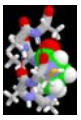
Ταξινόμηση Ι

Οι διαφάνειες στηρίζονται στο **P.-N. Tan, M.Steinbach, V. Kumar, «Introduction to Data Mining»**, Addison Wesley, 2006

Εισαγωγή

Ταξινόμηση (classification)
Το γενικό πρόβλημα της ανάθεσης ενός αντικειμένου σε μια ή περισσότερες *προκαθορισμένες* κατηγορίες (κλάσεις)



Παραδείγματα

- Εντοπισμός spam emails, με βάση πχ την επικεφαλίδα τους ή το περιεχόμενό τους
- Πρόβλεψη καρκινικών κυττάρων χαρακτηρίζοντας τα ως καλοήθη ή κακοήθη
- Κατηγοριοποίηση συναλλαγών με πιστωτικές κάρτες ως νόμιμες ή προϊόν απάτης
- Κατηγοριοποίηση δευτερευόντων δομών πρωτείνης ως alpha-helix, beta-sheet, ή random coil
- Χαρακτηρισμός ειδήσεων ως οικονομικές, αθλητικές, πολιτιστικές, πρόβλεψης καιρού, κλπ

Ορισμός

Είσοδος: συλλογή από εγγραφές
Κάθε εγγραφή περιέχει ένα σύνολο από γνωρίσματα (attributes)
Ένα από τα γνωρίσματα είναι η κλάση (class)

Βρες (έξοδος) ένα μοντέλο (model) για το γνώρισμα κλάση ως μια συνάρτηση των τιμών των άλλων γνωρισμάτων

Στόχος: νέες εγγραφές θα πρέπει να ανατίθενται σε μία από τις κλάσεις με τη μεγαλύτερη δυνατή ακρίβεια.

	κατηγορικό	κατηγορικό	συνεχές	κλάση
Tid	Επιστροφή	Οικογενειακή Κατάσταση	Φορολογητέο Εισόδημα	Απάτη
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

Ορισμός

Είσοδος: συλλογή από εγγραφές
Κάθε εγγραφή περιέχει ένα σύνολο από γνωρίσματα (attributes)
Ένα από τα γνωρίσματα είναι η κλάση (class)

Βρες ένα μοντέλο (model) για το γνώρισμα κλάση ως μια συνάρτηση των τιμών των άλλων γνωρισμάτων

Στόχος: νέες εγγραφές θα πρέπει να ανατίθενται σε μία κλάση με τη μεγαλύτερη δυνατή ακρίβεια.

Ταξινόμηση είναι η διαδικασία εκμάθησης μιας συνάρτησης στόχου (target function) f που απεικονίζει κάθε σύνολο γνωρισμάτων x σε μια από τις προκαθορισμένες ετικέτες κλάσεις y.

Συνήθως το σύνολο δεδομένων εισόδου χωρίζεται σε:
- ένα σύνολο εκπαίδευσης (training set) και
- ένα σύνολο ελέγχου (test test)

Το *σύνολο εκπαίδευσης* χρησιμοποιείται για να κατασκευαστεί το μοντέλο και
το *σύνολο ελέγχου* για να το επικυρώσει.

Σύνολο εγγραφών (x)

↓

Μοντέλο Ταξινόμησης

↓

Ετικέτα κλάσης (y)

Εισαγωγή

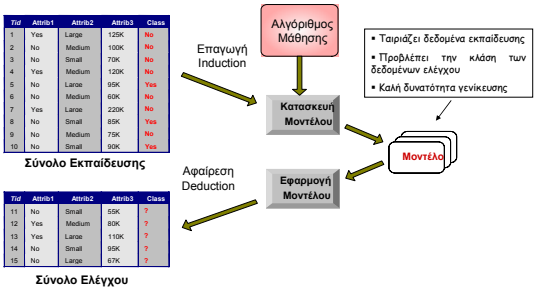
❑Χρησιμοποιείται ως:
- Περιγραφικό μοντέλο (descriptive modeling): ως επεξηγηματικό εργαλείο – πχ ποια χαρακτηριστικά κάνουν ένα ζώο να χαρακτηριστεί ως θηλαστικό
- Μοντέλο πρόβλεψης (predictive modeling): για τη πρόβλεψη της κλάσης άγνωστων εγγραφών – πχ δοσμένων των χαρακτηριστικών κάποιου ζώου να προβλέψουμε αν είναι θηλαστικό, πτηνό, ερπετό ή αμφίβιο

❑ Κατάλληλη κυρίως για:
- δυαδικές κατηγορίες ή κατηγορίες για τις οποίες δεν υπάρχει διάταξη διακριτές (nominal) vs διατεταγμένες (ordinal)
- για μη ιεραρχικές κατηγορίες

❑ Η τιμή (ετικέτα) της κλάσης – y – είναι διακριτή τιμή - Διαφορά από regression (οπισθοδρόμηση) όπου το γνώρισμα y παίρνει *συνεχείς τιμές*

Τεχνικές Ταξινόμησης

Βήματα Ταξινόμησης



Ορισμός

Τεχνικές ταξινόμησης

- Τεχνικές βασισμένες σε Δέντρα Απόφασης (decision trees)
- Τεχνικές βασισμένες σε Κανόνες (Rule-based Methods)
- Memory based reasoning
- Νευρωνικά Δίκτυα
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines

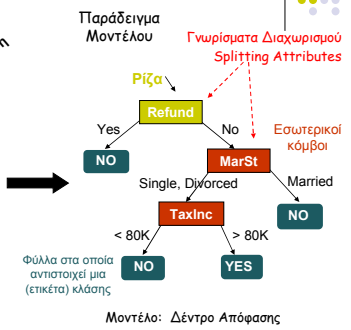
Δέντρα Απόφασης

Δέντρο Απόφασης: Παράδειγμα

Δεδομένα Εκπαίδευσης

κατηγορικό κατηγορικό συνεχές κλάση

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



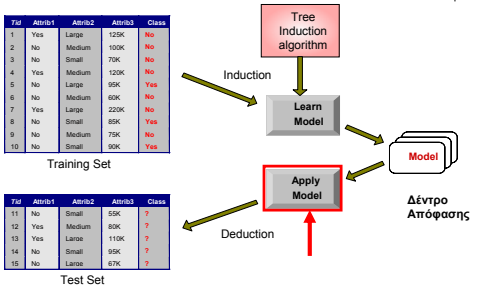
Μοντέλο: Δέντρο Απόφασης

Δέντρο Απόφασης: Παράδειγμα

Μοντέλο = Δέντρο Απόφασης

- Εσωτερικοί κόμβοι αντιστοιχούν σε κάποιο γνώρισμα
- Διαχωρισμός (split) ενός κόμβου σε παιδιά
  - η ετικέτα στην ακμή = συνθήκη/έλεγχος
- Φύλλα αντιστοιχούν σε κλάσεις

