy = 0.92	

Gini = 0.444

v4	
C1	3
C2	3
Entropy = 1.000	

Gini = 0.500

1. Η ελάχιστη τιμή του Entropy(t) είναι 0 (όταν όλα τα αντικείμενα ανήκουν στην ίδια κλάση - το οποίο δηλώνει την πιο ενδιαφέρουσα πληροφορία).
2. Η μέγιστη τιμή του Entropy(t) είναι $\log_2(c)$ (όταν τα αντικείμενα είναι ομοιόμορφα κατανεμημένα στις κλάσεις - το οποίο δηλώνει την λιγότερο ενδιαφέρουσα πληροφορία).

Βέλτιστη συνθήκη διαχωρισμού με βάση την εντροπία

Όταν ένας κόμβος t διασπάται σε k κόμβους (παιδιά) (το οποίο σημαίνει ότι το σύνολο των αντικειμένων του κόμβου t διαμερίζεται σε k υποσύνολα) η μη καθαρότητα του διαχωρισμού ορίζεται ως εξής:

$$\text{Info}(t)_{\text{split}} = \text{Entropy}_{\text{split}} = \text{Entropy}(\text{Children}) = \frac{1}{n}\sum_{i=1}^{k} n_i \text{Entropy}(i).$$

όπου
n_i ο αριθμός των αντικειμένων του παιδιού i.
n ο αριθμός των αντικειμένων του κόμβου t.

Και σ᾽ αυτήν την περίπτωση το **κέρδος (gain)** ενός διαχωρισμού ορίζεται

$$\Delta = \text{Gain}_{\text{split}} = \text{Entropy}(t) - \text{Entropy}(\text{children})$$

Στην περίπτωση της εντροπίας η διαφορά αυτή ονομάζεται **κέρδος πληροφορίας (information gain)**.

Βέλτιστη συνθήκη διαχωρισμού με βάση την εντροπία

Παράδειγμα

Κλάση

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

age	y_i	n_i	$I(y_i, n_i)$
<=30	2	3	0.971
31...40	4	0	0
>40	3	2	0.971

Συμβολισμοί: y_i # yes, n_i # no

$$I(x,y) = -\frac{x}{x+y}\log_2\frac{x}{x+y} - \frac{y}{x+y}\log_2\frac{y}{x+y}$$

Αρχικά

$$\text{Entropy} = I(9,5) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940.$$

Για το γνώρισμα age (με βάση τον πίνακα)

$$\text{Entropy}_{\text{age}} = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694.$$

Βέλτιστη συνθήκη διαχωρισμού με βάση την εντροπία

Παράδειγμα

Κλάση

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

age	y_i	n_i	$I(y_i, n_i)$
<=30	2	3	0.971
31...40	4	0	0
>40	3	2	0.971

Συμβολισμοί: y_i # yes, n_i # no

$$I(x,y) = -\frac{x}{x+y}\log_2\frac{x}{x+y} - \frac{y}{x+y}\log_2\frac{y}{x+y}$$

Αρχικά:

$$\text{Entropy} = 0.940.$$

Για το γνώρισμα age:

$$\text{Entropy}_{\text{age}} = 0.694.$$

Άρα,

$$\text{Gain(age)} = 0.940 - 0.694 = 0.246$$

Βέλτιστη συνθήκη διαχωρισμού με βάση την εντροπία

Παράδειγμα

Κλάση

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

age	y_i	n_i	$I(y_i, n_i)$
<=30	2	3	0.971
31...40	4	0	0
>40	3	2	0.971

Συμβολισμοί: y_i # yes, n_i # no

$$I(x,y) = -\frac{x}{x+y}\log_2\frac{x}{x+y} - \frac{y}{x+y}\log_2\frac{y}{x+y}$$

Αντίστοιχα έχουμε:

$$\text{Gain(age)} = 0.246$$

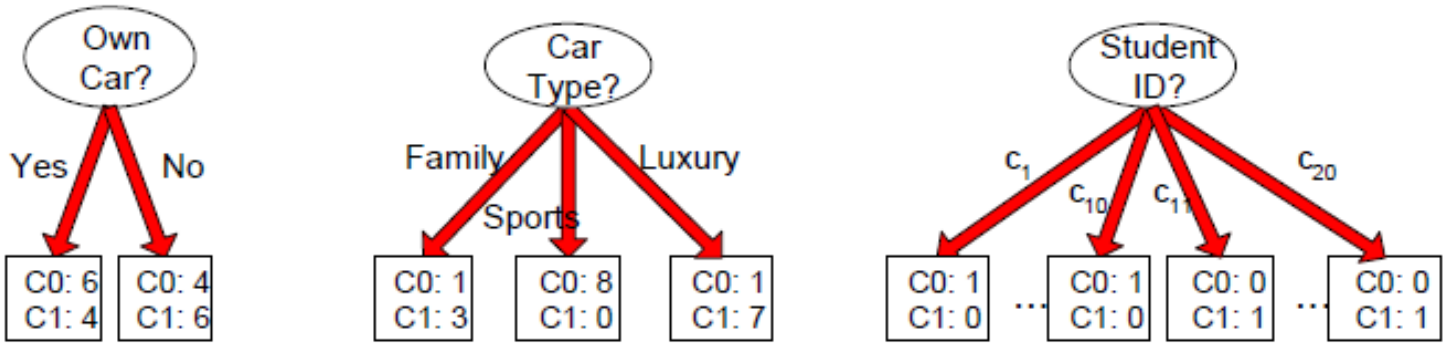
$$\text{Gain(income)} = 0.029, \qquad \text{Gain(student)} = 0.151$$

$$\text{Gain(credit\_rating)} = 0.048.$$

Άρα, επιλέγουμε να διαχωρίσουμε τα αντικείμενα με κριτήριο το age.

Βέλτιστη συνθήκη διαχωρισμού με βάση την εντροπία

Χρησιμοποιώντας συνθήκη διαχωρισμού με βάση την εντροπία μπορεί να καταλήξουμε σε κόμβους με πολύ λίγα αντικείμενα.



Στο παράδειγμα, το student_id είναι κλειδί και όχι χρήσιμο για προβλέψεις/κατηγοριοποίησεις.

Βέλτιστη συνθήκη διαχωρισμού με βάση την εντροπία

- Μια λύση είναι να έχουμε μόνο δυαδικές διασπάσεις.
- Μια άλλη λύση είναι να λάβουμε υπόψη μας και τον αριθμό των κόμβων. Συγκεκριμένα, ορίζουμε

$$\text{GainRATIO}_{\text{split}} = \frac{\text{Gain}_{\text{split}}}{\text{SplitInfo}}$$

όπου

$$\text{SplitInfo} = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

ο αριθμός SplitInfo ονομάζεται **εντροπία διάσπασης** και εξαρτάται από τον αριθμό των κόμβων της διάσπασης. Μεγάλος αριθμός μικρών διασπάσεων (υψηλή εντροπία) τιμωρείται. Η λύση αυτή χρησιμοποιείται στον αλγόριθμο C4.5.

Βέλτιστη συνθήκη διαχωρισμού με βάση την εντροπία

$$\text{SplitInfo} = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

Για παράδειγμα, αν έχουμε $6N$ αντικείμενα και τα χωρίσουμε σε 3 ομάδες από $2N$ αντικείμενα τότε

$$\text{SplitInfo} = \log_2 3$$

ενώ αν τα χωρίσουμε σε 2 ομάδες από $3N$ αντικείμενα τότε

$$\text{SplitInfo} = \log_2 2 = 1$$

δηλαδή οι 2 ομάδες ευνοούνται.

Αξιολόγηση συνθηκών διαχωρισμού

- **Κέρδος πληροφορίας** (Εντροπία): Δουλεύει καλύτερα σε γνωρίσματα με πολλαπλές τιμές.
- **Λόγος κέρδους**: Τείνει να ευνοεί διαχωρισμούς όπου μια διαμέριση είναι πολύ μικρότερη από μια άλλη.
- **Δείκτης Gini**: Δουλεύει καλύτερα σε γνωρίσματα με πολλαπλές τιμές. Δεν δουλεύει τόσο καλά όταν ο αριθμός των κλάσεων είναι μεγάλος. Τείνει να ευνοεί ελέγχους που οδηγούν σε ισομεγέθεις διαμερίσεις που και οι δύο είναι ομοιογενείς.

Και τα τρια παραπάνω μέτρα επιστρέφουν καλά αποτελέσματα.

Λάθος ταξινόμησης

Έστω c **ο συνολικός αριθμός των κλασεων** και $P(j|t)$ **ο αριθμός των αντικείμενων του κόμβου** t **που ανήκουν στην κλάση** j.

$$\text{Error}(t) = 1 - \max_j P(j|t)$$

Παραδείγματα

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Error = 1 – max (0, 1) = 1 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Error = 1 – max (1/6, 5/6) = 1 – 5/6 = 1/6

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Error = 1 – max (2/6, 4/6) = 1 – 4/6 = 1/3

