Αναγνώριση Προτύπων

Ταξινόμηση – Βασικές έννοιες

Ανδρέας Λ. Συμεωνίδης

Αν. Καθηγητής

Τμήμα Ηλεκτρολόγων Μηχ/κών

&

Μηχ/κών Υπολογιστών, Α.Π.Θ.

Email: asymeon@eng.auth.gr

issel
Intelligent Systems & Software Engineering Labgroup

ARISTOTLE UNIVERSITY OF THESSALONIKI

Διάρθρωση διάλεξης

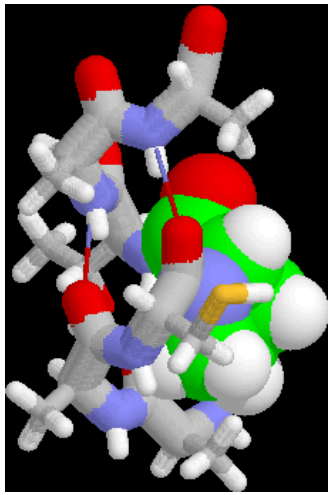
- Βασικές Έννοιες Ταξινόμησης
- Δένδρα απόφασης
- Μετρικές αξιολόγησης μοντέλων ταξινόμησης

Ταξινόμηση: Ορισμός

- Δεδομένης μιας συλλογής εγγραφών (*σετ εκπαίδευσης*)
  - Κάθε εγγραφή περιέχει ένα σετ από *χαρακτηριστικά*, όπου το ένα είναι η *κλάση*.
- Εύρεση ενός ***μοντέλου*** για το χαρακτηριστικό κλάση ως μια συνάρτηση των τιμών των άλλων χαρακτηριστικών.
- Στόχος: **πρότερα μη γνωστές** εγγραφές να μπορούν να ταξινομηθούν σε μια κλάση με όσο μεγαλύτερη ακρίβεια.
  - Ένα *σετ ελέγχου* χρησιμοποιείται η ακρίβεια του μοντέλου. Συνήθως, το αρχικό σετ χωρίζεται σε εκπαίδευσης και ελέγχου, με το σετ εκπαίδευσης να χρησιμοποιείται για τη δημιουργία του μοντέλου και με το σετ ελέγχου να χρησιμοποιείται για την επικύρωσή του.

Παραδείγματα ταξινόμησης

- Πρόγνωση ιστών ως καλοήθεις ή μη
- Ταξινόμηση συναλλαγών πιστωτικών καρτών ως απάτη ή κανονική δραστηριότητα
- Ταξινόμηση δευτεροταγών δομών πρωτεϊνών ως alpha-helix, beta-sheet, ή random coil
- Κατηγοριοποίηση της ειδησιογραφίας ως άρθρα οικονομικού, κοινωνικού, πολιτικού, αθλητικού κα περιεχομένου



Τεχνικές Ταξινόμησης

- Δένδρα απόφασης
- Πιθανοτικοί ταξινομητές
- Πιθανοτικά δίκτυα ταξινόμησης
- Νευρωνικά δίκτυα
- Support Vector Machines
- Κανόνες απόφασης

Παράδειγμα δένδρου απόφασης

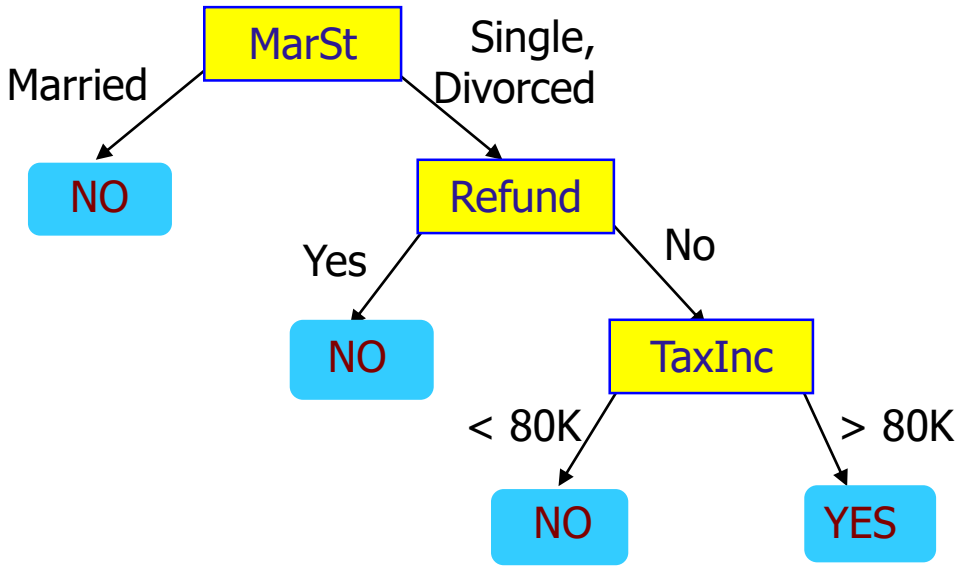
Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Σετ εκπαίδευσης