Δέντρο Απόφασης: Βήματα

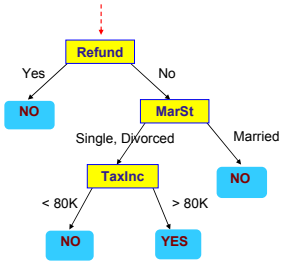


Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Ξεκίνα από τη ρίζα του δέντρου.

Δεδομένα Ελέγχου

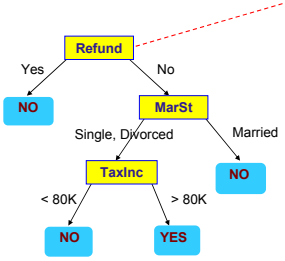
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Test Data

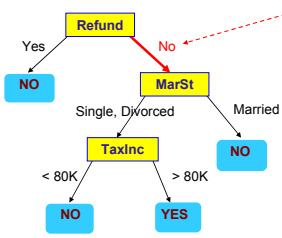
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Test Data

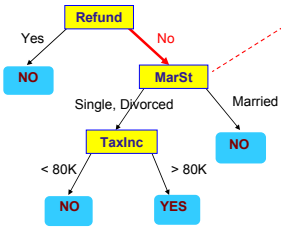
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Test Data

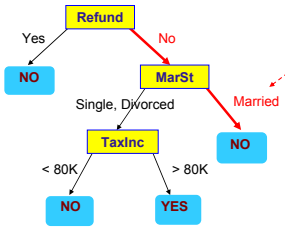
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Test Data

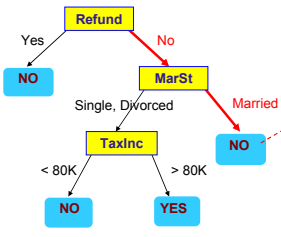
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

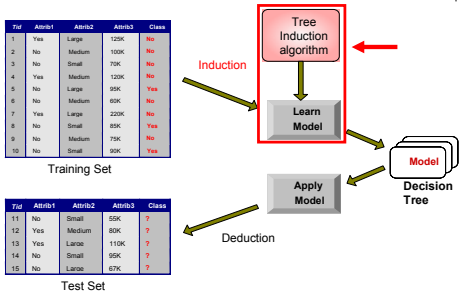
Είσοδος (δεδομένο ελέγχου)

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Ανάθεση στο Cheat “No”

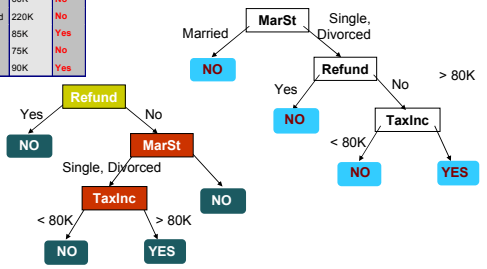
Δέντρο Απόφασης



Δέντρο Απόφασης: Παράδειγμα

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



Δέντρο Απόφασης: Κατασκευή

Ο αριθμός των πιθανών Δέντρων Απόφασης είναι εκθετικός.

Πολλοί αλγόριθμοι για την επαγωγή (induction) του δέντρου οι οποίοι ακολουθούν μια greedy στρατηγική: για να κτίσουν το δέντρο απόφασης παίρνοντας μια σειρά από *τοπικά βέλτιστες* αποφάσεις

- Hunt's Algorithm (από τους πρώτους)
- CART
- ID3, C4.5
- SLIQ, SPRINT

Δέντρο Απόφασης: Αλγόριθμος του Hunt

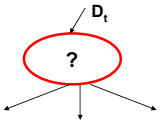
Κτίζει το δέντρο αναδρομικά

D_t: το σύνολο των εγγραφών εκπαίδευσης που έχουν φτάσει στον κόμβο t

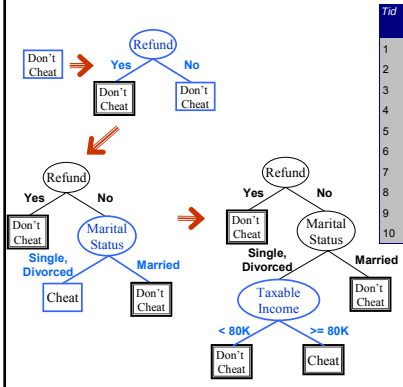
Γενική Διαδικασία:

- Αν το D_t περιέχει εγγραφές που ανήκουν στην *ίδια* κλάση y_t, τότε ο κόμβος t είναι κόμβος φύλλο με ετικέτα y_t
- Αν D_t είναι το *κενό σύνολο*, τότε ο κόμβος t είναι κόμβος φύλλο με ετικέτα την default κλάση, y_d
- 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.

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



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 D_t περιέχει εγγραφές που ανήκουν στην *ίδια* κλάση y_t, τότε ο κόμβος t είναι κόμβος φύλλο με ετικέτα y_t
- Αν D_t είναι το *κενό σύνολο*, αυτό σημαίνει ότι δεν υπάρχει εγγραφή στο σύνολο εκπαίδευσης με αυτό το συνδυασμό τιμών,
  τότε D_t γίνεται φύλλο με κλάση αυτή *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.

Το παραπάνω δεν είναι δυνατόν αν όλες οι εγγραφές έχουν τις ίδιες τιμές σε όλα τα γνωρίσματα (δηλαδή, ο ίδιος συνδυασμός αντιστοιχεί σε περισσότερες από μία κλάσεις)
τότε φύλλο με κλάση αυτής της πλειοψηφίας των εγγραφών εκπαίδευσης

Δέντρο Απόφασης: Κατασκευή Δέντρου

Πως θα γίνει η διάσπαση του κόμβου;

Greedy στρατηγική.