Λάθος ταξινόμησης

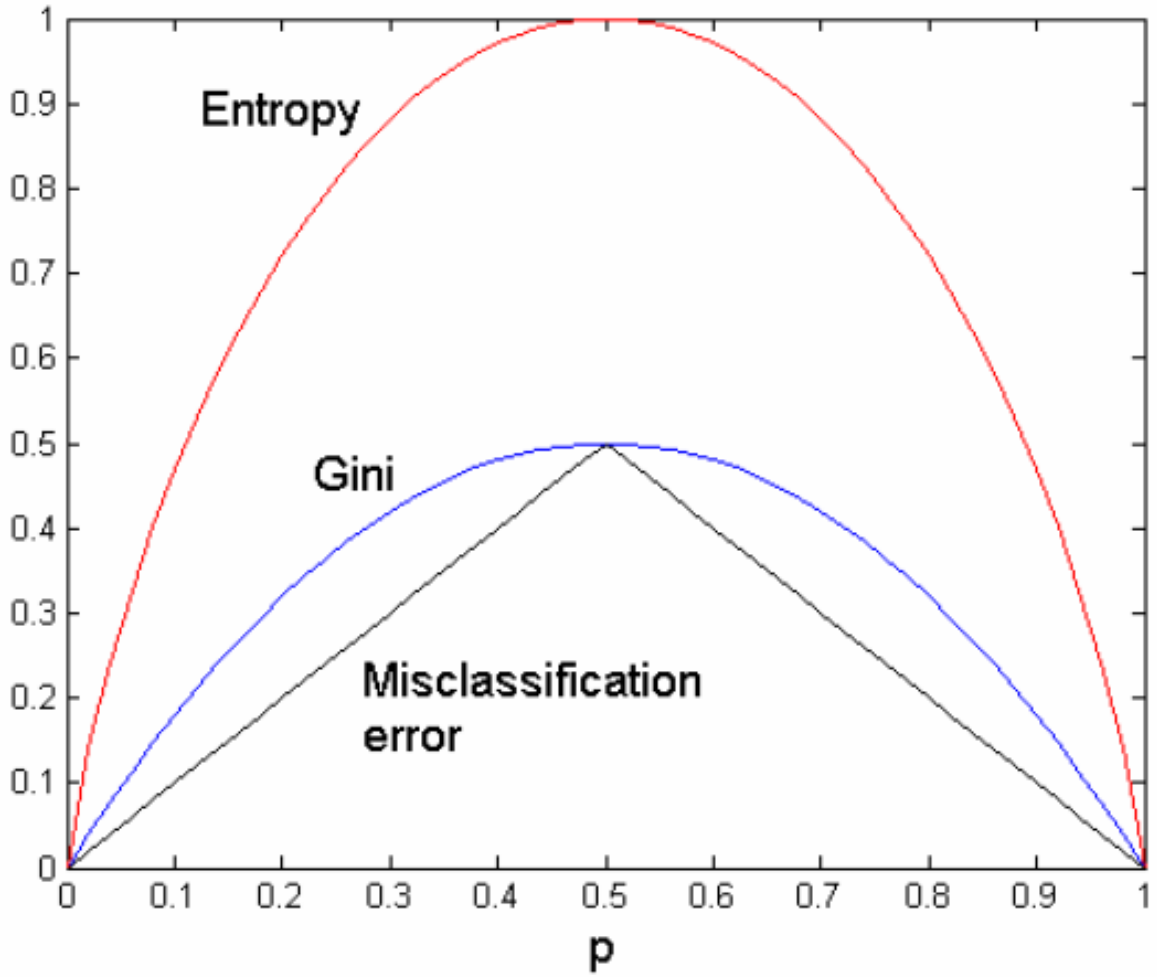
Έστω c **ο συνολικός αριθμός των κλασεων** και $P(j|t)$ **το ποσοστό των αντικείμενων του κόμβου** t **που ανήκουν στην κλάση** j.

$$\text{Error}(t) = 1 - \max_j P(j|t)$$

1. Η ελάχιστη τιμή του Error(t) είναι 0 (όταν όλα τα αντικείμενα ανήκουν στην ίδια κλάση).
2. Η μέγιστη τιμή του Error(t) είναι $\frac{c-1}{c}$ (όταν τα αντικείμενα είναι ομοιόμορφα κατανεμημένα στις κλάσεις).

Μια σύγκριση των συνθηκών διαχωρισμού

Για ένα πρόβλημα δύο κλάσεων



p ποσοστό εγγραφών που ανήκει σε μία από τις δύο κλάσεις

(p κλάση +, 1-p κλάση -)

Όλες την μεγαλύτερη τιμή για 0.5 (ομοιόμορφη κατανομή)

Όλες μικρότερη τιμή όταν όλες οι εγγραφές σε μία μόνο κλάση (0 και στο 1)

Μικρότερες τιμές μας δίνει το λάθος ταξινόμησης.

Μια σύγκριση των συνθηκών διαχωρισμού

Σε κατηγοριοποίηση δύο κλάσεων και οι τρεις συνθήκες διαχωρισμού είναι συνεπείς μεταξύ τους.

v1

C1	**0**
C2	**6**
Error=0.000	

Gini = 0.000

Entropy = 0.000

v2

C1	**1**
C2	**5**
Error=0.167	

Gini = 0.278

Entropy = 0.650

v3

C1	**2**
C2	**4**
Error = 0.333	

Gini = 0.444

Entropy = 0.920

v4

C1	**3**
C2	**3**
Error = 0.500	

Gini = 0.500

Entropy = 1.000

Κατηγοριοποίηση (classification)

Κατασκευή δένδρων απόφασης

Algorithm GenDecTree(Sample S, Attlist A)

1. create a node N
2. If all samples are of the same class C then label N with C; terminate;
3. If A is empty then label N with the most common class C in S (majority voting); terminate;
4. Select a $\in$ A, with the highest information gain (gini, error); Label N with a;
5. For each value v of a:
   a. Grow a branch from N with condition a=v;
   b. Let S_v be the subset of samples in S with a=v;
   c. If S_v is empty then attach a leaf labeled with the most common class in S;
   d. Else attach the node generated by GenDecTree(S_v, A-a)

Ερώτημα (κριτήρια τερματισμού)

Πότε θα σταματήσει ο διαχωρισμός ενός κόμβου;

- Σταματάμε όταν όλα τα αντικείμενα ανήκουν στην ίδια κλάση.
- Σταματάμε όταν όλα τα αντικείμενα έχουν τα ίδια γνωρίσματα.
- Σταματάμε με βάση τον αριθμό των αντικειμένων ή με το βάση το κέρδος. (**γρήγορος τερματισμός**)

Γιατί να μας ενδιαφέρει ο γρήγορος τερματισμός;

- Ο αριθμός των αντικειμένουν μειώνεται καθώς κατεβαίνουμε στο δένδρο.
- Αν ο αριθμός των αντικειμένων στα φύλλα είναι πολύ μικρός τότε δεν μπορούμε να πάρουμε οποιαδήποτε στατιστικά σημαντική απόφαση.
- Για να αποτρέψουμε το φαινόμενο αυτό (**διάσπαση δεδομένων** - **data framentation**) σταματάμε όταν ο αριθμός των αντικειμένων πέσει κάτω από κάποιο όριο.

Κατηγοριοποίηση (classification)

Πλεονεκτήματα δένδρων απόφασης

- Δεν στηρίζονται σε εκ των προτέρων γνωστές υποθέσεις για τον τύπος της κατανομής πιθανότητας που ακολουθούν τα αντικείμενα (**μη παραμετρική προσέγγιση**)
- Παρόλο που το πρόβλημα κατασκευής του βέλτιστου δένδρου απόφασης είναι NP-complete, οι διαθέσιμοι **ευρετικοί αλγόριθμοι** λειτουργούν πολύ καλά ακόμα και στην περίπτωση πολύ μεγάλου συνόλου δεδομένων.
- Είναι εύκολα στην **υλοποίηση**.
- Αφού το δένδρο κατασκευαστεί **η κατηγοριοποίηση των νέων αντικειμένων είναι πολύ γρήγορη**: $O(h)$ όπου h το ύψος του δένδρου.
- Είναι εύκολα στην **κατανόηση** (ιδιαίτερα τα μικρά δένδρα).
- Η **ακρίβειά** τους είναι συγκρίσιμη με άλλες τεχνικές για μικρά σύνολα δεδομένων.

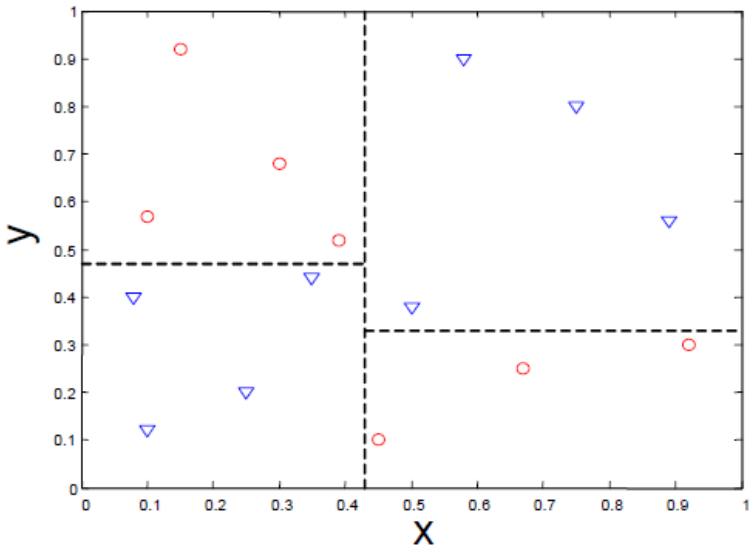
Κατηγοριοποίηση (classification)

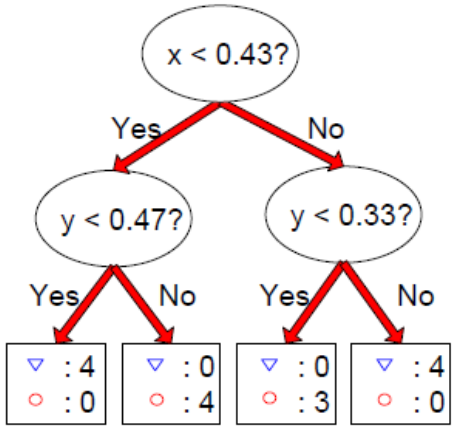
Πλεονεκτήματα δένδρων απόφασης (συνέχεια)

- Μπορούν να χειριστούν μεγάλο αριθμό γνωρισμάτων.
- Έχουν καλή συμπεριφορά στον θόρυβο.
- Η ύπαρξη πλεοναζόντων χαρακτηριστικών (γνώρισμα του οποίου η τιμή εξαρτάται από κάποιο άλλο) δεν είναι καταστροφική για την κατασκευή. Χρησιμοποιείται το ένα από τα δύο γνωρίσματα. Βέβαια, αν υπάρχουν πάρα πολλά τέτοια γνωρίσματα, μπορούν να οδηγήσουν σε δένδρα πιο μεγάλα από ότι χρειάζεται.
- Έχουν την δυνατότητα αναπαράστασης για γνωρίσματα διακριτών τιμών.
  Δεν έχουν πάντα καλή συμπεριφορά για συνεχείς μεταβλητές: Όταν η συνθήκη διαχωρισμού αφορά ένα συνεχές γνώρισμα δεν έχουμε καλά αποτελέσματα. Γι᾽ αυτό υπάρχουν παραλλαγές των αλγορίθμων που για τον διαχωρισμό χρησιμοποιούν **πολλά γνωρίσματα ταυτόχρονα** (βλέπε επόμενη διαφάνεια)

