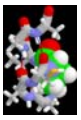
Ταξινόμηση

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 περιεχόμενο 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 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.

	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 χωρίζεται σε ένα σύνολο εκπαίδευσης (training set) και ένα σύνολο ελέγχου (test test)
Το σύνολο εκπαίδευσης χρησιμοποιείται για να κτιστεί 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

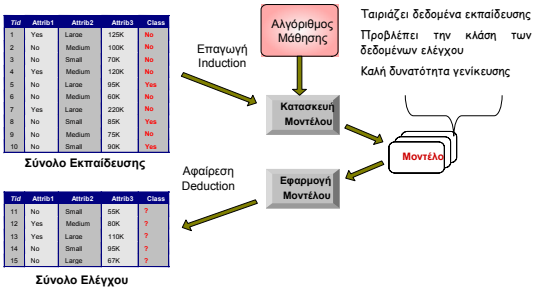
❑ Κατάλληλη κυρίως για:

- 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

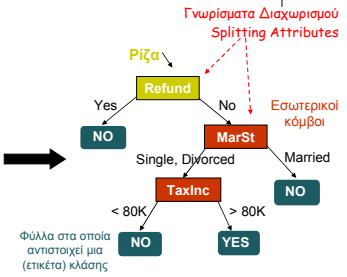
- Τεχνικές βασισμένες σε Δέντρα Απόφασης (decision trees)
- Τεχνικές βασισμένες σε Κανόνες - Rule-based Methods
- Memory based reasoning
- Νευρωνικά Δίκτυα
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines

Δέντρα Απόφασης

Δέντρο Απόφασης: Παράδειγμα

	categorical	categorical	continuous	class
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

Δεδομένα Εκπαίδευσης

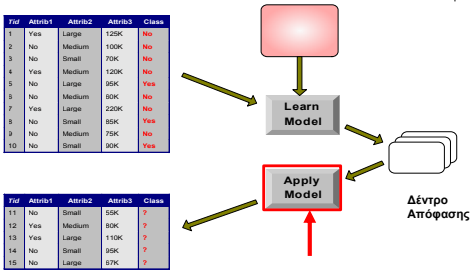


Μοντέλο: Δέντρο Απόφασης

Δέντρο Απόφασης: Παράδειγμα

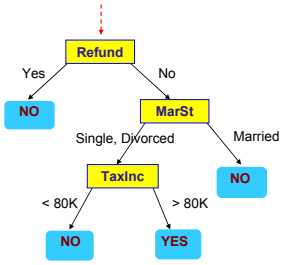
- Εσωτερικοί κόμβοι αντιστοιχούν σε κάποιο γνώρισμα
- Διαχωρισμός (split) ενός κόμβου σε παιδιά – η ετικέτα στην ακμή = συνθήκη/έλεγχος
- Φύλλα αντιστοιχούν σε κλάσεις

Δέντρο Απόφασης: Βήματα



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Ξεκίνα από τη ρίζα του δέντρου.



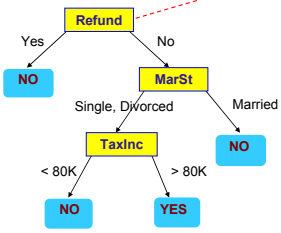
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Test Data

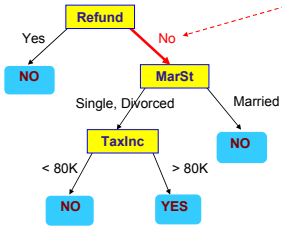
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Test Data

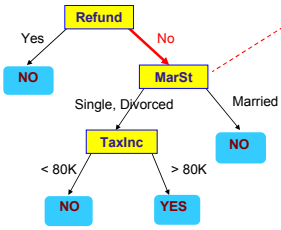
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Test Data

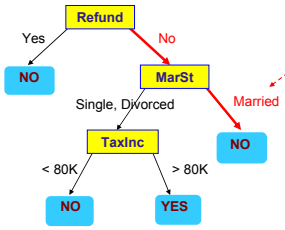
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Test Data

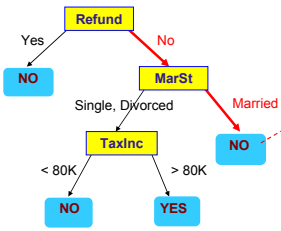
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

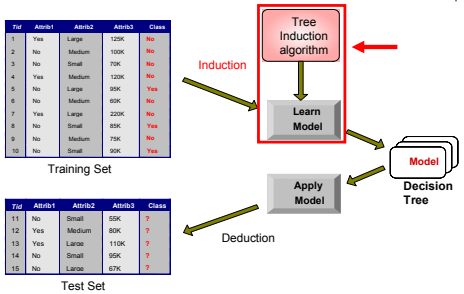
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Ανάθεση στο Cheat "No"

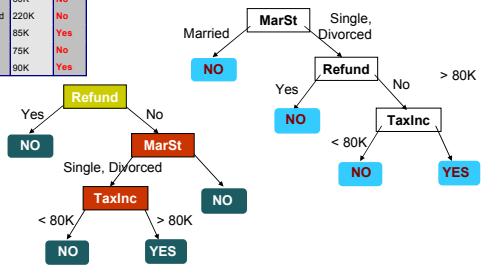
Δέντρο Απόφασης



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

Διαφορετικά δέντρα για το ίδιο σύνολο εκπαίδευσης



Δέντρο Απόφασης: Κατασκευή

Ο αριθμός των πιθανών Δέντρων Απόφασης είναι εκθετικός.