- Διάσπαση εγγραφών με βάση έναν έλεγχο γνωρίσματος που βελτιστοποιεί ένα συγκεκριμένο κριτήριο

- Θέματα
  - Καθορισμός του τρόπου διαχωρισμού των εγγραφών
    - Καθορισμός του ελέγχου γνωρίσματος
    - Καθορισμός του βέλτιστου διαχωρισμού
  - Πότε θα σταματήσει ο διαχωρισμός (συνθήκη τερματισμού)

Δέντρο Απόφασης: Κατασκευή Δέντρου

Καθορισμός των συνθηκών του ελέγχου για τα γνωρίσματα

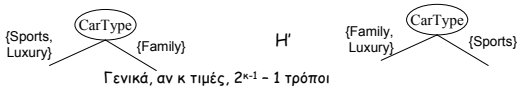
- Εξαρτάται από τον τύπο των γνωρισμάτων
  - Διακριτές - Nominal
  - Διατεταγμένες - Ordinal
  - Συνεχείς - Continuous
- Εξαρτάται από τον αριθμό των διαφορετικών τρόπων διάσπασης
  - 2-αδική διάσπαση - 2-way split
  - Πολλαπλή διάσπαση - Multi-way split

Δέντρο Απόφασης: Κατασκευή Δέντρου

Διαχωρισμός βασισμένος σε διακριτές τιμές

- Πολλαπλός διαχωρισμός: Χρησιμοποίησε τόσες διασπάσεις όσες οι διαφορετικές τιμές

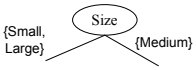
- Δυαδικός Διαχωρισμός: Χωρίζει τις τιμές σε δύο υποσύνολα. Πρέπει να βρει το *βέλτιστο* διαχωρισμό (partitioning).



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

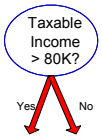
Όταν υπάρχει διάταξη, πρέπει οι διασπάσεις να μη την παραβιάζουν

Αυτός ο διαχωρισμός;

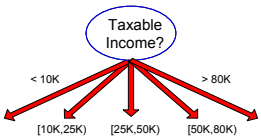


Δέντρο Απόφασης: Κατασκευή Δέντρου

Διαχωρισμός βασισμένος σε συνεχείς τιμές



Δυαδικός διαχωρισμός



Πολλαπλός διαχωρισμός

Δέντρο Απόφασης: Κατασκευή Δέντρου

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

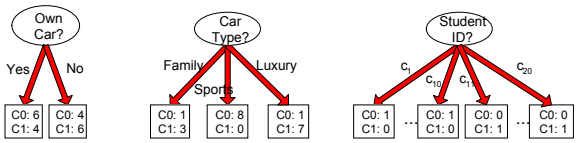
Δέντρο Απόφασης: Κατασκευή Δέντρου

- Greedy στρατηγική.
  - Διάσπαση εγγραφών με βάση έναν έλεγχο γνωρίσματος που βελτιστοποιεί ένα συγκεκριμένο κριτήριο
- Θέματα
  - Καθορισμός του τρόπου διαχωρισμού των εγγραφών
    - Καθορισμός του ελέγχου γνωρίσματος
    - Καθορισμός του βέλτιστου διαχωρισμού
  - Πότε θα σταματήσει ο διαχωρισμός (συνθήκη τερματισμού)

Δέντρο Απόφασης: Κατασκευή Δέντρου

Βέλτιστος Διαχωρισμός

Πριν το διαχωρισμό: **10 εγγραφές της κλάσης 0,**
10 εγγραφές της κλάσης 1



Ποια από τις 3 διασπάσεις να προτιμήσουμε;

Ποια συνθήκη ελέγχου είναι καλύτερη;

Δέντρο Απόφασης: Κατασκευή Δέντρου

Βέλτιστος Διαχωρισμός

Greedy προσέγγιση:
προτιμούνται οι κόμβοι με ομοιογενείς κατανομές κλάσεων (homogeneous class distribution)

Χρειαζόμαστε μία μέτρηση της μη καθαρότητας ενός κόμβου (node impurity)

«Καλός» κόμβος!!

C0: 5	Μη-ομοιογενής, Μεγάλος βαθμός μη καθαρότητας
C1: 5	

C0: 9	Ομοιογενής, Μικρός βαθμός μη καθαρότητας
C1: 1	

N1		N2		N3		N4	
C1	0	C1	1	C1	2	C1	3
C2	6	C2	5	C2	4	C2	3
Μη καθαρότητα ~ 0		ενδιάμεση		ενδιάμεση αλλά μεγαλύτερη		Μεγάλη μη καθαρότητα	

I(N1) < I(N2) < I(N3) < I(N4)

Δέντρο Απόφασης: Κατασκευή Δέντρου

Πως θα χρησιμοποιήσουμε τη μέτρηση καθαρότητας;

Κριτήριο για διάσπαση – Το τι κερδίζουμε από την διάσπαση:

Έστω ότι έχουμε ένα μέτρο για τη μέτρηση αυτής της καθαρότητας ενός κόμβου n: I(n)

Κοιτάμε την καθαρότητα του γονέα (πριν τη διάσπαση) και των παιδιών του (μετά τη διάσπαση)

Βάρος (εξαρτάται από τον αριθμό εγγραφών)

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

N είναι ο αριθμός των εγγραφών στο γονέα και $N(u_i)$ του i-οστού παιδιού

Διαλέγουμε την «καλύτερη» διάσπαση (μεγαλύτερο Δ)

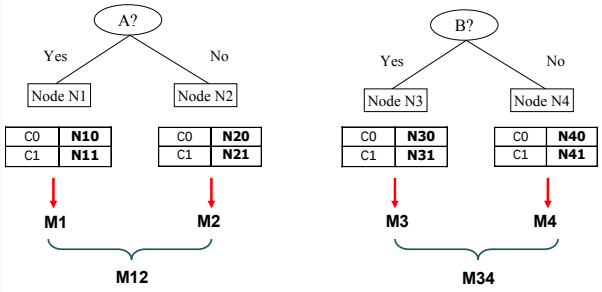
Δέντρο Απόφασης: Κατασκευή

Παράδειγμα

Πριν τη διάσπαση:

C0	N00
C1	N01

→ M0



Gain (κέρδος) = M0 – M12 vs M0 – M34

Δέντρο Απόφασης: Κατασκευή Δέντρου

Μέτρα μη Καθαρότητας

1. Ευρετήριο Gini - Gini Index
2. Εντροπία - Entropy
3. Λάθος ταξινομήσεις - Misclassification error

Δέντρο Απόφασης: GINI

Ευρετήριο Gini για τον κόμβο t :

$$GINI(t) = 1 - \sum_{j=1}^{c} [p(j \mid t)]^2$$

p(j | t) σχετική συχνότητα της κλάσης j στον κόμβο t (ποσοστό εγγραφών της κλάσης j στον κόμβο t)

c αριθμός κλάσεων

Παραδείγματα:

N1		N2		N3		N4	
C1	0	C1	1	C1	2	C1	3
C2	6	C2	5	C2	4	C2	3
Gini=0.000		Gini=0.278		Gini=0.444		Gini=0.500	

Δέντρο Απόφασης: GINI

Ευρετήριο Gini για τον κόμβο t :

$$GINI(t) = 1 - \sum_{j=1}^{c} [p(j \mid t)]^2$$

p(j | t) σχετική συχνότητα της κλάσης j στον κόμβο t

c αριθμός κλάσεων

Ελάχιστη τιμή (0.0) όταν όλες οι εγγραφές ανήκουν σε μία κλάση (που σημαίνει την πιο ενδιαφέρουσα πληροφορία)

Μέγιστη τιμή (1 - 1/c) όταν όλες οι εγγραφές είναι ομοιόμορφα κατανεμημένες στις κλάσεις (που σημαίνει τη λιγότερο ενδιαφέρουσα πληροφορία)

εξαρτάται από τον αριθμό των κλάσεων

Δέντρο Απόφασης: GINI

Χρήση του στην κατασκευή του δέντρου απόφασης

- Χρησιμοποιείται στα CART, SLIQ, SPRINT.

Όταν ένας κόμβος p διασπάται σε k κόμβους (παιδιά), (που σημαίνει ότι το σύνολο των εγγραφών του κόμβου χωρίζεται σε k υποσύνολα), η ποιότητα του διαχωρισμού υπολογίζεται ως:

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

όπου, n_i = αριθμός εγγραφών του παιδιού i,
n = αριθμός εγγραφών του κόμβου p.

Ψάχνουμε για:
- Πιο καθαρές
- Πιο μεγάλες (σε αριθμό) μικρές διασπάσεις

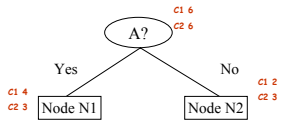
Δέντρο Απόφασης: GINI

Παράδειγμα Εφαρμογής

Περίπτωση 1: Δυαδικά Γνωρίσματα

Αρχικός κόμβος

	Parent
C1	6
C2	6
Gini = 0.500	



	N1	N2
C1	4	2
C2	3	3
Gini=0.486		

Gini(N1) = 1 – (4/7)² – (3/7)² = 0.49

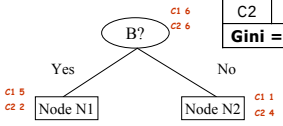
Gini(N2) = 1 – (2/5)² – (3/5)² = 0.48

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

Δέντρο Απόφασης: GINI

Παράδειγμα Εφαρμογής (συνέχεια)

	Parent
C1	6
C2	6
Gini = 0.500	



Υπενθύμιση: με βάση το A

	N1	N2
C1	4	2
C2	3	3
Gini=0.486		

	N1	N2
C1	5	1
C2	2	4
Gini=0.371		

Gini(N1) = 1 – (5/7)² – (2/7)² = 0.408

Gini(N2) = 1 – (1/5)² – (4/5)² = 0.32

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

Άρα διαλέγουμε το B

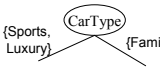
Δέντρο Απόφασης: GINI

Περίπτωση 2: Κατηγορικά Γνωρίσματα

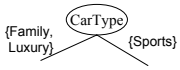
Για κάθε διαφορετική τιμή, μέτρησε τις τιμές στα δεδομένα που ανήκουν σε κάθε κλάση

Χρησιμοποίησε τον πίνακα με τους μετρητές για να πάρεις την απόφαση

Δυαδική διάσπαση
(βρες τον καλύτερο διαχωρισμό των τιμών)



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	

Δέντρο Απόφασης: GINI

Περίπτωση 2: Κατηγορικά Γνωρίσματα

Πολλαπλή Διάσπαση

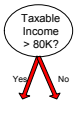
CarType	Family	Sports	Luxury
C1	1	8	1
C2	3	0	7
Gini	0.163		

Δέντρο Απόφασης: GINI

Συνεχή Γνωρίσματα

- 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
  - Για κάθε v, scan τα δεδομένα κατασκεύασε τον πίνακα και υπολόγισε το Gini ευρετήριο χρόνος O(N)
  - $O(N^2)$ Υπολογιστικά μη αποδοτικό! Επανάληψη υπολογισμού.