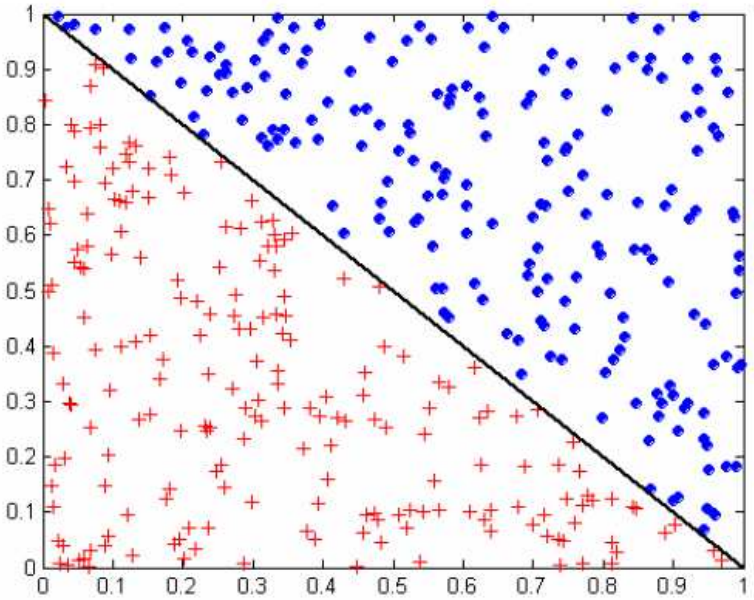
Κατηγοριοποίηση (classification)

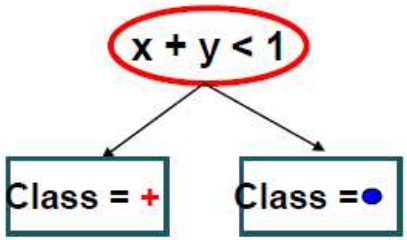
Πλεονεκτήματα δένδρων απόφασης (συνέχεια)





versus





Κατηγοριοποίηση (classification)

Μειονεκτήματα δένδρων απόφασης

- Δεν μπορούν να χειριστούν περίπλοκες σχέσεις μεταξύ των γνωρισμάτων
- Δημιουργούνται προβλήματα όταν λείπουν πολλά δεδομένα.

Κατηγοριοποίηση (classification)

Αξιολόγηση μοντέλων κατηγοριοποίησης

Η αξιολόγηση των μοντέλων κατηγοροποίησης (όπως π.χ. τα δένδρα απόφασης) γίνονται με βάση τον αριθμό των **σφαλμάτων (λαθών)** που κάνει το μοντέλο, δηλαδή όταν τοποθετεί ένα αντικείμενο σε λάθος κλάση (διαφορετική από αυτή που ανήκει πραγματικά).
Συνήθως μετράμε τις παρακάτω κατηγορίες λαθών:

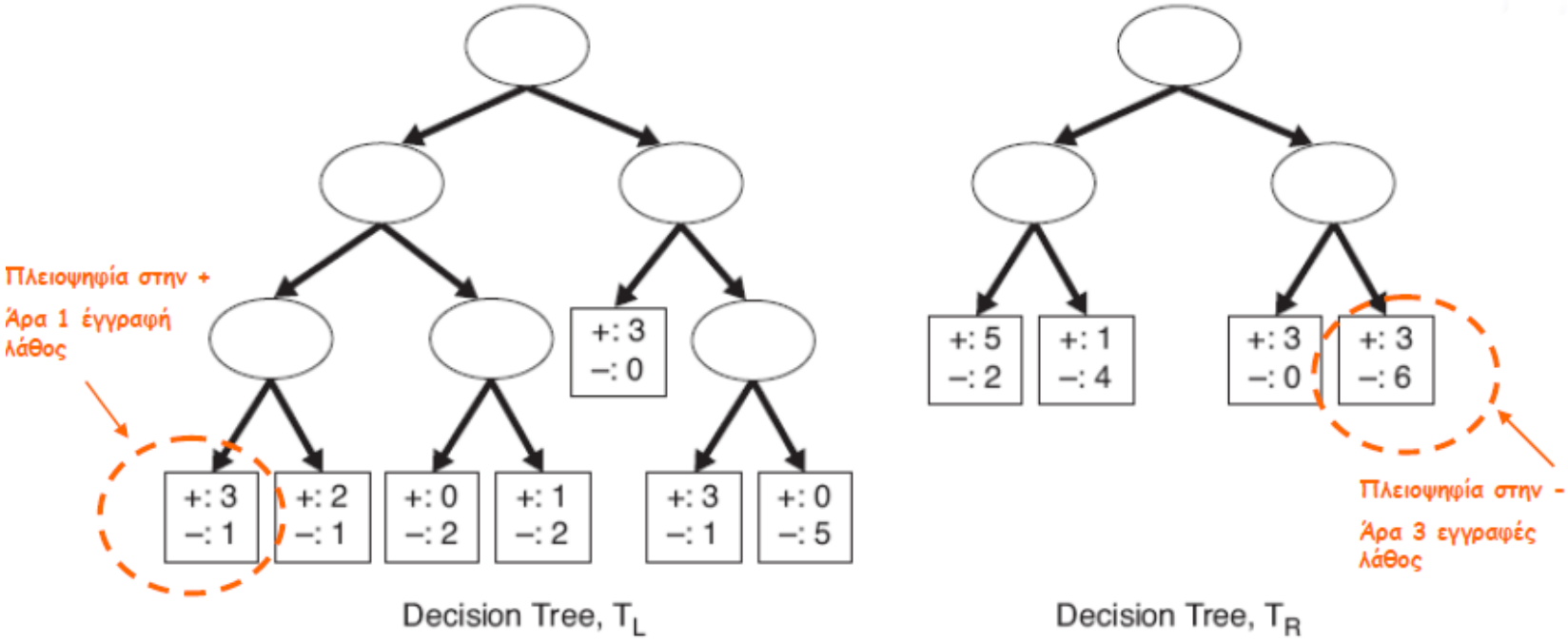
1. λάθη εκπαίδευσης: λάθη κατηγοριοποίησης του συνόλου εκπαίδευσης (ποσό αντικειμένων που κατηγοριοποιούνται σε λάθος κλάση - π.χ. λόγω γρήγορου τερματισμού).
2. λάθη ελέγχου: λάθη κατηγοριοποίησης του συνόλου ελέγχου.
3. λάθη γενίκευσης: τα αναμενόμενα λάθη κατηγοριοποίησης σε δεδομένα που δεν έχει δει το μοντέλο.

Κατηγοριοποίηση (classification)

Αξιολόγηση μοντέλων κατηγοριοποίησης

Λάθη και στα δεδομένα εκπαίδευσης, γιατί χρησιμοποιούμε την πλειοψηφία των εγγραφών σε ένα φύλλο για να αποδώσουμε κλάση

Εκτίμηση του Λάθους



Παράδειγμα δύο δέντρων για τα ίδια δεδομένα εκπαίδευσης

Με βάση το λάθος εκπαίδευσης

Αριστερό 4/24 = 0.167

Δεξί: 6/24 = 0.25

Κατηγοριοποίηση (classification)

Αξιολόγηση μοντέλων κατηγοριοποίησης

Η εκτίμηση του λάθους γενίκευσης γίνεται

- χρησιμοποιώντας τα δεδομένα εκπαίδευσης
- χρησιμοποιώντας τα δεδομένα ελέγχου

Κατηγοριοποίηση (classification)

Υπερπροσαρμογή δεδομένων

Προκειμένου να μειωθούν τα λάθη εκπαίδευσης μερικές φορές ένα μοντέλο ταιριάζει πολύ καλά με τα δεδομένα εκπαίδευσης (και λέμε ότι έχουμε **υπερπροσαρμογή** - overfitting) αλλά τότε ίσως να έχει μεγαλύτερο λάθος γενίκευσης.

Κατηγοριοποίηση (classification)

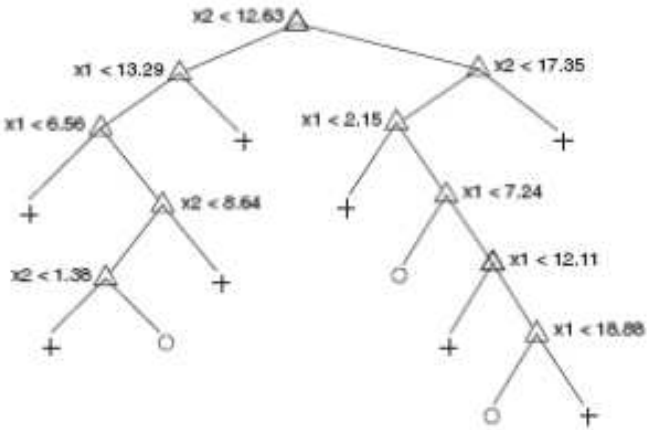
Υπερπροσαρμογή δεδομένων

Μπορούμε να διασπάμε το δέντρο μέχρι να φτάσουμε στο σημείο κάθε φύλλο να ταιριάζει απολύτως στα δεδομένα

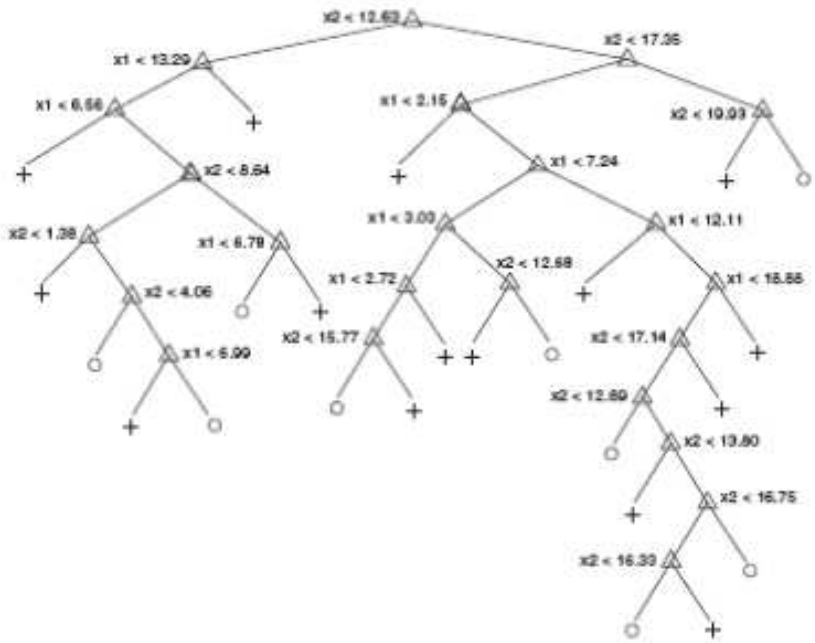
Overfitting

- Μικρό (μηδενικό) λάθος εκπαίδευσης
- Μεγάλο λάθος ελέγχου

Και το ανάποδο, μπορεί επίσης να ισχύει



(a) Decision tree with 11 leaf nodes.

