Ταξινόμηση ΙΙ

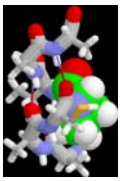
Οι διαφάνειες στηρίζονται στο **P.-N. Tan, M.Steinbach, V. Kumar, «Introduction to Data Mining»**, Addison Wesley, 2006

Σύντομη Επανάληψη

Εισαγωγή
Κατασκευή Δέντρου Απόφασης

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)

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.

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

Εισαγωγή

Βήματα Ταξινόμησης

1. Κατασκευή Μοντέλου

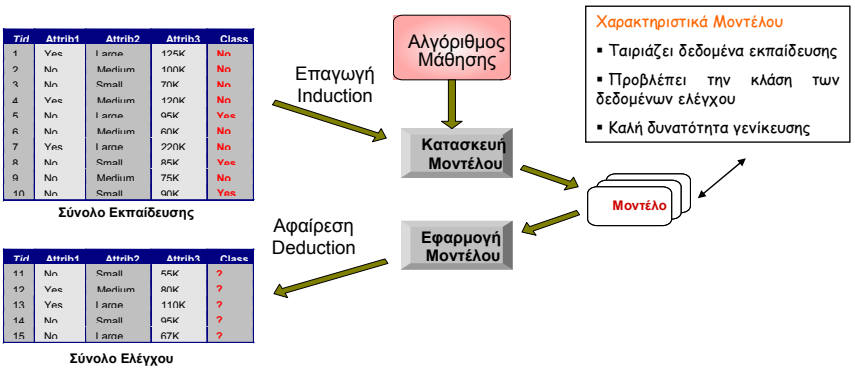
Χρησιμοποιώντας το σύνολο εκπαίδευσης (στις εγγραφές του το γνώρισμα της κλάσης είναι προκαθορισμένο)

Το μοντέλο μπορεί να είναι ένα δέντρο ταξινόμησης, κανόνες, μαθηματικοί τύποι κλπ)

2. Εφαρμογή Μοντέλου για την ταξινόμηση μελλοντικών ή άγνωστων αντικειμένων

Εκτίμηση της ακρίβειας του μοντέλου με χρήση συνόλου ελέγχου

Accuracy rate: το ποσοστό των εγγραφών του συνόλου ελέγχου που ταξινομούνται σωστά από το μοντέλο



Ορισμός

Τεχνικές ταξινόμησης βασισμένες σε

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

Δέντρο Απόφασης

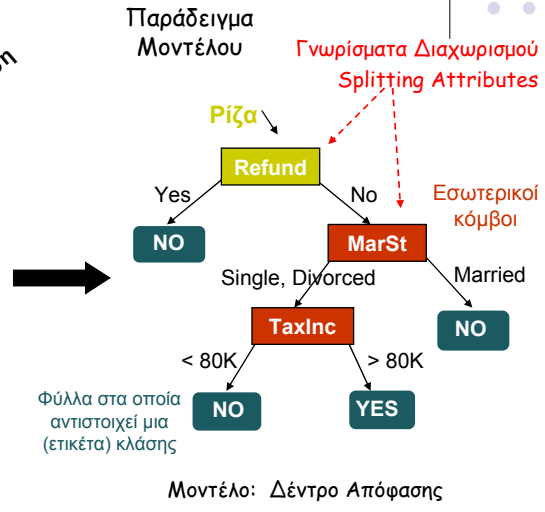
Μοντέλο = Δέντρο Απόφασης

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

Δέντρο Απόφασης: Παράδειγμα

Δεδομένα Εκπαίδευσης

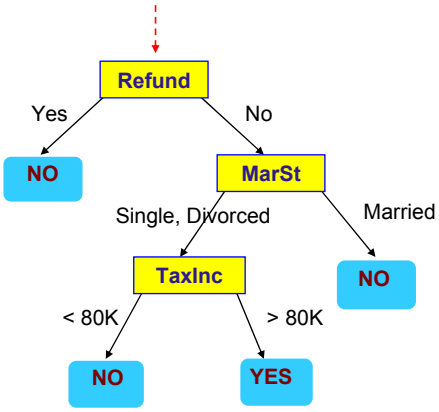
Tid	Refund (κατηγορικό)	Marital Status (κατηγορικό)	Taxable Income (συνεχές)	Cheat (κλάση)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Μοντέλο: Δέντρο Απόφασης

Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Ξεκίνα από τη ρίζα του δέντρου.



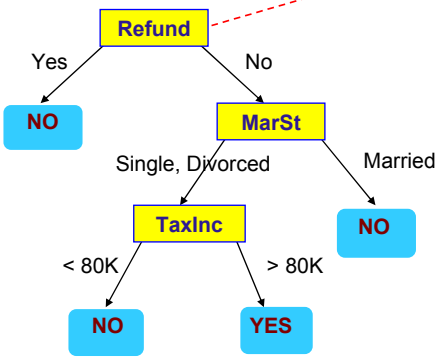
Δεδομένα Ελέγχου

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Test Data

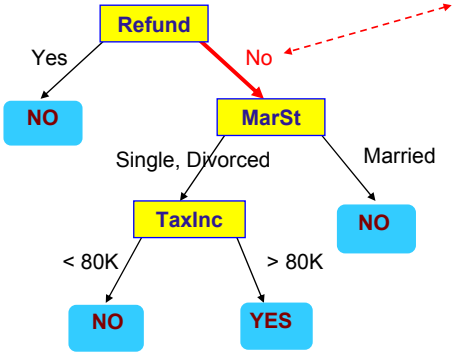
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Test Data

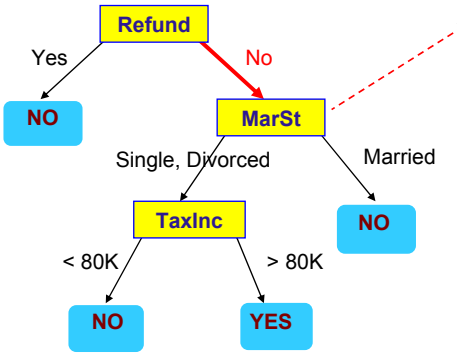
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Test Data

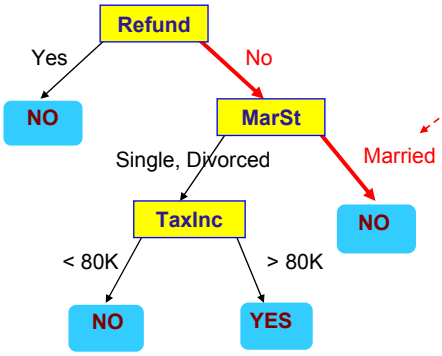
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Test Data

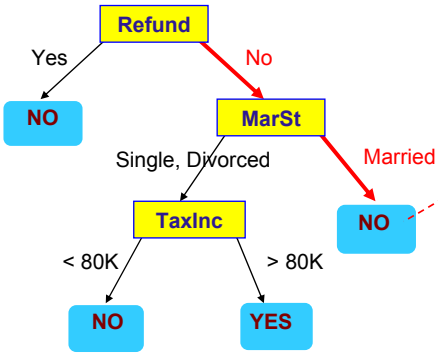
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Είσοδος (δεδομένο ελέγχου)

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Ανάθεση στο Cheat “No”

Δέντρο Απόφασης

Θα δούμε στη συνέχεια αλγορίθμους για την κατασκευή του (βήμα επαγωγής)

Κατασκευή του δέντρου (με λίγα λόγια):

1. ξεκίνα με *έναν* κόμβο που περιέχει *όλες* τις εγγραφές
2. *διαχωρισμός* του κόμβου (μοίρασμα των εγγραφών) με βάση μια συνθήκη-διαχωρισμού σε κάποιο από τα γνωρίσματα
3. Αναδρομική κλήση του 2 σε κάθε κόμβο
4. Αφού κατασκευαστεί το δέντρο, κάποιες βελτιστοποιήσεις (tree pruning)

Το βασικό θέμα είναι

Ποιο γνώρισμα-συνθήκη διαχωρισμού να χρησιμοποιήσουμε για τη διάσπαση των εγγραφών κάθε κόμβου

Δέντρο Απόφασης: Κατασκευή

Ο αριθμός των πιθανών Δέντρων Απόφασης είναι εκθετικός.

