ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Αποθήκες Δεδομένων και Εξόρυξη Γνώσης από Δεδομένα

Βασική Κατηγοριοποίηση

Μανώλης Μαραγκουδάκης

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

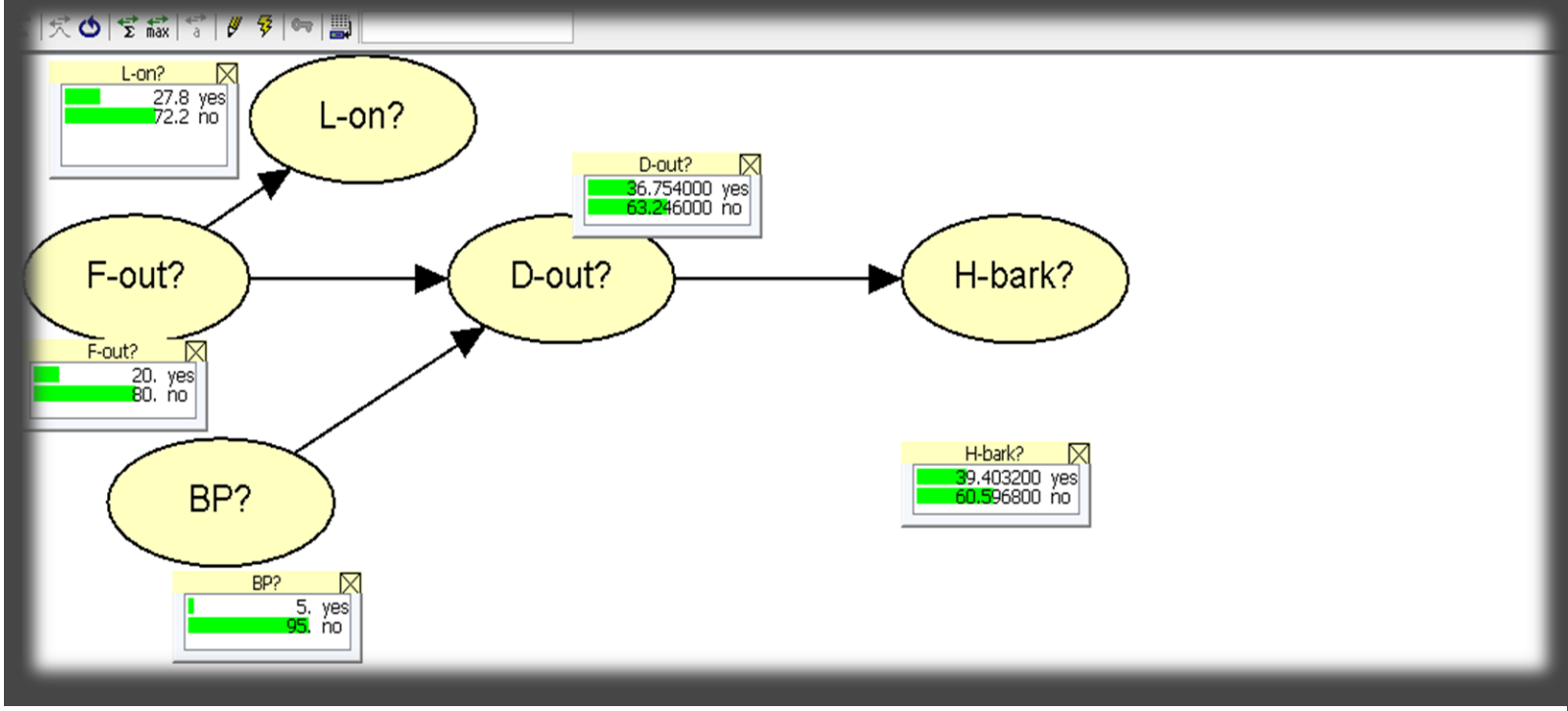
Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



Ενότητα 4: Βασική Κατηγοριοποίηση

Κατηγοριοποίηση: Ορισμός

- Δοσμένης μιας συλλογής εγγραφών (σώμα εκπαίδευσης-training set)
  - Κάθε εγγραφή περιέχει ένα σύνολο ιδιοτήτων-attributes, μιας εκ των οποίων είναι η κλάση-class
- Εύρεση ενός μοντέλου για την ιδιότητα της κλάσης ως συνάρτηση των τιμών των υπόλοιπων μεταβλητών
- Στόχος
  - Οι προηγουμένως αθέατες εγγραφές θα πρέπει να χαρακτηρισθούν με μια κλάση όσο ακριβέστερα γίνεται
- Ένα σύνολο αξιολόγησης-test set χρησιμοποιείται για να εξακριβωθεί η ακρίβεια του μοντέλου. Συνήθως, χωρίζουμε το σύνολο δεδομένων σε εκπαίδευσης και αξιολόγησης.

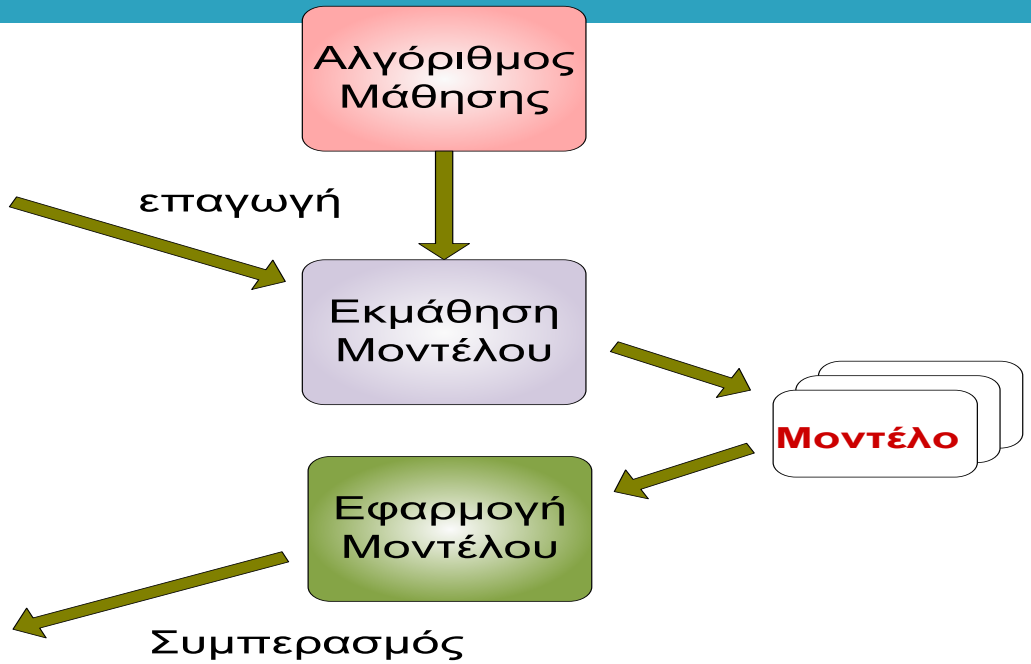
Γραφική απεικόνιση της κατηγοριοποίησης

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Σώμα Εκπαίδευσης

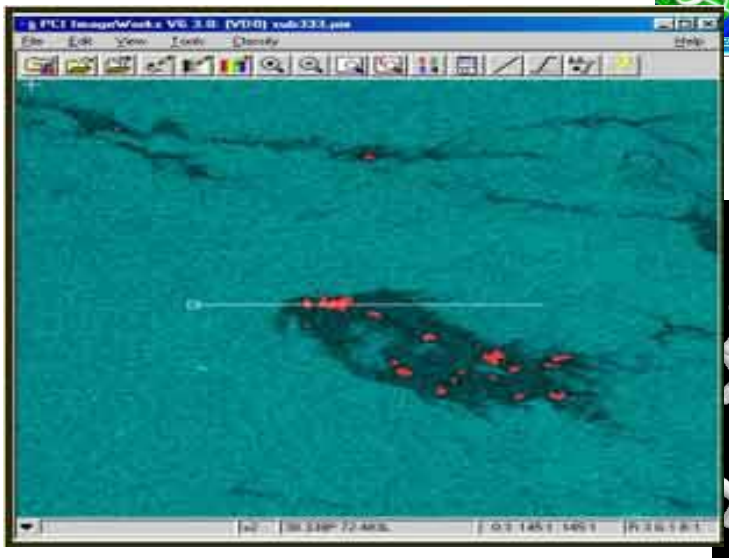
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

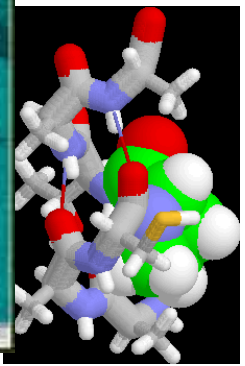
Σώμα Αξιολόγησης



Παραδείγματα εφαρμογών

- Πρόβλεψη καρκινικών κυττάρων για το αν είναι καλοήθη ή κακοήθη
- Κατηγοριοποίηση συναλλαγών με πιστωτική κάρτα για τον αν είναι νόμιμες ή μη
- Κατηγοριοποίηση των δομών πρωτεΐνης
- Κατηγοριοποίηση άρθρων εφημερίδων ως οικονομικά, αθλητικά, κοινωνικά, κτλ.
- Κατηγοριοποίηση δορυφορικών εικόνων θαλάσσης για το αν είναι πετρελαϊκή διαρροή ή φύκια





Τεχνικές

- Μέθοδοι με δέντρα αποφάσεων
- Μέθοδοι βασισμένοι σε κανόνες
- Μέθοδοι βασισμένοι στη μνήμη
- Νευρωνικά δίκτυα
- Δίκτυα Bayes και απλοϊκή μέθοδος Bayes
- Μηχανές διανυσμάτων υποστήριξης

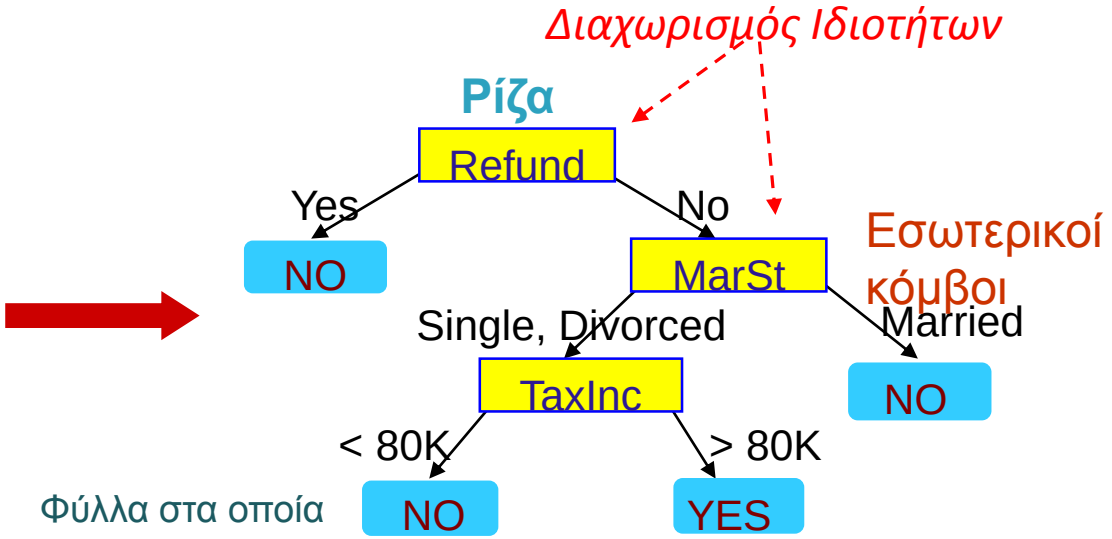
- Decision Tree based Methods
- Rule-based Methods
- Memory based reasoning
- Neural Networks
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines

Παράδειγμα δέντρου αποφάσεων

κατηγορικές κατηγορικές συνεχής κλάση

	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

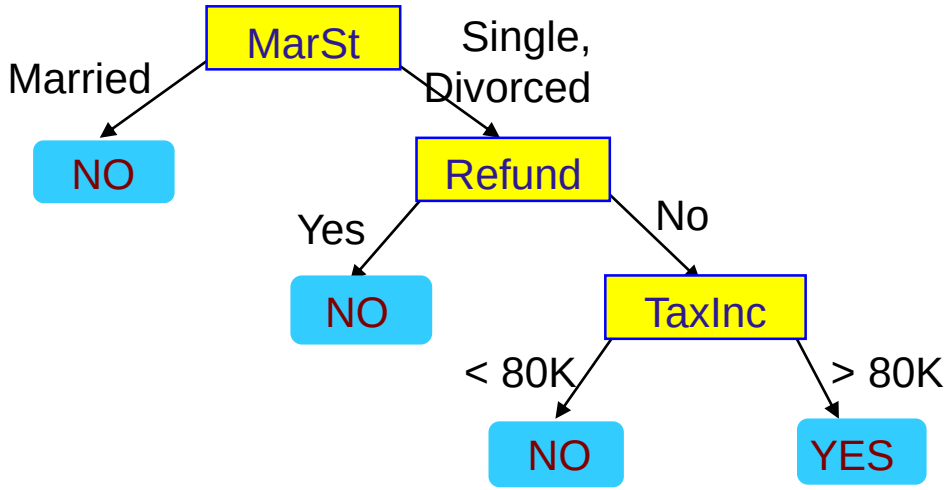
Δεδομένα εκπαίδευσης



Μοντέλο: Δέντρο Απόφασης

Άλλο παράδειγμα…

Tid	Refund (κατηγορικές)	Marital Status (κατηγορικές)	Taxable Income (συνεχής)	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



Μπορεί να υπάρχει και άλλο δέντρο που να χαρακτηρίζει τα δεδομένα

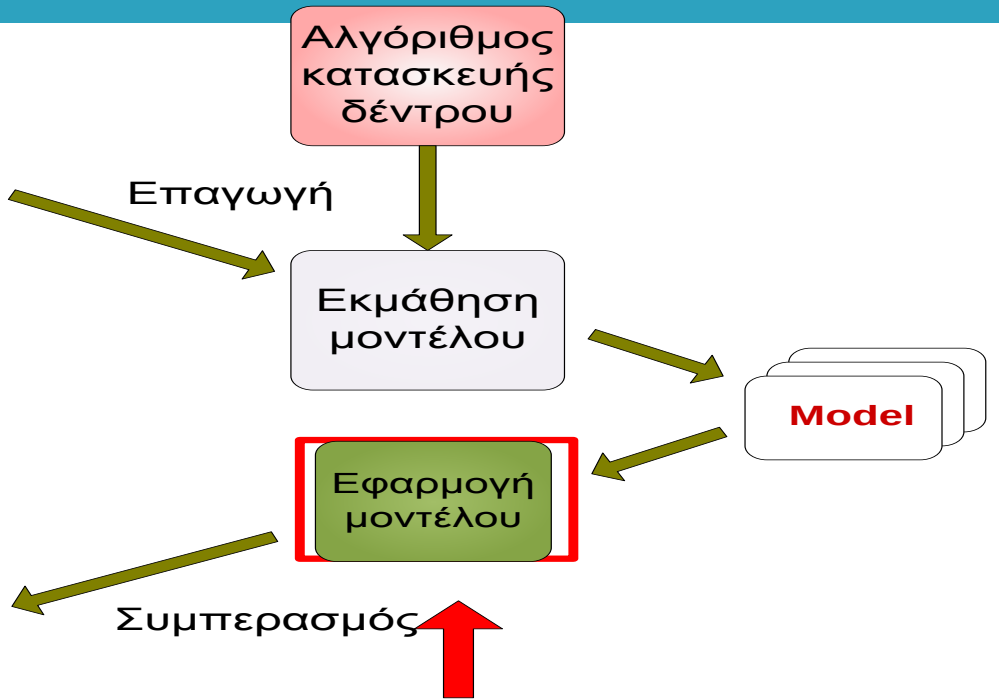
Η κατηγοριοποίηση με δέντρα αποφάσεων (εφαρμογή)