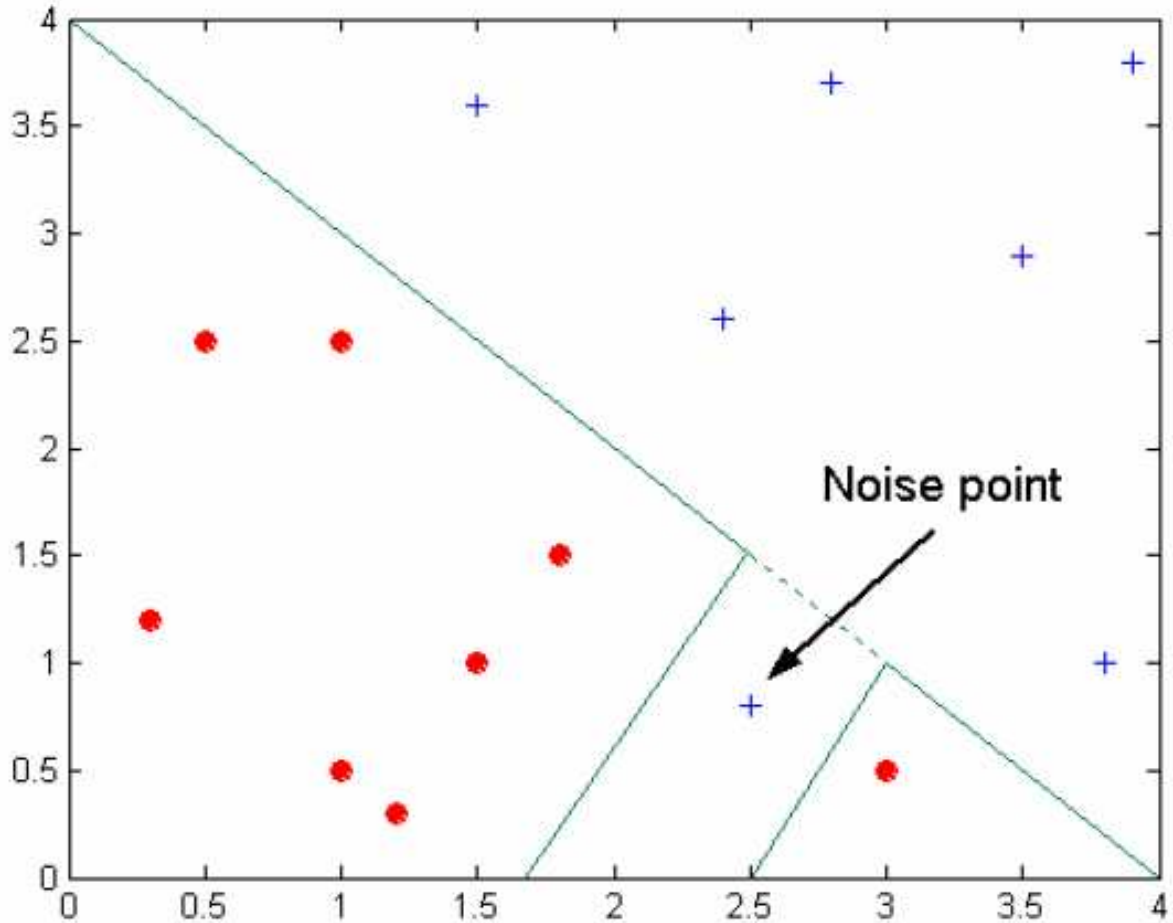


(b) Decision tree with 24 leaf nodes.

Κατηγοριοποίηση (classification)

Υπερπροσαρμογή δεδομένων

Overfitting εξαιτίας Θορύβου

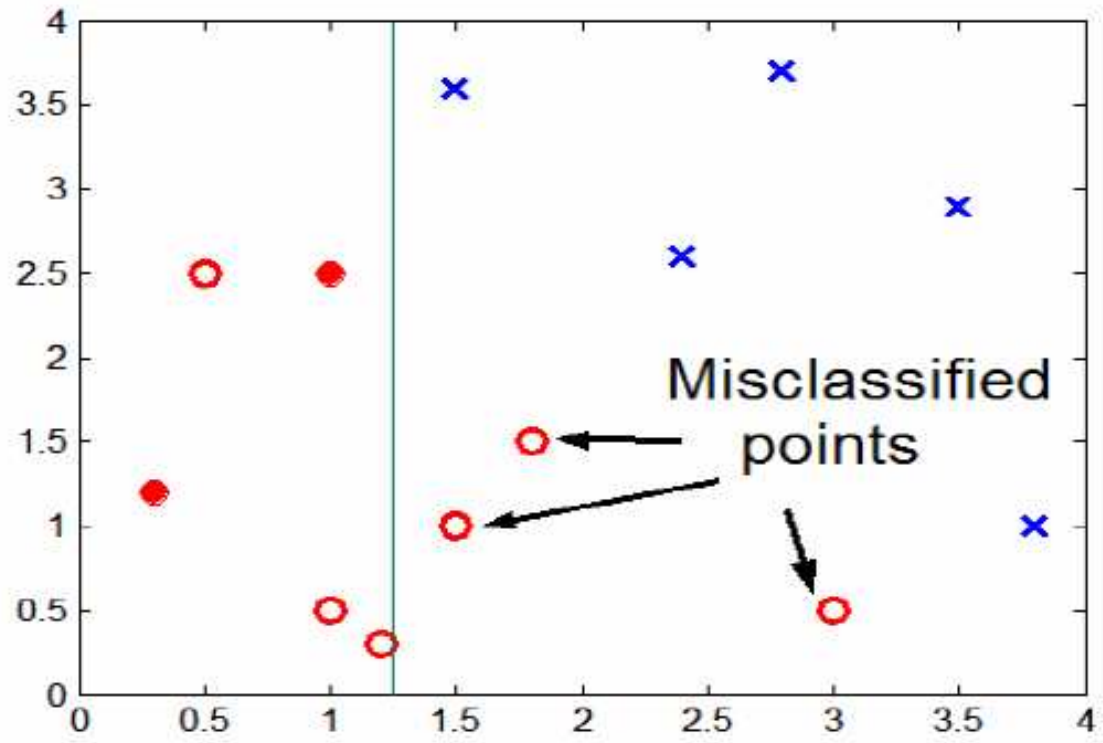


Το σημείο θορύβου επηρρεάζει τη συνθήκη διαχωρισμού.

Κατηγοριοποίηση (classification)

Υπερπροσαρμογή δεδομένων

Overfitting εξαιτίας μη Επαρκών Δειγμάτων



Κόκκινοι κύκλοι ανήκουν στην ίδια κλάση

Οι γεμάτοι είναι στο σύνολο εκπαίδευσης, οι άδειοι στο σύνολο ελέγχου

Η έλλειψη του συνόλου εκπαίδευσης σε κόκκινα σημεία στο κάτω μισό του διαγράμματος κάνει δύσκολη την πρόβλεψη των κλάσεων σ᾽ αυτή την περιοχή.

Κατηγοριοποίηση (classification)

Υπερπροσαρμογή δεδομένων

Μερικές φορές δημιουργείται πρόβλημα λόγω πολλαπλών επιλογών.
Επειδή σε κάθε βήμα εξετάζουμε πολλές διαφορετικές διασπάσεις, ενδέχεται κάποια διάσπαση να βελτιώνει το δένδρο **κατά τύχη**
Το πρόβλημα χειροτερεύει όταν αυξάνεται ο αριθμός των επιλογών και μειώνεται ο αριθμός των σημείων/αντικειμένων εκπαίδευσης.

Κατηγοριοποίηση (classification)

Υποπροσαρμογή δεδομένων

Μερικές φορές μπορεί να έχουμε το αντίθετο αποτέλεσμα όταν το μοντέλο είναι πολύ απλό (και λέμε ότι έχουμε **υποπροσαρμογή** - underfitting) και τότε τα λάθη εκπαίδευσης αλλά και ελέγχου είναι μεγάλα.
Everything should be made as simple as possible, but not simpler.

Occam's Razor

- Δοθέντων δυο μοντέλων με παρόμοια λάθη γενίκευσης, πρέπει να προτιμάται το απλούστερο από το πιο περίπλοκο

- Ένα πολύπλοκο μοντέλο είναι πιο πιθανό να έχει ταιριαστεί (Fitted) τυχαία λόγω λαθών στα δεδομένα

- Για αυτό η πολυπλοκότητα του μοντέλου θα πρέπει να αποτελεί έναν από τους παράγοντες της αξιολόγησής του

Bayesian κατηγοριοποίηση

Κατηγοριοποίηση (classification)

Bayesian κατηγοριοποίηση

Η Bayesian κατηγοροποίηση είναι μια πιθανοτική προσέγγιση για την επίλυση προβλημάτων κατηγοριοποίησης.

Κεντρική ιδέα

- Για κάθε αντικείμενο και για κάθε κλάση υπολογίζεται (στην συνέχεια θα δούμε πως) η πιθανότητα το αντικείμενο να ανήκει στην κλάση αυτή.
- Το αντικείμενο τοποθετείται στην κλάση εκείνη για την οποία υπολογίσαμε την μεγαλύτερη πιθανότητα να ανήκει σ' αυτή.

Αν A, B είναι (μη κενά) ενδεχόμενα ενός δειγματικού χώρου Ω τότε οι δεσμευμένες πιθανότητες $P(A|B)$ και $P(B|A)$ ορίζονται αντίστοιχα:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ και } P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Από τους ορισμούς αυτούς προκύπτει **ο τύπος του Bayes**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayesian κατηγοριοποίηση

Δεσμευμένη πιθανότητα - Τύπος του Bayes

Ο τύπος του Bayes συνδυάζεται και με το **θεώρημα ολικής πιθανότητας**:

Αν η οικογένεια $(B_i)_{i\in[n]}$ αποτελεί διαμέριση του Ω, τότε για κάθε $A \subseteq \Omega$ ισχύει

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

οπότε προκύπτει ότι

Αν η οικογένεια $(B_i)_{i\in[n]}$ αποτελεί διαμέριση του Ω, τότε για κάθε $A \subseteq \Omega$ ισχύει

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}.$$

Παράδειγμα

Ένας χρήστης έχει παρατηρήσει ότι από τα email που λαμβάνει καθημερινά ότι το 60% είναι γραμμένα στα Αγγλικά και το υπόλοιπο 40% στα Ελληνικά. Επίσης, έχει παρατηρήσει ότι από τα email γραμμένα στα Αγγλικά το 90% είναι ανεπιθύμητα (spam) και από τα email γραμμένα στα Ελληνικά το 30% είναι ανεπιθύμητα.
Να βρεθεί η πιθανότητα ένα email που διαβάζει ο χρήστης να είναι spam.
Να βρεθεί η πιθανότητα ένα spam email που διαβάζει ο χρήστης να είναι γραμμένο στα Ελληνικά.