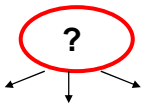
Πολλοί αλγόριθμοι για την **επαγωγή (induction)** του δέντρου οι οποίοι ακολουθούν μια greedy στρατηγική: για να κτίσουν το δέντρο απόφασης παίρνοντας μια σειρά από *τοπικά βέλτιστες* αποφάσεις

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

Δέντρο Απόφασης: Αλγόριθμος του Hunt

Κτίζει το δέντρο αναδρομικά, αρχικά όλες οι εγγραφές σε έναν κόμβο (ρίζα)

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



D_t

Γενική Διαδικασία (αναδρομικά σε κάθε κόμβο)

- 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*,
  τότε χρησιμοποίησε έναν έλεγχο-γνωρίσματος για το διαχωρισμό των δεδομένων σε μικρότερα υποσύνολα

  Σημείωση: ο διαχωρισμός δεν είναι δυνατός 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

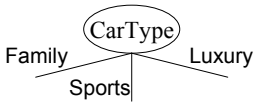
Καθορισμός των συνθηκών του ελέγχου για τα γνωρίσματα

- Εξαρτάται από τον τύπο των γνωρισμάτων
  - Διακριτές - Nominal
  - Διατεταγμένες - Ordinal
  - Συνεχείς - Continuous
- Είδη διαχωρισμού:
  - 2-αδικός διαχωρισμός - 2-way split
  - Πολλαπλός διαχωρισμός - Multi-way split

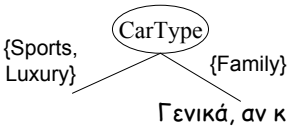
Δέντρο Απόφασης: Κατασκευή Δέντρου

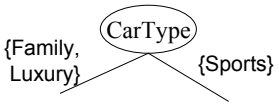
Διαχωρισμός βασισμένος σε διακριτές τιμές

- Πολλαπλός διαχωρισμός: Χρησιμοποίησε τόσες διασπάσεις όσες οι διαφορετικές τιμές



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

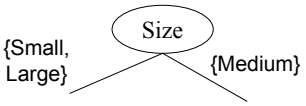




Γενικά, αν κ τιμές, $2^{κ-1} - 1$ τρόποι

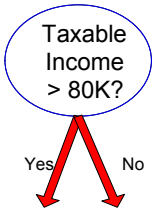
Όταν υπάρχει διάταξη, πρέπει οι διασπάσεις να μη την παραβιάζουν

Αυτός ο διαχωρισμός;

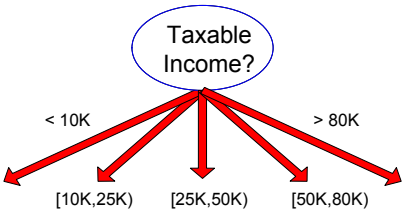


Δέντρο Απόφασης: Κατασκευή Δέντρου

Διαχωρισμός βασισμένος σε συνεχείς τιμές



Δυαδικός διαχωρισμός



Πολλαπλός διαχωρισμός

Δέντρο Απόφασης: Κατασκευή Δέντρου

Σε κάθε επίπεδο, πολλές διαφορετικές δυνατότητες για την διάσπαση. Ποια θα επιλέξουμε;

Ορίζουμε ένα κριτήριο για την «ποιότητα» ενός κόμβου

Έστω μια διάσπαση ενός κόμβου (parent) με N εγγραφές σε k παιδιά u_i

Έστω $N(u_i)$ ο αριθμός εγγραφών κάθε παιδιού ($\sum N(u_i) = N$)

Κοιτάμε το κέρδος, δηλαδή τη διαφορά μεταξύ της ποιότητας του γονέα (πριν τη διάσπαση) και το «μέσο όρο» της ποιότητας ων παιδιών του (μετά τη διάσπαση)

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

← Βάρος (εξαρτάται από τον αριθμό εγγραφών)

Διαλέγουμε τη διάσπαση με το μεγαλύτερο κέρδος (μεγαλύτερο Δ)

Δέντρο Απόφασης: Κατασκευή Δέντρου

Πότε ένας κόμβος είναι καλός;

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

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

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

«Καλός» κόμβος!!

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

N1	
C1	0
C2	6
Μη καθαρότητα ~ 0	

N2	
C1	1
C2	5
ενδιάμεση	

N3	
C1	2
C2	4
ενδιάμεση αλλά μεγαλύτερη	

N4	
C1	3
C2	3
Μεγάλη μη καθαρότητα	

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

Δέντρο Απόφασης: Αλγόριθμος του Hunt

Ψευδό-κώδικας

Algorithm GenDecTree(Sample S, Attlist A)

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

Δέντρο Απόφασης: Κατασκευή Δέντρου

Μέτρα μη Καθαρότητας

1. Ευρετήριο Gini (Gini Index)
2. Εντροπία (Entropy)
3. Λάθος ταξινόμησης (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	
C1	**0**
C2	**6**
Gini=0.000	

N2	
C1	**1**
C2	**5**
Gini=0.278	

N3	
C1	**2**
C2	**4**
Gini=0.444	

N4	
C1	**3**
C2	**3**
Gini=0.500	

- **Ελάχιστη τιμή** (0.0) όταν όλες οι εγγραφές ανήκουν σε μία κλάση
- **Μέγιστη τιμή** (1 - 1/c) όταν όλες οι εγγραφές είναι ομοιόμορφα κατανεμημένες στις κλάσεις

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

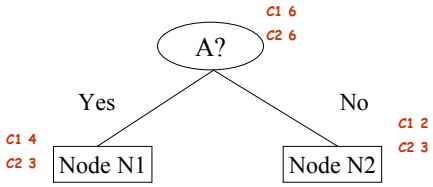
Δέντρο Απόφασης: GINI

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

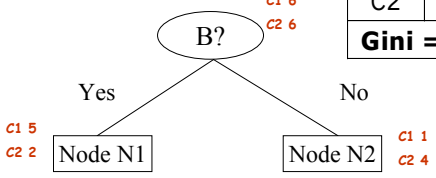
Δέντρο Απόφασης: GINI

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

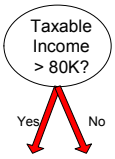
Δέντρο Απόφασης: GINI

Συνεχή Γνωρίσματα

Χρήση **δυαδικού διαχωρισμού** 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 διαφορετική τιμή 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

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 <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

Παράδειγμα – Διαχωρισμός στο γνώρισμα Income

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

βελτίωση βελτίωση βελτίωση

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

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

Δέντρο Απόφασης: Εντροπία

Εντροπία για τον κόμβο t : $$Entropy(t) = -\sum_{j=1}^{c} p(j \mid t) \log_2 p(j \mid t)$$

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

Μετράει την ομοιογένεια ενός κόμβου

N1	
C1	0
C2	6
Entropy=0.000	
Gini = 0.000	

N2	
C1	1
C2	5
Entropy=0.650	
Gini = 0.278	

N3	
C1	2
C2	4
Entropy = 0.92	
Gini = 0.444	

N4	
C1	3
C2	3
Entropy = 1.000	
Gini = 0.500	

- **Μέγιστη τιμή** log(c) όταν όλες οι εγγραφές είναι ομοιόμορφα κατανεμημένες στις κλάσεις
- **Ελάχιστη τιμή** (0.0) όταν όλες οι εγγραφές ανήκουν σε μία κλάση

Δέντρο Απόφασης: Εντροπία

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

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

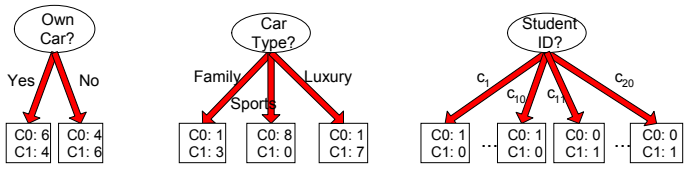
όπου, n_i = αριθμός εγγραφών του παιδιού i,
n = αριθμός εγγραφών του κόμβου p.

- 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: Λάθος Ταξινόμησης

Λάθος ταξινόμησης (classification error) για τον κόμβο t :

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

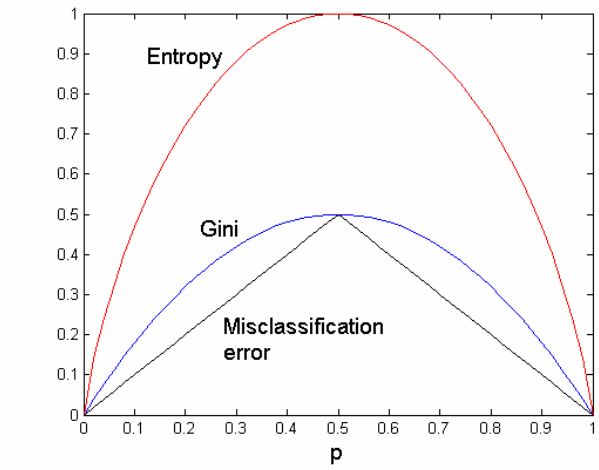
Μετράει το λάθος ενός κόμβου: επειδή δίνουμε στον κόμβο την κλάση της πλειοψηφίας (max p(i|t)), όλα τα άλλα (1-max) ταξινομούνται λάθος

N1		N2		N3		N4	
C1	0	C1	1	C1	2	C1	3
C2	6	C2	5	C2	4	C2	3
Error=0.000		Error=0.167		Error = 0.333		Error = 0.500	
Gini = 0.000		Gini = 0.278		Gini = 0.444		Gini = 0.500	
Entropy = 0.000		Entropy = 0.650		Entropy = 0.920		Entropy = 1.000	