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: GINI

- Για ποιο αποδοτικό υπολογισμό, για κάθε γνώρισμα
  - Ταξινόμησε το γνώρισμα - O(N logN)
  - Σειριακή διάσχιση των τιμών, ενημερώνοντας κάθε φορά των πίνακα με τους μετρητές και υπολογίζοντας το ευρετήριο Gini
  - Επιλογή του διαχωρισμού με το μικρότερο ευρετήριο Gini

Παράδειγμα – Διαχωρισμός στο γνώρισμα Income

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
Taxable Income	60	70	75	85	90	95	100	120	125	220
Ταξινόμηση Τιμών →										
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	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

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
Taxable Income	60	70	75	85	90	95	100	120	125	220
Ταξινομημένες Τιμές →										
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	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

Δέντρο Απόφασης: Κατασκευή Δέντρου

Μέτρα μη Καθαρότητας

1. Ευρετήριο Gini - Gini Index
2. Εντροπία - Entropy
3. Λάθος ταξινομήσεις - Misclassification error

Δέντρο Απόφασης: Εντροπία

Εντροπία για τον κόμβο t :

$$Entropy(t) = -\sum_{j=1}^{c} p(j \mid t) \log p(j \mid t)$$

p(j | t) σχετική συχνότητα της κλάσης j στον κόμβο t
c αριθμός κλάσεων

Μετράει την ομοιογένεια ενός κόμβου

Μέγιστη τιμή log(c) όταν όλες οι εγγραφές είναι ομοιόμορφα κατανεμημένες στις κλάσεις (που σημαίνει τη λιγότερο ενδιαφέρουσα πληροφορία)

Ελάχιστη τιμή (0.0) όταν όλες οι εγγραφές ανήκουν σε μία κλάση (που σημαίνει την πιο ενδιαφέρουσα πληροφορία)

Δέντρο Απόφασης: Εντροπία

Παραδείγματα

$$Entropy(t) = -\sum_{j} p(j \mid t) \log_2 p(j \mid t)$$

C1	0
C2	6

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1
Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

C1	1
C2	5

P(C1) = 1/6 P(C2) = 5/6
Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (5/6) = 0.65

C1	2
C2	4

P(C1) = 2/6 P(C2) = 4/6
Entropy = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

Δέντρο Απόφασης: Εντροπία

Και σε αυτήν την περίπτωση, όταν ένας κόμβος p διασπάται σε k σύνολα (παιδιά), η ποιότητα του διαχωρισμού υπολογίζεται ως:

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

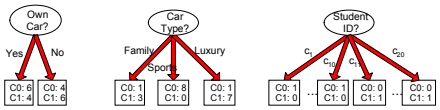
όπου, n_i = αριθμός εγγραφών του παιδιού i,
n = αριθμός εγγραφών του κόμβου p.

- Χρησιμοποιείται στα ID3 and C4.5
- Όταν χρησιμοποιούμε την εντροπία για τη μέτρηση της μη καθαρότητας τότε η διαφορά καλείται **κέρδος πληροφορίας (information gain)**

Δέντρο Απόφασης

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

Τείνει να ευνοεί διαχωρισμούς που καταλήγουν σε μεγάλο αριθμό από διασπάσεις που η κάθε μία είναι μικρή αλλά καθαρή



Μπορεί να καταλήξουμε σε πολύ μικρούς κόμβους (με πολύ λίγες εγγραφές) για αξιόπιστες προβλέψεις

Στο παράδειγμα, το student-id είναι κλειδί, όχι χρήσιμο για προβλέψεις

Δέντρο Απόφασης: Λόγος Κέρδους

- Μία λύση είναι να έχουμε μόνο δυαδικές διασπάσεις
- Εναλλακτικά, μπορούμε να λάβουμε υπό όψιν μας *τον αριθμό* των κόμβων

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

Όπου:

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

SplitINFO: εντροπία της διάσπασης
Μεγάλος αριθμός μικρών διασπάσεων (υψηλή εντροπία) τιμωρείται

Χρησιμοποιείται στο C4.5

Δέντρο Απόφασης: Λόγος Κέρδους

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

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

Παράδειγμα

Έστω N εγγραφές αν τις χωρίσουμε

Σε 3 κόμβους SplitINFO = - log(1/3) = log3

Σε 2 κόμβους SplitINFO = - log(1/2) = log2 = 1

Άρα οι 2 ευνοούνται

Δέντρο Απόφασης: Κατασκευή Δέντρου

Μέτρα μη Καθαρότητας

1. Ευρετήριο Gini - Gini Index
2. Εντροπία - Entropy
3. Λάθος ταξινομήσεις - Misclassification error

Δέντρο Απόφασης: Λάθος Ταξινόμησης

Λάθος ταξινόμησης (classification error) για τον κόμβο t :

$$Error(t) = 1 - \max_{class\ i} P(i \mid t)$$

Μετράει το λάθος ενός κόμβου

Παράδειγμα

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

Δέντρο Απόφασης: Λάθος Ταξινόμησης

Λάθος ταξινόμησης (classification error) για τον κόμβο t :

$$Error(t) = 1 - \max_{class\ i} P(i \mid t)$$

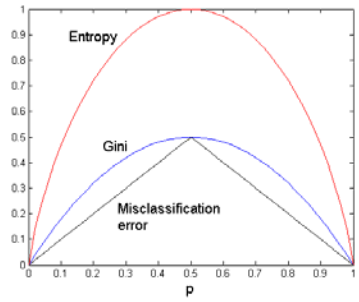
Μετράει το λάθος ενός κόμβου

Μέγιστη τιμή 1-1/c όταν όλες οι εγγραφές είναι ομοιόμορφα κατανεμημένες στις κλάσεις (που σημαίνει τη λιγότερο ενδιαφέρουσα πληροφορία)

Ελάχιστη τιμή (0.0) όταν όλες οι εγγραφές ανήκουν σε μία κλάση (που σημαίνει την πιο ενδιαφέρουσα πληροφορία)

Δέντρο Απόφασης: Σύγκριση

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)

Δέντρο Απόφασης: Σύγκριση

- Όπως είδαμε και στα παραδείγματα οι τρεις μετρήσεις είναι συνεπής μεταξύ τους, πχ N1 μικρότερη τιμή από το N2 και με τις τρεις μετρήσεις
- Ωστόσο το γνώρισμα που θα επιλεγεί για τη συνθήκη ελέγχου εξαρτάται από το ποια μέτρηση χρησιμοποιείται

Δέντρο Απόφασης: Κατασκευή Δέντρου

- Greedy στρατηγική.
  - Διάσπαση εγγραφών με βάση έναν έλεγχο γνωρίσματος που βελτιστοποιεί ένα συγκεκριμένο κριτήριο
- Θέματα
  - Καθορισμός του τρόπου διαχωρισμού των εγγραφών
    - Καθορισμός του ελέγχου γνωρίσματος
    - Καθορισμός του βέλτιστου διαχωρισμού
  - Πότε θα σταματήσει ο διαχωρισμός (συνθήκη τερματισμού)