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

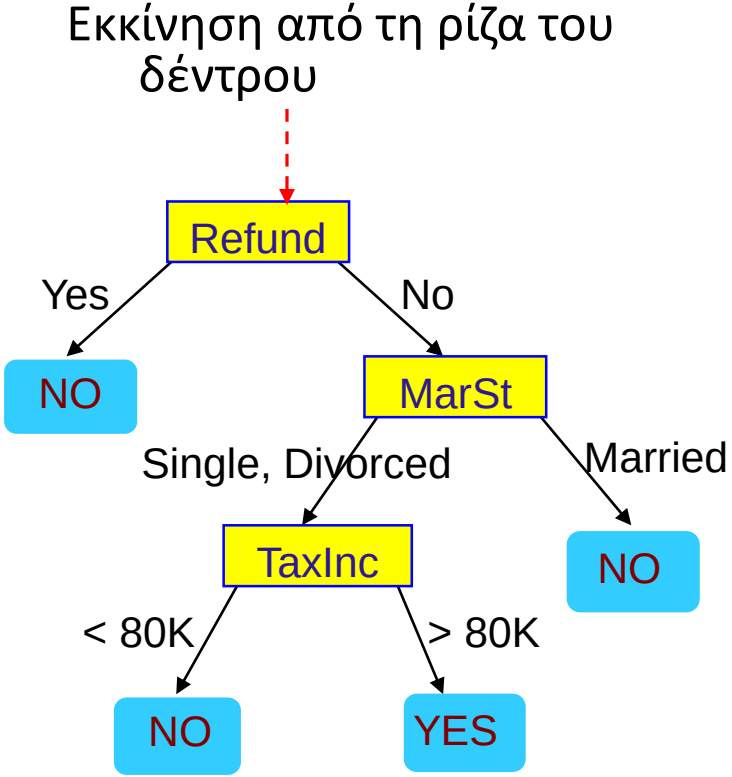
Σώμα εκπαίδευσης

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Σώμα αξιολόγησης



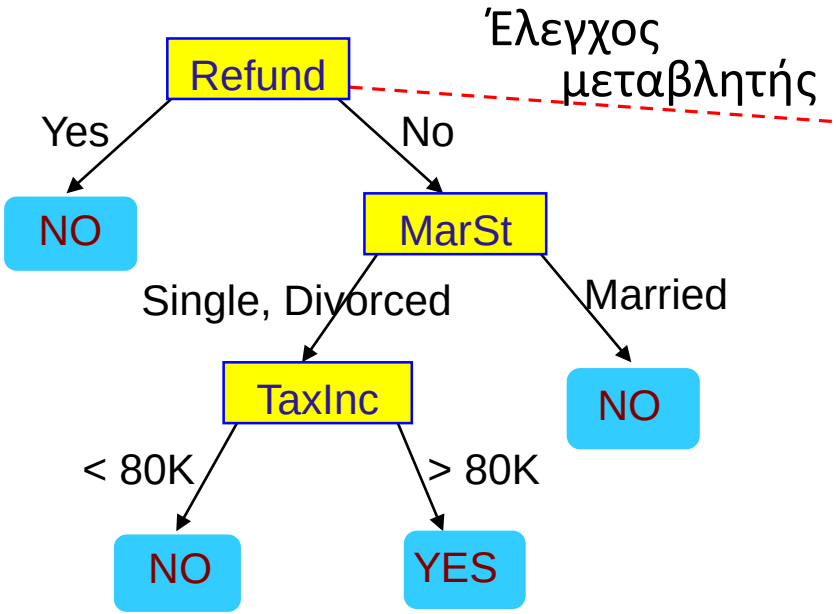
Εφαρμογή του μοντέλου



Σώμα αξιολόγησης

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

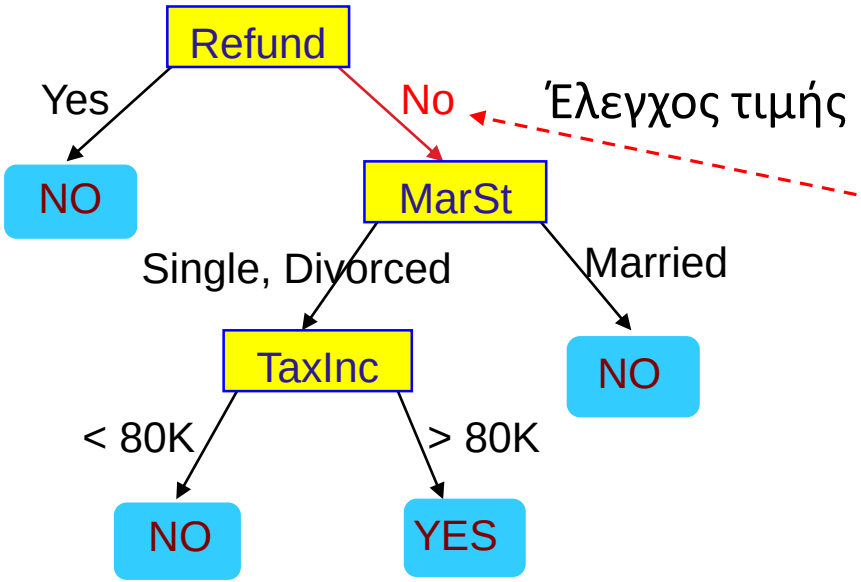
Εφαρμογή του μοντέλου



Σώμα αξιολόγησης

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Εφαρμογή του μοντέλου

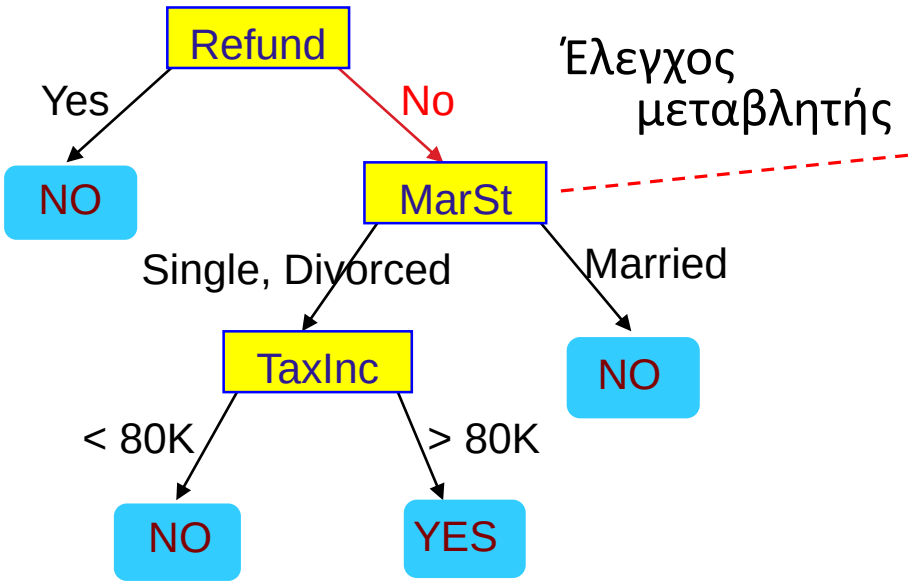


Έλεγχος τιμής

Σώμα αξιολόγησης

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Εφαρμογή του μοντέλου

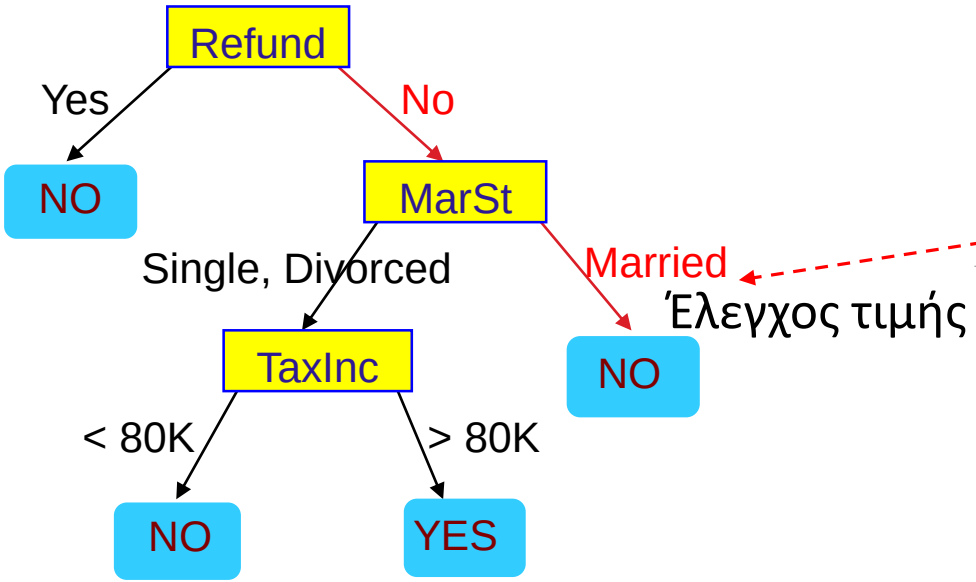


Έλεγχος μεταβλητής

Σώμα αξιολόγησης

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Εφαρμογή του μοντέλου

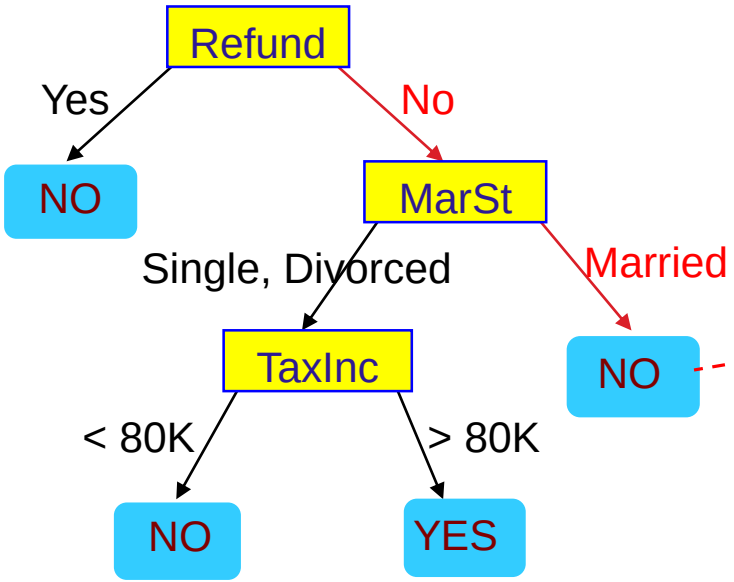


Σώμα αξιολόγησης

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Έλεγχος τιμής

Εφαρμογή του μοντέλου



Σώμα αξιολόγησης

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Ανάθεση της τιμής NO στην ιδιότητα Cheat

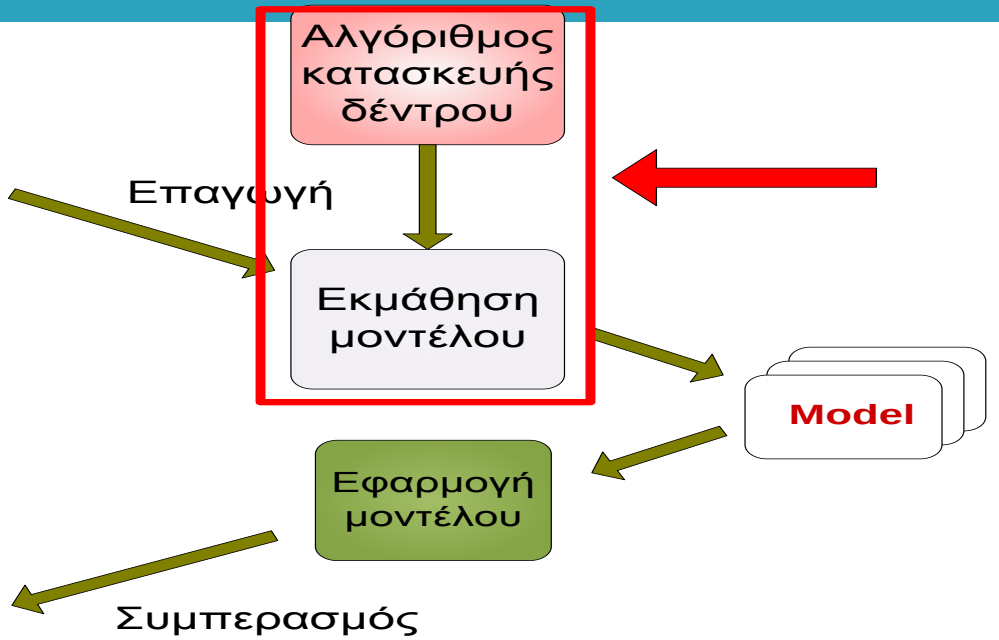
Η κατηγοριοποίηση με δέντρα αποφάσεων (εκμάθηση)

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Σώμα εκπαίδευσης

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Σώμα αξιολόγησης



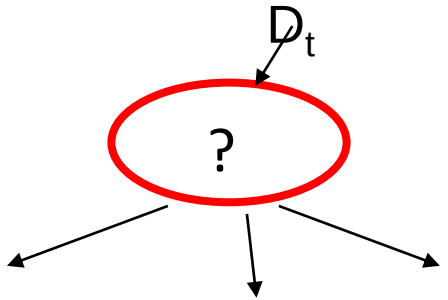
Επαγωγή δέντρου αποφάσεως

- Πολλοί αλγόριθμοι:
  - Ο αλγόριθμος του Hunt (από τις πρώτες μεθόδους)
  - CART
  - ID3, C4.5
  - SLIQ,SPRINT

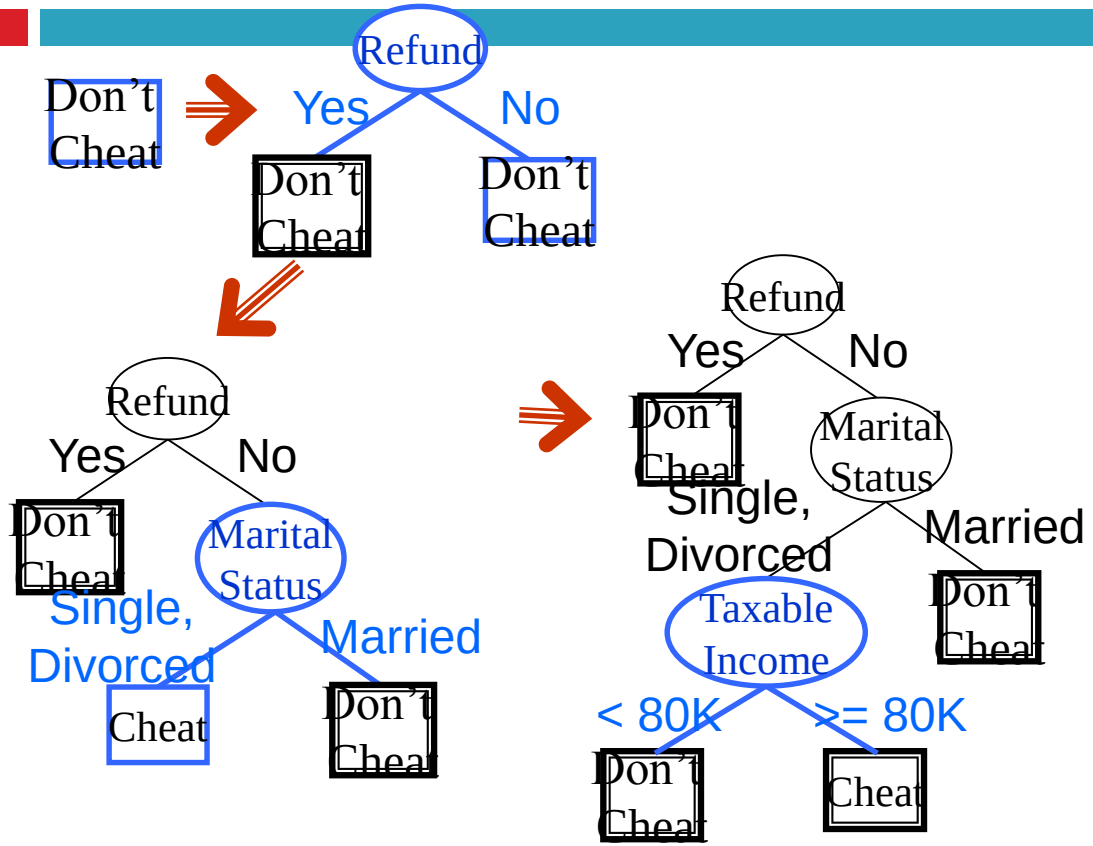
Γενική δομή του αλγόριθμου Hunt

- Έστω D_t ένα σύνολο δεδομένων εκπαίδευσης που καταφθάνουν σε ένα κόμβο t
- Διαδικασία
  - Αν το D_t περιέχει εγγραφές που ανήκουν στην ίδια κλάση y_t, τότε ο κόμβος παίρνει ως ετικέτα την κλάση y_t
  - Αν το D_t είναι ένα κενό σύνολο, τότε ο t παίρνει ως ετικέτα μια προκαθορισμένη κλάση y_d
  - Αν το D_t περιέχει εγγραφές που ανήκουν σε περισσότερες από μια κλάση, εφάρμοσε ένα τεστ ιδιοτήτων για να διαχωρίσεις τα δεδομένα σε υποσύνολα. Επανέλαβε τα προηγούμενα για κάθε υποσύνολο

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



Ο αλγόριθμος του Hunt στην πράξη



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

Δέντρο Απόφασης: Αλγόριθμος του Hunt

- Γενική Διαδικασία (πιο αναλυτικά):
  - Αν το Dt περιέχει εγγραφές που ανήκουν στην ίδια κλάση yt, τότε ο κόμβος t είναι κόμβος φύλλο με ετικέτα yt
  - Αν Dt είναι το κενό σύνολο, αυτό σημαίνει ότι δεν υπάρχει εγγραφή στο σύνολο εκπαίδευσης με αυτό το συνδυασμό τιμών,
    - τότε φύλλο με κλάση αυτής της πλειοψηφίας των εγγραφών εκπαίδευσης ή ανάθεση κάποιας default κλάσης
  - Αν το D_t περιέχει εγγραφές που ανήκουν σε περισσότερες από μία κλάσεις, χρησιμοποίησε έναν έλεγχο-γνωρίσματος για το διαχωρισμό των δεδομένων σε μικρότερα υποσύνολα
  - Εφάρμοσε την Διαδικασία αναδρομικά σε κάθε υποσύνολο.