Λύση

Έστω A το ενδεχόμενο το email να είναι γραμμένο στα Αγγλικά.
Έστω E το ενδεχόμενο το email να είναι γραμμένο στα Ελληνικά.
Έστω S το ενδεχόμενο το email να είναι spam.
$P(A) = 0.6$, $P(E) = 0.4$
$P(S|A) = 0.9$, $P(S|E) = 0.3$.
Από τον τύπο της ολικής πιθανότητας έχουμε ότι

$$P(S) = P(S|A)P(A) + P(S|E)P(E) = 0.9 \cdot 0.6 + 0.3 \cdot 0.4 = 0.66.$$

Από τον τύπο του Bayes έχουμε ότι

$$P(E|S) = \frac{P(S|E)P(E)}{P(S)} = \frac{0.9 \cdot 0.4}{0.66} = 0.5454 = 54.5\%.$$

Οι ίδιοι τύποι ισχύουν και τυχαίες μεταβλητές X, Y αντί για ενδεχόμενα A, B

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

και

Τύπος του Bayes

$$P(X = x|Y = y) = \frac{P(Y = y|X = x)P(X = x)}{P(Y = y)}$$

Bayesian κατηγοριοποίηση

Μοντελοποίηση κατηγοριοποίησης ως προβλήματος πιθανοτήτων

Έστω μια οικογένεια αντικειμένων της οποίας κάθε αντικείμενο κωδικοποιείται από ορισμένα χαρακτηριστικά $x_1, x_2, \ldots, x_n, y$, όπου

- $x_1, x_2, \ldots, x_n$ είναι τα γνωρίσματα/ανεξάρτητες μεταβλητές/είσοδος
- y είναι η κατηγορία/εξαρτημένη μεταβλητή/έξοδος

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Bayesian κατηγοριοποίηση

Μοντελοποίηση κατηγοριοποίησης ως προβλήματος πιθανοτήτων

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Στο παράδειγμα, τα γνωρίσματα του αντικείμενου είναι τα πεδία age, income, student, credit rating, buys computer.

Εδώ θέλουμε να προσδιορίσουμε την τιμή του πεδίου buys computer (Yes ή No) με βάση τις τιμές των πεδίων age, income, student, credit rating.

Bayesian κατηγοριοποίηση

Μοντελοποίηση κατηγοριοποίησης ως προβλήματος πιθανοτήτων

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Παρατήρηση: Σε κάθε αντικείμενο οι τιμές των γνωρισμάτων του μπορούν να θεωρηθούν ως τυχαίες μεταβλητές (τ.μ.).

Για το πεδίο age θεωρούμε την τ.μ. X_1 με τιμές ≤ 30, 31..40, > 40.

Για το πεδίο income θεωρούμε την τ.μ. X_2 με τιμές high, medium, low.

Για το πεδίο student θεωρούμε την τ.μ. X_3 με τιμές yes, no.

Για το πεδίο credit rating θεωρούμε την τ.μ. X_4 με τιμές fair, excellent.

Για το πεδίο buys computer θεωρούμε την τ.μ. Y με τιμές Yes, No.

Bayesian κατηγοριοποίηση

Μοντελοποίηση κατηγοριοποίησης ως προβλήματος πιθανοτήτων

Στην Bayesian κατηγοριοποίηση, προκειμένου να κατηγοριοποιήσουμε ένα αντικείμενο για το οποίο έχουμε ότι $X_1 = x_1$, $X_2 = x_2$, $X_3 = x_3$, $X_4 = x_4$ όπου x_1, x_2, x_3, x_4 είναι συγκεκριμένες τιμές των τ.μ. X_1, X_2, X_3, X_4 (π.χ. $x_1 = 31..40$, $x_2 = \text{high}$, $x_3 = \text{no}$, $x_4 = \text{fair}$) υπολογίζουμε (με τη βοήθεια των δεδομένων εκπαίδευσης) τις δεσμευμένες πιθανότητες

$$P(Y = \text{Yes} | X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$

και

$$P(Y = \text{No} | X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$

Επιλέγουμε ως τιμή της τ.μ. Y (Yes, No) αυτή που έχει την μεγαλύτερη πιθανότητα.

Οι δεσμευμένες πιθανότητες

$$P(Y = \text{Yes}|X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$

και

$$P(Y = \text{No}|X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$

υπολογίζονται με την βοήθεια του τύπου του Bayes απ' όπου προκύπτει

$$\begin{aligned} &P(Y = \text{Yes}|X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) \\ &\quad = \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4|Y = \text{Yes}) \cdot P(Y = \text{Yes})}{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)} \end{aligned}$$

και

$$\begin{aligned} &P(Y = \text{No}|X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) \\ &\quad = \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4|Y = \text{No}) \cdot P(Y = \text{No})}{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)} \end{aligned}$$

$$P(Y = \text{Yes}|X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$
$$= \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4|Y = \text{Yes}) \cdot P(Y = \text{Yes})}{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)}$$
$$P(Y = \text{No}|X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$
$$= \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4|Y = \text{No}) \cdot P(Y = \text{No})}{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)}$$

- Η πιθανότητα $P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$ είναι σταθερή και μπορούμε να την αγνοήσουμε στην σύγκριση. Αρκεί να συγκρίνουμε τους αριθμητές.
- Οι πιθανότητες $P(Y = \text{Yes})$ και $P(Y = \text{No})$ υπολογίζονται εύκολα από τα δεδομένα εκπαίδευσης: Είναι το ποσοστό των αντικειμένων που έχουν τιμή Yes και No αντίστοιχα στο πεδίο defaulted borrower.

$$
\begin{aligned}
&P(Y = \text{Yes} | X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) \\
&\quad = \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \text{Yes}) \cdot P(Y = \text{Yes})}{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)} \\
&P(Y = \text{No} | X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) \\
&\quad = \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \text{No}) \cdot P(Y = \text{No})}{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)}
\end{aligned}
$$

Πως όμως θα υπολογίσουμε τις πιθανότητες

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \textbf{Yes})$$

και

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \textbf{No})?$$

Πως όμως θα υπολογίσουμε τις πιθανότητες

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \textbf{Yes})$$

και

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \textbf{No})?$$

Υπάρχουν **δύο βασικοί τρόποι** υπολογισμού:

- **ο απλοϊκός τρόπος** (naive Bayes)
- **δίκτυο πεποίθησης** (Bayes belief network (BBN))

Πως όμως θα υπολογίσουμε τις πιθανότητες

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \textbf{Yes})$$

και

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \textbf{No})?$$

Βασική παραδοχή:
Για να απλοποίησουμε τους υπολογισμούς θεωρούμε ότι οι τ.μ. X_1, X_2, X_3, X_4 είναι **υπο συνθήκη ανεξάρτητες δοθέντος του** Y δηλαδή αν γνωρίζουμε π.χ. την τιμή της τ.μ. X_1 με δεδομένο ότι γνωρίζουμε την κλάση Y δεν μπορούμε να πούμε τίποτα για την τιμή της τ.μ. X_2. και ισχύει η σχέση

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | Y = \text{yes}) =$$
$$P(X_1 = x_1 | Y = \text{yes}) \cdot P(X_2 = x_2 | Y = \text{yes}) \cdot P(X_3 = x_3 | Y = \text{yes}) \cdot P(X_4 = x_4 | Y = \text{yes})$$

Γενικότερα οι τ.μ. $X_1, X_2, \ldots, X_n$ ονομάζονται **υπο συνθήκη ανεξάρτητες δοθέντος του** Y αν ισχύει η σχέση:

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | Y = y) \\ &= P(X_1 = x_1 | Y = y) \cdot P(X_2 = x_2 | Y = y) \cdots P(X_n = x_n | Y = y) \end{aligned}$$

Προσοχή!
Η παραδοχή για την ανεξαρτησία των γνωρισμάτων σχεδόν ποτέ δεν ισχύει (!!!) αλλά η υπόθεση αυτή λειτουργεί στην πράξη διότι η κατηγοριοποίηση Bayes δεν απαιτεί ακριβείς εκτιμήσεις πιθανοτήτων αλλά αρκεί η μέγιστη πιθανότητα να αντιστοιχεί στην σωστή κλάση.

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πεδίο age θεωρούμε την τ.μ. X_1 με τιμές ≤ 30, 31..40, > 40.
Για το πεδίο income θεωρούμε την τ.μ. X_2 με τιμές high, medium, low.
Για το πεδίο student θεωρούμε την τ.μ. X_3 με τιμές yes, no.
Για το πεδίο credit rating θεωρούμε την τ.μ. X_4 με τιμές fair, excellent.
Για το πεδίο buys computer θεωρούμε την τ.μ. Y με τιμές Yes, No.

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Που θα κατηγοριοποίησουμε ένα άτομο με τα εξής χαρακτηριστικά;

age	income	student	credit rating	buys computer
> 40	high	no	excellent	?