- **Μέγιστη τιμή** 1-1/c όταν όλες οι εγγραφές είναι ομοιόμορφα κατανεμημένες στις κλάσεις
- **Ελάχιστη τιμή** (0.0) όταν όλες οι εγγραφές ανήκουν σε μία κλάση

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

N1	
C1	0
C2	6
Error=0.000	

Gini = 0.000
Entropy = 0.000

N2	
C1	1
C2	5
Error=0.167	

Gini = 0.278
Entropy = 0.650

N3	
C1	2
C2	4
Error = 0.333	

Gini = 0.444
Entropy = 0.920

N4	
C1	3
C2	3
Error = 0.500	

Gini = 0.500
Entropy = 1.000

Δέντρο Απόφασης: Αλγόριθμος του Hunt

Ψευδό-κώδικας (πάλι)

Algorithm GenDecTree(Sample S, Attlist A)

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

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
  - με βάση τις εγγραφές
  - με βάση το λάθος

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

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

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

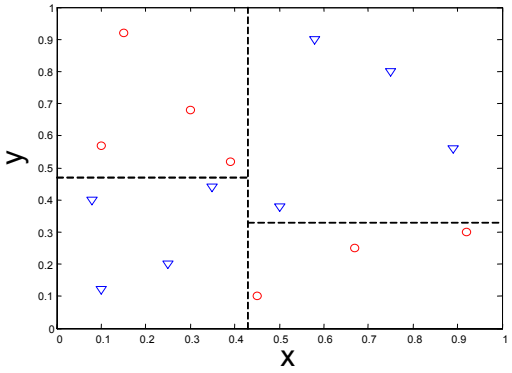
Decision Boundary

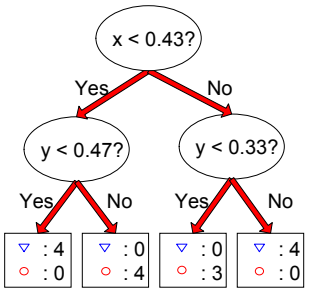
Μέχρι στιγμής είδαμε ελέγχους που αφορούν μόνο ένα γνώρισμα τη φορά, μπορούμε να δούμε τη διαδικασία ως τη διαδικασία *διαμερισμού του χώρου* των γνωρισμάτων σε ξένες περιοχές μέχρι κάθε περιοχή να περιέχει εγγραφές που να ανήκουν στην ίδια κλάση

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

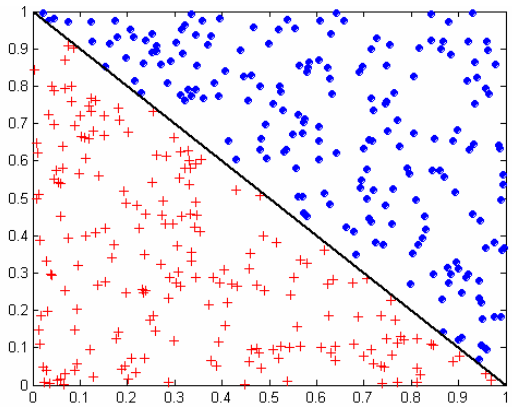
Δέντρο Απόφασης

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

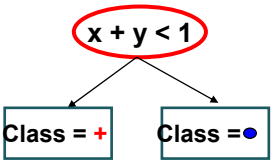




Δέντρο Απόφασης



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

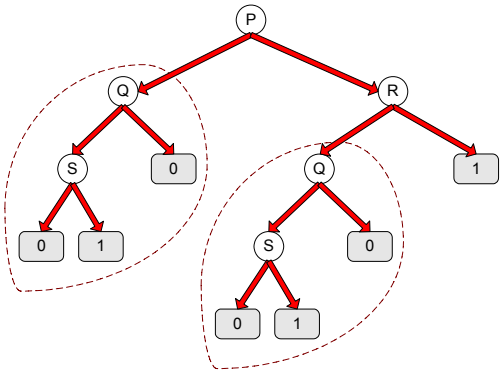
Δέντρο Απόφασης

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

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: 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

Δέντρο Απόφασης - Περίληψη

- Προτερήματα - Pros
  - + Λογικός χρόνος εκπαίδευσης
  - + Γρήγορη εφαρμογή
  - + Ευκολία στην κατανόηση
  - + Εύκολη υλοποίηση
  - + Μπορεί να χειριστεί μεγάλο αριθμό γνωρισμάτων
- Μειονεκτήματα - Cons
  - – Δεν μπορεί να χειριστεί περίπλοκες σχέσεις μεταξύ των γνωρισμάτων
  - – Απλά όρια απόφασης (decision boundaries)
  - – Προβλήματα όταν λείπουν πολλά δεδομένα

Θέματα στην Ταξινόμηση

Θέματα Ταξινόμησης

- Underfitting and Overfitting
- Εκτίμηση Λάθους
- Τιμές που λείπουν

Εκτίμηση του Λάθους

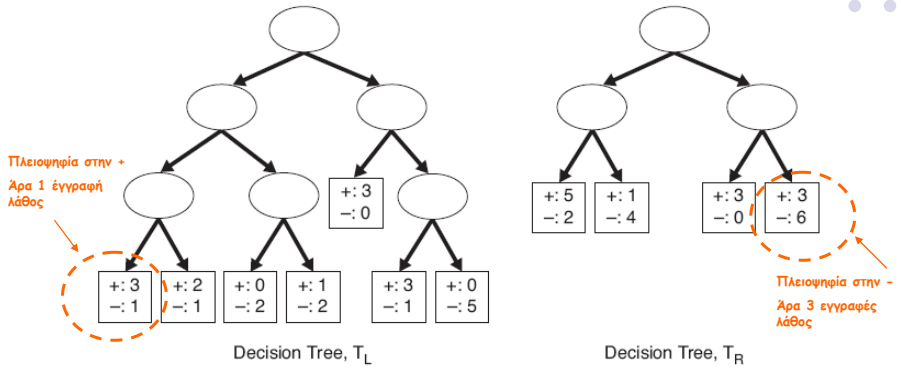
Ως λάθος μετράμε τις εγγραφές που ο ταξινομητής τοποθετεί σε λάθος κλάση

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

Εκτίμηση του Λάθους

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



Παράδειγμα δύο δέντρων για τα ίδια δεδομένα εκπαίδευσης

Με βάση το λάθος εκπαίδευσης

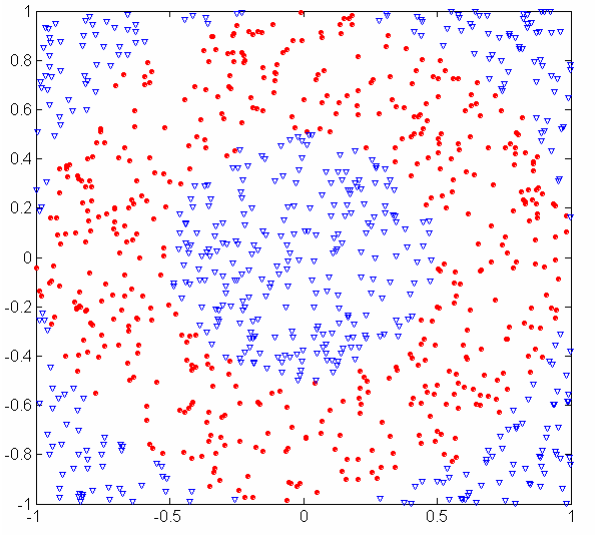
Αριστερό 4/24 = 0.167 Δεξί: 6/24 = 0.25

Overfitting

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):

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

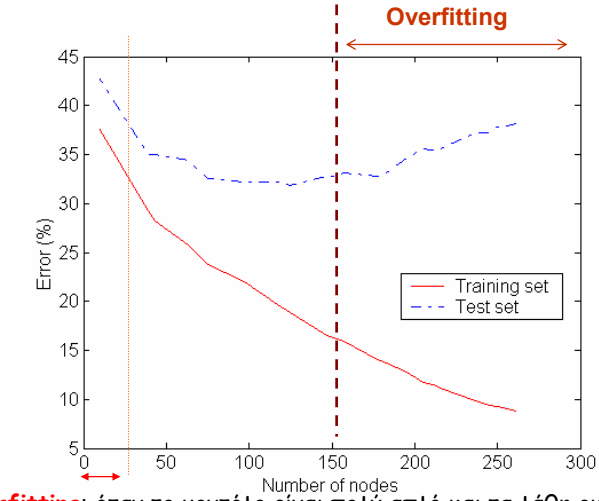
Για τα σημεία της κλάσης 2 (τριγωνικά σημεία):

$sqrt(x_1^2+x_2^2) > 0.5$ or

$sqrt(x_1^2+x_2^2) < 1$

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

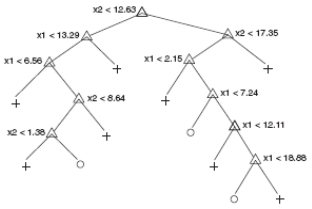
Overfitting

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

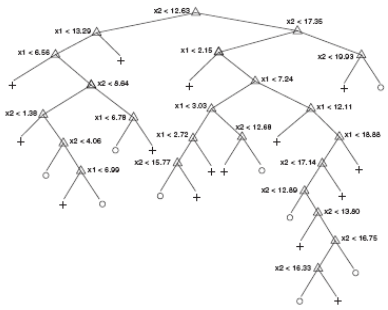
Μικρό (μηδενικό) λάθος εκπαίδευσης

Μεγάλο λάθος ελέγχου

Και το ανάποδο, μπορεί επίσης να ισχύει



(a) Decision tree with 11 leaf nodes.