Επαγωγή του δέντρου

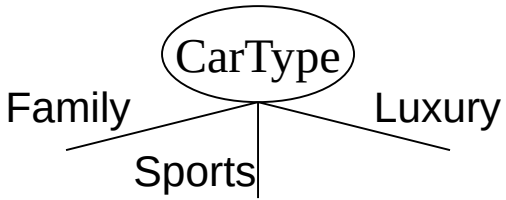
- Αδηφάγος (Greedy) στρατηγική
  - Διαχωρισμός των εγγραφών με βάση ένα τεστ ιδιότητας που βελτιστοποιεί ένα συγκεκριμένο κριτήριο
- Ζητήματα
  - Προσδιορισμός του τρόπου διάσπασης ή διαχωρισμού (splitting)
    - Πως προσδιορίζουμε τη συνθήκη ελέγχου ιδιοτήτων;
    - Πως καθορίζεται ο καλύτερος διαχωρισμός;
  - Προκαθορισμός του πότε θα σταματήσει η διάσπαση

Πως προσδιορίζουμε τη συνθήκη ελέγχου ιδιοτήτων;

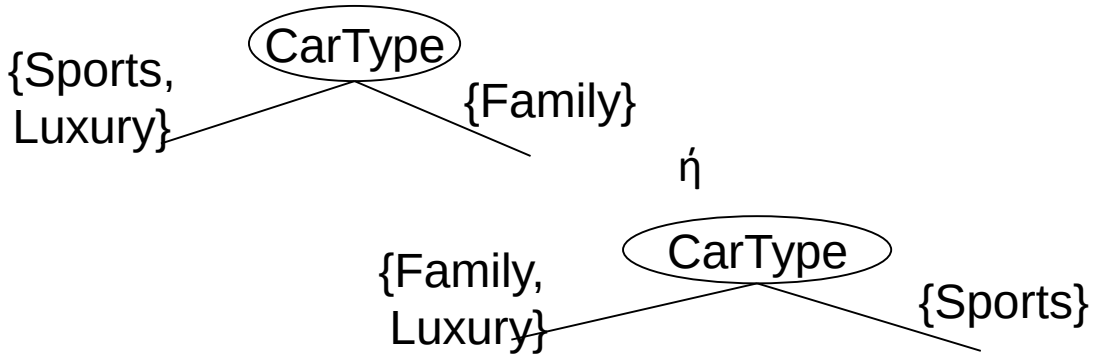
- Εξαρτάται από τον τύπο των ιδιοτήτων
  - Ονοματικές
  - Ταξινομημένες
  - Συνεχείς
- Εξαρτάται από τον αριθμό των τρόπων διάσπασης
  - Δυαδική διάσπαση
  - Πολλαπλή διάσπαση

Διαχωρίζοντας ονοματικές ιδιότητες

- Πολλαπλός διαχωρισμός
  - Χρησιμοποιούνται τόσοι κλάδοι όσοι και οι ξεχωριστές τιμές της ιδιότητας

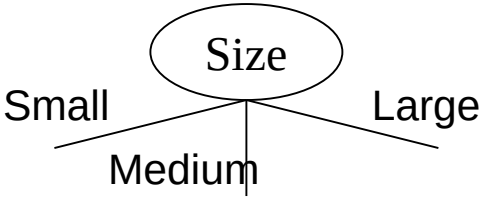


- Δυαδικός διαχωρισμός
  - Παίρνουμε δυο υποσύνολα τιμών

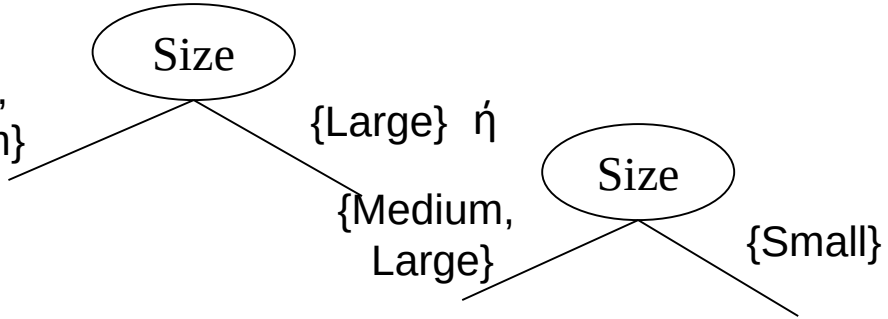


Διαχωρίζοντας ταξινομημένες ιδιότητες

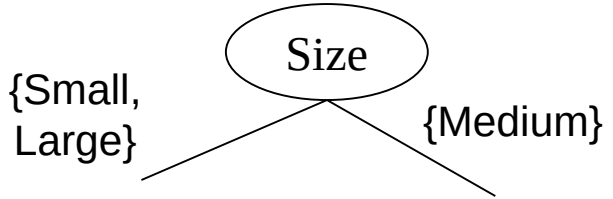
- Πολλαπλός διαχωρισμός
  - Χρησιμοποιούνται τόσοι κλάδοι όσοι και οι ξεχωριστές τιμές της ιδιότητας



- Δυαδικός διαχωρισμός
  - Παίρνουμε δυο υποσύνολα τιμών

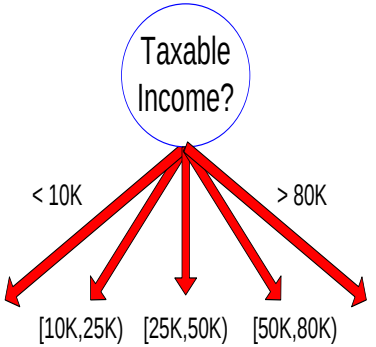


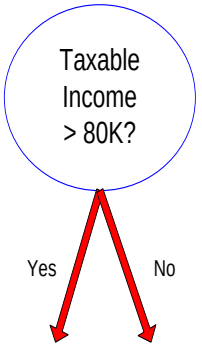
ή μήπως;



Διαχωρίζοντας συνεχείς ιδιότητες

- Διαφορετικοί τρόποι προσέγγισης
  - Διακριτοποίηση για το σχηματισμός μιας ταξινομημένης ιδιότητας
    - Στατική: γίνεται μια φορά στην αρχή
    - Δυναμική: συσταδοποίηση, percentiles κτλ
  - Δυαδική απόφαση: $(A < v)$ ή $(A \geq v)$
    - Θεωρούμε όλους τους πιθανούς διαχωρισμούς και βρίσκουμε την καλύτερη τομή
    - Είναι υπολογιστικά δαπανηρή διαδικασία





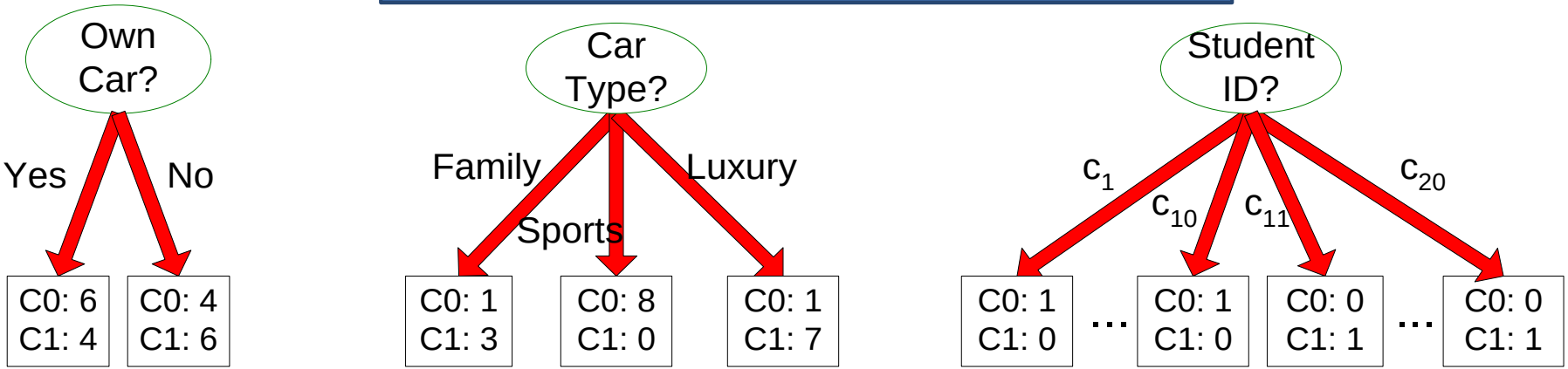
Επαγωγή του δέντρου

- Άπληστη (Greedy) στρατηγική
  - Διαχωρισμός των εγγραφών με βάση ένα τεστ ιδιότητας που βελτιστοποιεί ένα συγκεκριμένο κριτήριο
- Ζητήματα
  - Προσδιορισμός του τρόπου διαχωρισμού
    - Πως προσδιορίζουμε τη συνθήκη ελέγχου ιδιοτήτων;
    - *Πως καθορίζεται ο καλύτερος διαχωρισμός;*
  - Προκαθορισμός του πότε θα σταματήσει ο διαχωρισμός

Πως καθορίζεται ο καλύτερος διαχωρισμός

Πριν το διαχωρισμό: 10 εγγραφές της κλάσης 0

10 εγγραφές της κλάσης 1



Ποια είναι η καλύτερη συνθήκη ελέγχου ιδιοτήτων;

Πως καθορίζεται ο καλύτερος διαχωρισμός

C0: 5
C1: 5

Μη ομοιογενής,

Μικρός βαθμός καθαρότητας

- Άπληστη προσέγγιση
  - Οι κόμβοι με ομοιογενή κατανομή κλάσης είναι προτιμητέοι
- Χρειαζόμαστε ένα μέτρο καθαρότητας ενός κόμβου
  - Δείκτης Gini
  - Εντροπία
  - Σφάλμα ταξινόμησης (classification error)

C0: 9
C1: 1

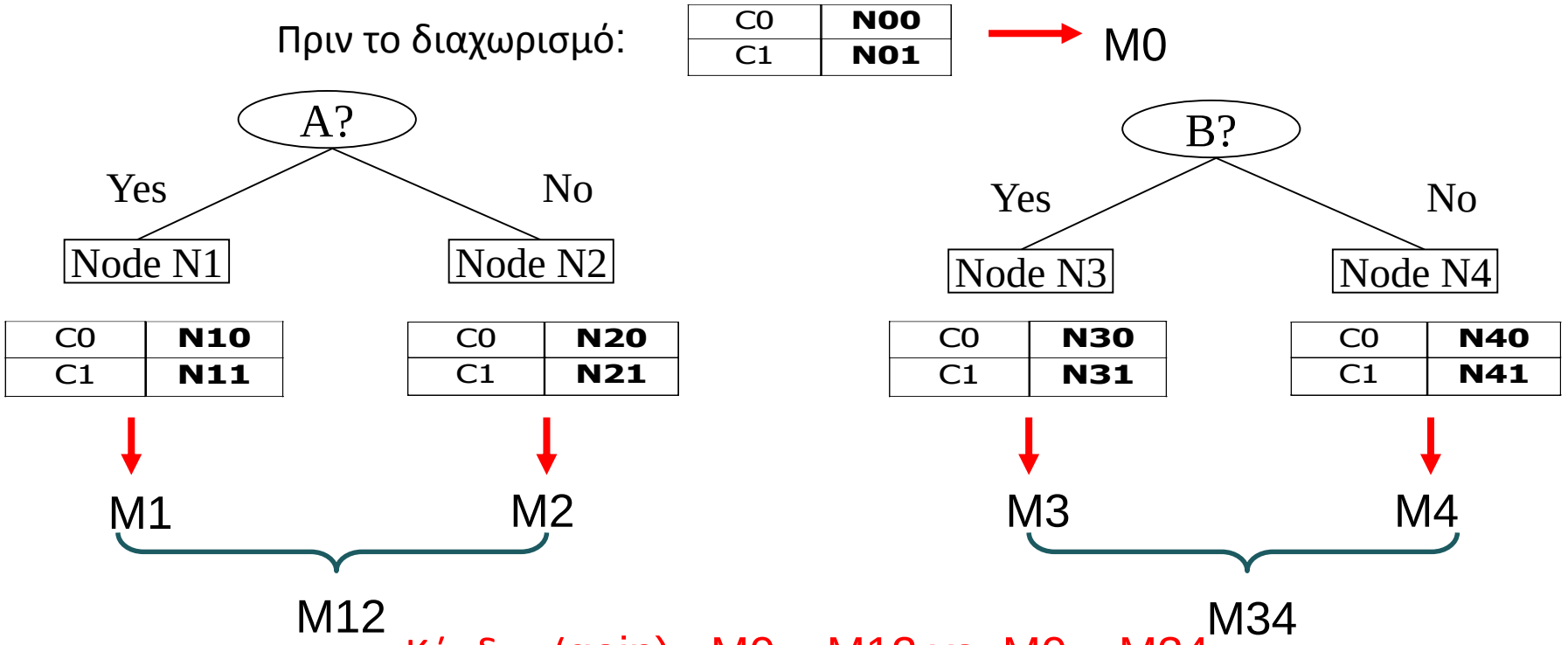
Ομοιογενής,

Μεγάλος βαθμός καθαρότητας

$$\Delta = I(parent) - \sum_{i=1}^{k} \frac{N(u_i)}{N} I(u_i)$$

Ν είναι ο αριθμός των εγγραφών στο γονέα και $N(u_j)$ του j-οστού παιδιού

Πως βρίσκεται ο καλύτερος διαχωρισμός;



Δείκτης GINI

- Έστω t ένας κόμβος

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

όπου p(j|t) η πιθανότητα εμφάνισης της κλάσης j στον κόμβο t

- Μέγιστο: $(1 - 1/n_c)$ όταν όλες οι εγγραφές είναι εξίσου κατανεμημένες σε κάθε κλάση
  - Λιγότερο σημαντική πληροφορία
- Ελάχιστο: 0, όταν όλες οι εγγραφές έχουν μια κοινή κλάση
  - Περισσότερο σημαντική πληροφορία

C1	**0**
C2	**6**
Gini=0.000	

C1	**1**
C2	**5**
Gini=0.278	

C1	**2**
C2	**4**
Gini=0.444	

C1	**3**
C2	**3**
Gini=0.500	

Παραδείγματα υπολογισμού δείκτη GINI

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

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

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

Gini = 1 – P(C1)^2 – P(C2)^2 = 1 – 0 – 1 = 0

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

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

Gini = 1 – (1/6)^2 – (5/6)^2 = 0.278

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

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

Gini = 1 – (2/6)^2 – (4/6)^2 = 0.444

Διάσπαση με βάση το δείκτη GINI

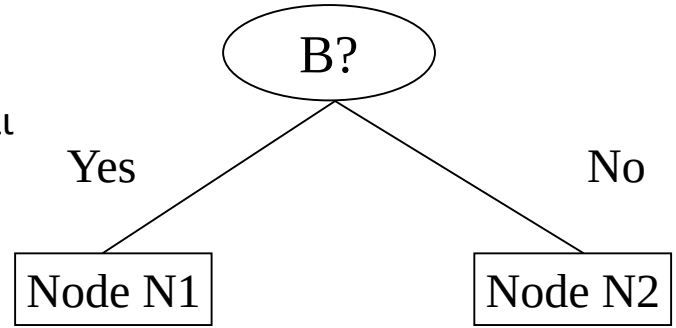
- Χρησιμοποιείται από τους αλγόριθμους CART, SLIQ, SPRINT
- Όταν ένας κόμβος p διαχωρίζεται σε k μέρη (παιδιά), η ποιότητα του διαχωρισμού μετριέται ως:

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

- όπου n_i = αριθμός των εγγραφών στο παιδί i
- n= αριθμός των εγγραφών στον κόμβο P

Δυαδικές ιδιότητες: Υπολογισμός του δείκτη GINI

- Γίνεται διαχωρισμός σε 2 τμήματα
- Επίδραση στα βάρη κάθε τμήματος
  - Επιδιώκουμε μεγαλύτερα και πιο ομοιογενή τμήματα



	Πατρικός
C1	**6**
C2	**6**
Gini = 0.500	

	N1	N2
C1	**5**	**1**
C2	**2**	**4**
Gini=0.333		

Gini(N1)
$= 1 - (5/6)^2 - (2/6)^2$
$= 0.194$

Gini(N2)
$= 1 - (1/6)^2 - (4/6)^2$
$= 0.528$

Gini(απογόνων)
$= 7/12 * 0.194 +$
$5/12 * 0.528$
$= 0.333$

Κατηγορικές ιδιότητες: Υπολογισμός του δείκτη GINI

- Για κάθε διακεκριμένη τιμή, συγκέντρωσε τα αθροίσματα κάθε κλάσης στο σύνολο δεδομένων
- Χρησιμοποίησε τον πίνακα αθροισμάτων για να λάβεις αποφάσεις