Πολλοί αλγόριθμοι για την επαγωγή (induction) του δέντρου η οποίοι ακολουθούν μια greedy στρατηγική κτίζοντας 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

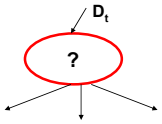
Κτίζει το δέντρο αναδρομικά

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

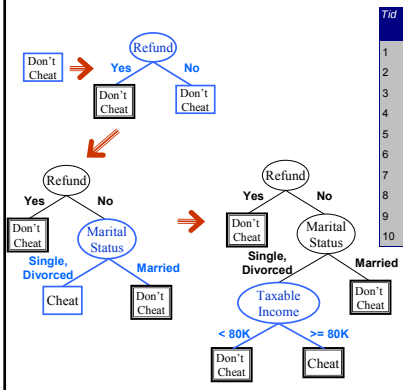
Γενική Διαδικασία:

- 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 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.

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

- 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

- 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 κατηγορήματα

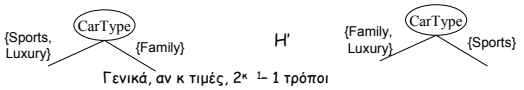
- Εξαρτάται από τον τύπο των γνωρισμάτων
  - Διακριτές - Nominal
  - Διατεταγμένες - Ordinal
  - Συνεχείς - Continuous
- Εξαρτάται από τον αριθμό των διαφορετικών τρόπων διάσπασης
  - 2-αδική διάσπαση - 2-way split
  - Πολλαπλή διάσπαση - Multi-way split

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

Αυτός ο διαχωρισμός;

Όταν υπάρχει διάταξη, πρέπει οι διασπάσεις να μη την παραβιάζουν

Δέντρο Απόφασης: Κατασκευή Δέντρου

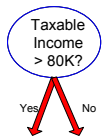
Διαχωρισμός βασισμένος σε συνεχείς τιμές

Τρόποι χειρισμού

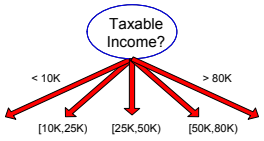
- Discretization (διακριτοποίηση) ώστε να προκύψει ένα διατεταγμένο κατηγορικό γνώρισμα
  - Ταξινόμηση των τιμών και χωρισμός τους σε περιοχές καθορίζοντας n – 1 σημεία διαχωρισμού, απεικόνιση όλων των τιμών μιας περιοχής στην ίδια κατηγορική τιμή
    - **Στατικό** - μια φορά στην αρχή
    - **Δυναμικό** - εύρεση των περιοχών πχ έτσι ώστε οι περιοχές να έχουν το ίδιο διάστημα ή τις ίδιες συχνότητες εμφανίσεις ή με χρήση συσταδοποίησης
- Δυαδική Απόφαση: $(A < v)$ or $(A \geq v)$
  εξετάζει όλους τους δυνατούς διαχωρισμούς και επιλέγει τον καλύτερο – υπολογιστικά βαρύ

Δέντρο Απόφασης: Κατασκευή Δέντρου

Διαχωρισμός βασισμένος σε συνεχείς τιμές



Δυαδικός διαχωρισμός



Πολλαπλός διαχωρισμός

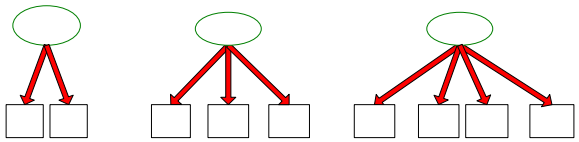
Δέντρο Απόφασης: Κατασκευή Δέντρου

- 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 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)

p(i|t) ποσοστό εγγραφών της κλάσης i στον κόμβο t

Αν δύο κλάσεις p0 και p1, p1 = 1 – p0

Μη-ομοιογενής,
Μεγάλος βαθμός μη καθαρότητας

Ομοιογενής,
Μικρός βαθμός μη καθαρότητας

Δέντρο Απόφασης: Κατασκευή Δέντρου

Κριτήριο για διάσπαση (τι κερδίζουμε):

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)

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

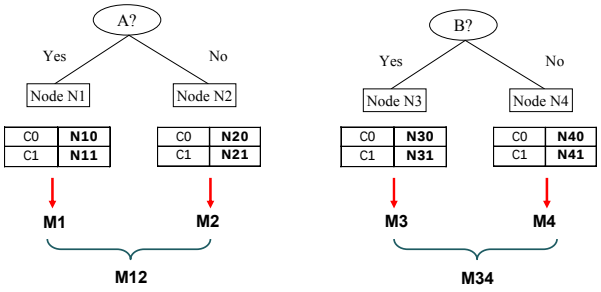
N είναι ο αριθμός των εγγραφών στο γονέα και $N(u_i)$ του j-οστού παιδιού

Δέντρο Απόφασης: Κατασκευή

Πριν τη διάσπαση:

C0	N00
C1	N01

→ **M0**



Gain (κέρδος) = M0 – M12 vs M0 – M34

Own Car?

Δέντρο Απόφασης: Κατασκευή Δέντρου

Μετρήσεις για την επιλογή της καλύτερης διάσπασης – μέτρα μη καθαρότητας:

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

Car Type?