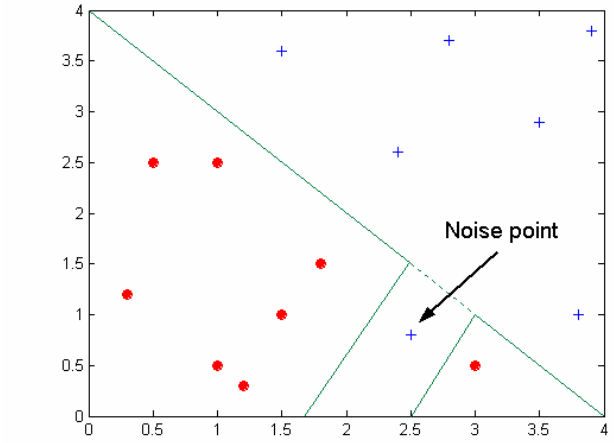


(b) Decision tree with 24 leaf nodes.

Δέντρα απόφασης με διαφορετική πολυπλοκότητα

Overfitting

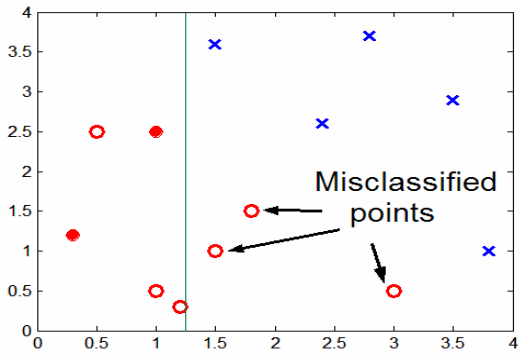
Overfitting εξαιτίας Θορύβου



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

Αντιμετώπιση Overfitting

Δύο βασικές προσεγγίσεις:

Pre-pruning

Σταμάτημα της ανάπτυξης του δέντρου μετά από κάποιο σημείο

Post-pruning

Η κατασκευή του δέντρου χωρίζεται σε δύο φάσεις:

1. Φάση Ανάπτυξης
2. Φάση Ψαλιδίσματος

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)

Χρησιμοποιείται πιο συχνά

Χρήση άλλων δεδομένων για τον υπολογισμό του καλύτερου δέντρου (δηλαδή του λάθους γενίκευσης)

Εκτίμηση του Λάθους Ταξινόμησης

Εκτίμηση Λάθους Γενίκευσης

Ως λάθος μετράμε τις εγγραφές που ο ταξινομητής τοποθετεί σε λάθος κλάση

- Χρήση Δεδομένων Εκπαίδευσης
  - αισιόδοξη εκτίμηση
  - απαισιόδοξη εκτίμηση
- 3. Χρήση Δεδομένων Ελέγχου

Εκτίμηση του Λάθους Γενίκευσης

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

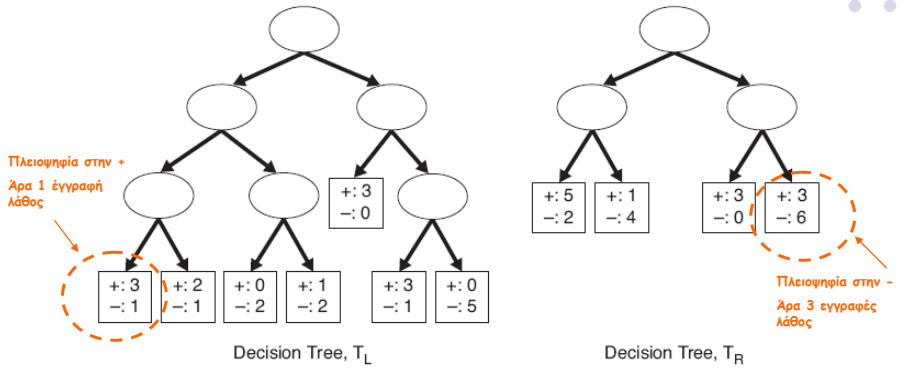
Ως λάθος μετράμε το ποσοστό των εγγραφών που ο ταξινομητής τοποθετεί σε λάθος κλάση

Μέθοδοι εκτίμησης του λάθους γενίκευσης:

1. Optimistic approach – Αισιόδοξη προσέγγιση:

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

Εκτίμηση του Λάθους Γενίκευσης



Παράδειγμα δύο δέντρων για τα ίδια δεδομένα – Το δέντρο στο δεξί (T_R) μετά από ψαλίδισμα του δέντρου στα αριστερά (T_L) – *sub-tree raising*

Με βάση το λάθος εκπαίδευσης

Αριστερό 4/24 = 0.167 Δεξί: 6/24 = 0.25

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

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, V(t) = 0.5: $e'(t) = e(t) + 0.5$

Συνολικό λάθος: $e'(T) = e(T) + k \times 0.5$ (*k*: αριθμός φύλλων)

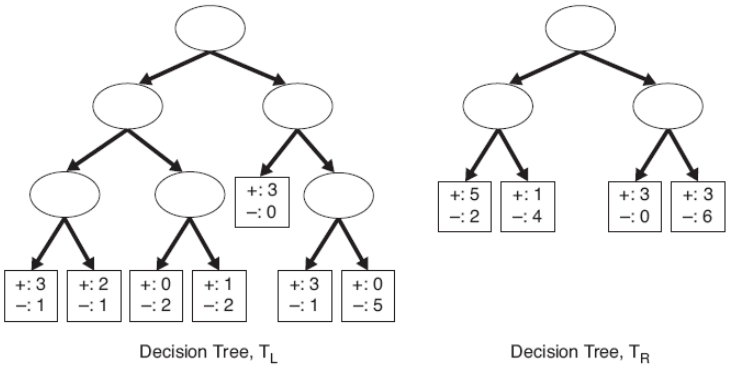
Το 0.5 σημαίνει ότι διαχωρισμός ενός κόμβου δικαιολογείται αν βελτιώνει τουλάχιστον μία εγγραφή

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%

Εκτίμηση του Λάθους Γενίκευσης



Decision Tree, T_L Decision Tree, T_R

Παράδειγμα δύο δέντρων για τα ίδια δεδομένα

Με βάση το λάθος εκπαίδευσης

Αριστερό (4 + 7*0.5)/24 = 0.3125

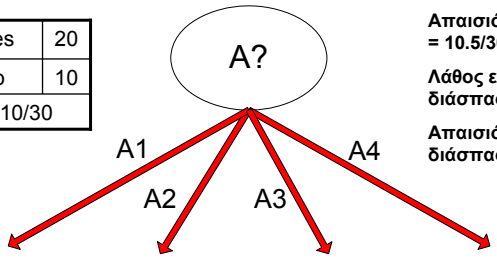
Δεξί: (6 + 4*0.5)/24 = 0.3333

Αν αντί για 0.5, κάτι μεγαλύτερο;

Overfitting

Παράδειγμα Post-Pruning

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



Class = Yes	8
Class = No	4

Class = Yes	4
Class = No	1

Class = Yes	5
Class = No	1

Class = Yes	3
Class = No	4

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 × 0.5)/30 = 11/30

PRUNE!

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

Overfitting

Παράδειγμα post-pruning

- Αισιόδοξη προσέγγιση?

  Όχι διάσπαση

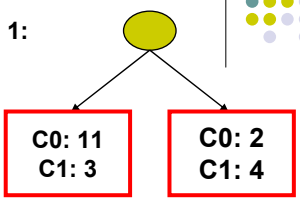
- Απαισιόδοξη προσέγγιση?

  όχι case 1, ναι case 2

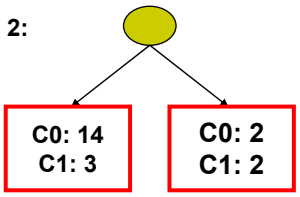
- REP?