Πολλαπλός διαχωρισμός

	CarType		
	Family	Sports	Luxury
C1	1	2	1
C2	4	1	1
Gini	0.393		

Δυαδικός διαχωρισμός
(εύρεση των καλύτερων ομαδοποιήσεων)

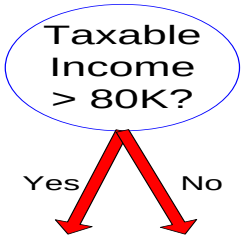
	CarType	
	{Sports, Luxury}	{Family}
C1	3	1
C2	2	4
Gini	0.400	

	CarType	
	{Sports}	{Family, Luxury}
C1	2	2
C2	1	5
Gini	0.419	

Συνεχείς ιδιότητες: Υπολογισμός του δείκτη GINI

- Χρησιμοποίησε δυαδικές αποφάσεις με βάση μια τιμή
- Πληθώρα επιλογών της τιμής διάσπασης
  - Αριθμός πιθανών τιμών = αριθμός των διακεκριμένων τιμών
- Κάθε τιμή διάσπασης έχει ένα πίνακα αθροισμάτων
  - Σε κάθε διαχωρισμό, τα αθροίσματα της κλάσης $A < v$ και $A \geq v$
- Απλή μέθοδος επιλογής της καλύτερης τιμής v
  - Για κάθε v, σάρωση της βάσης και υπολογισμός του δείκτη GINI
  - Υπολογιστικά ΑΝΑΠΟΤΕΛΕΣΜΑΤΙΚΟ (ΕΠΑΝΑΛΗΨΗ ΥΠΟΛΟΓΙΣΜΩΝ)

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



Συνεχείς ιδιότητες: Υπολογισμός του δείκτη GINI

- Για αποτελεσματικό υπολογισμό: για κάθε ιδιότητα,
  - Ταξινόμησε τις τιμές
  - Γραμμική σάρωση των τιμών με ανανέωση του πίνακα αθροισμάτων και υπολογισμός του δείκτη GINI
  - Επιλογή της θέσης διάσπασης με το μικρότερο δείκτη GINI

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
Taxable Income										
Ταξινομημένες →	60	70	75	85	90	95	100	120	125	220
Θέσεις διάσπασης →	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

Εναλλακτικά κριτήρια διάσπασης με βάση την πληροφορία

- Εντροπία ενός κόμβου t:

$$\text{Εντροπία}(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

- Μετρά την ομοιογένεια ενός κόμβου
  - Μέγιστη :($\log n_c$), όταν όλες οι εγγραφές είναι εξίσου κατανεμημένες σε κάθε κλάση
    - Λιγότερο σημαντική πληροφορία
  - Ελάχιστη:0, όταν όλες οι εγγραφές έχουν μια κοινή κλάση
    - Περισσότερο σημαντική πληροφορία
- Οι υπολογισμοί με βάση την εντροπία είναι ισότιμοι με αυτούς του δείκτη GINI

Παραδείγματα υπολογισμού εντροπίας

$$\text{Εντροπία}(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

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

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

Εντροπία= – 0 log 0 – 1 log 1 = – 0 – 0 = 0

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

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

Εντροπία = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (1/6) = 0.65

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

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

Εντροπία = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

Εναλλακτικά κριτήρια διάσπασης με βάση την πληροφορία…

- Κέρδος πληροφορίας (Information Gain):

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

  - Ο πατρικός κόμβος p χωρίζεται σε k τμήματα
  - n_i είναι ο αριθμός εγγραφών στο τμήμα i
- Μετρά την μείωση της εντροπίας που επιτυγχάνεται με τη διάσπαση. Επιλέγουμε το διαχωρισμό με το μεγαλύτερο κέρδος πληροφορίας
- Χρησιμοποιείται στους αλγόριθμους ID3 και C4.5
- Μειονέκτημα: τείνει να προτιμά διασπάσεις που αποφέρουν μεγάλο αριθμό τμημάτων, κάθε ένα μικρό αλλά ομοιογενές

Εναλλακτικά κριτήρια διάσπασης με βάση την πληροφορία…

- Λόγος κέρδους (GAIN ratio)

$$GainRATIO_{split} = \frac{GAIN_{Split}}{SplitINFO}$$

$$SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$

  - Ρυθμίζει το κέρδος πληροφορίας με βάση την εντροπία της διάσπασης (SplitINFO). Η μεγαλύτερη εντροπία διαχωρισμού (μεγάλος αριθμός μικρών τμημάτων) τιμωρείται!

- Προτάθηκε για να αντιμετωπίσει το μειονέκτημα του κέρδους πληροφορίας
- Χρησιμοποιείται από τον αλγόριθμο C4.5

Εναλλακτικά κριτήρια διάσπασης με βάση το σφάλμα ταξινόμησης…

- Σφάλμα ταξινόμησης σε ένα κόμβο t

$$Error(t) = 1 - \max_i P(i \mid t)$$

- Μετράει το σφάλμα στην κατηγοριοποίηση που κάνει ένας κόμβος
  - Μέγιστο: $(1 - 1/n_c)$ όταν όλες οι εγγραφές είναι εξίσου κατανεμημένες σε κάθε κλάση
    - Λιγότερο σημαντική πληροφορία
  - Ελάχιστο: 0, όταν όλες οι εγγραφές έχουν μια κοινή κλάση
    - Περισσότερο σημαντική πληροφορία

Παραδείγματα υπολογισμού του σφάλματος

$$Error(t) = 1 - \max_i P(i \mid t)$$

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

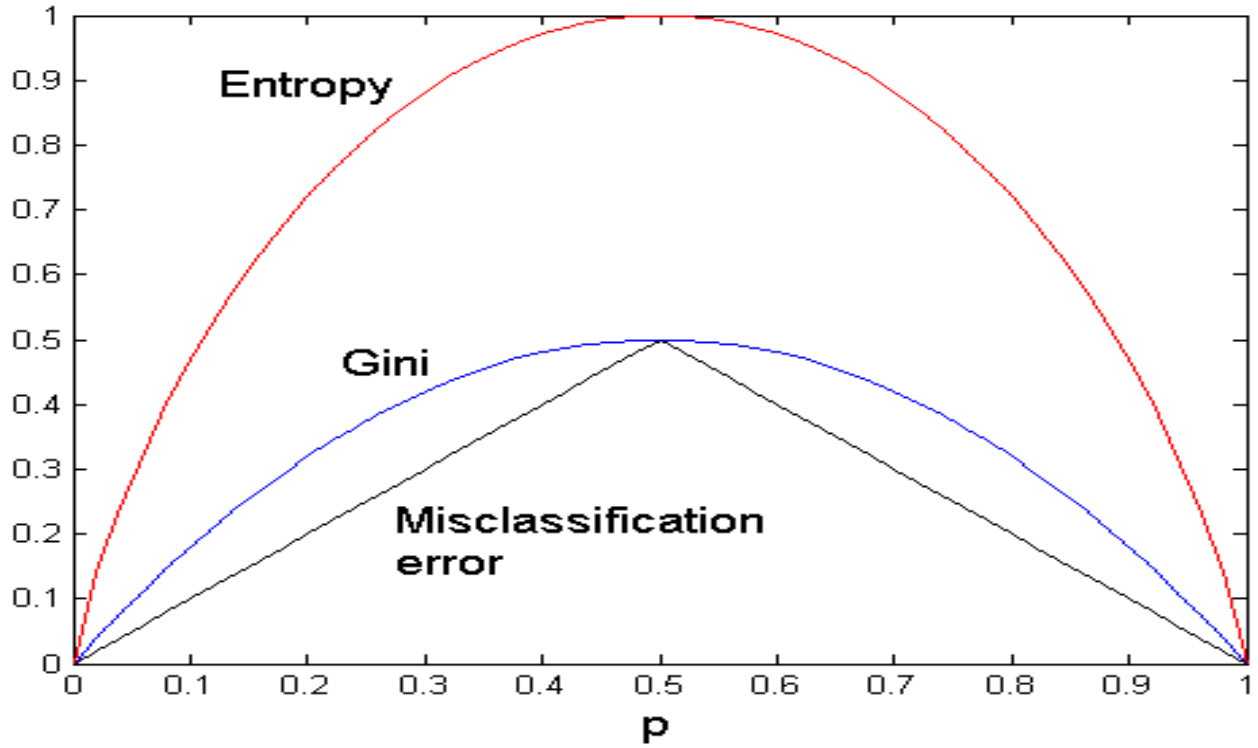
C1	**2**
C2	**4**

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

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

Σύγκριση κριτηρίων διάσπασης

Για πρόβλημα 2 τιμών κλάσης



Επαγωγή του δέντρου

- Άπληστη (Greedy) στρατηγική
  - Διαχωρισμός των εγγραφών με βάση ένα τεστ ιδιότητας που βελτιστοποιεί ένα συγκεκριμένο κριτήριο
- Ζητήματα
  - Προσδιορισμός του τρόπου διαχωρισμού
    - Πως προσδιορίζουμε τη συνθήκη ελέγχου ιδιοτήτων;
    - Πως καθορίζεται ο καλύτερος διαχωρισμός;
  - *Προκαθορισμός του πότε θα σταματήσει ο διαχωρισμός*

Κριτήρια τερματισμού στην επαγωγή

- Τερματισμός επέκτασης ενός κόμβου όταν όλες οι εγγραφές ανήκουν στην ίδια κλάση
- Τερματισμός όταν όλες οι εγγραφές έχουν παρόμοιες τιμές ιδιοτήτων
- Πρόωρος τερματισμός (θα αναλυθεί αργότερα)

Κατηγοριοποίηση με δέντρα αποφάσεων

- Πλεονεκτήματα
  - Κατασκευάζονται εύκολα
  - Ιδιαίτερα γρήγορα στην κατηγοριοποίηση άγνωστων εγγραφών
  - Η ακρίβεια είναι συγκρίσιμη με αυτή άλλων τεχνικών κατηγοριοποίησης για απλά σύνολα δεδομένων

Παράδειγμα: C4.5

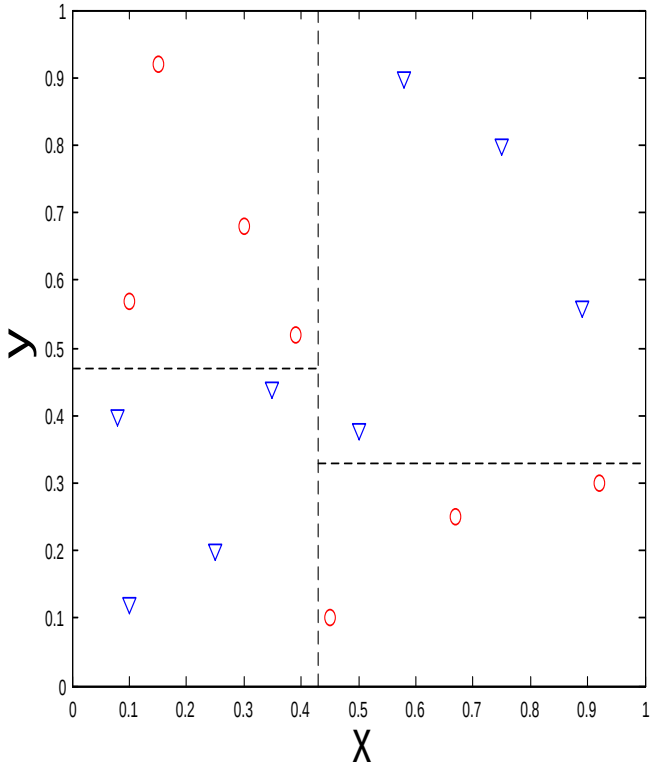
- Απλή κατά βάθος (depth-first) κατασκευή
- Χρησιμοποιεί το κέρδος πληροφορίας (information gain)
- Ταξινομεί τις συνεχείς ιδιότητες σε κάθε κόμβο
- Φορτώνει όλα τα δεδομένα στη μνήμη
  - Δεν είναι κατάλληλος για μεγάλα σώματα δεδομένων
- Μπορείτε να το κατεβάσετε από την ιστοσελίδα:
  - http://www.cse.unsw.edu.au/~quinlan/c4.5r8.tar.gz

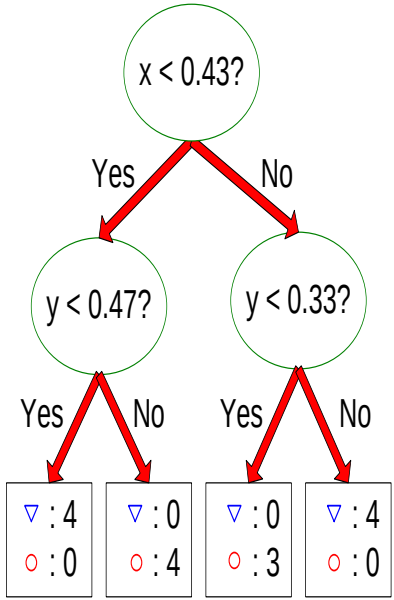
Όριο απόφασης

- Μέχρι στιγμής είδαμε ελέγχους που αφορούν μόνο ένα γνώρισμα τη φορά
  - μπορούμε να δούμε τη διαδικασία ως τη διαδικασία *διαμερισμού του χώρου* των γνωρισμάτων σε ξένες περιοχές μέχρι κάθε περιοχή να περιέχει εγγραφές που να ανήκουν στην ίδια κλάση
- Η οριακή γραμμή (Border line) μεταξύ δυο γειτονικών περιοχών που ανήκουν σε διαφορετικές κλάσεις ονομάζεται και **decision boundary (όριο απόφασης)**

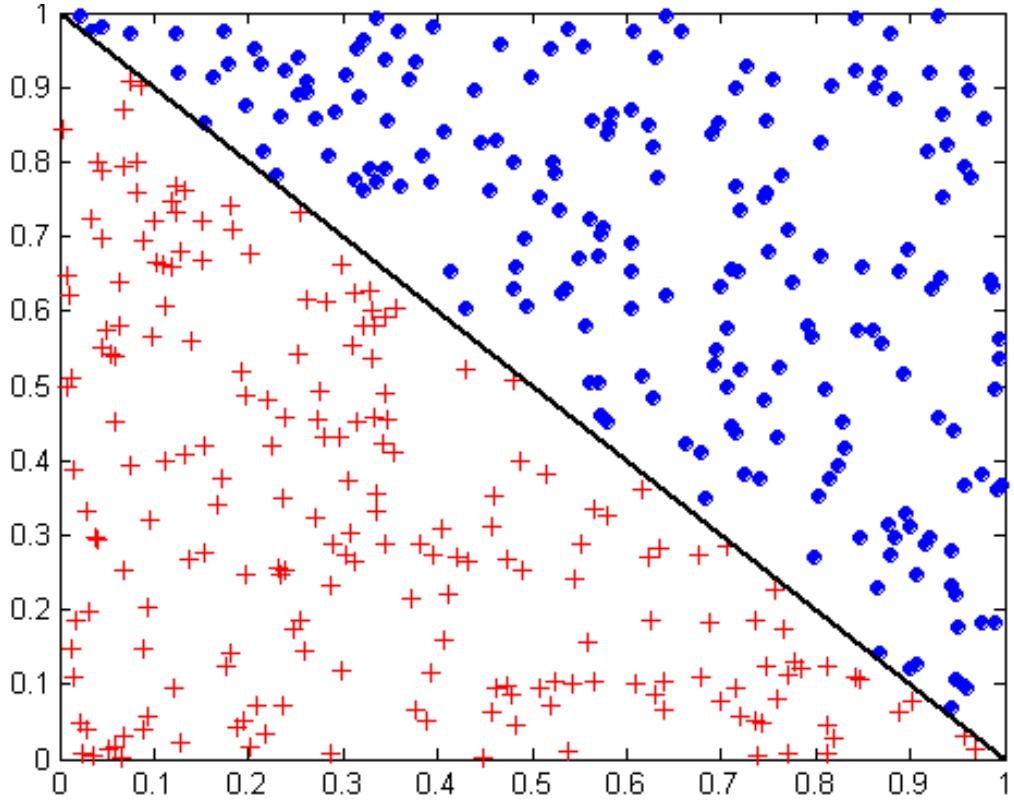
Όριο απόφασης

- Όταν η συνθήκη ελέγχου περιλαμβάνει μόνο ένα γνώρισμα τη φορά τότε το Decision boundary είναι παράλληλη στους άξονες
  - (τα decision boundaries είναι ορθογώνια παραλληλόγραμμα)