Δέντρο Απόφασης: Κριτήρια Τερματισμού

- Σταματάμε την επέκταση ενός κόμβου όταν όλες οι εγγραφές του ανήκουν στην ίδια κλάση
- Σταματάμε την επέκταση ενός κόμβου όταν όλα τα γνωρίσματα έχουν τις ίδιες τιμές
- Γρήγορος τερματισμός

Δέντρο Απόφασης

Πλεονεκτήματα Δέντρων Απόφασης

- Μη παραμετρική προσέγγιση: Δε στηρίζεται σε υπόθεση εκ των προτέρων γνώσης σχετικά με τον τύπο της κατανομής πιθανότητας που ικανοποιεί η κλάση ή τα άλλα γνωρίσματα
- Η κατασκευή του βέλτιστου δέντρου απόφασης είναι ένα NP-complete πρόβλημα. Ευριστικοί: Αποδοτική κατασκευή ακόμα και στην περίπτωση πολύ μεγάλου συνόλου δεδομένων
- Αφού το δέντρο κατασκευαστεί, η ταξινόμηση νέων εγγραφών πολύ γρήγορη O(h) όπου h το μέγιστο ύψος του δέντρου
- Εύκολα στην κατανόηση (ιδιαίτερα τα μικρά δέντρα)
- Η ακρίβεια τους συγκρίσιμη με άλλες τεχνικές για μικρά σύνολα δεδομένων

Δέντρο Απόφασης

Πλεονεκτήματα

- Καλή συμπεριφορά στο θόρυβο
- Η ύπαρξη πλεοναζόντων γνωρισμάτων (γνωρίσματα των οποίων η τιμή εξαρτάται από κάποιο άλλο) δεν είναι καταστροφική για την κατασκευή. Χρησιμοποιείται ένα από τα δύο.
  Αν πάρα πολλά, μπορεί να οδηγήσουν σε δέντρα πιο μεγάλα από ότι χρειάζεται

Δέντρο Απόφασης

Στρατηγική αναζήτησης

- Ο αλγόριθμος που είδαμε χρησιμοποιεί μια greedy, top-down, αναδρομική διάσπαση για να φτάσει σε μια αποδεκτή λύση
- Άλλες στρατηγικές?
  - Bottom-up (από τα φύλλα, αρχικά κάθε εγγραφή και φύλλο)
  - Bi-directional

Δέντρο Απόφασης

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 γνωρίσματα
- Όχι καλή συμπεριφορά για συνεχείς μεταβλητές Ιδιαίτερα όταν η συνθήκη ελέγχου αφορά ένα γνώρισμα τη φορά

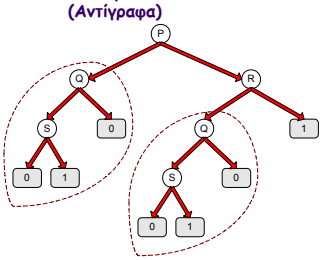
Δέντρο Απόφασης

Data Fragmentation – Διάσπαση Δεδομένων

- Ο αριθμός των εγγραφών μειώνεται όσο κατεβαίνουμε στο δέντρο
- Ο αριθμός των εγγραφών στα φύλλα μπορεί να είναι πολύ μικρός για να πάρουμε οποιαδήποτε στατιστικά σημαντική απόφαση
- Μπορούμε να αποτρέψουμε την περαιτέρω διάσπαση όταν ο αριθμός των εγγραφών πέσει κάτω από ένα όριο

Δέντρο Απόφασης

Tree Replication (Αντίγραφα)



Το ίδιο υπο-δέντρο να εμφανίζεται πολλές φορές σε ένα δέντρο απόφασης

Αυτό κάνει το δέντρο πιο περίπλοκο και πιθανών δυσκολότερο στην κατανόηση

Σε περιπτώσεις διάσπασης ενός γνωρίσματος σε κάθε εσωτερικό κόμβο – ο ίδιος έλεγχος σε διαφορετικά σημεία

Δέντρο Απόφασης

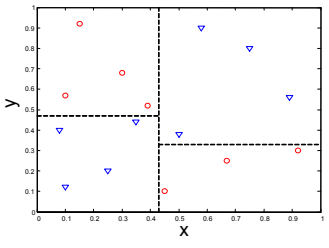
Decision Boundary

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

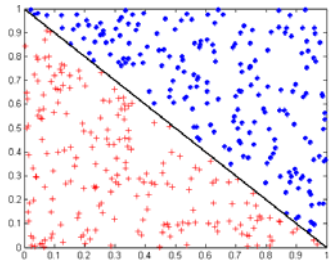
Η οριακή γραμμή (Border line) μεταξύ δυο γειτονικών περιοχών που ανήκουν σε διαφορετικές κλάσεις ονομάζεται και decision boundary (όριο απόφασης)

Δέντρο Απόφασης

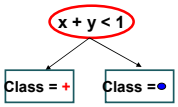
Όταν η συνθήκη ελέγχου περιλαμβάνει μόνο ένα γνώρισμα τη φορά τότε το Decision boundary είναι παράλληλη στους άξονες (τα decision boundaries είναι ορθογώνια παραλληλόγραμμα)



Δέντρο Απόφασης



Oblique (πλάγιο) Δέντρο Απόφασης



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

Δέντρο Απόφασης

Constructive induction

Κατασκευή σύνθετων γνωρισμάτων ως αριθμητικών ή λογικών συνδυασμών άλλων γνωρισμάτων

Δέντρο Απόφασης: C4.5

Simple depth-first construction.
Uses Information Gain
Sorts Continuous Attributes at each node.
Needs entire data to fit in memory.
Unsuitable for Large Datasets.
Needs out-of-core sorting.