Θα συγκρίνουμε τις πιθανότητες

$$P(Y = \text{Yes} | X_1 = \text{">40"}, X_2 = \text{high}, X_3 = \text{no}, X_4 = \text{excellent})$$

και

$$P(Y = \text{No} | X_1 = \text{">40"}, X_2 = \text{high}, X_3 = \text{no}, X_4 = \text{excellent})$$

Ισοδύναμα, από τον τύπο του Bayes αρκεί να συγκρίνουμε τα γινόμενα

$$P(X_1 = \text{">40"}, X_2 = \text{high}, X_3 = \text{no}, X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes})$$

και

$$P(X_1 = \text{">40"}, X_2 = \text{high}, X_3 = \text{no}, X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No})$$

Χρησιμοποιώντας την υπόθεση της υπό συνθήκη ανεξαρτησίας των X_1, X_2, X_3, X_4 δεδομένου του Y αρκεί να συγκρίνουμε τα γινόμενα

- $P(X_1 = \text{" } > 40\text{"} | Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes})$
- $P(X_1 = \text{" } > 40\text{"} | Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No})$

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πρώτο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) =$$

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πρώτο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot$$
$$P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) = \frac{3}{9}$$

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πρώτο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot$$
$$P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) = \frac{3}{9} \cdot \frac{2}{9}$$

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πρώτο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot$$
$$P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) = \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{3}{9}$$

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πρώτο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{Yes}) \cdot P(X_2 = \text{high} \,|\, Y = \text{Yes}) \cdot P(X_3 = \text{no} \,|\, Y = \text{Yes}) \cdot$$
$$P(X_4 = \text{excellent} \,|\, Y = \text{Yes}) \cdot P(Y = \text{Yes}) = \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9}$$

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πρώτο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) = \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{9}{14}$$

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το πρώτο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} | Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot$$
$$P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) = \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{9}{14} = \frac{1}{189} = 0.005$$

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το δεύτερο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot$$

$$P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No}) =$$

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το δεύτερο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot$$
$$P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No}) = \frac{2}{5}$$

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το δεύτερο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{No}) \cdot P(X_2 = \text{high} \,|\, Y = \text{No}) \cdot P(X_3 = \text{no} \,|\, Y = \text{No}) \cdot$$
$$P(X_4 = \text{excellent} \,|\, Y = \text{No}) \cdot P(Y = \text{No}) = \frac{2}{5} \cdot \frac{2}{5}$$

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το δεύτερο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{No}) \cdot P(X_2 = \text{high} \,|\, Y = \text{No}) \cdot P(X_3 = \text{no} \,|\, Y = \text{No}) \cdot$$
$$P(X_4 = \text{excellent} \,|\, Y = \text{No}) \cdot P(Y = \text{No}) = \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5}$$

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το δεύτερο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot$$
$$P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No}) = \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{3}{5}$$

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το δεύτερο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot$$
$$P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No}) = \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} \cdot \frac{5}{14}$$

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Για το δεύτερο γινόμενο έχουμε ότι

$$P(X_1 = \text{" } > 40\text{"} \,|\, Y = \text{No}) \cdot P(X_2 = \text{high} \,|\, Y = \text{No}) \cdot P(X_3 = \text{no} \,|\, Y = \text{No}) \cdot$$
$$P(X_4 = \text{excellent} \,|\, Y = \text{No}) \cdot P(Y = \text{No}) = \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} \cdot \frac{5}{14} = \frac{24}{875} = 0.027.$$

Συνοψίζοντας:

$$P(X_1 = \text{" } > 40\text{"} | Y = \text{Yes}) \cdot P(X_2 = \text{high} | Y = \text{Yes}) \cdot P(X_3 = \text{no} | Y = \text{Yes}) \cdot P(X_4 = \text{excellent} | Y = \text{Yes}) \cdot P(Y = \text{Yes}) = \frac{1}{189} = 0.005$$

$$P(X_1 = \text{" } > 40\text{"} | Y = \text{No}) \cdot P(X_2 = \text{high} | Y = \text{No}) \cdot P(X_3 = \text{no} | Y = \text{No}) \cdot P(X_4 = \text{excellent} | Y = \text{No}) \cdot P(Y = \text{No}) = \frac{24}{875} = 0.027.$$

Άρα, η πιθανότητα

$$P(Y = \text{Yes} | X_1 = \text{" } > 40\text{"}, X_2 = \text{high}, X_3 = \text{no}, X_4 = \text{excellent})$$

είναι μικρότερη από την πιθανότητα

$$P(Y = \text{No} | X_1 = \text{" } > 40\text{"}, X_2 = \text{high}, X_3 = \text{no}, X_4 = \text{excellent})$$

Επομένως, το άτομο με τα χαρακτηριστικά:

age	income	student	credit rating	buys computer
> 40	high	no	excellent	?

θα κατηγοριοποιηθεί με No στο γνώρισμα buys computer.

Σχόλιο

Οι πιθανότητες

$$P(Y = \text{Yes} | X_1 = \text{"} > 40\text{"}, X_2 = \text{high}, X_3 = \text{no}, X_4 = \text{excellent})$$

και

$$P(Y = \text{No} | X_1 = \text{"} > 40\text{"}, X_2 = \text{high}, X_3 = \text{no}, X_4 = \text{excellent})$$

αθροίζουν στο 1, οπότε μοιάζει περιττό το ότι υπολογίσαμε και τις δύο για να τις συγκρίνουμε.
Αυτό δεν είναι ακριβές, στην πραγματικότητα βρήκαμε μόνο τους αριθμητές των αντίστοιχων κλασμάτων (και όχι τον κοινό παρονομαστή). Με βάση τον αριθμητή δεν μπορούμε να γνωρίζουμε αν κάποια είναι μικρότερη ή ίση από $1/2$.

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Σε κάποιες περιπτώσεις μπορεί να συναντήσουμε το εξής πρόβλημα:
Που θα κατηγοριοποιήσουμε ένα άτομο με τα εξής χαρακτηριστικά;

age	income	student	credit rating	buys computer
31..40	high	no	excellent	?

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Όπως και πριν θα υπολογίσουμε δύο γινόμενα.

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Κατά τον υπολογισμό του γινόμενου
$P(X_1 = 31..40|Y = \text{No}) \cdot P(X_2 = \text{high}|Y = \text{No}) \cdot P(X_3 = \text{no}|Y = \text{No}) \cdot P(X_4 = \text{excellent}|Y = \text{No}) \cdot P(Y = \text{No})$
παρατηρούμε ότι σύμφωνα με τα δεδομένα, η πιθανότητα $P(X_1 = 31..40|Y = \text{No})$ είναι 0, οπότε **μηδενίζεται** ολόκληρο το γινόμενο αυτό!

Για περιπτώσεις αυτές όπου τα δεδομένα εκπαίδευσης δεν καλύπτουν όλες τις κατηγορίες εφαρμόζουμε **μια διαδικασία διόρθωσης** εισάγοντας στο δείγμα εικονικές εγγραφές ή ισοδύναμα τροποποιούμε τις δεσμευμένες πιθανότητες

$$P(X_i = x_i | Y = y)$$

ως εξής:

Αν αρχικά

$$P(X_i = x_i | Y = y) = \frac{k}{n}$$

όπου
k ο αριθμός των αντικειμένων με $Y = y$ για τα οποία $X_i = x_i$
n ο συνολικός αριθμός των αντικειμένων με $Y = y$.
Η διορθωμένη πιθανότητα (**σύμφωνα με τον Laplace**) ορίζεται ως

$$P(X_i = x_i | Y = y) = \frac{k + 1}{n + c}$$

όπου
c το πλήθος των διαφορετικών κλάσεων (το πλήθος των διαφορετικών τιμών της Y)
Το αποτέλεσμα τώρα είναι ότι οι πιθανότητες δεν είναι ποτέ μηδέν!

Αν αρχικά

$$P(X_i = x_i | Y = y) = \frac{k}{n}$$

όπου
k ο αριθμός των αντικειμένων με $Y = y$ για τα οποία $X_i = x_i$
n ο συνολικός αριθμός των αντικειμένων με $Y = y$.
Η διορθωμένη πιθανότητα (**σύμφωνα με την m-εκτίμηση**) ορίζεται ως

$$P(X_i = x_i | Y = y) = \frac{k + pm}{n + m}$$

όπου
m το πλήθος των εικονικών εγγραφών με $Y = y$ που εισάγαμε στο δείγμα εκπαίδευσης
p το ποσοστό των εικονικών εγγραφών με $Y = y$ και $X_i = x_i$.
Πρέπει $m > 1$, $p > 0$. Το αποτέλεσμα τώρα είναι ότι οι πιθανότητες δεν είναι ποτέ μηδέν!

Bayesian κατηγοριοποίηση

Naive Bayes

age	income	student	credit rating	buys computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31..40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31..40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
> 40	medium	no	excellent	no

Άσκηση

Που θα κατηγοριοποιήσουμε δύο άτομα με τα εξής χαρακτηριστικά;

age	income	student	credit rating	buys computer
$<= 30$	low	yes	fair	?
> 40	medium	no	fair	?

Άσκηση

Με βάση το σύνολο εκπαίδευσης

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

που θα κατηγοριοποιηθεί ένα ζώο με τα εξής χαρακτηριστικά

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

Bayesian κατηγοριοποίηση

Naive Bayes

Μια άλλη δυσκολία είναι ο χειρισμός των χαρακτηριστικών που λαμβάνουν **συνεχείς τιμές**. Για παράδειγμα, αν έχουμε ως δεδομένα εκπαίδευσης

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Που θα κατηγοριοποίησουμε ένα αντικείμενο με τα εξής χαρακτηριστικά;

Home owner	Marital status	Annual income	Defaulted borrower
No	Single	82K	?

Bayesian κατηγοριοποίηση

Naive Bayes

Πως θα υπολογίσουμε τις αντίστοιχες δεσμευμένες πιθανότητες

$$P(X_i = x_i | Y = y)$$

όταν η τ.μ. X_i λαμβάνει τιμές σε ένα συνεχές διάστημα;

Για παράδειγμα, πως θα εκτιμηθεί η πιθανότητα

$$P(\text{Annual Income} = 82\text{K} | \text{Defaulted Borrower} = \text{No})?$$

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Bayesian κατηγοριοποίηση

Naive Bayes

Πως θα υπολογίσουμε τις αντίστοιχες δεσμευμένες πιθανότητες

$$P(X_i = x_i | Y = y)$$

όταν η τ.μ. X_i λαμβάνει τιμές σε ένα συνεχές διάστημα;

Υπάρχουν δύο προσεγγίσεις για τον χειρισμό **συνεχών** γνωρισμάτων:

- Διακριτοποίηση των τιμών
- Χρήση κάποιας συνεχούς κατανομής

Χειρισμός **συνεχών** γνωρισμάτων:
Στην περίπτωση της **διακριτοποίησης**

- διαμερίζουμε το σύνολο τιμών της τ.μ. σε διαστήματα
- και ο υπολογισμός της πιθανότητας γίνεται με βάση το ποσοστό των αντικειμένων που έχουν γνώρισμα με τιμή στο αντίστοιχο διάστημα.

Εδώ πρέπει να έχουμε υπόψη ότι

- πολλά διαστήματα θα έχουν ως συνέπεια λίγα αντικείμενα σε κάθε διάστημα
- λίγα διαστήματα θα έχουν ως συνέπεια πολλά αντικείμενα σε κάθε διάστημα τα οποία πιθανόν να ανήκουν σε διαφορετικές κατηγορίες

Bayesian κατηγοριοποίηση

Naive Bayes

Χειρισμός **συνεχών** γνωρισμάτων:

- Υποθέτουμε κάποια συγκεκριμένη μορφή κατανομής πιθανοτήτων. Συνήθως κανονική (Gaussian) κατανομή.
- Μια κατανομή χαρακτηρίζεται από την συνάρτηση πυκνότητας πιθανότητας $f(x) = P(X = x)$ αυτής. Εδώ μας ενδιαφέρει η συνάρτηση πυκνότητας πιθανότητας (ς.π.π.)

$$f(x|y) = P(X = x|Y = y).$$

- Αν $Y = y$, για την κανονική κατανομή η σ.π.π. $f(x|y)$ έχει την μορφή

$$P(X = x|Y = y) = \frac{1}{\sqrt{2\pi\sigma^2_{x|y}}} \exp(-\frac{(x - \mu_{x|y})^2}{2\sigma^2_{x|y}})$$

όπου $\mu_{x|y}$ είναι η αναμενόμενη τιμή της τ.μ. X δεδομένου ότι $Y = y$ και $\sigma^2_{x|y}$ είναι η διακύμανση της τ.μ. X δεδομένου ότι $Y = y$.