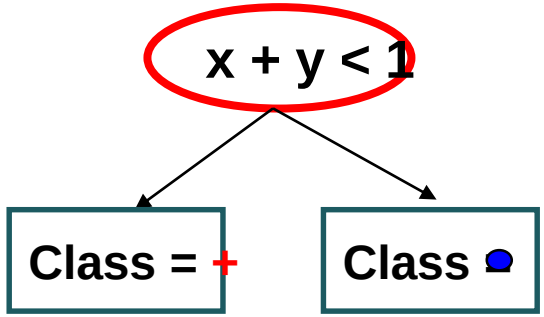




Oblique (πλάγιο) Δέντρο Απόφασης



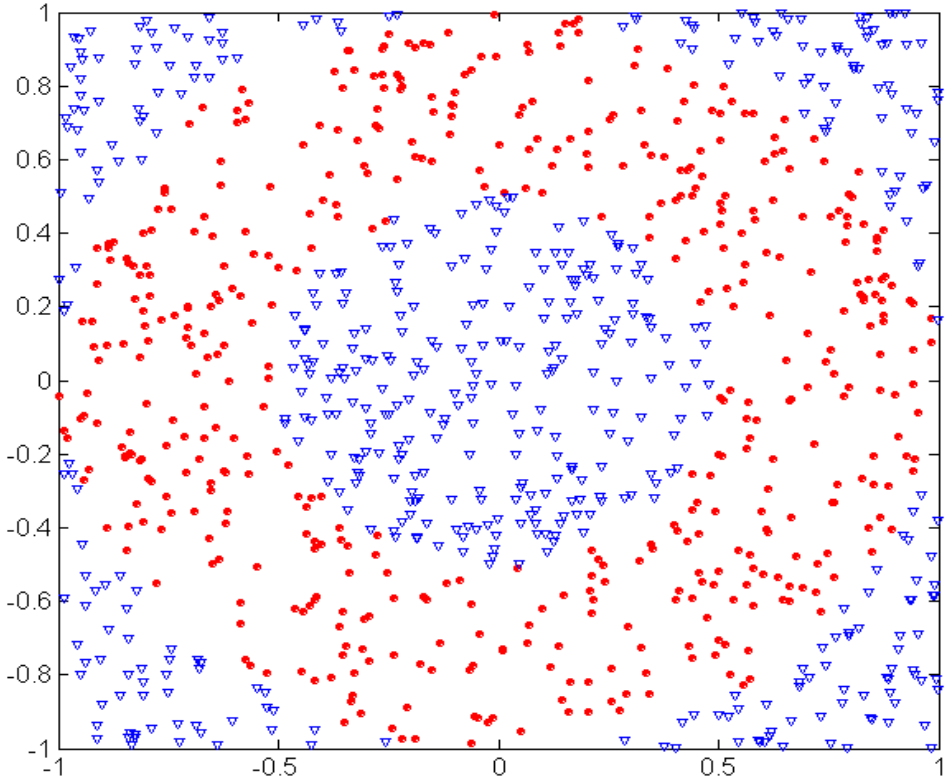
- Οι συνθήκες ελέγχου μπορούν να περιλαμβάνουν περισσότερα από ένα γνωρίσματα
- Μεγαλύτερη εκφραστικότητα
- Η εύρεση βέλτιστων συνθηκών ελέγχου είναι υπολογιστικά ακριβή



Πρακτικά ζητήματα κατηγοριοποίησης

- Υπερβολικό ή καθόλου ταίριασμα δεδομένων (overfitting & underfitting)
- Ελλιπείς τιμές (missing values)
- Κόστος κατηγοριοποίησης

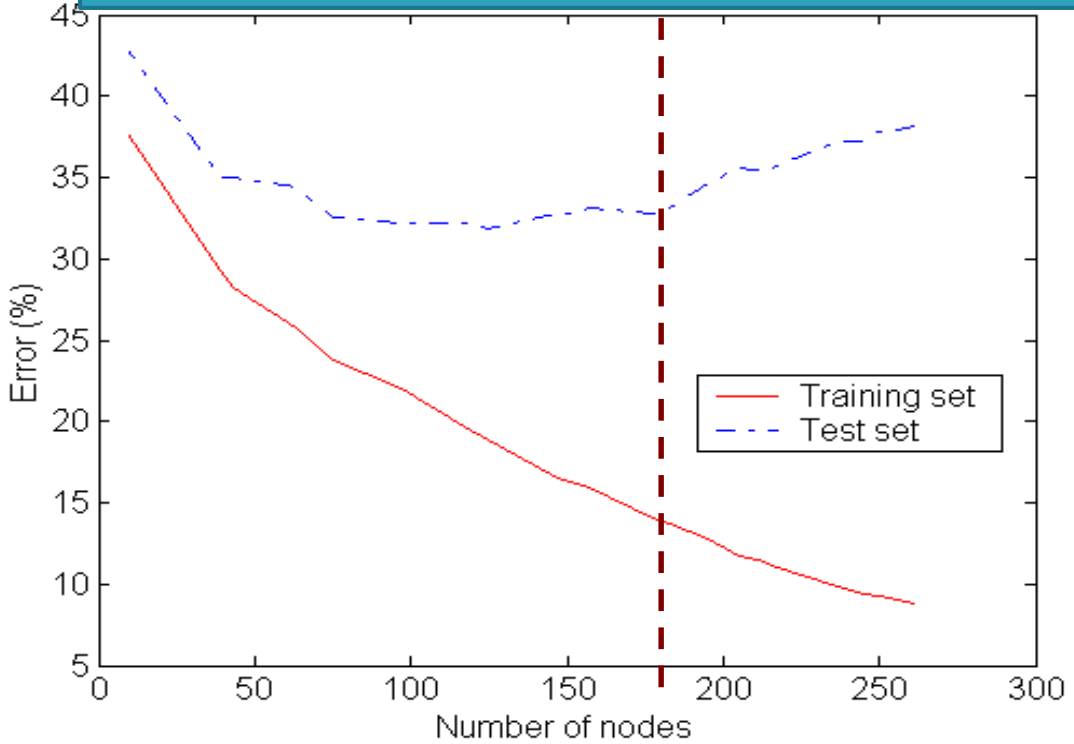
Παράδειγμα underfitting & overfitting:



- 500 κυκλικά και 500 τριγωνικά σημεία
- Κυκλικά σημεία:
  - $0.5 \leq \text{sqrt}(x_1^2+x_2^2) \leq 1$
- Τριγωνικά σημεία:
  - $\text{sqrt}(x_1^2+x_2^2) > 0.5$ ή
  - $\text{sqrt}(x_1^2+x_2^2) < 1$

Παράδειγμα underfitting & overfitting:

Το δέντρο απόφασης για το προηγούμενα δεδομένα 30% εκπαίδευση 70% έλεγχο. Εφαρμογή δείκτη GINI

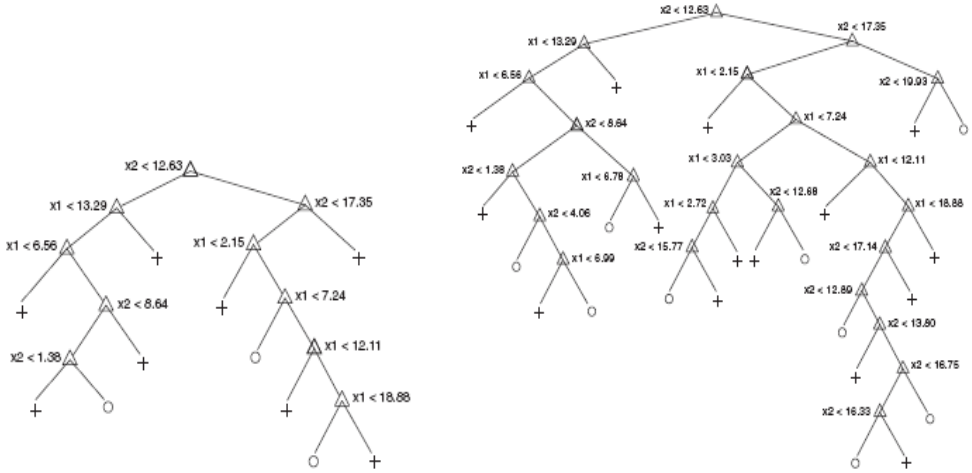


Overfitting: αποδίδει καλά όταν αξιολογείται στο σώμα εκπαίδευσης αλλά όχι και στο σώμα αξιολόγησης

Underfitting: σε πολύ απλά μοντέλα, το σφάλμα είναι πολύ μεγάλο και στα δυο σώματα (εκπαίδευσης και αξιολόγησης)

Overfitting

- Μπορούμε να διασπάσουμε το δέντρο μέχρι να φτάσουμε στο σημείο κάθε φύλλο να *ταιριάζει* απολύτως στα δεδομένα
  - Μικρό (μηδενικό) λάθος εκπαίδευσης
  - Μεγάλο λάθος ελέγχου
- Συχνά ισχύει και το αντίθετο !!!

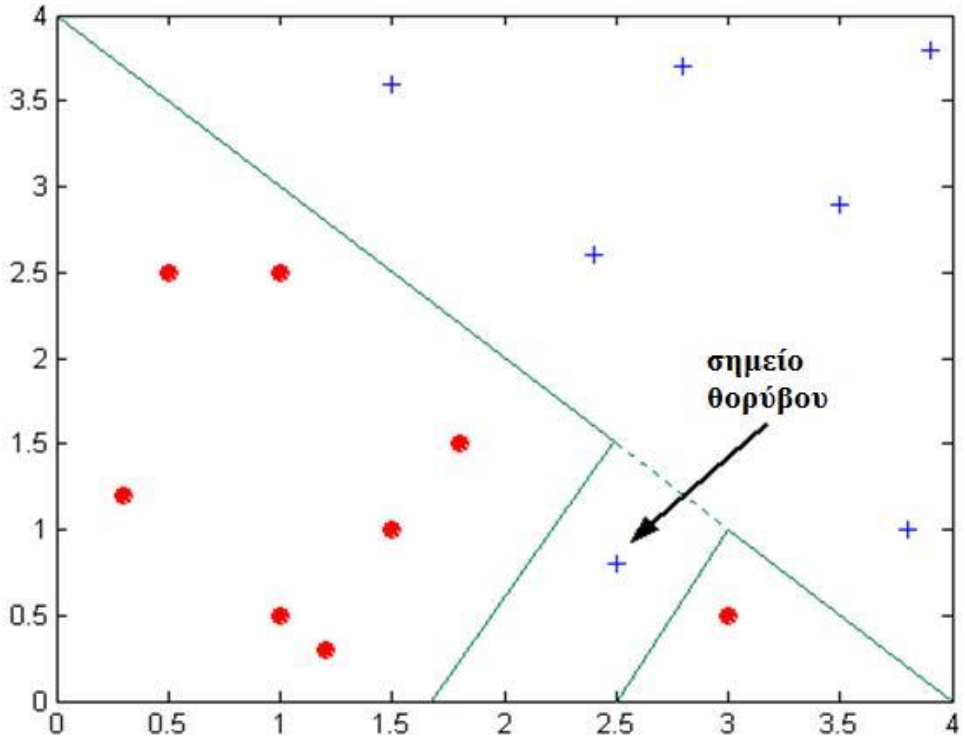


(a) Decision tree with 11 leaf nodes.

(b) Decision tree with 24 leaf nodes.

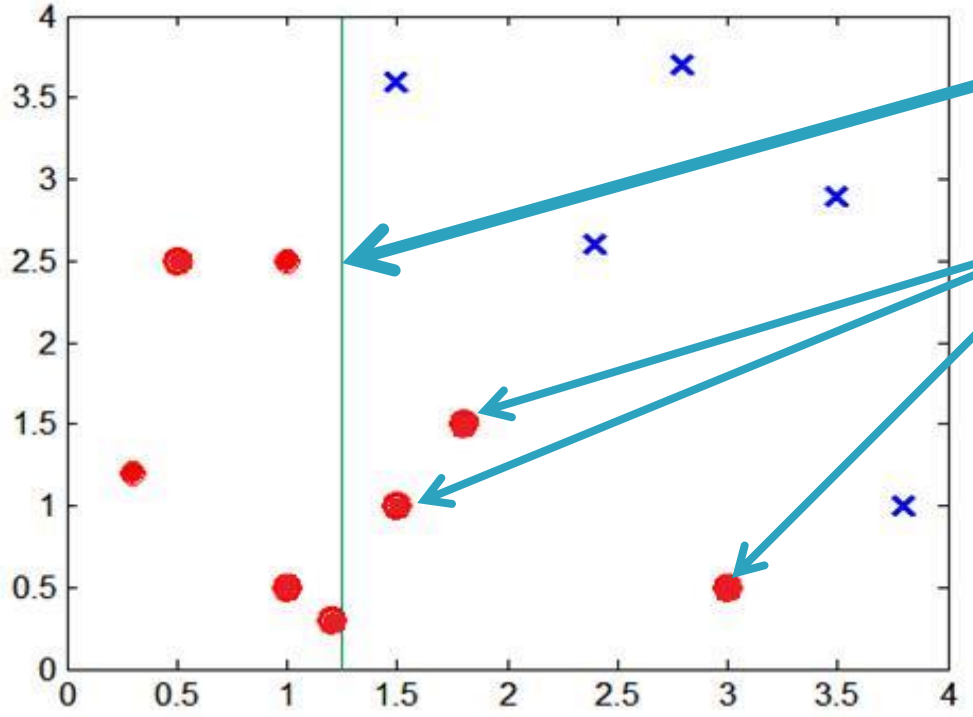
Figure 4.24. Decision trees with different model complexities.

Overfitting λόγω θορύβου



Το όριο απόφασης παραμορφώνεται από το σημείο θορύβου

Overfitting λόγω έλλειψης δεδομένων



Όριο απόφασης

Εσφαλμένη κατηγοριοποίηση

Η έλλειψη δεδομένων δεν καθορίζει με ακρίβεια το όριο απόφασης

Overfitting: Γενικές παρατηρήσεις

- Το υπερβολικό ταίριασμα έχει ως συνέπεια δέντρα αποφάσεων που είναι πιο περίπλοκα από ότι χρειάζεται
- Το σφάλμα στο σώμα εκπαίδευσης δεν είναι πλέον καλή ένδειξη για το πόσο καλά θα αποδώσει σε άγνωστα παραδείγματα
- Χρειαζόμαστε νέους τρόπους για να εκτιμήσουμε τα λάθη κατηγοριοποίησης

Εκτιμώντας το σφάλμα γενίκευσης

- Σφάλμα επανα-αντικαστάστασης (re-substitution error)
  - Σφάλμα στα δεδομένα εκπαίδευσης
    - Σ e(t)
- Σφάλμα γενίκευσης (generalization error)
  - Σφάλμα στα δεδομένα αξιολόγησης
    - Σ e'(t)

- Μέθοδοι εκτίμησηςs:
  - Αισιόδοξη προσέγγιση: e'(t) = e(t)
  - Απαισιόδοξη προσέγγιση:
    - για κάθε κόμβο: e'(t) = (e(t)+0.5)
    - Συνολικά σφάλματα
      - e'(T) = e(T) + N $\times$ 0.5 (N: αριθμός κόμβων)
    - για ένα δέντρο με 30 κόμβους και training error= 10/1000 = 1%
      - Generalization error = (10 + 30$\times$0.5)/1000 = 2.5%

Reduced error pruning (REP)

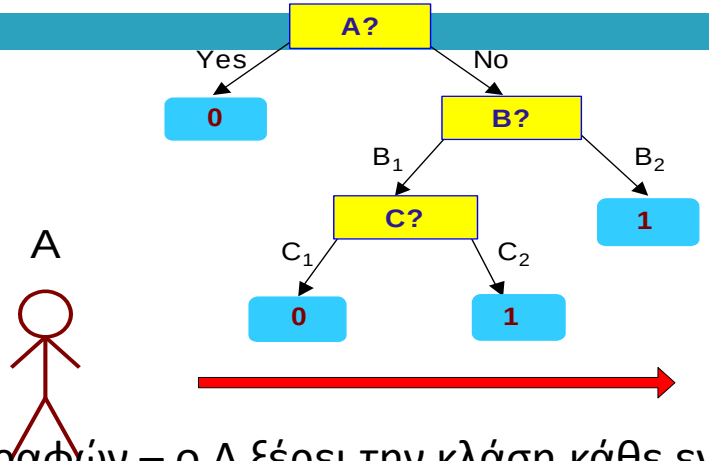
- χρήση ενός συνόλου επαλήθευσης για την εκτίμηση του λάθους γενίκευσης
  - Χώρισε τα δεδομένα εκπαίδευσης:
    - 2/3 εκπαίδευση
    - 1/3 (σύνολο επαλήθευσης – validation set) για υπολογισμό λάθους
  - Χρήση για εύρεση του κατάλληλου μοντέλου

Occam's Razor

- Δεδομένων δυο μοντέλων παρόμοιων σφαλμάτων γενίκευσης, προτιμούμε το απλούστερο από το πιο σύνθετο
- Ένα πολύπλοκο μοντέλο είναι πιο πιθανό να έχει ταιριαστεί (Fitted) τυχαία λόγω λαθών στα δεδομένα
- Επομένως στην αξιολόγηση ενός μοντέλου θα πρέπει να λαμβάνεται υπόψη και η πολυπλοκότητα

Πολυπλοκότητα μοντέλου (Minimum Description Length)

X	y
X_1	1
X_2	0
X_3	0
X_4	1
…	…
X_n	1



X	y
X_1	?
X_2	?
X_3	?
X_4	?
…	…
X_n	?