Μοντέλο: Δένδρο απόφασης

Παράδειγμα δένδρου απόφασης

	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**

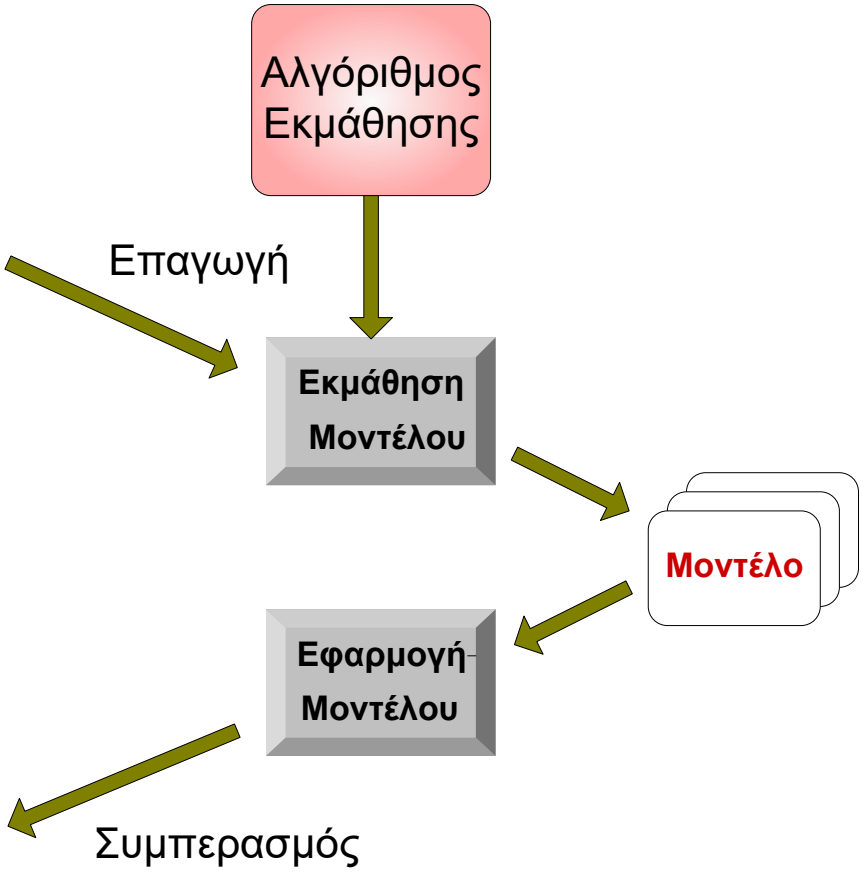


Μπορεί να υπάρχουν περισσότερα του ενός μοντέλα που ταιριάζουν στα δεδομένα

Επίδειξη Ταξινόμησης

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

Σετ εκπαίδευσης