Εξαρτάται από το σύνολο επαλήθευσης

Case 1:



Case 2:



Τιμές που λείπουν

Οι τιμές που λείπουν (missing values) 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

Κόμβος 1

Κόμβος 2

3
7

0
3

2
4

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

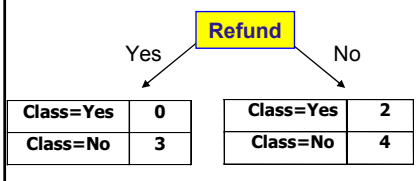
Tid	Refund	Marital Status	Taxable Income	Class
10	?	Single	90K	Yes

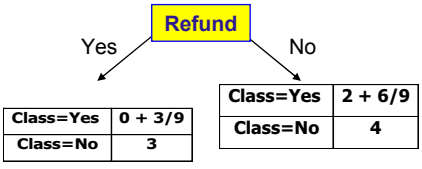
Σε ποιο φύλλο;

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

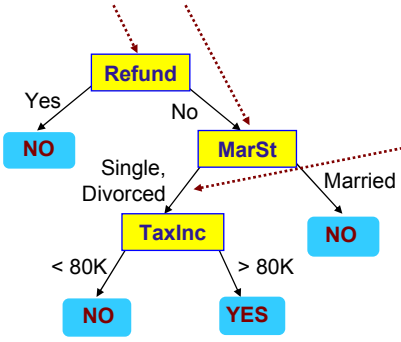




Τιμές που λείπουν

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

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

Αποτίμηση Μοντέλου

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

Ισχύουν για όλα τα μοντέλα ταξινόμησης

Μέτρα Εκτίμησης

Αφού κατασκευαστεί ένα μοντέλο, θα θέλαμε να αξιολογήσουμε/εκτιμήσουμε την ποιότητα του/την ακρίβεια της ταξινόμησης που πετυχαίνει

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

Μέτρα Εκτίμησης

Confusion Matrix (Πίνακας Σύγχυσης)

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

	πρόβλεψη PREDICTED CLASS		
πραγματική ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	f_{11} TP	f_{10} FN
	Class=No	f_{01} FP	f_{00} TN

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

Μέτρα Εκτίμησης

Πιστότητα - Accuracy

Πιστότητα (ακρίβεια;) (accuracy)
Το πιο συνηθισμένο μέτρο

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

= 0

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

Λόγος Λάθους

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

ErrorRate(C) = 1 – Accuracy(C)

Αποτίμηση Μοντέλου

Μπορούμε να χρησιμοποιήσουμε τα λάθη εκπαίδευσης/γενίκευσης (αισιόδοξη ή απαισιόδοξη προσέγγιση)

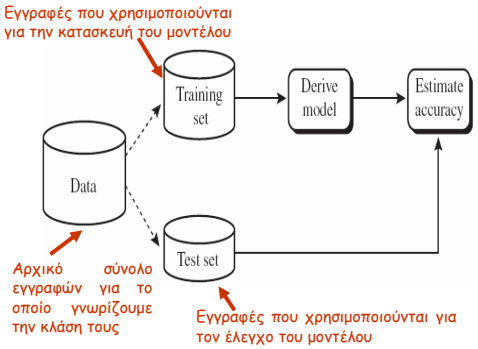
Δεν είναι κατάλληλα γιατί βασίζονται στα δεδομένα εκπαίδευσης μόνο

Συνήθως, σύνολο ελέγχου

Μέθοδοι Αποτίμησης Μοντέλου

Μέθοδος Holdout

Διαμέριση του αρχικού συνόλου σε δύο ξένα σύνολα:
Σύνολο εκπαίδευσης (2/3) – Σύνολο Ελέγχου (1/3)



- Κατασκευή μοντέλου με βάση το σύνολο εκπαίδευσης
- Αποτίμηση μοντέλου με βάση το σύνολο ελέγχου

Μέθοδοι Αποτίμησης Μοντέλου

Μέθοδος Holdout

(-) 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 (υποσύνολα του ίδιου συνόλου - 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 βελτίωσή της
έστω k επαναλήψεις, παίρνουμε το μέσο όρο της ακρίβειας

(-) Πάλι αφαιρούμε δεδομένα από το σύνολο εκπαίδευσης

(-) Ένα ακόμα πρόβλημα είναι ότι μια εγγραφή μπορεί να χρησιμοποιείται (επιλέγεται) ως εγγραφή εκπαίδευσης πιο συχνά από κάποια άλλη

Μέθοδοι Αποτίμησης Μοντέλου

Cross validation

Κάθε εγγραφή χρησιμοποιείται τον ίδιο αριθμό φορών στην εκπαίδευση και ακριβώς μια φορά για έλεγχο

- Διαμοίραση των δεδομένων σε *k* ίσα διαστήματα
- Κατασκευή του μοντέλου αφήνοντας κάθε φορά ένα διάστημα ως σύνολο ελέγχου και χρησιμοποιώντας όλα τα υπόλοιπα ως σύνολα εκπαίδευσης
- Επανάληψη *k* φορές

2-fold (δύο ίσα υποσύνολα, το ένα μια φορά για έλεγχο – το άλλο για εκπαίδευση και μετά ανάποδα)

Αν k = N, (N ο αριθμός των εγγραφών) leave-one-out

Μέθοδοι Αποτίμησης Μοντέλου

Bootstrap

Sample with replacement – δειγματοληψία με επανένταξη

Μια εγγραφή που επιλέχθηκε ως δεδομένο εκπαίδευσης, *ξαναμπαίνει στο αρχικό σύνολο*
Οι υπόλοιπες εγγραφές (όσες δεν επιλεγούν στο σύνολο εκπαίδευσης) – εγγραφές ελέγχου

Αν N δεδομένα, ένα δείγμα N στοιχείων 63.2% των αρχικών

Πιθανότητα ένα δεδομένο να επιλεγεί $1 - (1-1/N)^N$

Για μεγάλο N, η πιθανότητα επιλογής τείνει ασυμπτωτικά στο $1-e^{-1} = 0.632$, πιθανότητα μη επιλογής 0.368

.632 boostrap

$$acc_{boot} = \frac{1}{b}\sum_{i=1}^{b}(0.6328 * errortest_i + 0.368 * acc_s)$$

b: αριθμός επαναλήψεων

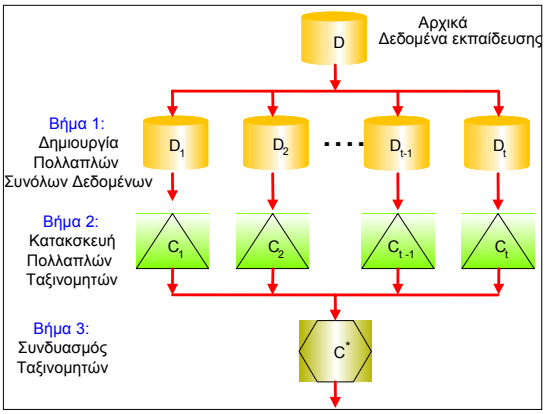
acc_s ακρίβεια όταν όλα τα δεδομένα ως σύνολο εκπαίδευσης

Βελτίωση Απόδοσης

Ensemble Methods – Σύνολο Μεθόδων

Κατασκευή ενός συνόλου από ταξινομητές από τα δεδομένα εκπαίδευσης C_1, C_2, ... C_t -> C^*

Υπολογισμός της κλάσης των δεδομένων συναθροίζοντας (aggregating) τις προβλέψεις των t ταξινομητών
Πως: πχ με πλειοψηφικό σύστημα (Voting majority)



Βελτίωση Απόδοσης

- Έστω t = 25 βασικοί ταξινομητές
  - Αν ο καθένας λάθος, ε = 0.35
  - Έστω ότι ανεξάρτητοι και μόνο 2 κλάσεις
  - Πιθανότητα λανθασμένης πρόβλεψης του συνόλου:

$$\sum_{i=13}^{25} \binom{25}{i} \varepsilon^i (1-\varepsilon)^{25-i} = 0.06$$

Βελτίωση Απόδοσης

Bagging (Bootstarp + Aggregation)

- Δειγματοληψία με επανένταξη (Sampling with replacement)
- Κατασκευή ταξινομητή για κάθε δείγμα
- Κάθε δείγμα έχει πιθανότητα $(1 - 1/n)^n$ να επιλεγεί

Original Data	1	2	3	4	5	6	7	8	9	10
Bagging (Round 1)	7	8	10	8	2	5	10	10	5	9
Bagging (Round 2)	1	4	9	1	2	3	2	7	3	2
Bagging (Round 3)	1	8	5	10	5	5	9	6	3	7

Boosting

Δε δίνουμε το ίδιο βάρος σε όλους τους ταξινομητές, αλλά παίρνουμε υπόψη μας την ακρίβειά τους -- C* βάρος με βάση την ακρίβεια του

- Βασική ιδέα:

Έστω C_i, ο C_{i+1} μεγαλύτερο λάθος στις πλειάδες που ταξινόμησε λάθος ο C_i

Πως; «πειράζουμε» την πιθανότητα επιλογής τους στο σύνολο εκπαίδευσης

σωστά, πιθανότητα επιλογής -

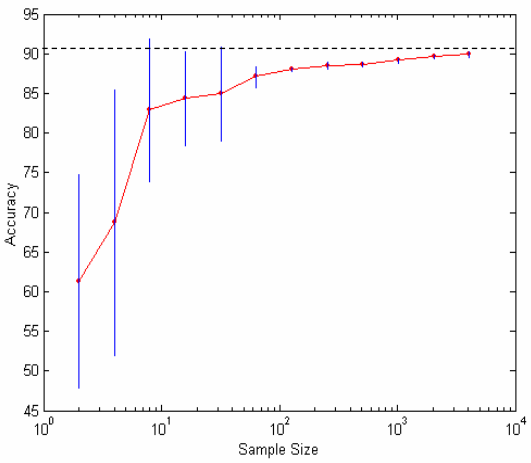
λάθος, πιθανότητα επιλογής +

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 (accuracy) με την αύξηση του μεγέθους του δείγματος
- Επίδραση δείγματος μικρού μεγέθους:
  - Bias in the estimate
  - Variance of estimate

Άλλα Μέτρα Εκτίμησης πέραν της Πιστότητας

Πίνακας σύγχυσης

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

Πιστότητα (accuracy) -- υπενθύμιση --

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

Λόγος Λάθους

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

Άλλα Μέτρα Εκτίμησης πέραν της Πιστότητας

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 = +**	**Class = -**
ACTUAL CLASS	**Class = +**	C(+, +)	C(+, -)
	Class = -	C(-, +)	C(-, -)

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

		PREDICTED CLASS	
	C(i|j)	**Class = +**	**Class = -**
ACTUAL CLASS	**Class = +**	**TP** F_{11}	**FN** F_{10}
	Class = -	**FP** F_{01}	**TN** F_{00}

Αρνητική τιμή κόστους σημαίνει επιπρόσθετη «επιβράβευση» σωστής πρόβλεψης

$C(M) = TP \times C(+, +) + FN \times C(+, -) + FP\ C(-, +) + TN\ C(-, -)$

Στα προηγούμενα, είχαμε

$C(+, +) = C(-, -) = 0$ -> όχι επιβράβευση

$C(+, -) = C(-, +) = 1$ -> κάθε λάθος μετρά 1

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

Κανονικά, ως ετικέτα ενός φύλλου την πλειοψηφούσα κλάση,

Έστω p(j) τον ποσοστό των εγγραφών του κόμβου που ανήκουν στην κλάση j

Τότε,

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

Τώρα, δίνουμε την κλάση i στον κόμβο που έχει το ελάχιστο:

$$\sum_j p(j)C(j,i)$$

Για όλες τις κλάσεις

Μέτρα Εκτίμησης

Έστω 2 κλάσεις: + και -

Αν όχι κόστος, ετικέτα +, ανν, p(+) > 0.5

Τώρα, αυτήν με το μικρότερο κόστος:

κόστος της κλάσης - : p(+) x C(+, +) + p(+) x C(+, -)

κόστος της κλάσης + : p(-) x C(-, -) + p(-) x C(-, +)

Αν C(-, -) = C(+, +) = 0 (όχι κόστος (επιβράβευση) στα σωστά)

Δίνουμε +, αν

p(+) C(+, -) > p(-) x C(-, +) =>

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

p(-) = 1 – p(+)

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

Μέτρα Εκτίμησης

Κόστος vs Πιστότητας (Accuracy)

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

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

Η πιστότητα είναι ανάλογη του κόστους αν:

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

N = a + b + c + d

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 ικανοποιητικά 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 [TruePositiveRate] (στον άξονα των y) προς τα FPR [FalsePositiveRate] (στον άξονα των x)
- Η απόδοση κάθε ταξινομητή αναπαρίσταται ως ένα σημείο στην καμπύλη ROC

Πόσα από τα θετικά βρίσκει

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

Πόσα από τα αρνητικά βρίσκει

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

Αποτίμηση Μοντέλου: ROC

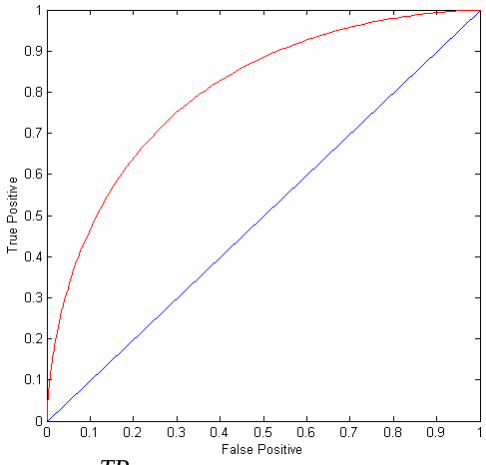
(TP,FP):

- (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

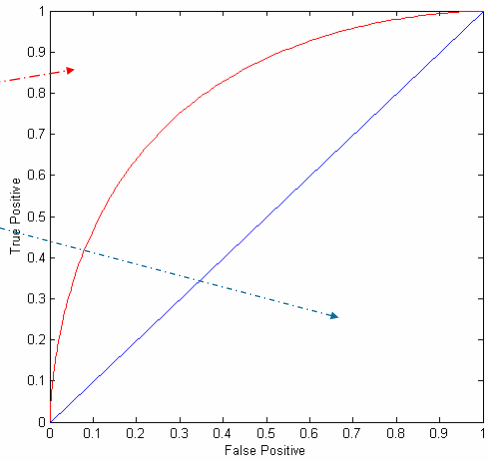


$$\text{TPR} = \frac{TP}{TP + FN} \qquad \text{FPR} = \frac{FP}{TN + FP}$$

Αποτίμηση Μοντέλου: ROC

Καλοί ταξινομητές κοντά στην αριστερή πάνω γωνία του διαγράμματος

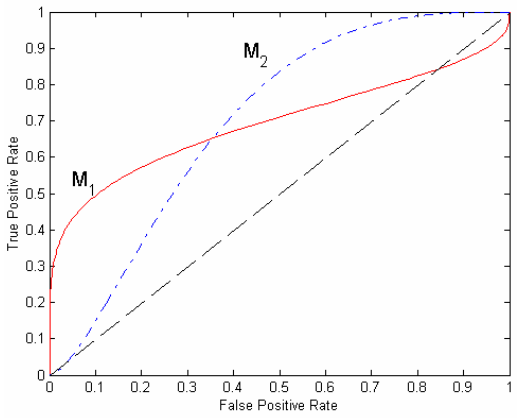
Κάτω από τη διαγώνιο Πρόβλεψη είναι το αντίθετο της πραγματικής κλάσης



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

Αποτίμηση Μοντέλου: ROC

Σύγκριση δύο μοντέλων



- Κανένα μοντέλο δεν είναι πάντα καλύτερο του άλλου
  - M_1 καλύτερο για μικρό FPR
  - M_2 καλύτερο για μεγάλο FPR
- Η περιοχή κάτω από την καμπύλη ROC
  - 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 Μ1: ακρίβεια = 85%, έλεγχος σε 30 εγγραφές
  - Μοντέλο Μ2: ακρίβεια = 75%, έλεγχος σε 5000 εγγραφές
- Είναι το Μ1 καλύτερο από το Μ2;
  - Πόση εμπιστοσύνη (confidence) μπορούμε να έχουμε για την πιστότητα του Μ1 και πόση για την πιστότητα του Μ2;
  - Μπορεί η διαφορά στην απόδοση να αποδοθεί σε τυχαία διακύμανση του συνόλου ελέγχου;

Αποτίμηση Μοντέλου

Διάστημα Εμπιστοσύνης για την Ακρίβεια (Confidence Interval)

Η πρόβλεψη μπορεί να θεωρηθεί σε ένα πείραμα Bernoulli

- Ένα Bernoulli πείραμα έχει δύο πιθανά αποτελέσματα
- Πιθανά αποτελέσματα πρόβλεψης: σωστό ή λάθος
- Μια συλλογή από πειράματα έχει δυωνυμική κατανομή Binomial distribution:
  - $x \sim Bin(N, p)$ x: αριθμός σωστών προβλέψεων
  - Πχ: ρίξιμο τίμιου νομίσματος (κορώνα/γράμματα) 50 φορές, αριθμός κεφαλών;
    Expected number of heads = $N \times p = 50 \times 0.5 = 25$

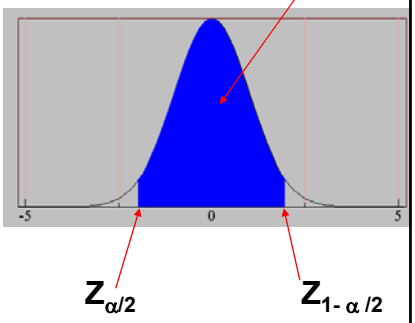
Δοθέντος του x (# σωστών προβλέψεων) ή ισοδύναμα, acc=x/N, και του N (# εγγραφών ελέγχου),