Έστω X ένα σύνολο εγγραφών – ο A ξέρει την κλάση κάθε εγγραφής – μετάδοση στον B

Cost(Model,Data) = Cost(Data|Model) + Cost(Model)

Cost=αριθμός των Bit που χρειάζονται για την κωδικοποίηση. Επιθυμούμε το μικρότερο κόστος.

Cost(Data|Model) = λάθη ταξινόμησης.

Cost(Model) = αριθμός παιδιών + κωδικοποίηση διάσπασης

Διευθετώντας την περίπτωση overfitting

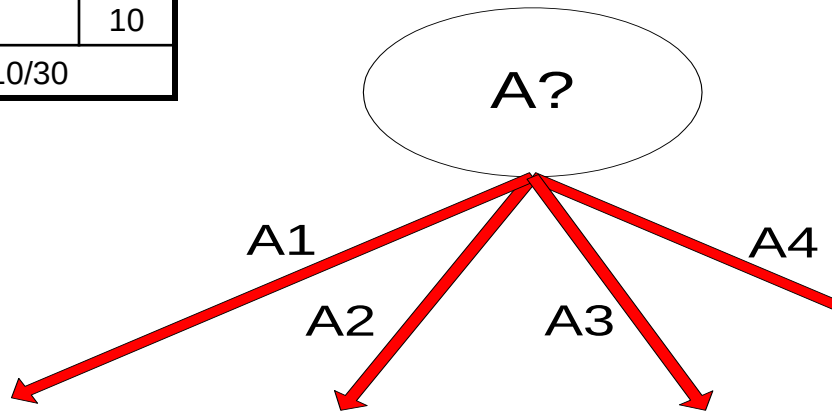
- Pre-Pruning
  - Κανόνας πρόωρου τερματισμού
- Σταμάτα τον αλγόριθμο πριν σχηματιστεί ένα πλήρες δέντρο
- Συνηθισμένες συνθήκες τερματισμού για έναν κόμβο:
  - Σταμάτα όταν όλες οι εγγραφές ανήκουν στην ίδια κλάση
  - Σταμάτα όταν όλες οι τιμές των γνωρισμάτων είναι οι ίδιες
- Ποιο περιοριστικές συνθήκες:
  - Σταμάτα όταν ο αριθμός των εγγραφών είναι μικρότερος από κάποιο προκαθορισμένο κατώφλι
  - Σταμάτα όταν η επέκταση ενός κόμβου δεν βελτιώνει την καθαρότητα (π.χ., Gini ή information gain) ή το λάθος γενίκευσης περισσότερο από κάποιο κατώφλι. (-) δύσκολος ο καθορισμός του κατωφλιού, (-) αν και το κέρδος μικρό, κατοπινοί διαχωρισμοί μπορεί να καταλήξουν σε καλύτερα δέντρα

Διευθετώντας την περίπτωση overfitting…

- Post-Pruning
  - Ανάπτυξε το δέντρο πλήρως
  - Ψαλίδισε (Trim) τους κόμβους με λογική bottom-up
  - Αν το λάθος γενίκευσης μειώνεται με το ψαλίδισμα, αντικατέστησε το υποδέντρο με ένα φύλλο
    - οι ετικέτες κλάσεις των φύλων καθορίζονται από την πλειοψηφία των κλάσεων των εγγραφών του υποδέντρου (subtree replacement)
    - Αντικατέστησε το υποδέντρο με ένα από τα κλαδιά του (Branch), αυτό που χρησιμοποιείται συχνότερα (subtree raising)
  - Πιθανή χρήση του MDL

Παράδειγμα Post-Pruning

Class = Yes	20
Class = No	10
Error = 10/30	



Class = Yes	8
Class = No	4

Class = Yes	4
Class = No	1

Class = Yes	5
Class = No	1

Class = Yes	3
Class = No	4

Σφάλμα εκπαίδευσης (πριν το διαχωρισμό) = 10/30

Απαισιόδοξη εκτίμηση σφάλματος = (10 + 0.5)/30 = 10.5/30

Σφάλμα εκπαίδευσης (μετά το διαχωρισμό) = 9/30

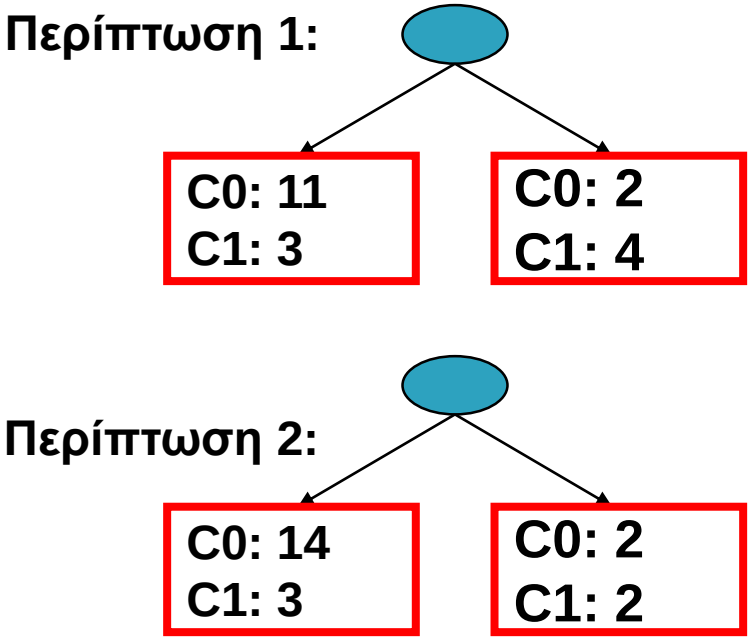
Απαισιόδοξη εκτίμηση σφάλματος

$= (9 + 4 \times 0.5)/30 = 11/30$

PRUNE!

Παράδειγμα Post-Pruning

- Optimistic?
  - Χωρίς pruning
- Pessimistic?
  - Την περίπτωση 2 μόνο, όχι την περίπτωση 1
- Reduced error pruning?
  - Εξαρτάται από το σύνολο επαλήθευσης



Ελλιπείς τιμές (missing values)

- Οι τιμές που λείπουν επηρεάζουν την κατασκευή του δέντρου με τρεις τρόπους
  - Πως υπολογίζονται τα μέτρα καθαρότητας
  - Πως κατανέμονται στα φύλλα οι εγγραφές με τιμές που λείπουν
  - Πως ταξινομείται μια εγγραφή εκπαίδευσης στην οποία λείπει μια τιμή

Υπολογισμός μέτρου καθαρότητας

Tid	Refund	Marital Status	Taxable Income	Class
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	?	Single	90K	Yes

Τιμή που λείπει

Πριν το διαχωρισμό: Entropy(Parent) = -0.3 log(0.3)-(0.7)log(0.7) = 0.8813

	Class = Yes	Class = No
Refund=Yes	**0**	**3**
Refund=No	**2**	**4**
Refund=?	**1**	**0**

Διαχωρισμός στην ιδιότητα Refund:

- Entropy(Refund=Yes) = 0
- Entropy(Refund=No)
  = -(2/6)log(2/6) – (4/6)log(4/6) = 0.9183
- Entropy(Children)
  = **0.3** (0) + **0.6** (0.9183) = 0.551

•Gain = **0.9** × (0.8813 – 0.551) = 0.3303

Ελλιπείς τιμές

Σε ποιο φύλλο θα τοποθετηθεί η εγγραφή 10;

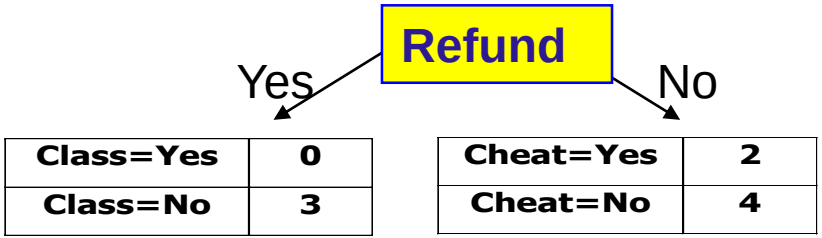
Tid	Refund	Marital Status	Taxable Income	Class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No

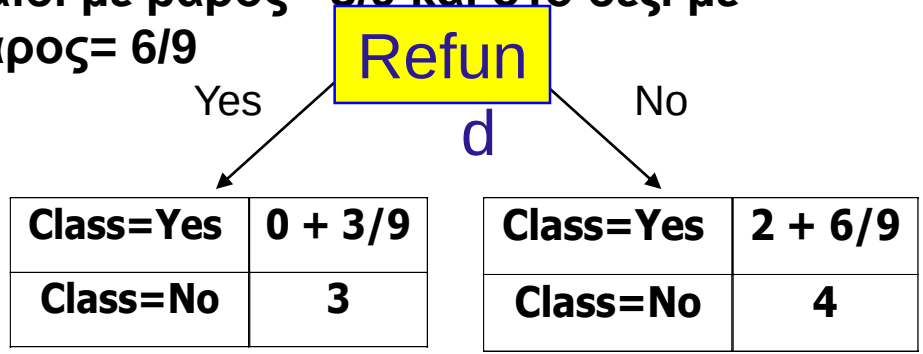
Tid	Refund	Marital Status	Taxable Income	Class
10	?	Single	90K	Yes

Πιθανότητα Refund=Yes→3/9 (3 από τις 9 εγγραφές έχουν refund=Yes)

Πιθανότητα Refund=No → 6/9

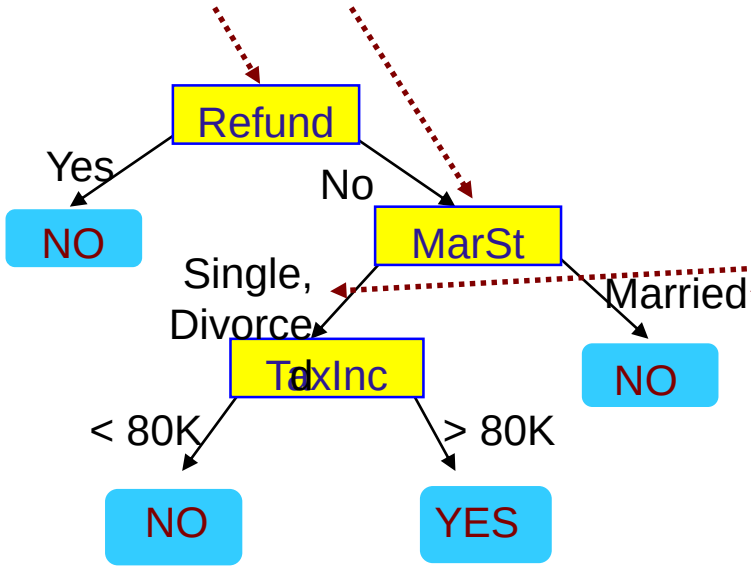
Η εγγραφή τοποθετείται στο αριστερό παιδί με βάρος= 3/9 και στο δεξί με βάρος= 6/9





Κατηγοριοποίηση εγγραφών με ελλιπείς τιμές

Tid	Refund	Marital Status	Taxable Income	Class
11	No	?	85K	?



Refund=NO	Married	Single	Divorced	Σύνολο
Class=No	3	1	0	4
Class=Yes	6/9	1	1	2.67
Σύνολο	3.67	2	1	6.67

Πιθανότητα Marital Status = Married → 3.67/6.67

Πιθανότητα Marital Status ={Single,Divorced} → 3/6.67

Άλλα ζητήματα-κατακερματισμός δεδομένων

- Ο αριθμός των παραδειγμάτων γίνεται μικρότερος όσο διανύουμε το δέντρο προς τα κάτω
- Ο αριθμός των παραδειγμάτων στους κόμβους με ετικέτες κλάσης μπορεί να είναι πολύ μικρός για να πραγματοποιήσουμε στατιστικά σημαντικές αποφάσεις

Άλλα ζητήματα-στρατηγική αναζήτησης

- Η εύρεση ενός βέλτιστου δέντρου απόφασης είναι NP-hard
- Ο αλγόριθμος που παρουσιάστηκε χρησιμοποιεί μια top-down αναδρομική στρατηγική κατακερματισμού
- Άλλες στρατηγικές
  - Bottom-up
  - Bi-directional

Άλλα ζητήματα - εκφραστικότητα

- Δυνατότητα αναπαράστασης για συναρτήσεις διακριτών τιμών, αλλά δε δουλεύουν σε κάποια είδη δυαδικών προβλημάτων –
  - πχ, parity 0(1) αν υπάρχει μονός (ζυγός) αριθμός από δυαδικά γνωρίσματα 2^d κόμβοι για d ιδιότητες
- Όχι καλή συμπεριφορά για συνεχείς μεταβλητές
  - Ιδιαίτερα όταν η συνθήκη ελέγχου αφορά ένα γνώρισμα τη φορά

Αποτίμηση Μοντέλου

- Επιλογή Μοντέλου (model selection): το μοντέλο που έχει την απαιτούμενη πολυπλοκότητα χρησιμοποιώντας την εκτίμηση του λάθους γενίκευσης
- Αφού κατασκευαστεί μπορεί να χρησιμοποιηθεί στα δεδομένα ελέγχου για να προβλέψει σε ποιες κλάσεις ανήκουν
- Για να γίνει αυτό πρέπει να ξέρουμε τις κλάσεις των δεδομένων ελέγχου

Αποτίμηση Μοντέλου

- Μέτρα (metrics) για την εκτίμηση της απόδοσης του μοντέλου
  - Πως να εκτιμήσουμε την απόδοση ενός μοντέλου
- Μέθοδοι για την εκτίμηση της απόδοσης
  - Πως μπορούνε να πάρουμε αξιόπιστες εκτιμήσεις
- Μέθοδοι για την σύγκριση μοντέλων
  - Πως να συγκρίνουμε τη σχετική απόδοση δύο ανταγωνιστικών μοντέλων

Μέτρα εκτίμησης

Έμφαση στην ικανότητα πρόβλεψης του μοντέλου παρά στην αποδοτικότητα (πόσο γρήγορα κατασκευάζει το μοντέλο ή ταξινομεί μια εγγραφή, κλιμάκωση κλπ.)

Confusion Matrix (Πίνακας Σύγχυσης)

f_{ij}: αριθμός των εγγραφών της κλάσης i που προβλέπονται ως κλάση j

	ΠΡΟΒΛΕΥΘΕΙΣΑ ΚΛΑΣΗ		
ΠΡΑΓΜΑΤΙΚΗ ΚΛΑΣΗ		Class=Yes	Class=No
	Class=Yes	f_{11}	f_{10}
	Class=No	f_{01}	f_{00}

a: TP (true positive) f_{11}

b: FN (false negative) f_{10}

c: FP (false positive) f_{01}

d: TN (true negative) f_{00}

Μέτρα εκτίμησης - πιστότητα

Πιστότητα (accuracy) Το πιο συνηθισμένο μέτρο

	ΠΡΟΒΛΕΥΘΕΙΣΑ ΚΛΑΣΗ		
ΠΡΑΓΜΑΤΙΚΗ ΚΛΑΣΗ		Class=Yes	Class=No
	Class=Yes	TP	FN
	Class=No	FP	TN

$$\text{Accuracy} = \frac{f_{11} + f_{00}}{f_{11} + f_{00} + f_{01} + f_{10}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Error rate: $$\rho\upsilon\theta\mu\acute{o}\varsigma\sigma\phi\acute{\alpha}\lambda\mu\alpha\tau\varsigma = \frac{f_{01} + f_{10}}{f_{11} + f_{00} + f_{01} + f_{10}}$$

Μειονεκτήματα της πιστότητας

- Θεωρείστε ένα πρόβλημα με 2 κλάσεις
  - Αριθμός παραδειγμάτων της κλάσης 0 = 9990
  - Αριθμός παραδειγμάτων της κλάσης 1 = 10
- Αν ένα μοντέλο προβλέπει οτιδήποτε ως κλάση 0 τότε πιστότητα = 9990/10000 = 99.9 %
  - Η πιστότητα είναι παραπλανητική γιατί το μοντέλο δεν προβλέπει κανένα παράδειγμα της κλάσης 1

Πίνακας κόστους

	ΠΡΟΒΛΕΥΘΕΙΣΑ ΚΛΑΣΗ		
ΠΡΑΓΜΑΤΙΚΗ ΚΛΑΣΗ	C(i|j)	**Class=Yes**	**Class=No**
	Class=Yes	C(Yes|Yes)	C(Yes|No)
	Class=No	C(No|Yes)	C(No|No)

C(i|j): **κόστος** λανθασμένης ταξινόμησης ενός παραδείγματος της κλάσης i ως κλάση j

C(M) = TP x C(Yes|Yes) + FN x C(Yes|No) + FP C(No|Yes) + TN C(No|No)

Υπολογισμός του κόστους ταξινόμησης

Cost Matrix	ΠΡΟΒΛΕΥΘΕΙΣΑ ΚΛΑΣΗ		
ΠΡΑΓΜΑΤΙΚΗ ΚΛΑΣΗ	C(i|j)	+	-
	+	-1	100
	-	1	0

C(i|j): κόστος λανθασμένης ταξινόμησης ενός παραδείγματος της κλάσης i ως κλάση j