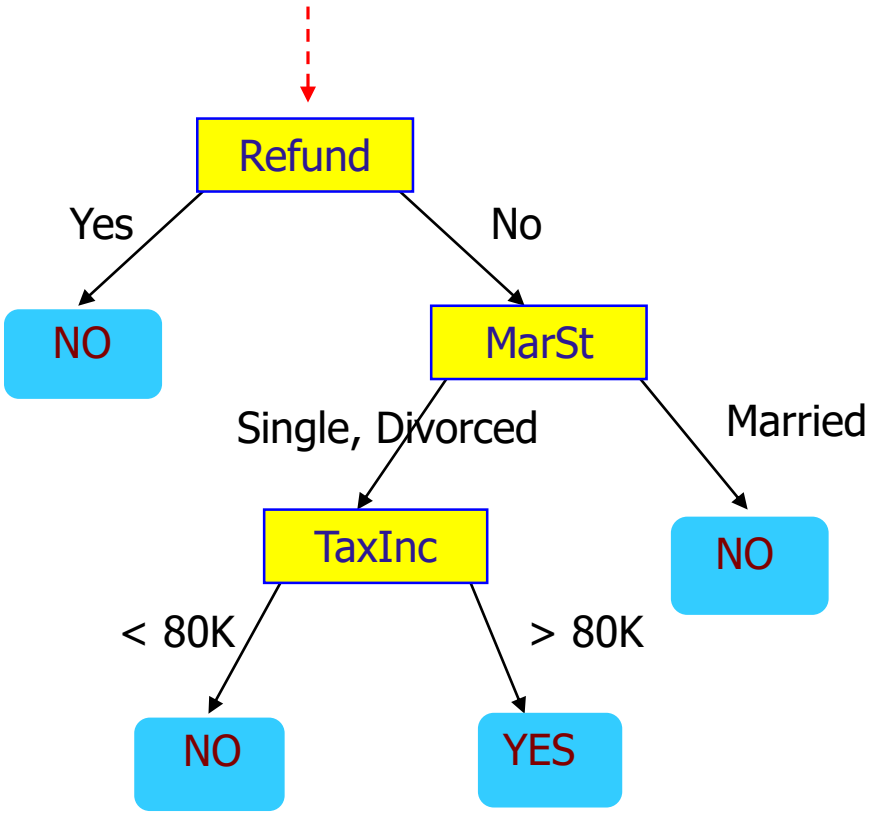


Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Σετ ελέγχου

Εφαρμογή του μοντέλου στο σετ Ελέγχου

Ξεκινήστε από τη ρίζα του δένδρου



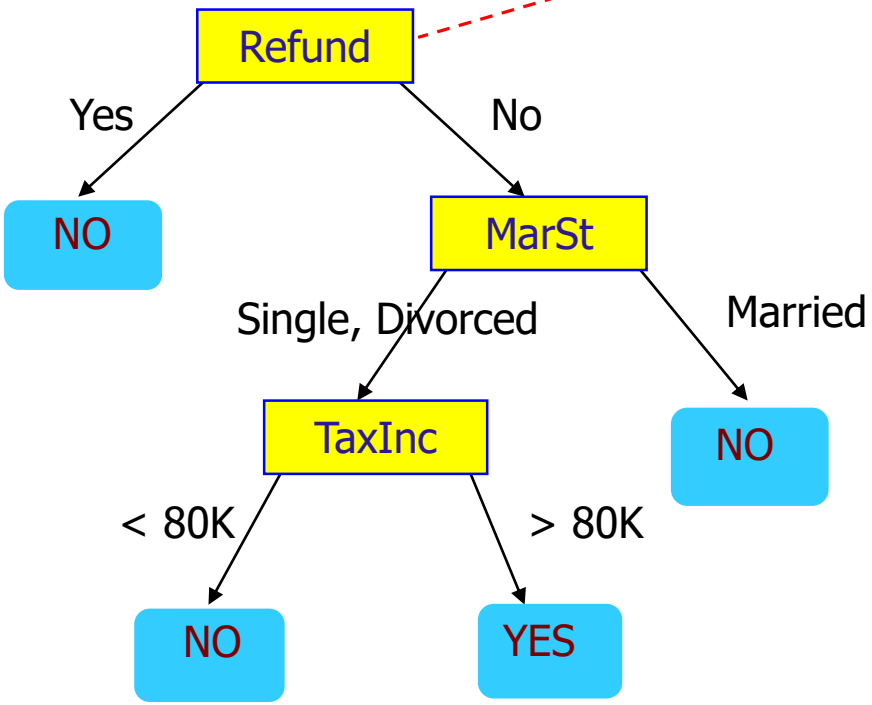
Δεδομένα ελέγχου

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Εφαρμογή του μοντέλου στο σετ Ελέγχου