You can download the software from:
http://www.cse.unsw.edu.au/~quinlan/c4.5r8.tar.gz

Θέματα στην Ταξινόμηση

Θέματα Ταξινόμησης

- Underfitting and Overfitting
- Εκτίμηση Λάθους
- Τιμές που λείπουν

Overfitting

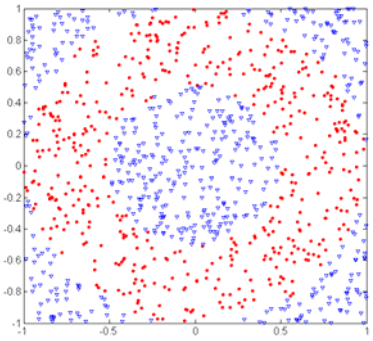
Λάθη

- *Εκπαίδευσης* (training, resubstitution, apparent): λάθη ταξινόμησης στα δεδομένα του συνόλου εκπαίδευσης (ποσοστό δεδομένων εκπαίδευσης που ταξινομούνται σε λάθος κλάση)
- *Γενίκευσης* (generalization): τα αναμενόμενα λάθη ταξινόμησης του μοντέλου σε δεδομένα που δεν έχει δει

Overfitting

Μπορεί ένα μοντέλο που ταιριάζει πολύ καλά με τα δεδομένα εκπαίδευσης να έχει μεγαλύτερο λάθος γενίκευσης από ένα μοντέλο που ταιριάζει λιγότερο καλά στα δεδομένα εκπαίδευσης

Overfitting



Δύο κλάσεις: κλάση 1 (500 κυκλικά σημεία) και κλάση 2 (500 τριγωνικά σημεία)

Για τα σημεία της κλάσης 1 (κυκλικά σημεία):

$0.5 \le \text{sqrt}(x_1^2+x_2^2) \le 1$

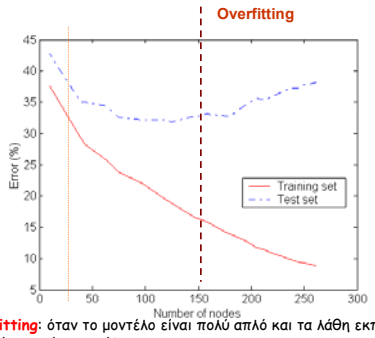
Για τα σημεία της κλάσης 2 (τριγωνικά σημεία):

$\text{sqrt}(x_1^2+x_2^2) > 0.5$ or

$\text{sqrt}(x_1^2+x_2^2) < 1$

Overfitting

"Everything should be made as simple as possible, but not simpler", Einstein



Το δέντρο απόφασης για το προηγούμενα δεδομένα

30% εκπαίδευση

70% έλεγχο

Gini

Στη συνέχεια, pruning

Underfitting: όταν το μοντέλο είναι πολύ απλό και τα λάθη εκπαίδευσης και τα λάθη ελέγχου είναι μεγάλα

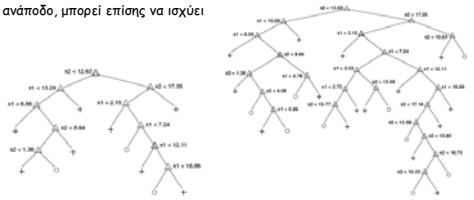
Overfitting

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

Μικρό (μηδενικό) λάθος εκπαίδευσης

Μεγάλο λάθος ελέγχου

Και το ανάποδο, μπορεί επίσης να ισχύει



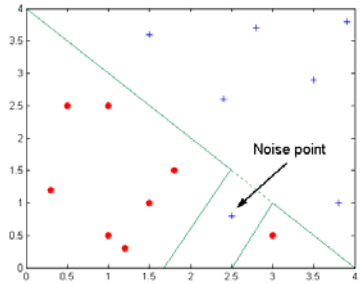
(a) Decision tree with 11 leaf nodes.

(b) Decision tree with 24 leaf nodes.

Figure 4.24. Decision trees with different model complexities.

Overfitting

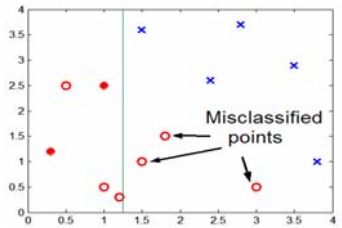
Overfitting εξαιτίας Θορύβου



Decision boundary is distorted by noise point

Overfitting

Overfitting εξαιτίας μη Επαρκών Δειγμάτων



Κόκκινοι κύκλοι ανήκουν στην ίδια κλάση

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

Η έλλειψη κόκκινων σημείων στο κάτω μισό του διαγράμματος κάνει δύσκολη την πρόβλεψη των κλάσεων σε αυτήν την περιοχή

Μη επαρκής αριθμός εγγραφών εκπαίδευσης έχει ως αποτέλεσμα το δέντρο απόφασης να κάνει πρόβλεψη για τα σημεία αυτής της περιοχής χρησιμοποιώντας εγγραφές εκπαίδευσης μη σχετικές με το έργο της ταξινόμησης

Overfitting

Πρόβλημα λόγω πολλαπλών επιλογών

- Επειδή σε κάθε βήμα εξετάζουμε πάρα πολλές διαφορετικές διασπάσεις,
  - κάποια διάσπαση βελτιώνει το δέντρο *κατά τύχη*

Το πρόβλημα χειροτερεύει όταν αυξάνει ο αριθμός των επιλογών και μειώνεται ο αριθμός των δειγμάτων

Overfitting

- Το overfitting έχει ως αποτέλεσμα δέντρα απόφασης που είναι πιο περίπλοκα από ό,τι χρειάζεται
- Τα λάθη εκπαίδευσης δεν αποτελούν πια μια καλή εκτίμηση για τη συμπεριφορά του δέντρου σε εγγραφές που δεν έχει δει ξανά
- Νέοι μέθοδοι για την εκτίμηση του λάθους

Πολυπλοκότητα Μοντέλου

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

Εκτίμηση του Λάθους Γενίκευσης

- Re-substitution errors: Λάθος στην εκπαίδευση (Σ e(t))
- Generalization errors: Λάθος στον έλεγχο (Σ e'(t))

Ως λάθος μετράμε το ποσοστό των εγγραφών που ο ταξινομητής τοποθετεί σε λάθος κλάση

Μέθοδοι εκτίμησης του λάθους γενίκευσης:

1. Optimistic approach – Αισιόδοξη προσέγγιση:

$$e'(t) = e(t)$$

Εκτίμηση του Λάθους Γενίκευσης

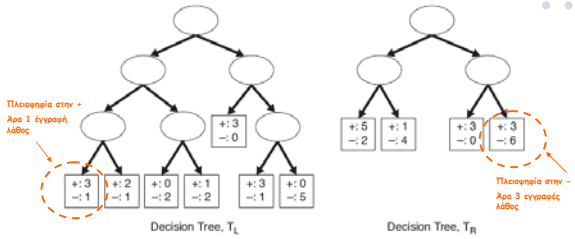


Figure 4.27. Example of two decision trees generated from the same training data.