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

Παραδείγματα:

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	

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: GINI

- Χρησιμοποιείται στα CART, SLIQ, SPRINT.

Όταν ένας κόμβος p διασπάται σε k σύνολα (παιδιά), η ποιότητα του διαχωρισμού υπολογίζεται ως:

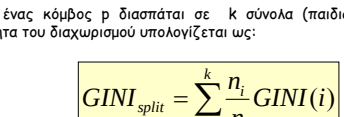
$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

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

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

Δέντρο Απόφασης: GINI

Δυαδικά Γνωρίσματα



	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

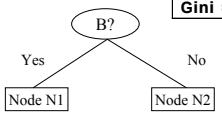
Gini(N2)
= 1 – (2/5)² – (3/5)² = 0.48

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

Δέντρο Απόφασης: GINI

Δυαδικά Γνωρίσματα

	Parent
C1	6
C2	6
Gini = 0.500	



Με βάση το A

	N1	N2
C1	4	2
C2	3	3
Gini=0.486		

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

	N1	N2
C1	5	1
C2	2	4
Gini=0.371		

Άρα διαλέγουμε το B

Δέντρο Απόφασης: GINI

Κατηγορικά Γνωρίσματα

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

Χρησιμοποίησε τον πίνακα με τους μετρητές για να πάρεις την απόφαση

Πολλαπλή Διάσπαση

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

Δυαδική διάσπαση
(βρες τον καλύτερο διαχωρισμό των τιμών)

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

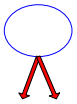
	CarType	
	{Sports}	{Family, Luxury}
C1	8	2
C2	0	10
Gini	0.167	

Δέντρο Απόφασης: GINI

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 συσχετίζεται με έναν πίνακα μετρητών
  - Μετρητές των κλάσεων για κάθε μια από τις δύο διασπάσεις, A < v and A ≥ 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 (N²) 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

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
Sorted Values →	Taxable Income									
	60	70	75	85	90	95	100	120	125	220
Split Positions →	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

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, ΝΟ 0->1, 7->6 YES δεν αλλάζει

Για <72, κοιτάμε το μικρότερο το 70, ΝΟ 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										
Sorted Values →	60	70	75	85	90	95	100	120	125	220
Split Positions →	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

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

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) $\log_2$ (1/6) = 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

Υπενθύμιση, 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.

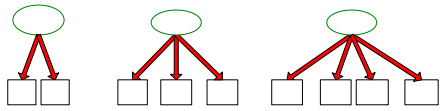
- Χρησιμοποιείται στα ID3 and C4.5.
- Όταν χρησιμοποιούμε για τη μέτρηση της μη καθαρότητας την εντροπία τότε η διαφορά καλείται **κέρδος πληροφορίας (information gain)**

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

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: 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: Λάθος Ταξινόμησης

Παραδείγματα

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

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

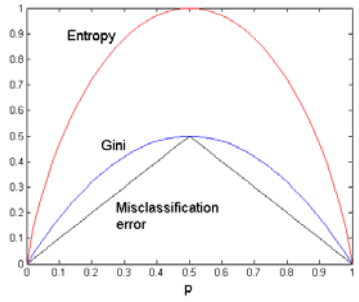
P(C1) = 1/6 P(C2) = 5/6
Error = 1 – max (1/6, 5/6) = 1 – 5/6 = 1/6

C1	2
C2	4

P(C1) = 2/6 P(C2) = 4/6
Error = 1 – max (2/6, 4/6) = 1 – 4/6 = 1/3

Δέντρο Απόφασης: Σύγκριση

Για ένα πρόβλημα δύο κλάσεων



p ποσοστό εγγραφών που ανήκει σε μία από τις δύο κλάσεις

Όλες 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 Ν1 μικρότερη τιμή από το Ν2 και με τις τρεις μετρήσεις
- Ωστόσο το γνώρισμα που θα επιλεγεί για τη συνθήκη ελέγχου εξαρτάται από το μια 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

- Μη παραμετρική προσέγγιση: Δε στηρίζεται σε υπόθεση εκ των προτέρων γνώση 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
  - 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

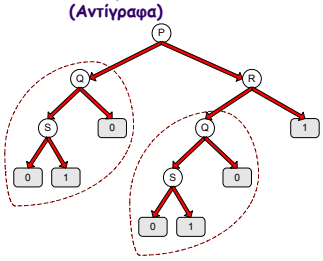
Δέντρο Απόφασης

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

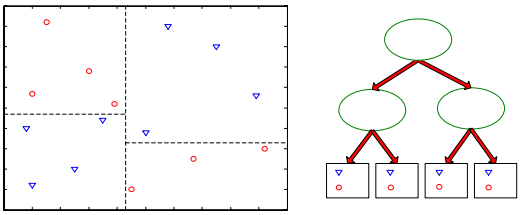
Decision Boundary

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

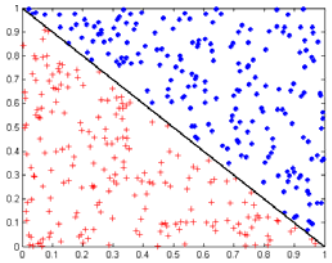
Η οριακή γραμμή (Border line) μεταξύ δυο γειτονικών περιοχών που ανήκουν σε διαφορετικές κλάσεις ονομάζεται και **decision boundary (όριο απόφασης)**