Bayesian κατηγοριοποίηση

Naive Bayes

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$P(\text{Annual Income} = 82\text{K} | \text{Defaulted Borrower} = \text{No}) = ?$$

Οι τιμές της τ.μ. Annual Income όταν Defaulted Borrower = No έχουν μέσο όρο

$$\mu = \frac{125 + 100 + 70 + 120 + 60 + 220 + 75}{7} = 110.$$

Bayesian κατηγοριοποίηση

Naive Bayes

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$P(\text{Annual Income} = 82\text{K} | \text{Defaulted Borrower} = \text{No}) = ?$$

$\mu = 110$ και διακύμανση σ^2.

$$\sigma^2 = \frac{(125-110)^2 + (100-110)^2 + (70-110)^2 + (60-110)^2 + (220-110)^2 + (75-110)^2}{7}$$

$$= 2550$$

Σχόλιο: Αμερόληπτη (δια 6) v.s. μη αμερόληπτη εκτιμήτρια (δια 7).

Bayesian κατηγοριοποίηση

Naive Bayes

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Επομένως, μπορούμε να θεωρήσουμε ότι

$$P(\text{Annual Income} = x | \text{Defaulted Borrower} = \text{No})$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi \cdot 2550}} \exp\left(-\frac{(x-110)^2}{2 \cdot 2550}\right)$$

Bayesian κατηγοριοποίηση

Naive Bayes

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Οπότε

$$P(\text{Annual Income} = 82\text{K} | \text{Defaulted Borrower} = \text{No})$$
$$= \frac{1}{\sqrt{2\pi \cdot 2550}} \exp\left(-\frac{(82-110)^2}{2 \cdot 2550}\right)$$
$$= 0.0067$$

Bayesian κατηγοριοποίηση

Naive Bayes

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Αντίστοιχα, υπολογίζουμε την συνάρτηση πυκνότητας πιθανότητας

$$P(\text{Annual Income} = x | \text{Defaulted Borrower} = \text{Yes})$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi \cdot 16.67}} \exp\left(-\frac{(x-90)^2}{2 \cdot 16.67}\right)$$

όπου εδώ $\mu = 90$ και $\sigma^2 = \frac{50}{3} = 16.67$

Bayesian κατηγοριοποίηση

Naive Bayes

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Επομένως,

$$P(\text{Annual Income} = 82\text{K} | \text{Defaulted Borrower} = \text{Yes})$$
$$= \frac{1}{\sqrt{2\pi \cdot 16.67}} \exp\left(-\frac{(82-90)^2}{2 \cdot 16.67}\right) = 0.014.$$

Άσκηση

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Που θα κατηγοριοποίησουμε ένα άτομο με τα εξής χαρακτηριστικά;

Home owner	Marital status	Annual income	Defaulted borrower
No	Single	82K	?

$P(\text{Annual Income} = 82\text{K}|\text{Defaulted Borrower} = \text{No}) = 0.0067$

$P(\text{Annual Income} = 82\text{K}|\text{Defaulted Borrower} = \text{Yes}) = 0.014.$

Άσκηση

Υποθέτωντας ότι οι τιμές της παραμέτρου T Temperature ακολουθούν την κανονική κατανομή να βρεθούν οι συναρτήσεις πυκνότητας πιθανότητας

$$P(T = x|\text{Alarm} = 0) \text{ και } P(T = x|\text{Alarm} = 1)$$

χρησιμοποιώντας το επόμενο σύνολο εκπαίδευσης

Temperature	Pressure	Alarm 1
95	1105	0
85	1040	1
103	1090	1
97	1084	1
80	1038	0
100	1080	1
83	1025	1
86	1030	1
101	1100	1

Bayesian κατηγοριοποίηση

Naive Bayes

Η κατηγοριοποίηση με τη χρήση Naive Bayes έχει τα εξής χαρακτηριστικά:

- Ανθεκτικότητα στον θόρυβο
- Δεν επηρρεάζεται από τιμές που λείπουν στα δεδομένα. (Απλά τις αγνοούμε όταν υπολογίζουμε πιθανότητες).
- Εφαρμόζεται σε μεγάλο όγκο δεδομένων. (Αρκεί μια ανάγνωση των δεδομένων εκπαίδευσης)
- Έχει προβλήματα όταν υπάρχουν ισχυρές εξαρτήσεις μεταξύ των γνωρισμάτων. Σ' αυτές τις περιπτώσεις χρησιμοποιούμε άλλες τεχνικές όπως **Bayesian δίκτυα πεποίθησης** (Bayesian belief networks BBN).

Bayesian κατηγοριοποίηση

Bayesian belief networks

Τα λεγόμενα **Bayesian belief networks** βασίζονται στην αναπαράσταση των σχέσεων εξάρτησης μεταξύ των γνωρισμάτων χρησιμοποιώντας ένα προσανατολισμένο γράφημα.

Παράδειγμα

Έστω ότι μας ενδιαφέρει να δούμε να ένα άτομο έχει κάποιο καρδιακό νόσημα (heart disease).
Έχουμε μια συλλογή δεδομένων εκπαίδευσης που αφορούν n άτομα και για κάθε άτομο έχουμε τα παρακάτω γνωρίσματα:

- Exercise (Yes, No)
- Diet (Healthy, Unhealthy)
- Chest pain (Yes, No)
- Blood pressure (Yes, No)
- Heart disease (Yes, No)

Bayesian κατηγοριοποίηση

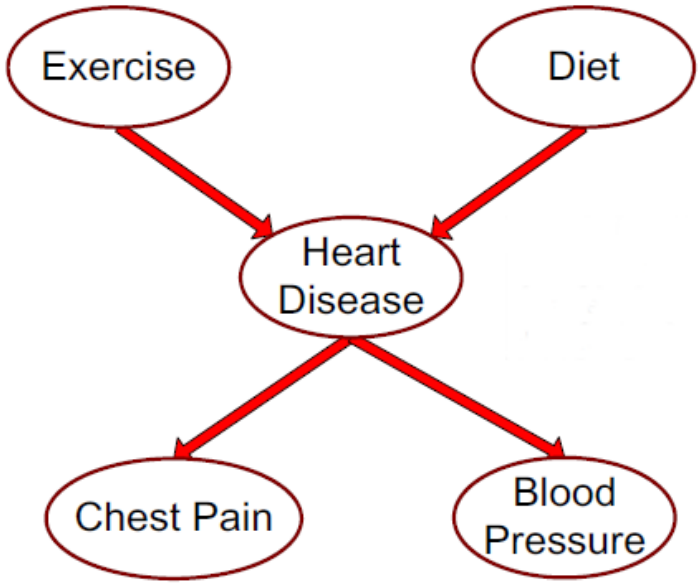
Bayesian belief networks

- Οι τιμές των γνωρίσματων Exercise και Diet επηρρεάζουν τις τιμές του γνωρίσματος Heart disease (π.χ. η ανθυγεινή διατροφή και η απουσία άσκησης μπορεί να δημιουργήσουν προβλήματα στην καρδιά). Υπάρχει μεταξύ τους μια **σχέση αιτίου – αποτελέσματος**.
- Από την άλλη, οι τιμές του γνωρίσματος Heart disease επηρρεάζουν τις τιμές των γνωρισμάτων Blood pressure και Heart disease (π.χ. προβλήματα στην καρδιά μπορεί να έχουν ως συνέπεια υψηλή πίεση και πόνο στο στήθος). Υπάρχει μεταξύ τους μια **σχέση αιτίου – αποτελέσματος**.

Bayesian κατηγοριοποίηση

Bayesian belief networks

Μπορούμε να απεικονίσουμε τις προηγούμενες εξαρτήσεις με ένα προσανατολισμένο γράφημα



- Τα γνωρίσματα Exercise και Diet έχουν ανεξάρτητες τιμές.
- Το γνώρισμα Heart Disease εξαρτάται από τα Exercise και Diet
- Τα γνωρίσματα Chest Pain και Blood Pressure εξαρτώνται άμεσα από τις τιμές του Heart Disease (π.χ. είναι συμπτώματα). Επίσης, είναι ανεξάρτητα μεταξύ τους.

Έστω ότι θέλουμε να κατηγοροποιήσουμε ένα άτομο στην κλάση Yes ή No του γνωρίσματος Heart Disease όταν γνωρίζουμε ότι έχει τα παρακάτω χαρακτηριστικά:

Exercise	Diet	Chest pain	Blood Pressure
Yes	Unhealthy	No	Yes

Θεωρούμε τις τ.μ. E (Exercise), D (Diet), C (Chest Pain), B (Blood Pressure) και H (Heart Disease)

Επίσης, θεωρούμε τις συντομογραφίες $Y = \text{Yes}, \text{Healthy}$ και $N = \text{No}, \text{Unhealthy}$.

Όπως και στον Naive Bayes, για να αποφασίσουμε, θα συγκρίνουμε τις πιθανότητες

- $P(H = Y | E = Y, D = N, C = N, B = Y)$
- $P(H = N | E = Y, D = N, C = N, B = Y)$

Bayesian κατηγοριοποίηση

Bayesian belief networks

Εδώ, λόγω των εξαρτήσεων δεν μπορούμε να χρησιμοποιήσουμε την παραδοχή της υπό συνθήκη ανεξαρτησίας. Γι' αυτό θα ακολουθήσουμε την εξής μέθοδο για την σύγκριση:
Από τον ορισμό της δεσμευμένης πιθανότητας $P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$ έχουμε αντίστοιχα να συγκρίνουμε τις πιθανότητες

$$\frac{P(E = Y, D = N, C = N, B = Y, H = Y)}{P(E = Y, D = N, C = N, B = Y)}$$

και

$$\frac{P(E = Y, D = N, C = N, B = Y, H = N)}{P(E = Y, D = N, C = N, B = Y)}$$

Επειδή τα κλάσματα έχουν τους ίδιους παρονομαστές αρκεί να συγκρίνουμε τους αριθμητές.