ΜΟΝΤΕΛΟ M_1	ΠΡΟΒΛΕΥΘΕΙΣΑ ΚΛΑΣΗ		
ΠΡΑΓΜΑΤΙΚΗ ΚΛΑΣΗ		+	-
	+	150	40
	-	60	250

Πιστότητα= 80%

Κόστος= 3910

ΜΟΝΤΕΛΟ M_2	ΠΡΟΒΛΕΥΘΕΙΣΑ ΚΛΑΣΗ		
ΠΡΑΓΜΑΤΙΚΗ ΚΛΑΣΗ		+	-
	+	250	45
	-	5	200

Πιστότητα= 90%

Κόστος = 4255

Κόστος - Πιστότητα

Μετρήσεις	ΠΡΟΒΛΕΥΘΕΙΣΑ		
ΠΡΑΓΜΑΤΙΚΗ		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

Κόστη	ΠΡΟΒΛΕΥΘΕΙΣΑ		
ΠΡΑΓΜΑΤΙΚΗ		Class=Yes	Class=No
	Class=Yes	p	q
	Class=No	q	p

Η πιστότητα είναι ανάλογη του κόστους αν:

1. C(Yes|No)=C(No|Yes) = q
2. C(Yes|Yes)=C(No|No) = p

N = a + b + c + d

πιστότητα= (a + d)/N

κόστος= p (a + d) + q (b + c)

= p (a + d) + q (N – a – d)

= q N – (q – p)(a + d)

= N [q – (q-p) × πιστότητα]

Άλλες μετρήσεις με βάση τον πίνακα σύγχυσης

	ΠΡΟΒΛΕΥΘΕΙΣΑ ΚΛΑΣΗ		
ΠΡΑΓΜΑΤΙΚΗ ΚΛΑΣΗ		Class=Yes	Class=No
	Class=Yes	TP	FN
	Class=No	FP	TN

True positive rate or sensitivity: Το ποσοστό των θετικών παραδειγμάτων που ταξινομούνται σωστά

$$\mathrm{TPR} = \frac{TP}{TP + FN}$$

True negative rate or specificity: Το ποσοστό των αρνητικών παραδειγμάτων που ταξινομούνται σωστά

$$\mathrm{TNR} = \frac{TN}{TN + FP}$$

Άλλες μετρήσεις με βάση τον πίνακα σύγχυσης

	ΠΡΟΒΛΕΥΘΕΙΣΑ ΚΛΑΣΗ		
ΠΡΑΓΜΑΤΙΚΗ ΚΛΑΣΗ		Class=Yes	Class=No
	Class=Yes	TP	FN
	Class=No	FP	TN

False positive rate: Το ποσοστό των αρνητικών παραδειγμάτων που ταξινομούνται λάθος (δηλαδή, ως θετικά)

$$\mathrm{FPR} = \frac{FP}{TN + FP}$$

False negative rate: Το ποσοστό των θετικών παραδειγμάτων που ταξινομούνται λάθος (δηλαδή, ως αρνητικά)

$$\mathrm{FNR} = \frac{FN}{TP + FN}$$

Ακρίβεια (precision) – Ανάκληση (recall)

	ΠΡΟΒΛΕΥΘΕΙΣΑ ΚΛΑΣΗ		
ΠΡΑΓΜΑΤΙΚΗ ΚΛΑΣΗ		Class=Yes	Class=No
	Class=Yes	TP	FN
	Class=No	FP	TN

Precision $p = \frac{TP}{TP + FP}$

Πόσα από τα παραδείγματα που ο ταξινομητής έχει ταξινομήσει ως θετικά είναι πραγματικά θετικά Όσο πιο μεγάλη η ακρίβεια, τόσο μικρότερος ο αριθμός των FP

Recall $r = \frac{TP}{TP + FN}$

Πόσα από τα θετικά παραδείγματα κατάφερε ο ταξινομητής να βρει Όσο πιο μεγάλη η ανάκληση, τόσο λιγότερα θετικά παραδείγματα έχουν ταξινομεί λάθος (=TPR)

Ακρίβεια (precision) – Ανάκληση (recall)

Precision

$$\mathrm{p} = \frac{TP}{TP + FP}$$

Πόσα από τα παραδείγματα που ο ταξινομητής έχει ταξινομήσει ως θετικά είναι πραγματικά θετικά

Recall

$$\mathrm{r} = \frac{TP}{TP + FN}$$

Πόσα από τα θετικά παραδείγματα κατάφερε ο ταξινομητής να βρει

Συχνά το ένα καλό και το άλλο όχι

Πχ, ένας ταξινομητής που όλα τα ταξινομεί ως θετικά, την καλύτερη ανάκληση με τη χειρότερη ακρίβεια

Πώς να τα συνδυάσουμε;

F-measure

$$\mathrm{F} = \frac{2rp}{r + p} = \frac{2TP}{2TP + FP + FN}$$

$$\mathrm{F} = \frac{2}{1/r + 1/p}$$

Αρμονικό μέσο (Harmonic mean)

Τείνει να είναι πιο κοντά στο μικρότερο από τα δύο

Υψηλή τιμή σημαίνει ότι και τα δύο είναι ικανοποιητικά μεγάλα

Μέθοδοι αποτίμησης μοντέλου-Hold out

- Διαμέριση του αρχικού συνόλου σε δύο ξένα σύνολα:
- Σύνολο εκπαίδευσης – Σύνολο Ελέγχου
- Κατασκευή μοντέλου με βάση το σύνολο εκπαίδευσης
- Αποτίμηση μοντέλου με βάση το σύνολο ελέγχου

Μέθοδοι αποτίμησης μοντέλου-Hold out

- Λιγότερες εγγραφές για εκπαίδευση – πιθανόν όχι τόσο καλό μοντέλο, όσο αν χρησιμοποιούνταν όλες
- Τα σύνολα ελέγχου και εκπαίδευσης δεν είναι ανεξάρτητα μεταξύ τους (πχ μια κλάση που έχει πολλά δείγματα στο ένα, θα έχει λίγα στο άλλο και το ανάποδο)
- Το μοντέλο εξαρτάται από τη σύνθεση των συνόλων εκπαίδευσης και ελέγχου – όσο μικρότερο το σύνολο εκπαίδευσης, τόσο μεγαλύτερη η variance του μοντέλου – όσο μεγαλύτερο το σύνολο εκπαίδευσης, τόσο λιγότερο αξιόπιστη η πιστότητα του μοντέλου που υπολογίζεται με το σύνολο ελέγχου – wide confidence interval

Μέθοδοι αποτίμησης- Τυχαία Λήψη Δειγμάτων – Random Subsampling

- Επανάληψη της μεθόδου για τη βελτίωση του μοντέλου
- Cross-validation
- Διαμοίραση των δεδομένων σε k διαστήματα
- Κατασκευή του μοντέλου αφήνοντας κάθε φορά ένα διάστημα ως σύνολο ελέγχου και χρησιμοποιώντας τα υπόλοιπα ως σύνολα εκπαίδευσης
- Αν k = N, leave-one-out

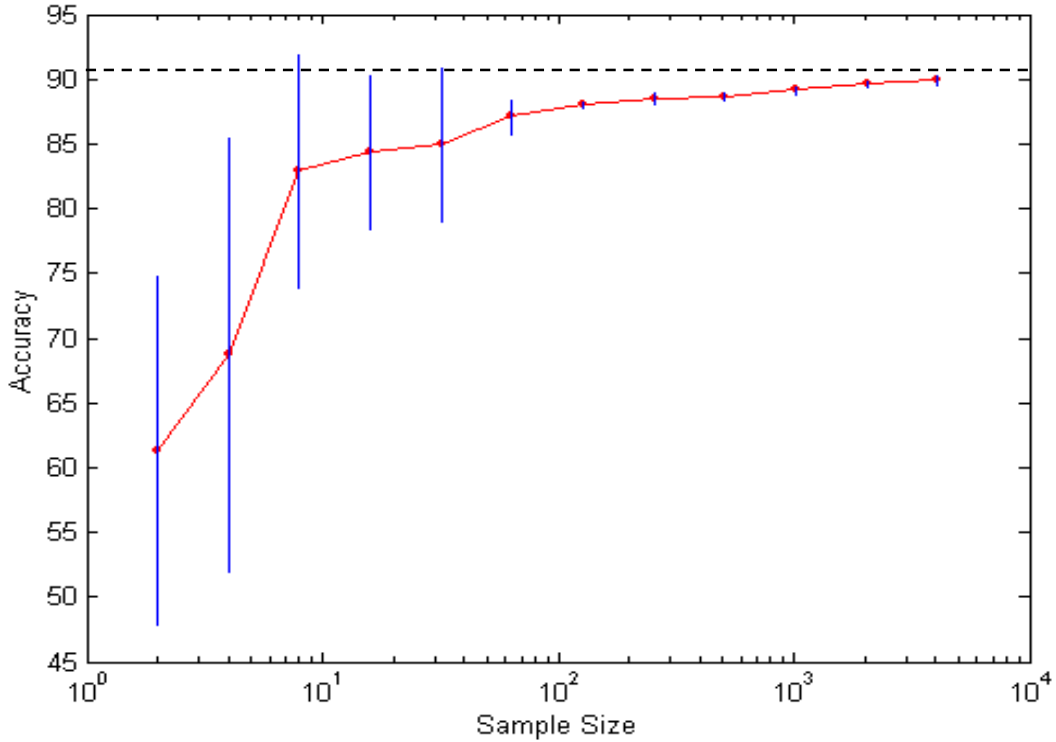
Μέθοδοι αποτίμησης μοντέλου-Bootstrap

- Δειγματοληψία με αντικατάσταση (sample with replacement)
- Μια εγγραφή που επιλέχθηκε ως δεδομένο εκπαίδευσης, ξαναμπαίνει στο αρχικό σύνολο
- Αν N δεδομένα, ένα δείγμα N στοιχείων 63.2% των αρχικών
- Πιθανότητα ένα δεδομένο να επιλεγεί $1 – (1-1/N)^N$
- Για μεγάλο N, τείνει ασυμπτωτικά στο $1-e^{-1} = 0.632$
- Οι υπόλοιπες εγγραφές – εγγραφές ελέγχου

Μέθοδοι αποτίμησης μοντέλου

- Πως μπορούμε να πάρουμε αξιόπιστες εκτιμήσεις της απόδοσης
- Η απόδοση ενός μοντέλου μπορεί να εξαρτάται από πολλούς παράγοντες εκτός του αλγορίθμου μάθησης:
  - Κατανομή των κλάσεων
  - Το κόστος της λανθασμένης ταξινόμησης
  - Το μέγεθος του συνόλου εκπαίδευσης και του συνόλου ελέγχου

Καμπύλη μάθησης (learning curve)



- Η καμπύλη μάθησης δείχνει πως μεταβάλλεται η πιστότητα με την αύξηση του μεγέθους του δείγματος
- Απαιτεί μια στρατηγική δειγματοληψίας
  - Αριθμητική (langley et al)
  - Γεωμετρική (provost et al)
- Επίδραση μικρού δείγματος
  - Πόλωση στην εκτίμηση
  - Διασπορά της εκτίμησης

ROC (Receiver Operating Characteristic Curve)

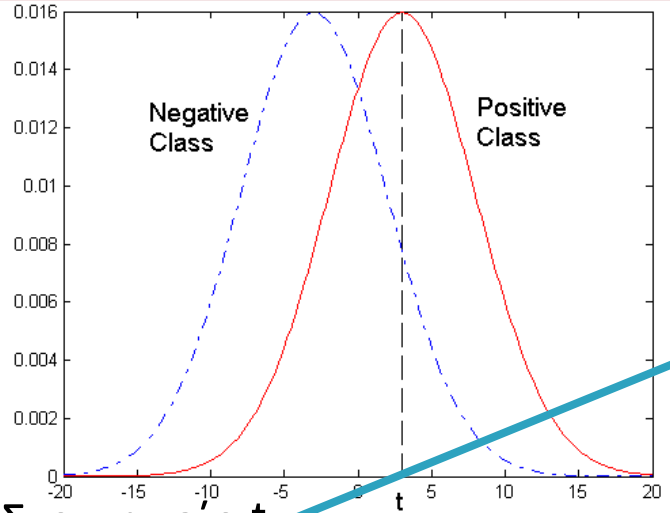
- Αναπτύχθηκε στη δεκαετία 1950 για την ανάλυση θορύβου στα σήματα
  - Χαρακτηρίζει το trade-off μεταξύ positive hits και false alarms
- Η καμπύλη ROC δείχνει τα TPR (στον άξονα των y) προς τα FPR (στον άξονα των x)
- Η απόδοση κάθε ταξινομητή αναπαρίσταται ως ένα σημείο στην καμπύλη ROC

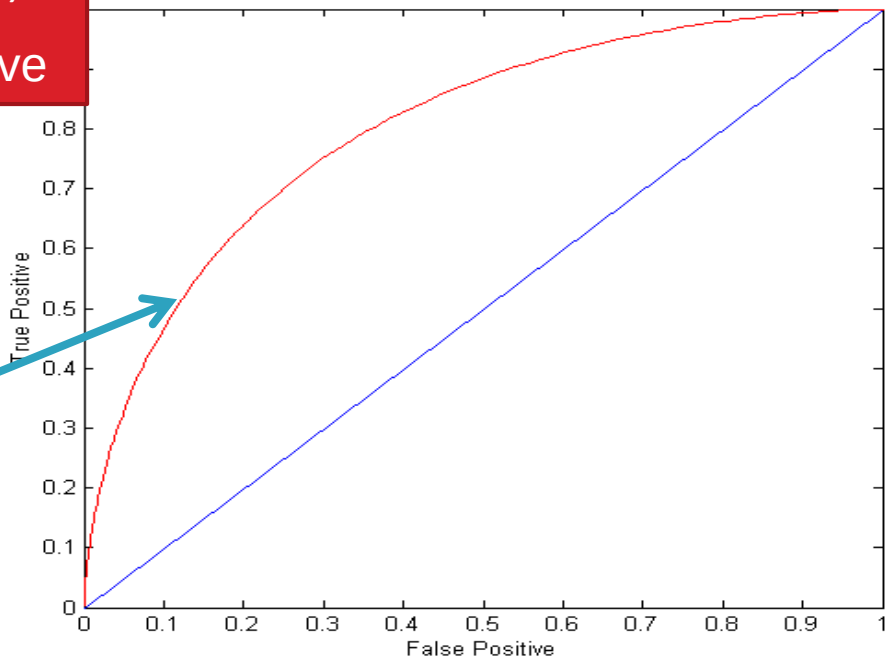
$$\mathrm{TPR} = \frac{TP}{TP + FN}$$

$$\mathrm{FPR} = \frac{FP}{TN + FP}$$

ROC (Receiver Operating Characteristic Curve)

- Μονοδιάσταστο σώμα δεδομένων με 2 κλάσεις
- Όποιο σημείο έχει x > t ταξινομείται ως Positive





Στο σημείο t

TP=0.5, FN=0.5, FP=0.12,

Αποτίμηση Μοντέλου: ROC

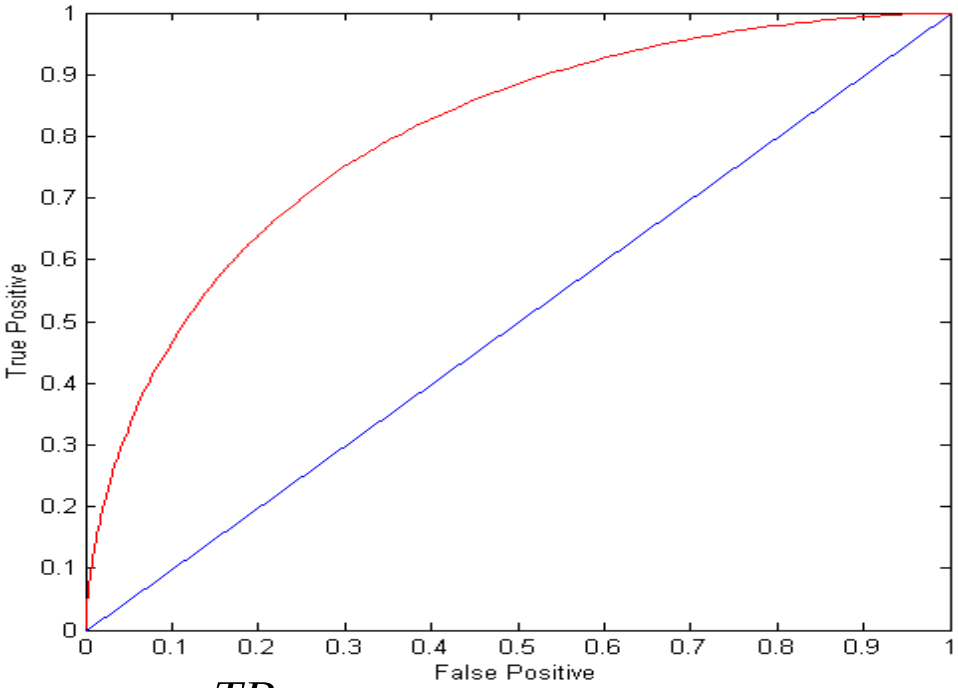
(TP,FP):