Δέντρο Απόφασης

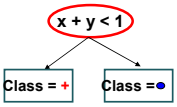
Όταν η συνθήκη ελέγχου περιλαμβάνει μόνο ένα γνώρισμα τη φορά τότε το Decision boundary είναι παράλληλη στους άξονες (τα decision boundaries είναι ορθογώνια παραλληλόγραμμα)



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

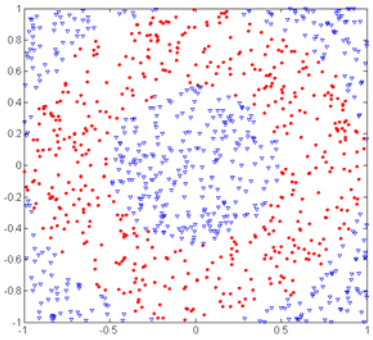
Λάθη

- *Εκπαίδευσης* (training, resubstitution, apparent): λάθη ταξινόμησης στα δεδομένα του συνόλου εκπαίδευσης
- *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: 500 circular and 500 triangular data points.

Circular points:

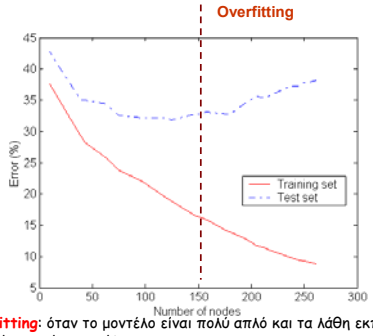
$0.5 \leq \text{sqrt}(x_1^2+x_2^2) \leq 1$

Triangular points:

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

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

Overfitting



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

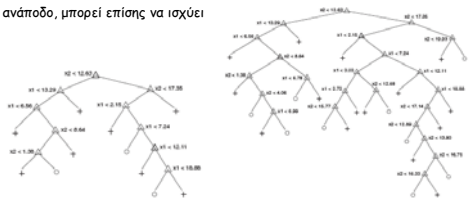
Overfitting

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



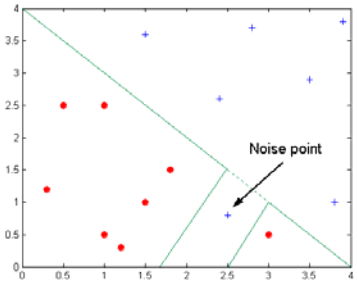
(a) Decision tree with 11 leaf nodes.

(b) Decision tree with 24 leaf nodes.

Figure 4.24. Decision trees with different model complexities.

Overfitting

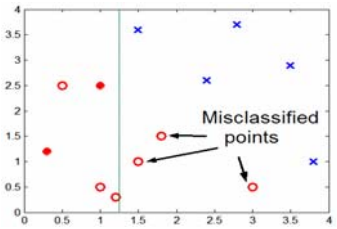
Overfitting εξαιτίας Θορύβου



Decision boundary is distorted by noise point

Overfitting

Overfitting εξαιτίας μη Επαρκών Δειγμάτων



Η έλλειψη 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

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 ότι 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

Εκτίμηση του Λάθους Γενίκευσης

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

Μέθοδοι εκτίμησης του λάθους γενίκευσης:

1. Optimistic approach – Αισιόδοξη προσέγγιση:

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

Εκτίμηση του Λάθους Γενίκευσης

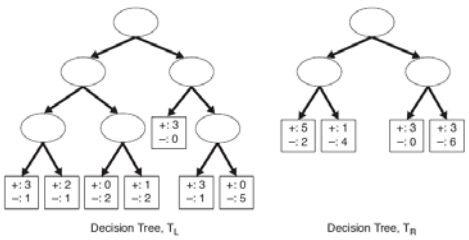


Figure 4.27. Example of two decision trees generated from the same training data.

Με βάση το λάθος εκπαίδευσης

Αριστερό 4/24 = 0.167

Δεξί: 6/24 = 0.25

Εκτίμηση του Λάθους Γενίκευσης

2. Pessimistic approach - Απαισιόδοξη προσέγγιση:

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

Για κάθε φύλλο: e'(t) = (e(t)+0.5)
Συνολικό λάθος: e'(T) = e(T) + N × 0.5 (N: 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

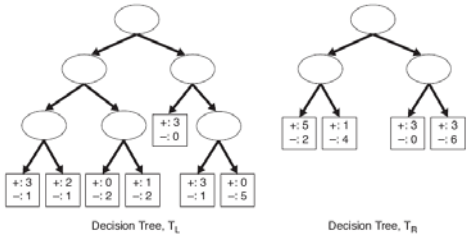


Figure 4.27. Example of two decision trees generated from the same training data.

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

Πολυπλοκότητα Μοντέλου

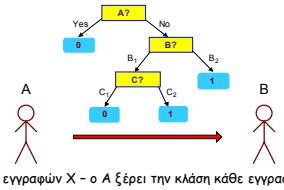
Occam's Razor

- 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 αξιολόγησης του

Πολυπλοκότητα Μοντέλου

Minimum Description Length (MDL)

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	?

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

- Cost(Model,Data) = Cost(Data|Model) + Cost(Model)
  - Cost is the number of bits needed for encoding.
  - Search for the least costly model.
- Cost(Data|Model) encodes the misclassification errors.
- Cost(Model) uses node encoding (number of children) plus splitting condition encoding.