Bayesian κατηγοριοποίηση

Bayesian belief networks

Για να συγκρίνουμε τους αριθμητές

- $P(E = Y, D = N, C = N, B = Y, H = Y)$
- $P(E = Y, D = N, C = N, B = Y, H = N)$

μπορούμε να χρησιμοποιήσουμε τον τύπο

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n) \\ &= P(X_1 = x_1) \cdot P(X_2 = x_2 | X_1 = x_1) \cdot P(X_3 = x_3 | X_1 = x_1, X_2 = x_2) \cdot \\ &\quad \cdots P(X_n = x_n | X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}) \end{aligned}$$

(ο οποίος προκύπτει με διαδοχική εφαρμογή του τύπου του ορισμού της δεσμευμένης πιθανότητας

$$P(X = x, Y = y) = P(Y = y)P(X = x | Y = y).)$$

Παρατηρήστε ότι στον τύπο

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n) \\ &= P(X_1 = x_1) \cdot P(X_2 = x_2 | X_1 = x_1) \cdot P(X_3 = x_3 | X_1 = x_1, X_2 = x_2) \cdot \\ &\quad \cdots P(X_n = x_n | X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}) \end{aligned}$$

κάθε όρος του γινομένου είναι της μορφής

$$P(X_k = x_k | X_1 = x_1, X_2 = x_2, \ldots, X_{k-1} = x_{k-1})$$

όπου $X_1, X_2, \ldots, X_{k-1}$ είναι οι τ.μ. των όρων του γινομένου που προηγούνται (από τα αριστερά προς τα δεξιά).
Επομένως, όταν εφαρμόζουμε τον τύπο ορίζουμε έμμεσα και κάποια σειρά στις τ.μ. $X_1, X_2, \ldots, X_n$.

Έδω έχουμε να τις πιθανότητες

$$P(E = Y, D = N, C = N, B = Y, H = Y)$$

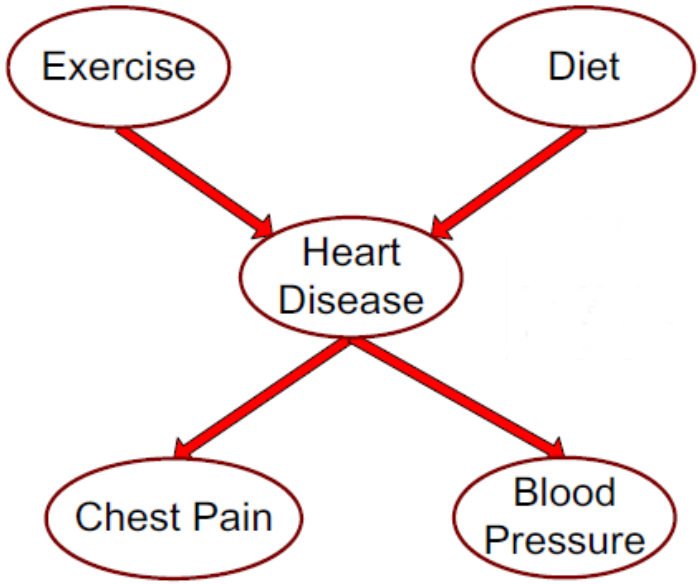
$$P(E = Y, D = N, C = N, B = Y, H = N)$$

Ποια σειρά μας συμφέρει να χρησιμοποιήσουμε για τις τ.μ. E, D, C, B, H;

Bayesian κατηγοριοποίηση

Bayesian belief networks

Θα βρούμε μια σειρά με την βοήθεια του γραφήματος εξαρτήσεων που κατασκευάσαμε:



Γνωρίζουμε ότι η τ.μ. H εξαρτάται από τις τ.μ. E, D, ενώ οι τ.μ. B και C εξαρτώνται από την τ.μ. H. Δεν υπάρχουν άλλες εξαρτήσεις.
Θα εφαρμόσουμε τον τύπο πρώτα για τις E, D που δεν έχουν άλλες εξαρτήσεις. Μετά για την H που εξαρτάται μόνο από τις E, D και μετά για τις B, C που εξαρτώνται μόνο από την H.

Κανόνας

Γενικά, με βάση το γράφημα των εξαρτήσεων, εφαρμόζουμε τον τύπο

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n) \\ &= P(X_1 = x_1) \cdot P(X_2 = x_2 | X_1 = x_1) \cdot P(X_3 = x_3 | X_1 = x_1, X_2 = x_2) \cdot \\ &\quad \cdots P(X_n = x_n | X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}) \end{aligned}$$

για μια τ.μ. Y όταν έχουμε ήδη εφαρμόσει τον τύπο για όλες τις τ.μ. $Y_1, Y_2, \ldots, Y_k$ από τις οποίες η Y έχει εξαρτήσεις.

Με βάση τη σειρά E, D, H, C, B έχουμε ότι

$$\begin{aligned} P(E = Y, D = N, C = N, B = Y, H = Y) = \\ P(E = Y) \\ \cdot P(D = N|E = Y) \\ \cdot P(H = Y|E = Y, D = N) \\ \cdot P(C = N|E = Y, D = N, H = Y) \\ \cdot P(B = Y|E = Y, D = N, H = Y, C = N) \end{aligned}$$

Όμως,

- $P(D = N|E = Y) = P(D = N)$ (D, E ανεξάρτητες)
- $P(C = N|E = Y, D = N, H = Y) = P(C = N|H = Y)$ (C ανεξάρτητη από τις E, D, Εξαρτάται μόνο από την H)
- $P(B = Y|E = Y, D = N, H = Y, C = N) = P(B = Y|H = Y)$ (B ανεξάρτητη από τις E, D, C. Εξαρτάται μόνο από την H)

Επομένως, με βάση τις προηγούμενες απλοποίησεις

$$\begin{aligned} &P(E = Y, D = N, C = N, B = Y, H = Y) \\ &\quad = P(E = Y) \cdot P(D = N) \cdot P(H = Y | E = Y, D = N) \\ &\qquad \cdot P(C = N | H = Y) \cdot P(B = Y | H = Y) \end{aligned}$$

Εντελώς, ανάλογα

$$\begin{aligned} &P(E = Y, D = N, C = N, B = Y, H = N) \\ &\quad = P(E = Y) \cdot P(D = N) \cdot P(H = N | E = Y, D = N) \\ &\qquad \cdot P(C = N | H = N) \cdot P(B = Y | H = N) \end{aligned}$$

Τελικά, λόγω των κοινών παραγόντων $P(E = Y) \cdot P(D = N)$, αρκεί να συγκρίνουμε τα γινόμενα

- $P(H = Y | E = Y, D = N) \cdot P(C = N | H = Y) \cdot P(B = Y | H = Y)$
- $P(H = N | E = Y, D = N) \cdot P(C = N | H = N) \cdot P(B = Y | H = N)$.

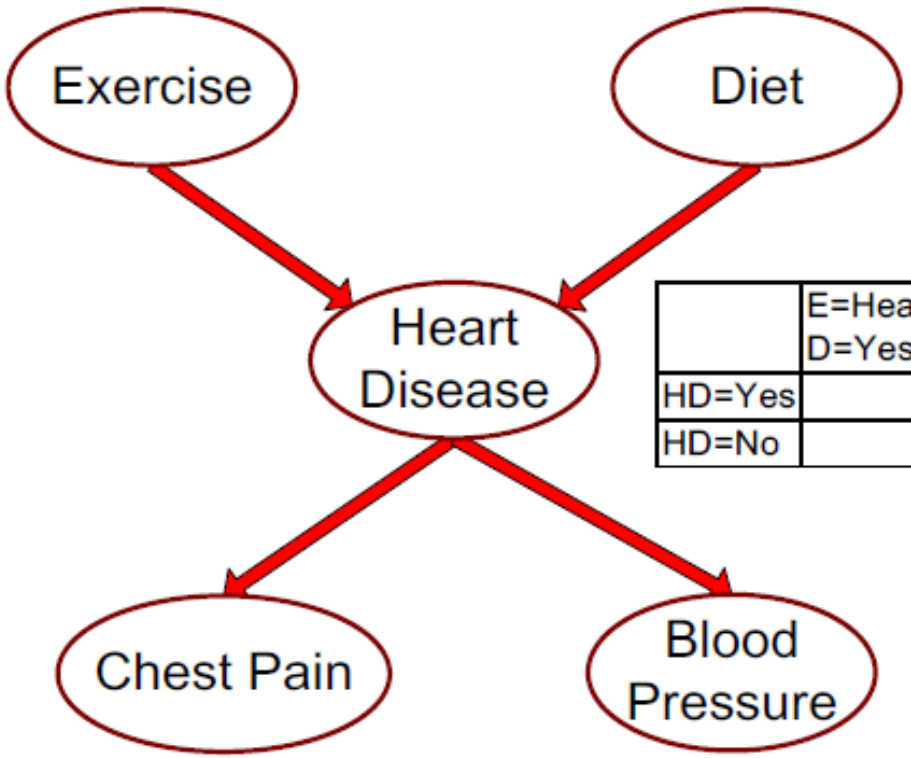
Bayesian κατηγοριοποίηση

Bayesian belief networks

Έστω ότι από τα δεδομένα υπολογίσαμε τις παρακάτω πιθανότητες:

Exercise=Yes	0.7
Exercise=No	0.3

Diet=Healthy	0.25
Diet=Unhealthy	0.75



	E=Healthy D=Yes	E=Healthy D=No	E=Unhealthy D=Yes	E=Unhealthy D=No
HD=Yes	0.25	0.45	0.55	0.75
HD=No	0.75	0.55	0.45	0.25

	HD=Yes	HD=No
CP=Yes	0.8	0.01
CP=No	0.2	0.99

	HD=Yes	HD=No
BP=High	0.85	0.2
BP=Low	0.15	0.8

Τότε, έχουμε ότι

- $P(H = Y|E = Y, D = N) \cdot P(C = N|H = Y) \cdot P(B = Y|H = Y) = 0.55 \cdot 0.2 \cdot 0.85 = 0.0935$
- $P(H = N|E = Y, D = N) \cdot P(C = N|H = N) \cdot P(B = Y|H = N) = 0.44 \cdot 0.01 \cdot 0.2 = 0.00088.$

Άρα, το άτομο με τα παρακάτω χαρακτηριστικά

Exercise	Diet	Chest pain	Blood Pressure
Yes	Unhealty	No	Yes

θα το κατηγοριοποίησουμε με Yes στο γνώρισμα Heart disease.

Πηγές διαφανειών

- P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar, Introduction to Data Mining, 2nd edition
- Μαθήματα εξόρυξης δεδομένων, Παν. Θεσσαλίας.