Μπορούμε να προβλέψουμε το p (την πραγματική πιστότητα του μοντέλου);

Αποτίμηση Μοντέλου

Για μεγάλα σύνολα ελέγχου (N > 30), acc έχει κανονική κατανομή με μέσο 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
(Διάστημα εμπιστοσύνης για το p):

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

Αποτίμηση Μοντέλου

Έστω ένα μοντέλο που έχει accuracy 80% όταν αποτιμάται σε 100 στιγμιότυπα ελέγχου: Ποιο είναι το διάστημα εμπιστοσύνης για την πραγματική του πιστότητα (p) με επίπεδο εμπιστοσύνης (1-a) 95%

N=100, acc = 0.8
1-α = 0.95 (95% confidence)
Από τον πίνακα, $Z_{\alpha/2}$ =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

- Δοσμένων δύο μοντέλων, έστω M1 και M2, ποιο είναι καλύτερο;
  - M1 ελέγχεται στο D1 (size=n1), error rate = e_1
  - M2 ελέγχεται στο D2 (size=n2), error rate = e_2
  - Έστω D1 and D2 είναι ανεξάρτητα

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

Αν τα n1 και n2 είναι αρκετά are μεγάλα, τότε:

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

Approximate

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

Αποτίμηση Μοντέλου

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

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

Η εκτιμώμενη variance της διαφοράς στα error rates

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

Για 95% confidence level, Zα/2=1.96

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

=> Το διάστημα περιέχει το 0 => η διαφορά μπορεί να είναι στατιστικά μη σημαντική

Άλλοι Ταξινομητές
Ταξινομητές με κανόνες

Ταξινομητές με Κανόνες

Ταξινόμηση των εγγραφών με βάση ένα σύνολο από κανόνες της μορφής "if…then…"

Κανόνας: $(Συνθήκη) \rightarrow y$

όπου

Συνθήκη (Condition) είναι σύζευξη συνθηκών στα γνωρίσματα

y η ετικέτα της κλάσης

LHS: rule antecedent (πρότερο) ή condition (συνθήκη)
RHS: rule consequent (επακόλουθο ή απότοκο)

Παραδείγματα κανόνων ταξινόμησης:

(Blood Type=Warm) $\wedge$ (Lay Eggs=Yes) $\rightarrow$ Birds
(Taxable Income < 50K) $\wedge$ (Refund=Yes) $\rightarrow$ Evade=No

Ταξινομητές με Κανόνες

Παράδειγμα

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

R1: (Give Birth = no) $\wedge$ (Can Fly = yes) $\rightarrow$ Birds
R2: (Give Birth = no) $\wedge$ (Live in Water = yes) $\rightarrow$ Fishes
R3: (Give Birth = yes) $\wedge$ (Blood Type = warm) $\rightarrow$ Mammals
R4: (Give Birth = no) $\wedge$ (Can Fly = no) $\rightarrow$ Reptiles
R5: (Live in Water = sometimes) $\rightarrow$ Amphibians

Ταξινομητές με Κανόνες

Εφαρμογή Ταξινομητών με Κανόνες

Ένας κανόνας r καλύπτει (covers) ένα στιγμιότυπο (εγγραφή) αν τα γνωρίσματα του στιγμιότυπου ικανοποιούν τη συνθήκη του κανόνα

R1: (Give Birth = no) $\wedge$ (Can Fly = yes) $\rightarrow$ Birds
R2: (Give Birth = no) $\wedge$ (Live in Water = yes) $\rightarrow$ Fishes
R3: (Give Birth = yes) $\wedge$ (Blood Type = warm) $\rightarrow$ Mammals
R4: (Give Birth = no) $\wedge$ (Can Fly = no) $\rightarrow$ Reptiles
R5: (Live in Water = sometimes) $\rightarrow$ Amphibians

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
hawk	warm	no	yes	no	?
grizzly bear	warm	yes	no	no	?

Ο κανόνας R1 καλύπτει το hawk (ή αλλιώς το hawk ενεργοποιεί (trigger) τον κανόνα) => Bird

Ο κανόνας R3 καλύπτει το grizzly bear => Mammal

Ταξινομητές με Κανόνες

Κάλυψη Κανόνα - Coverage:
Το ποσοστό των εγγραφών που ικανοποιούν το LHS του κανόνα

Πιστότητα Κανόνα - Accuracy:
Το ποσοστό των κανόνων που καλύπτουν και το LHS και το RHS του κανόνα

(Status=Single) → No

Coverage = 40%, Accuracy = 50%

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	No	**Single**	90K	**Yes**

Ταξινομητές με Κανόνες

Χαρακτηριστικά Ταξινομητών με Κανόνες

- Αμοιβαία αποκλειόμενοι κανόνες (Mutually exclusive rules)
  Ένας ταξινομητής περιέχει αμοιβαία αποκλειόμενους κανόνες, αν οι κανόνες είναι ανεξάρτητοι ο ένας από τον άλλο

  Κάθε εγγραφή καλύπτεται από το *πολύ έναν* κανόνα

- Εξαντλητικοί κανόνες (Exhaustive rules)
  Ένας ταξινομητής έχει εξαντλητική κάλυψη (coverage) αν καλύπτει όλους τους πιθανούς συνδυασμούς τιμών γνωρισμάτων

  Κάθε εγγραφή καλύπτεται από *τουλάχιστον έναν* κανόνα

Ταξινομητές με Κανόνες

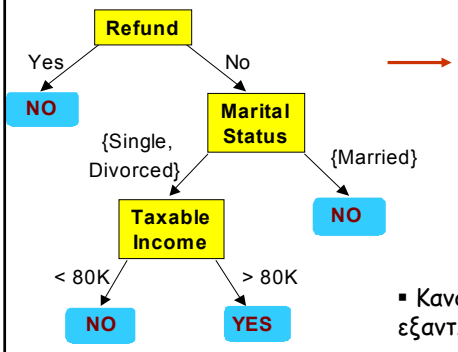
Κατασκευή Ταξινομητών με Κανόνες

- Άμεση Μέθοδος:
  - Εξαγωγή κανόνων απευθείας από τα δεδομένα
  - Π.χ.: RIPPER, CN2, Holte's 1R
- Έμμεση Μέθοδος:
  - Εξαγωγή κανόνων από άλλα μοντέλα ταξινομητών (πχ από δέντρα απόφασης)
  - Π.χ.: C4.5 κανόνες

Ταξινομητές με Κανόνες

Έμμεση Μέθοδος: Από Δέντρα Απόφασης σε Κανόνες

Ένας κανόνας για κάθε μονοπάτι από τη ρίζα σε φύλλο
Κάθε ζευγάρι γνώρισμα-τιμή στο μονοπάτι αποτελεί ένα όρο στη σύζευξη και το φύλλο αφορά την κλάση (RHS)



Κανόνες Ταξινόμησης (Classification Rules)

(Refund=Yes) ==> No

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

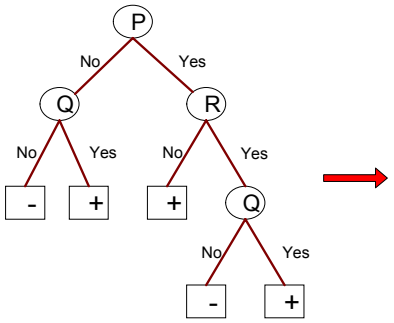
(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes

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

- Κανόνες αμοιβαία αποκλειόμενοι και εξαντλητικοί
- Το σύνολο κανόνων περιέχει όση πληροφορία περιέχει και το δέντρο

Ταξινομητές με Κανόνες

Από Δέντρα Απόφασης σε Κανόνες (Παράδειγμα)

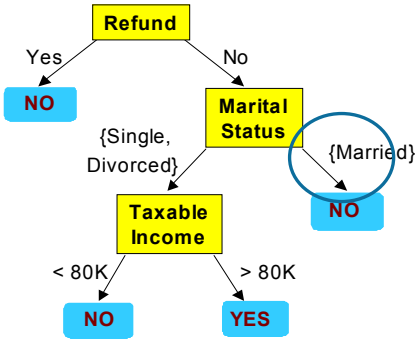


Rule Set

r1: (P=No,Q=No) ==> -
r2: (P=No,Q=Yes) ==> +
r3: (P=Yes,R=No) ==> +
r4: (P=Yes,R=Yes,Q=No) ==> -
r5: (P=Yes,R=Yes,Q=Yes) ==> +

Από Δέντρα Απόφασης σε Κανόνες

Οι κανόνες μπορεί να απλοποιηθούν (απαλοιφή κάποιων όρων στο LHS αν δεν αλλάζει πολύ το λάθος)