Αντιμετώπιση Overfitting

Pre-Pruning (Early Stopping Rule)

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

Overfitting

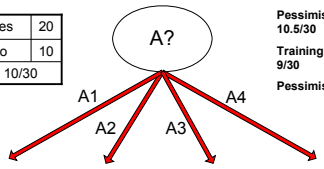
Post-pruning

- 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 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

Overfitting

Παράδειγμα Post-Pruning

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



Training Error (Before splitting) = 10/30

Pessimistic error = (10 + 0.5)/30 = 10.5/30

Training Error (After splitting) = 9/30

Pessimistic error (After splitting) = (9 + 4 x 0.5)/30 = 11/30

PRUNE!

Class = Yes	8
Class = No	4

Class = Yes	4
Class = No	1

Class = Yes	5
Class = No	1

Class = Yes	3
Class = No	4

Overfitting

Παράδειγμα post-pruning

- Optimistic error?

  Don't prune for both cases

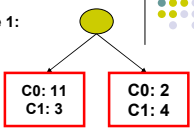
- Pessimistic error?

  Don't prune case 1, prune case 2

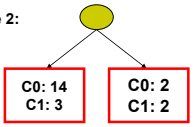
- Reduced error pruning?

  Depends on validation set

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

Before Splitting:

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

Split on 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

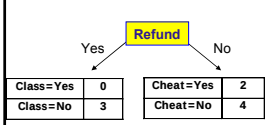
Tid	Refund	Marital Status	Taxable Income	Class
10	?	Single	90K	Yes

Σε ποιο φύλλο;

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

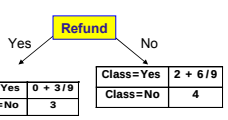
Probability that Refund=No is 6/9

Assign record to the left child with weight = 3/9 and to the right child with weight = 6/9



Class=Yes	0
Class=No	3

Cheat=Yes	2
Cheat=No	4



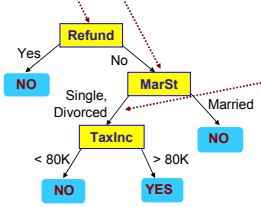
Class=Yes	0 + 3/9
Class=No	3

Class=Yes	2 + 6/9
Class=No	4

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

Probability that Marital Status = Married is 3.67/6.67

Probability that Marital Status ={Single,Divorced} is 3/6.67

Αποτίμηση Μοντέλου

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

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

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	f_{11}	f_{10}
	Class=No	f_{01}	f_{00}

TP (true positive) f_{11}
FN (false negative) f_{10}
FP (false positive) f_{01}
TN (true negative) f_{00}

Μέτρα Εκτίμησης

Πιστότητα

Πιστότητα (accuracy) Το πιο συνηθισμένο μέτρο

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	TP	FN
	Class=No	FP	TN

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

Λόγος Λάθους

$$\text{Error rate} = \frac{f_{01}+f_{10}}{f_{11}+f_{00}+f_{01}+f_{10}}$$

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

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

(-) Λιγότερες εγγραφές για εκπαίδευση – πιθανόν όχι τόσο καλό μοντέλο, όσο αν χρησιμοποιούνταν όλες
(-) 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
(-) 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)

Μέθοδοι Αποτίμησης Μοντέλου

Τυχαία Λήψη Δειγμάτων – Random Subsampling

Επανάληψη της μεθόδου για τη βελτίωση της μεθόδου

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

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, τίνει ασυμπτωτικά στο $1-e^{-1} = 0.632$

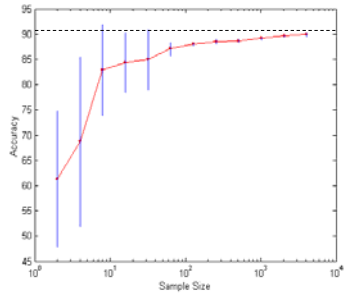
Οι υπόλοιπες εγγραφές – εγγραφές ελέγχου

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

Μέθοδοι Αποτίμησης Μοντέλου

Καμπύλη Μάθησης (Learning Curve)



- Η καμπύλη μάθησης δείχνει πως μεταβάλλεται η πιστότητα με την αύξηση του μεγέθους του δείγματος
- Effect of small sample size:
  - Bias in the estimate
  - Variance of estimate

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

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	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

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	TP	FN
	Class=No	FP	TN

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

Λόγος Λάθους

$$\text{Error rate} = \frac{f_{01}+f_{10}}{f_{11}+f_{00}+f_{01}+f_{10}}$$

Μέτρα Εκτίμησης

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

Μέτρα Εκτίμησης

Πίνακας Κόστους

		PREDICTED CLASS	
	C(i|j)	**Class=Yes**	**Class=No**
ACTUAL CLASS	**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

$C(M) = TP \times C(Yes|Yes) + FN \times C(Yes|No) + FP\ C(No|Yes) + TN\ C(No|No)$

Μέτρα Εκτίμησης

Υπολογισμός του Κόστους της Ταξινόμησης

Cost Matrix	PREDICTED CLASS		
	C(i|j)	+	-
ACTUAL CLASS	+	-1	100
	-	1	0

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

Model M_1	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	150	40
	-	60	250

Accuracy = 80%

Cost = 3910

Model M_2	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	250	45
	-	5	200

Accuracy = 90%

Cost = 4255

Μέτρα Εκτίμησης

Ταξινόμηση που λαμβάνει υπό όψιν της το κόστος

Κατασκευή Δέντρου Ταξινόμησης

- Επιλογή γνωρίσματος στο οποίο θα γίνει η διάσπαση
- Στην απόφαση αν θα ψαλιδιστεί κάποιο υπο-δέντρο
- Στον καθορισμό της κλάσης του φύλλου