Με βάση το λάθος εκπαίδευσης

Αριστερό 4/24 = 0.167

Δεξί: 6/24 = 0.25

Εκτίμηση του Λάθους Γενίκευσης

2. Pessimistic approach - Απαισιόδοξη προσέγγιση:

k: αριθμός φύλλων,

για κάθε φύλλο t_i προσθέτουμε ένα κόστος $V(t_i)$

$$e'(T) = \frac{\sum_{i=1}^{k}[e(t_i)+V(t_i)]}{\sum_{i=1}^{k} n(t_i)}$$

Αν για κάθε φύλλο t: e'(t) = e(t) + 0.5
Συνολικό λάθος: e'(T) = e(T) + k × 0.5 (k: αριθμός φύλλων)

Για ένα δέντρο με 30 φύλλα και 10 λάθη στο σύνολο εκπαίδευσης (από σύνολο 1000 εγγραφών):
Training error = 10/1000 = 1%
Generalization error = (10 + 30×0.5)/1000 = 2.5%

Το 0.5 σημαίνει ότι διαχωρισμός ενός κόμβου δικαιολογείται αν βελτιώνει τουλάχιστον μία εγγραφή

Εκτίμηση του Λάθους Γενίκευσης

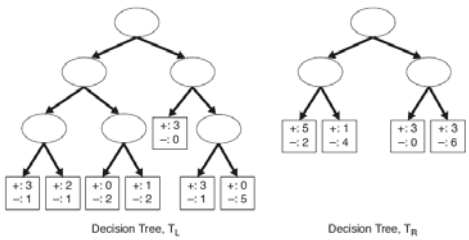


Figure 4.27. Example of two decision trees generated from the same training data.

Με βάση το λάθος εκπαίδευσης

Αριστερό (4 + 7*0.5)/24 = 0.3125

Δεξί: (6 + 4*0.5)/24 = 0.3333

Αν αντί για 0.5, κάτι μεγαλύτερο;

Εκτίμηση του Λάθους Γενίκευσης

3. 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

Αντιμετώπιση Overfitting

Pre-Pruning (Early Stopping Rule)

Σταμάτα τον αλγόριθμο πριν σχηματιστεί ένα πλήρες δέντρο

Συνήθεις συνθήκες τερματισμού για έναν κόμβο:
- Σταμάτα όταν όλες οι εγγραφές ανήκουν στην ίδια κλάση
- Σταμάτα όταν όλες οι τιμές των γνωρισμάτων είναι οι ίδιες

Πιο 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

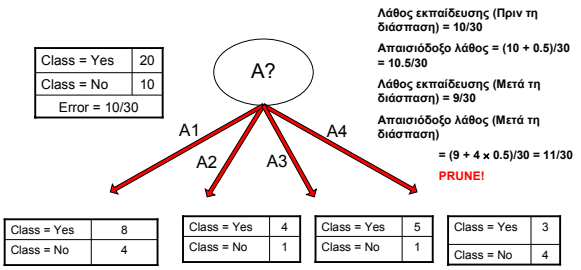
Overfitting

Post-pruning

- Ανάπτυξε το δέντρο πλήρως
- Trim – ψαλίδισε τους κόμβους bottom-up
- Αν το λάθος γενίκευσης μειώνεται με το ψαλίδισμα, αντικατέστησε το υποδέντρο με
  - ένα φύλλο - οι ετικέτες κλάσεις του φύλλου καθορίζεται από την πλειοψηφία των κλάσεων των εγγραφών του υποδέντρου (subtree replacement)
  - ένα από τα κλαδιά του (Branch), αυτό που χρησιμοποιείται συχνότερα (subtree raising)

Overfitting

Παράδειγμα Post-Pruning



Λάθος εκπαίδευσης (Πριν τη διάσπαση) = 10/30

Απαισιόδοξο λάθος = (10 + 0.5)/30 = 10.5/30

Λάθος εκπαίδευσης (Μετά τη διάσπαση) = 9/30

Απαισιόδοξο λάθος (Μετά τη διάσπαση)

= (9 + 4 x 0.5)/30 = 11/30

PRUNE!

Overfitting

Παράδειγμα post-pruning

- Αισιόδοξη προσέγγιση?

  Όχι διάσπαση

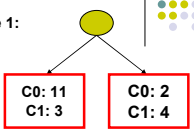
- Απαισιόδοξη προσέγγιση?

  όχι case 1, ναι case 2

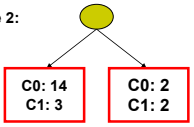
- REP?

Εξαρτάται από το σύνολο επαλήθευσης

Case 1:



Case 2:



Τιμές που λείπουν

- 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

Τιμές που λείπουν

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

Missing value

Υπολογισμό μέτρων καθαρότητας

Πριν τη διάσπαση:

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

Διάσπαση στο 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

Τιμές που λείπουν

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

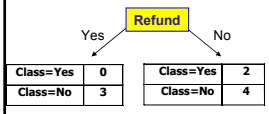
Tid	Refund	Marital Status	Taxable Income	Class
10	?	Single	90K	Yes

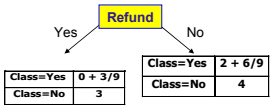
Σε ποιο φύλλο;

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

Πιθανότητα Refund=No is 6/9

Ανάθεση εγγραφής στο αριστερό παιδί με βάρος 3/9 και στο δεξί παιδί με βάρος 6/9



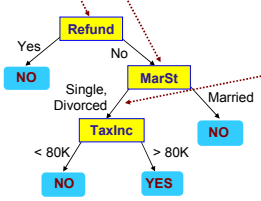


Τιμές που λείπουν

Νέα εγγραφή

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

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



Πιθανότητα οικογενειακή κατάσταση (MarSt) = Married is 3.67/6.67

Πιθανότητα οικογενειακή κατάσταση (MarSt) ={Single,Divorced} is 3/6.67