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

Αρχικός Κανόνας: (Refund=No) $\wedge$ (Status=Married) $\rightarrow$ No

Απλοποιημένος Κανόνας: (Status=Married) $\rightarrow$ No

Από Δέντρα Απόφασης σε Κανόνες

Αν γίνει απλοποίηση (κλάδεμα):

- Οι κανόνες δεν είναι πια αμοιβαία αποκλειόμενοι

  Μια εγγραφή μπορεί να ενεργοποιήσει παραπάνω από έναν κανόνα

  Λύση (conflict resolution)

  (1) Διάταξη του συνόλου κανόνων (αν μια εγγραφή ενεργοποιεί πολλούς κανόνες, της ανατίθεται αυτός με τη μεγαλύτερη προτεραιότητα) (decision list) ή (2) ο κανόνας με τις πιο πολλές απαιτήσεις (πχ με το μεγαλύτερο αριθμό όρων) (size ordering) ή (3) διάταξη των κλάσεων (αν μια εγγραφή ενεργοποιεί πολλούς κανόνες, της ανατίθεται η τάξη με τη μεγαλύτερη προτεραιότητα) (misclassification cost)

  Χωρίς διάταξη του συνόλου κανόνων – χρήση σχήματος ψηφοφορίας

- Οι κανόνες δεν είναι πια εξαντλητικοί

  Μια εγγραφή μπορεί να μην ενεργοποιεί κάποιον κανόνα

  Λύση

  Χρήση default κλάσης

Άλλοι Ταξινομητές

Ταξινομητές στιγμιοτύπου

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 βασισμένοι σε στιγμιότυπα)

Μην κατασκευάσεις μοντέλο αν δε χρειαστεί

Ταξινομητές βασισμένοι σε Στιγμιότυπα

Σύνολο Αποθηκευμένων Περιπτώσεων

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

- Αποθήκευσε τις εγγραφές του συνόλου εκπαίδευσης
- Χρησιμοποίησε τις αποθηκευμένες εγγραφές για την εκτίμηση της κλάσης των νέων περιπτώσεων

Unseen Case

Atr1	.....	AtrN

Ταξινομητές βασισμένοι σε Στιγμιότυπα

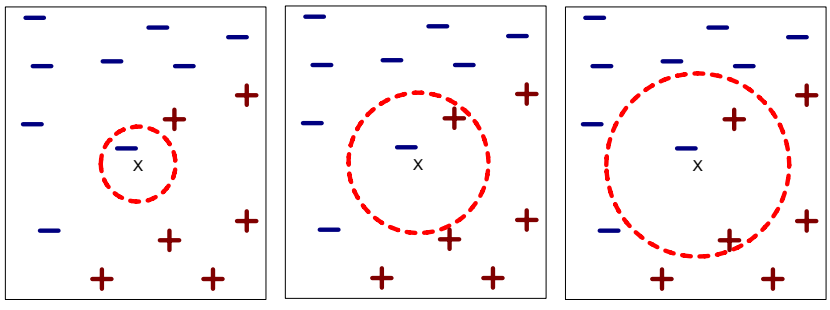
Παραδείγματα:

- Rote-learner
  - Κρατά (Memorizes) όλο το σύνολο των δεδομένων εκπαίδευσης και ταξινομεί μια εγγραφή αν ταιριάζει πλήρως με κάποιο από τα δεδομένα εκπαίδευσης

- Nearest neighbor – Κοντινότερος Γείτονας
  - Χρήση των *k* κοντινότερων "closest" σημείων (nearest neighbors) για την ταξινόμηση

Ταξινομητές Κοντινότερου Γείτονα

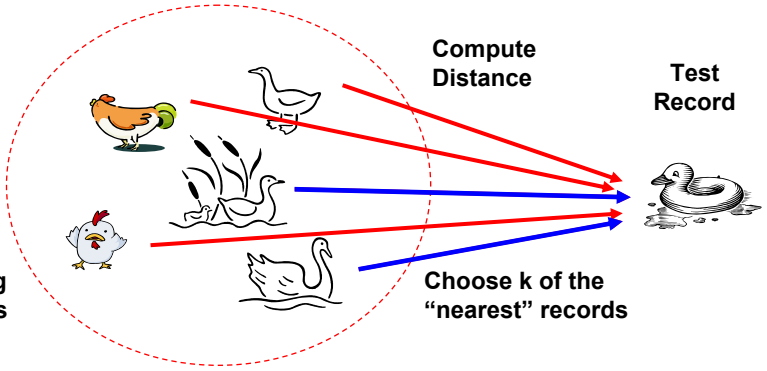
k-κοντινότεροι γείτονες μιας εγγραφής x είναι τα σημεία που έχουν την *k*-οστή μικρότερη απόσταση από το x



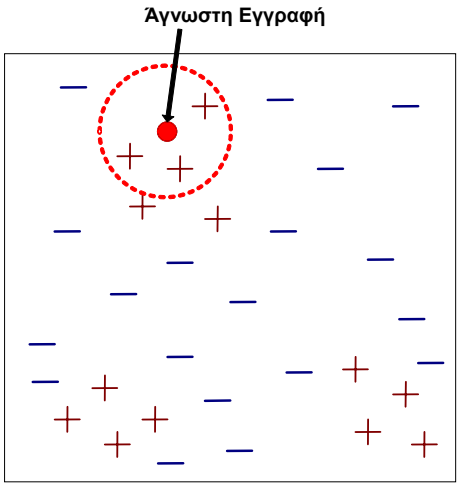
(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

Ταξινομητές Κοντινότερου Γείτονα

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



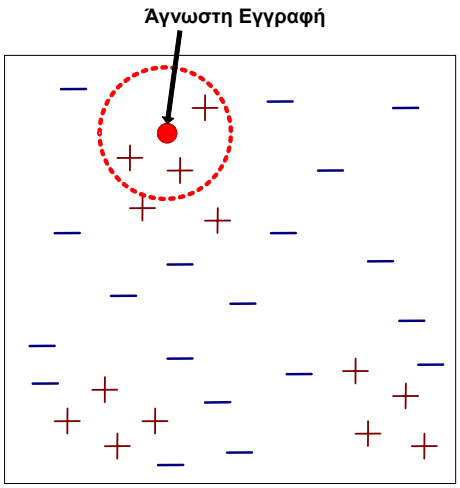
Ταξινομητές Κοντινότερου Γείτονα



Χρειάζεται

1. Το σύνολο των αποθηκευμένων εγγραφών
2. Distance Metric Μετρική απόστασης για να υπολογίσουμε την απόσταση μεταξύ εγγραφών
3. Την τιμή του *k*, δηλαδή τον αριθμό των κοντινότερων γειτόνων που πρέπει να ανακληθούν

Ταξινομητές Κοντινότερου Γείτονα



Για να ταξινομηθεί μια άγνωστη εγγραφή:

- Υπολογισμός της απόστασης από τις εγγραφές του συνόλου
- Εύρεση των *k* κοντινότερων γειτόνων
- Χρήση των κλάσεων των κοντινότερων γειτόνων για τον καθορισμό της κλάσης της άγνωστης εγγραφής - π.χ., με βάση την πλειοψηφία (majority vote)

Ταξινομητές Κοντινότερου Γείτονα

- Απόσταση μεταξύ εγγραφών:
  - Πχ ευκλείδεια απόσταση

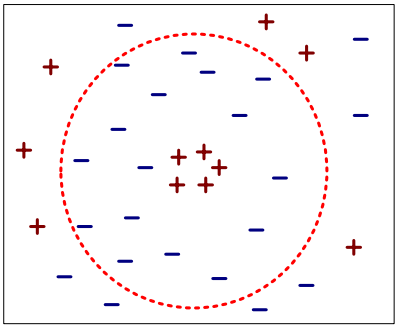
$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- Καθορισμός τάξης
  - Απλά τη πλειοψηφική κλάση
  - Βάρος σε κάθε ψήφο με βάση την απόσταση
    - weight factor, w = $1/d^2$

Ταξινομητές Κοντινότερου Γείτονα

Επιλογή της τιμής του k:

- k πολύ μικρό, ευαισθησία στα σημεία θορύβου
- k πολύ μεγάλο, η γειτονιά μπορεί να περιέχει σημεία από άλλες κλάσεις



Ταξινομητές Κοντινότερου Γείτονα

- Θέματα Κλιμάκωσης
  - Τα γνωρίσματα ίσως πρέπει να κλιμακωθούν ώστε οι αποστάσεις να μην κυριαρχηθούν από κάποιο γνώρισμα
  - Παράδειγμα:
    - 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

- Δεν κατασκευάζεται μοντέλο, μεγάλο κόστος για την ταξινόμηση
- Πολλές διαστάσεις (κατάρα των διαστάσεων)
- Θόρυβο (ελάττωση μέσω k-γειτόνων)

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
- Ταξινομητές Στιγμιότυπου (k-κοντινότεροι γείτονες)