Μέτρα Εκτίμησης

Καθορισμός κλάσης

Leaf-label = max p(i), το ποσοστό των εγγραφών της κλάσης i που έχουν ανατεθεί στον κόμβο

Για δύο κλάσεις, p(+) > 0.5

Στην κλάση που ελαχιστοποιεί το:

$$\text{leaf label} = \sum_j p(j)C(j,i)$$

Για δύο κλάσεις: p(+) × C(+, +) + p(+) × C(+, -)
p(-) × C(-, -) + p(-) × C(-, +)

Αν C(-, -) = C(+, +) = 0

$$p(+) > \frac{C(-,+)}{C(-,+)+C(+,-)}$$

p(+) C(+, -) > p(-) × C(-, +) =>

Αν C(-, +) < C(+, -), τότε λιγότερο του 0.5

Μέτρα Εκτίμησης

Κόστος vs Πιστότητας

Count	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

Η πιστότητα είναι ανάλογη του κόστους αν:

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

Cost	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	p	q
	Class=No	q	p

N = a + b + c + d

Accuracy = (a + d)/N

Cost = p (a + d) + q (b + c)
= p (a + d) + q (N – a – d)
= q N – (q – p)(a + d)
= N [q – (q-p) × Accuracy]

Μέτρα Εκτίμησης

Άλλες μετρήσεις με βάση τον πίνακα σύγχυσης

	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	TP	FN
	Class=No	FP	TN

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

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

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

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

Μέτρα Εκτίμησης

Άλλες μετρήσεις με βάση τον πίνακα σύγχυσης

	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	TP	FN
	Class=No	FP	TN

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

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

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

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

Μέτρα Εκτίμησης

Recall (ανάκληση) – Precision (ακρίβεια)

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	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)

Μέτρα Εκτίμησης

Recall (ανάκληση) – Precision (ακρίβεια)

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

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

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

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;

Μέτρα Εκτίμησης

F_1 measure

$$F_1 = \frac{2rp}{r + p} = \frac{2TP}{2TP + FP + FN}$$

$$F_1 = \frac{2}{1/r + 1/p}$$

Αρμονικό μέσο (Harmonic mean)

Τείνει να είναι πιο κοντά στο μικρότερο από τα δύο

Υψηλή τιμή σημαίνει ότι και τα δύο είναι είναι ικαναποιητικά μεγάλα

Μέτρα Εκτίμησης

Αρμονικά, Γεωμετρικά και Αριθμητικά Μέσα

Παράδειγμα

α=1, b=5

Μέτρα Εκτίμησης

$$\text{Precision (p)} = \frac{TP}{TP + FP}$$

$$\text{Recall (r)} = \frac{TP}{TP + FN}$$

$$\text{F - measure (F)} = \frac{2rp}{r + p} = \frac{2TP}{2TP + FN + FP}$$

- Precision - C(Yes|Yes) & C(Yes|No)
- Recall - C(Yes|Yes) & C(No|Yes)
- F-measure όλα εκτός του C(No|No)

$$\text{Weighted Accuracy} = \frac{w_1 TP + w_4 TN}{w_1 TP + w_2 FP + w_3 FN + w_4 TN}$$

	w1	w2	w3	w4
Recall	1	1	0	0
Precision	1	0	1	1
F1	2	1	1	0
Accuracy	1	1	1	1

Αποτίμηση Μοντέλου: ROC

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

$$\text{TPR} = \frac{TP}{TP + FN} \qquad \text{FPR} = \frac{FP}{TN + FP}$$

Αποτίμηση Μοντέλου: ROC

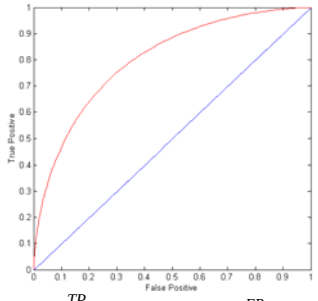
(TP,FP):

- (0,0): declare everything to be negative class
- (1,1): declare everything to be positive class
- (1,0): ideal

Diagonal line:

- Random guessing

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

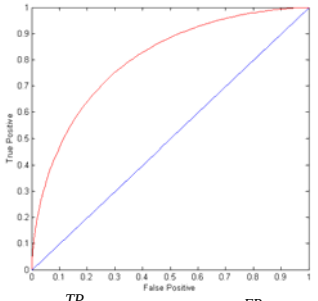


$$\text{TPR} = \frac{TP}{TP + FN} \qquad \text{FPR} = \frac{FP}{TN + FP}$$

Αποτίμηση Μοντέλου: ROC

Καλοί ταξινομητές κοντά στην αριστερή πάνω γωνία του διαγράμματος

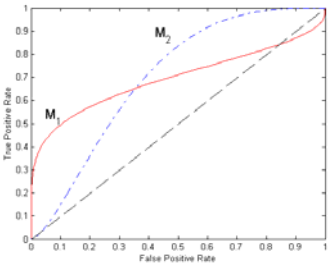
Κάτω από τη διαγώνιο Πρόβλεψη είναι το αντίθετο της πραγματικής κλάσης



$$\text{TPR} = \frac{TP}{TP + FN} \qquad \text{FPR} = \frac{FP}{TN + FP}$$

Αποτίμηση Μοντέλου: ROC

Σύγκριση δύο μοντέλων



- No model consistently outperform the other
  - M_1 is better for small FPR
  - M_2 is better for large FPR
- Area Under the ROC curve
  - Ideal:
    - Area = 1
  - Random guess:
    - Area = 0.5