- (0,0): όλα σημειώνονται ως αρνητική κλάση
- (1,1): όλα σημειώνονται ως θετική κλάση
- (1,0): ιδανική ταξινόμηση

Διαγώνιος:

- Τυχαία εκτίμηση κλάσης

  Μια εγγραφή θεωρείται θετική με καθορισμένη πιθανότητα p ανεξάρτητα από τις τιμές των γνωρισμάτων της



$$\mathrm{TPR} = \frac{TP}{TP + FN}$$

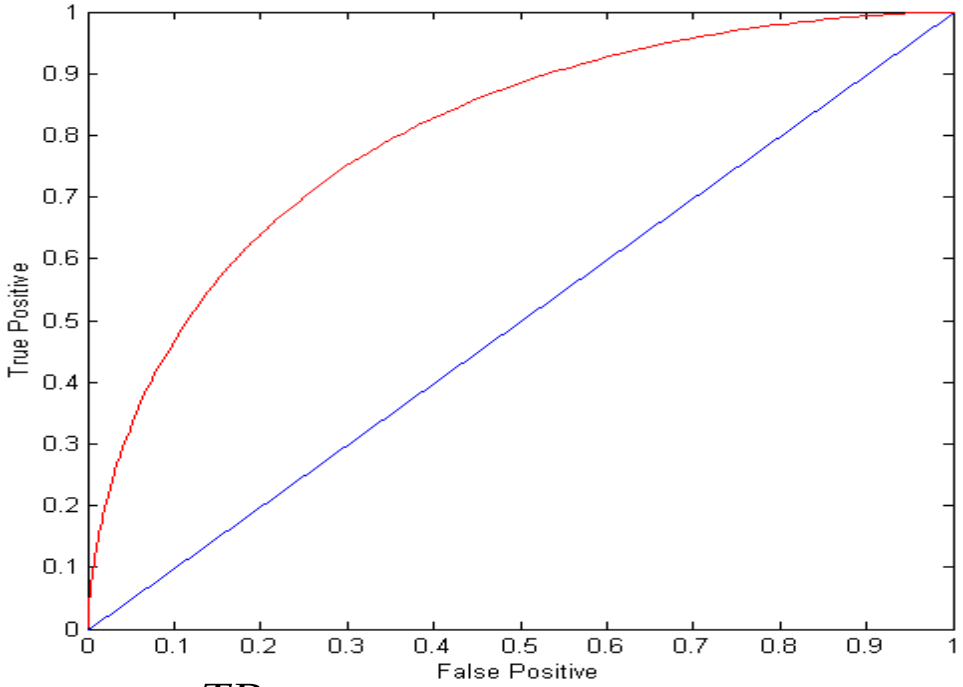
$$\mathrm{FPR} = \frac{FP}{TN + FP}$$

Αποτίμηση Μοντέλου: ROC

Καλοί ταξινομητές κοντά στην αριστερή πάνω γωνία του διαγράμματος

Κάτω από τη διαγώνιο

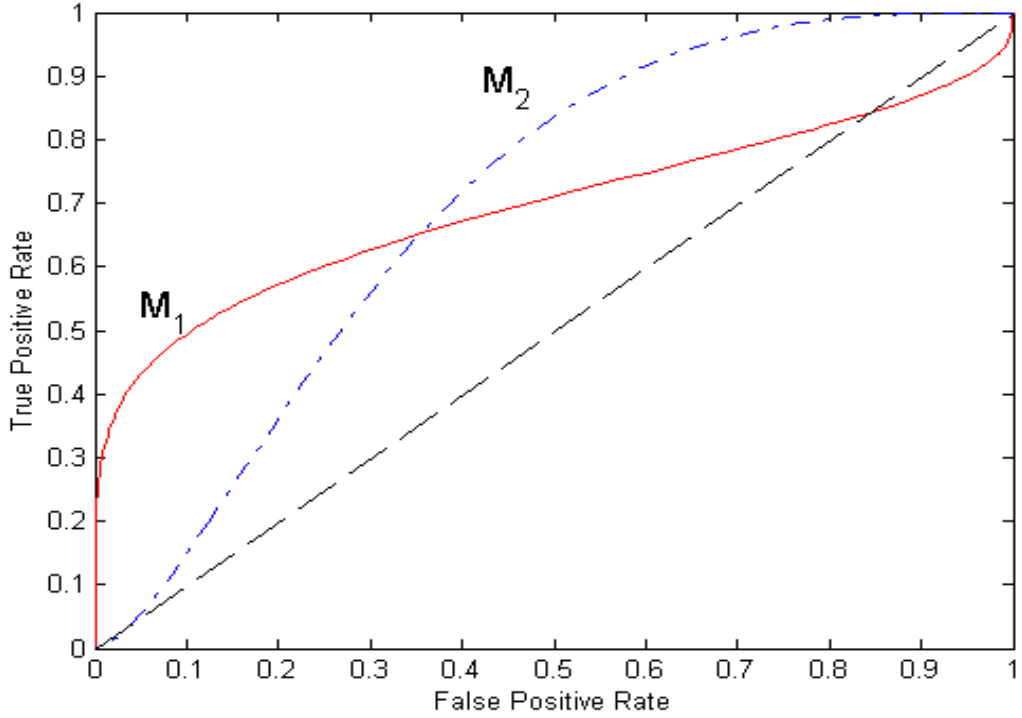
Πρόβλεψη είναι το αντίθετο της πραγματικής κλάσης



$$\mathrm{TPR} = \frac{TP}{TP + FN}$$

$$\mathrm{FPR} = \frac{FP}{TN + FP}$$

ROC: Σύγκριση 2 μοντέλων ταξινόμησης



- Κανένα μοντέλο δεν υπερτερεί του άλλου
  - M_1 είναι καλύτερο για μικρό FPR
  - M_2 είναι καλύτερο για μεγάλο FPR

Κατασκευή ROC

Διάνυσμα	P(+|A)	Πραγματική κλάση
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

- Χρησιμοποίηση ενός ταξινομητή που επιστρέφει τη μεταγενέστερη πιθανότητα p(+|A)
- Ταξινόμηση με φθίνουσα σειρά
- Για κάθε τιμή p(+|A)
  - Υπολογισμός TP, TN, FP, FN
  - TPR=TP/(TP+FN)
  - FPR=FP/(FP+TN)

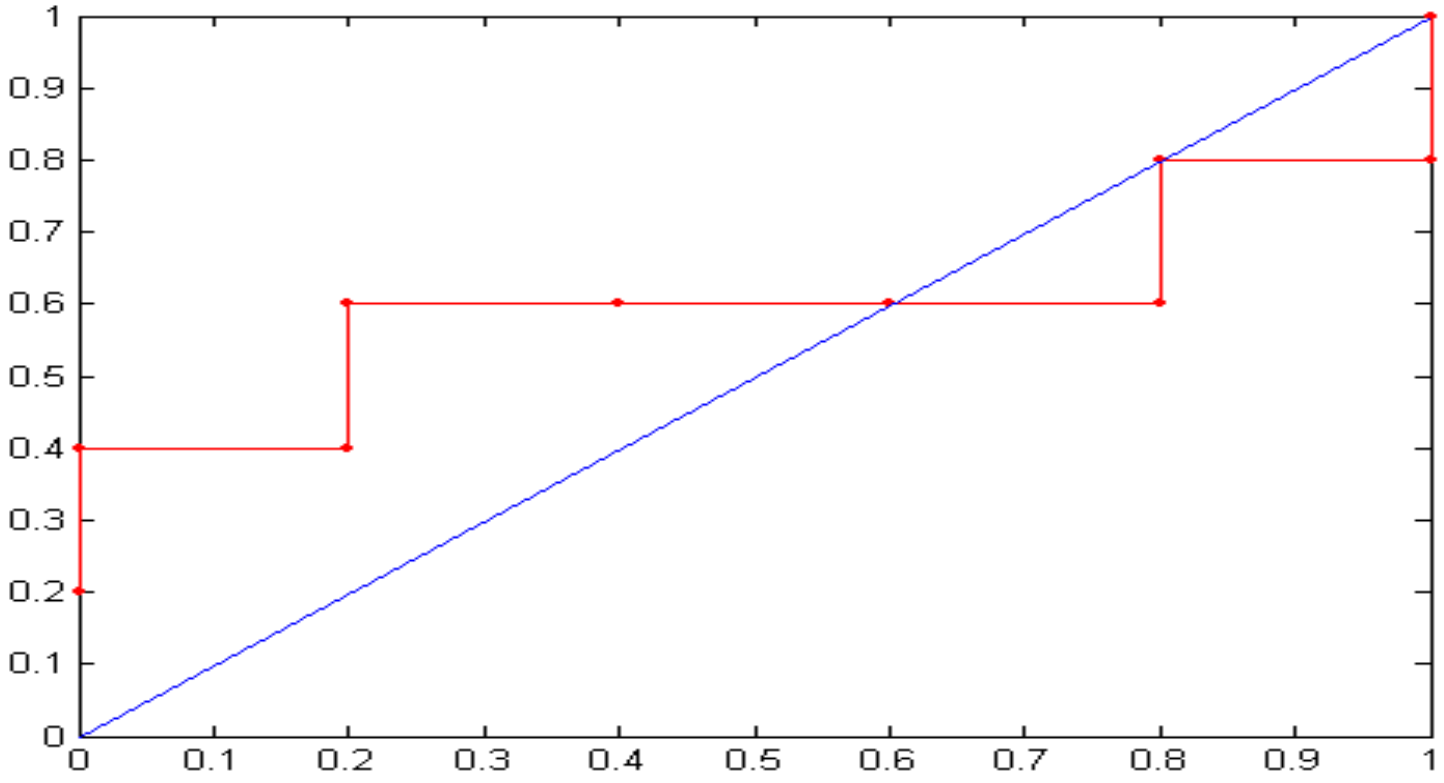
Κατασκευή ROC

1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

Class	+	-	+	-	-	-	+	-	+	+	
	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

Κατασκευή ROC

Η καμπύλη ROC που προκύπτει



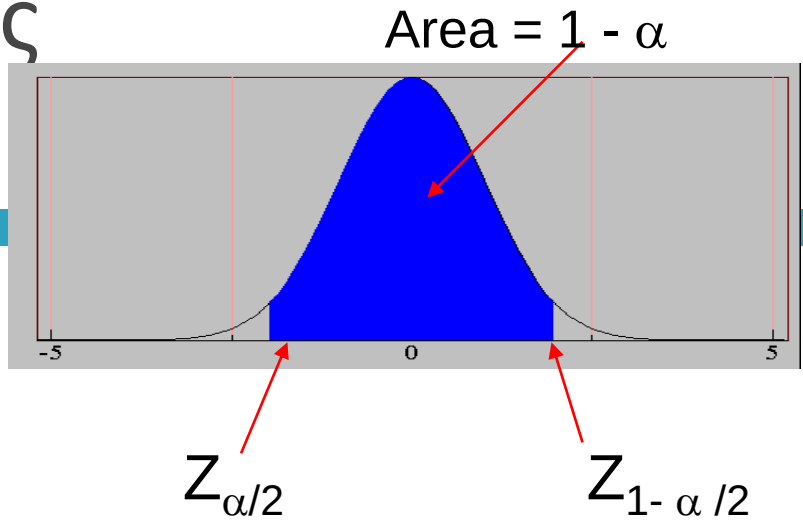
Έλεγχος για τη σύγκριση μοντέλων

- Έστω δύο μοντέλα:
  - Μοντέλο Μ1: ακρίβεια = 85%, έλεγχος σε 30 εγγραφές
  - Μοντέλο Μ2: ακρίβεια = 75%, έλεγχος σε 5000 εγγραφές
- Είναι το Μ1 καλύτερο από το Μ2?
  - Πόση εμπιστοσύνη (confidence) μπορούμε να έχουμε για την πιστότητα του Μ1 και πόση για την πιστότητα του Μ2;
  - Μπορεί η διαφορά στην απόδοση να αποδοθεί σε τυχαία διακύμανση του συνόλου ελέγχου;

Διάστημα εμπιστοσύνης ακρίβειας

- Η πρόβλεψη μπορεί να θεωρηθεί ως μια δοκιμή Bernoulli
  - Μια δοκιμή έχει 2 εξόδους
  - Πιθανές έξοδοι για την πρόβλεψη
    - Σωστό, λάθος
- Μια συλλογή δοκιμών Bernoulli ακολουθεί τη διωνυμική κατανομή
  - $x \sim Bin(N, p)$ x: αριθμός σωστών προβλέψεων
  - e.g: αν ρίξουμε ένα νόμισμα 50 φορές, πόσες θα έρθει κορώνα;
    - $N \times p = 50 \times 0.5 = 25$

Διάστημα εμπιστοσύνης ακρίβειας



- Δοθέντος του x (# σωστών προβλέψεων) ή ισοδύναμα, acc=x/N, και του N (# εγγραφών ελέγχου),
  - Μπορούμε να προβλέψουμε το p (την πραγματική πιστότητα του μοντέλο);
- Για μεγάλα σύνολα ελέγχου (N>30)
  - Η ακρίβεια ακολουθεί κανονική κατανομή με μέση τιμή p και διασπορά p(1-p)/N

$$P(Z_{\alpha/2} < \frac{acc - p}{\sqrt{p(1-p)/N}} < Z_{1-\alpha/2}) = 1-\alpha$$

Διάστημα εμπιστοσύνης

$$p = \frac{2 \times N \times acc + Z^2_{\alpha/2} \pm \sqrt{Z^2_{\alpha/2} + 4 \times N \times acc - 4 \times N \times acc^2}}{2(N + Z^2_{\alpha/2})}$$

Διάστημα εμπιστοσύνης ακρίβειας

- Έστω ένα μοντέλο με πιστότητα 80% όταν αξιολογήθηκε με 100 παραδείγματα:
  - N=100, acc = 0.8
  - Έστω 1-α = 0.95 (95% εμπιστοσύνη)
  - Από τον πίνακα πιθανοτήτων. $Z_{\alpha/2}=1.96$

1-α	Z
0.99	2.58
0.98	2.33
0.95	1.96
0.90	1.65

N	50	100	500	1000	5000
p(lower)	0.670	0.711	0.763	0.774	0.789
p(upper)	0.888	0.866	0.833	0.824	0.811

Σύγκριση απόδοσης 2 μοντέλων

- Δοσμένων δυο μοντέλων M1 και M2, ποιο είναι καλύτερο;
  - M1 αξιολογείται στο D1 (μέγεθος=n1), με ρυθμό λάθους= e_1
  - M2 αξιολογείται στο D2 (μέγεθος=n2), με ρυθμό λάθους= e_2

- Έστω D1 και D2 ανεξάρτητα
- Αν τα n1 και n2 είναι επαρκώς μεγάλα

$$e_1 \sim N(\mu_1, \sigma_1)$$

$$e_2 \sim N(\mu_2, \sigma_2)$$

- Προσέγγιση

$$\hat{\sigma}_i = \frac{e_i(1-e_i)}{n_i}$$

Σύγκριση απόδοσης 2 μοντέλων

- Για να ελέγξουμε αν η απόδοση διαφέρει στατιστικά σημαντικά: d=e1-e2
  - $d \sim N(d_t, \sigma_t)$ όπου d_t είναι η πραγματική διαφορά
  - Αφού τα D1 και D2 είναι ανεξάρτητα, η διασπορά τους προστίθεται

$$\sigma_t^2 = \sigma_1^2 + \sigma_2^2 \cong \hat{\sigma}_1^2 + \hat{\sigma}_2^2$$

$$= \frac{e1(1-e1)}{n1} + \frac{e2(1-e2)}{n2}$$

- Στο επίπεδο εμπιστοσύνης (1-α)
  - $d_t = d \pm Z_{\alpha/2}\hat{\sigma}_t$

Ένα παράδειγμα

- Έστω: M1: n1 = 30, e1 = 0.15
  M2: n2 = 5000, e2 = 0.25
- d = |e2 – e1| = 0.1

$$\hat{\sigma}_d = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

- Σε επίπεδο εμπιστοσύνης 95%, $Z_{\alpha/2}$=1.96

$$d_t = 0.100 \pm 1.96 \times \sqrt{0.0043} = 0.100 \pm 0.128$$

=> το διάστημα περιέχει το 0 => η διαφορά δεν είναι στατιστικά σημαντική

Σύγκριση απόδοσης 2 αλγόριθμων

- Κάθε αλγόριθμος μάθησης παράγει k μοντέλα
  - Ο L1 τα M11 , M12, …, M1k
  - Ο L2 τα M21 , M22, …, M2k
- Αν τα μοντέλα παρήχθησαν από τα ίδια σώματα αξιολόγησης D1,D2, …, Dk (π.χ. με cross-validation)
  - Για κάθε σύνολο: $d_j = e_{1j} - e_{2j}$
- d_j έχει μέση τιμή d_t και διασπορά σ_t
  - Υπολόγισε:

$$\hat{\sigma}_t^2 = \frac{\sum_{j=1}^{k} (d_j - \bar{d})^2}{k(k-1)}$$

$$d_t = d \pm t_{1-\alpha,k-1} \hat{\sigma}_t$$