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

Αποτίμηση Μοντέλου

Έλεγχος Σημαντικότητας (Test of Significance)

- Έστω δύο μοντέλα:
  - Μοντέλο M1: ακρίβεια = 85%, έλεγχος σε 30 εγγραφές
  - Μοντέλο M2: ακρίβεια = 75%, έλεγχος σε 5000 εγγραφές
- Είναι το M1 καλύτερο από το M2?
  - Πόση εμπιστοσύνη (confidence) μπορούμε να έχουμε για την πιστότητα του M1 και πόση για την πιστότητα του M2;
  - Μπορεί η διαφορά στην απόδοση να αποδοθεί σε τυχαία διακύμανση του συνόλου ελέγχου;

Αποτίμηση Μοντέλου

Διάστημα Εμπιστοσύνης για την Ακρίβεια (Confidence Interval)

Prediction can be regarded as a Bernoulli trial

- A Bernoulli trial has 2 possible outcomes
- Possible outcomes for prediction: correct or wrong
- Collection of Bernoulli trials has a Binomial distribution:
  - x ~ Bin(N, p) x: number of correct predictions
  - e.g: Toss a fair coin 50 times, how many heads would turn up?

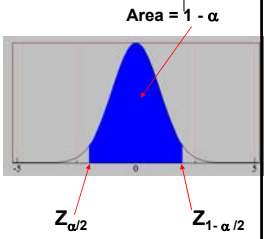
Expected number of heads = $N \times p = 50 \times 0.5 = 25$

Δοθέντος του x (# σωστών προβλέψεων) ή ισοδύναμα, acc=x/N, και του N (# εγγραφών ελέγχου),
Μπορούμε να προβλέψουμε το p (την πραγματική πιστότητα του μοντέλου);

Αποτίμηση Μοντέλου

For large test sets (N > 30), acc has a normal distribution with mean p and variance p(1-p)/N

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



Confidence Interval for p:

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

Αποτίμηση Μοντέλου

Consider a model that produces an accuracy of 80% when evaluated on 100 test instances: Ποιο είναι το διάστημα εμπιστοσύνης για την πραγματική του πιστότητα με επίπεδο εμπιστοσύνης 95%
- N=100, acc = 0.8
- Let 1-α = 0.95 (95% confidence)
- From probability table, $Z_{\alpha/2}$ =1.96
- Κάνοντας τις πράξεις 71.1% - 86.7%

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

Πλησιάζει το 80% όσο το N μεγαλώνει

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

Αποτίμηση Μοντέλου

- Given two models, say M1 and M2, which is better?
  - M1 is tested on D1 (size=n1), found error rate = e_1
  - M2 is tested on D2 (size=n2), found error rate = e_2
  - Assume D1 and D2 are independent

Θέλουμε να εξετάσουμε αν η διαφορά d = e_1-e_2 είναι στατιστικά σημαντική

If n1 and n2 are sufficiently large, then

$$e_1 \sim N(\mu_1, \sigma_1)$$
$$e_2 \sim N(\mu_2, \sigma_2)$$

Approximate

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

Αποτίμηση Μοντέλου

To test if performance difference is statistically significant: d = $e_1 - e_2$
- d ~ $N(d_t, \sigma_t)$ όπου d_t είναι η πραγματική διαφορά
- Since D1 and D2 are independent, their variance adds up:

$$\sigma_t^2 = \sigma_1^2 + \sigma_2^2 \cong \hat{\sigma}_1^2 + \hat{\sigma}_2^2$$
$$= \frac{e1(1-e1)}{n1} + \frac{e2(1-e2)}{n2}$$

At (1-α) confidence level, $d_t = d \pm Z_{\alpha/2}\hat{\sigma}_t$

Αποτίμηση Μοντέλου

Παράδειγμα

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

The estimated variance of the observed difference in error rates

$$\hat{\sigma}_d^2 = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

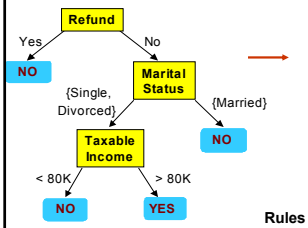
At 95% confidence level, Zα/2=1.96

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

=> Interval contains 0 => difference may not be statistically significant

Άλλοι Ταξινομητές

Από Δέντρα Απόφασης σε Κανόνες



Κανόνες Ταξινόμησης (Classification Rules)

(Refund=Yes) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income<80K) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes

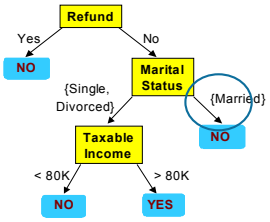
(Refund=No, Marital Status={Married}) ==> No

Rules are mutually exclusive and exhaustive

Rule set contains as much information as the tree

Από Δέντρα Απόφασης σε Κανόνες

Οι κανόνες μπορεί να απλοποιηθούν



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

Initial Rule: **(Refund=No) ∧ (Status=Married) → No**

Simplified Rule: **(Status=Married) → No**

Από Δέντρα Απόφασης σε Κανόνες

- Rules are no longer mutually exclusive
  - A record may trigger more than one rule
  - Solution?
    - Ordered rule set
    - Unordered rule set – use voting schemes
- Rules are no longer exhaustive
  - A record may not trigger any rules
  - Solution?
    - Use a default class

Ταξινομητές βασισμένοι σε Στιγμιότυπα

Μέχρι στιγμής

Ταξινόμηση βασισμένη σε δύο βήματα

Βήμα 1: Induction Step – Κατασκευή Μοντέλου Ταξινομητή

Βήμα 2: Deduction Step – Εφαρμογή του μοντέλου για έλεγχο παραδειγμάτων

Eager Learners vs Lazy Learners

πχ Instance Based Classifiers (ταξινομητές βασισμένη σε στιγμιότυπα)

Μην κατασκευάσεις μοντέλο αν δε χρειαστεί

Ταξινομητές βασισμένοι σε Στιγμιότυπα

Set of Stored Cases

Atr1	.....	AtrN	Class
			A
			B
			B
			C
			A
			C
			B

- **Store the training records**
- **Use training records to predict the class label of unseen cases**

Unseen Case

Atr1	.....	AtrN

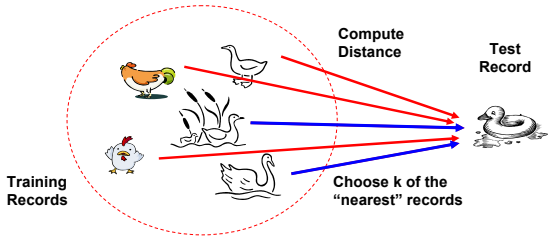
Ταξινομητές βασισμένοι σε Στιγμιότυπα

Παραδείγματα:

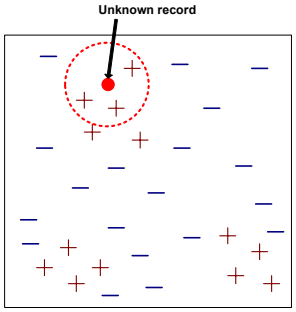
- Rote-learner
  - Memorizes entire training data and performs classification only if attributes of record match one of the training examples exactly
- Nearest neighbor – Κοντινότερος Γείτονας
  - Uses k "closest" points (nearest neighbors) for performing classification

Ταξινομητές Κοντινότερου Γείτονα

Basic idea: If it walks like a duck, quacks like a duck, then it's probably a duck



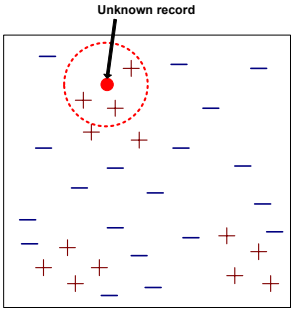
Ταξινομητές Κοντινότερου Γείτονα



Requires three things

1. The set of stored records
2. Distance Metric to compute distance between records
3. The value of *k*, the number of nearest neighbors to retrieve

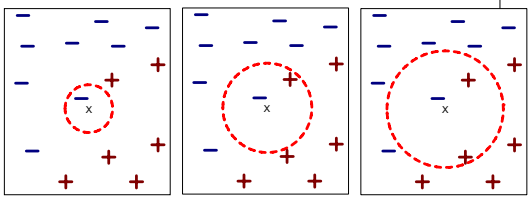
Ταξινομητές Κοντινότερου Γείτονα



To classify an unknown record:

- Compute distance to other training records
- Identify *k* nearest neighbors
- Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

Ταξινομητές Κοντινότερου Γείτονα



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

Ταξινομητές Κοντινότερου Γείτονα

- Compute distance between two points:
  - Euclidean distance

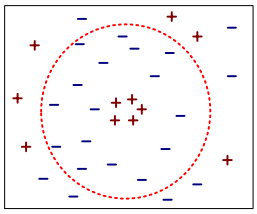
$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- Determine the class from nearest neighbor list
  - take the majority vote of class labels among the k-nearest neighbors
  - Weigh the vote according to distance
    - weight factor, $w = 1/d^2$

Ταξινομητές Κοντινότερου Γείτονα

Επιλογή της τιμής του k:

- If k is too small, sensitive to noise points
- If k is too large, neighborhood may include points from other classes



Ταξινομητές Κοντινότερου Γείτονα

- Scaling issues
  - Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes
  - Example:
    - height of a person may vary from 1.5m to 1.8m
    - weight of a person may vary from 90lb to 300lb
    - income of a person may vary from $10K to $1M

Ταξινομητές Κοντινότερου Γείτονα

- k-NN classifiers are lazy learners
  - It does not build models explicitly
  - Unlike eager learners such as decision tree induction and rule-based systems
  - Classifying unknown records are relatively expensive

Περίληψη

- Ορισμός Προβλήματος Ταξινόμησης
- Μια Κατηγορία Ταξινομητών: Δέντρο Απόφασης
- Μέθοδοι ορισμού της μη καθαρότητας ενός κόμβου
- Θέματα στην Ταξινόμηση: over and under-fitting, missing values, εκτίμηση λάθους
- Αποτίμηση μοντέλου
- Lazy Classifiers

X