Δεδομένα ελέγχου

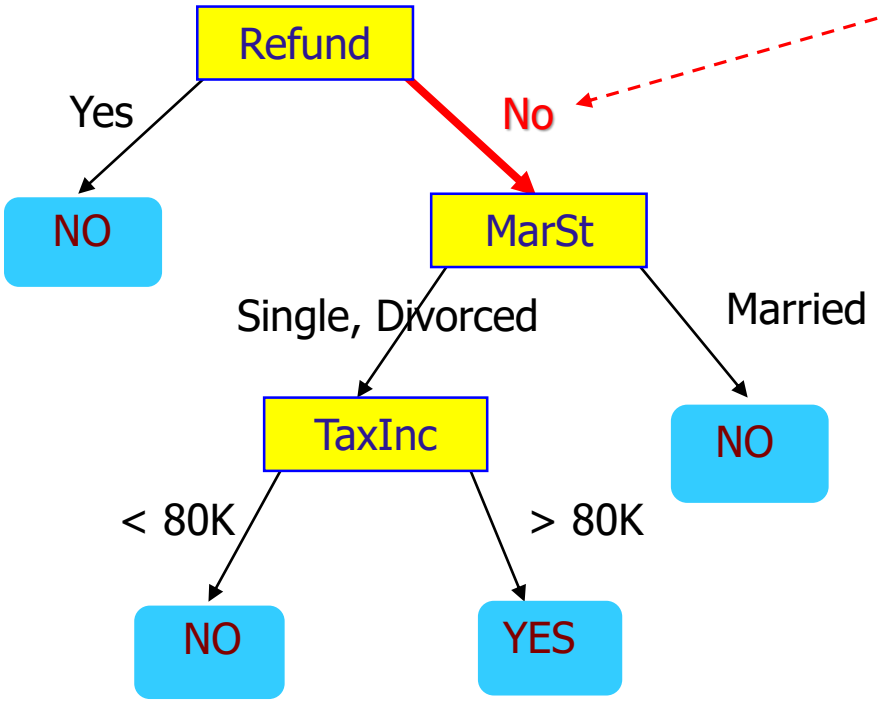
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Εφαρμογή του μοντέλου στο σετ Ελέγχου

Δεδομένα ελέγχου

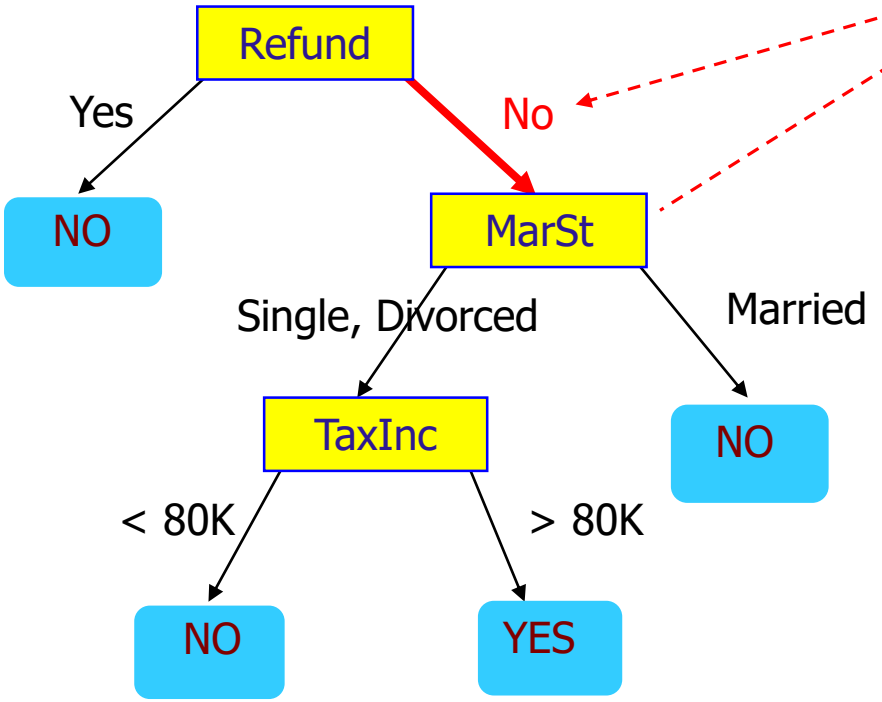
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Εφαρμογή του μοντέλου στο σετ Ελέγχου

Δεδομένα ελέγχου

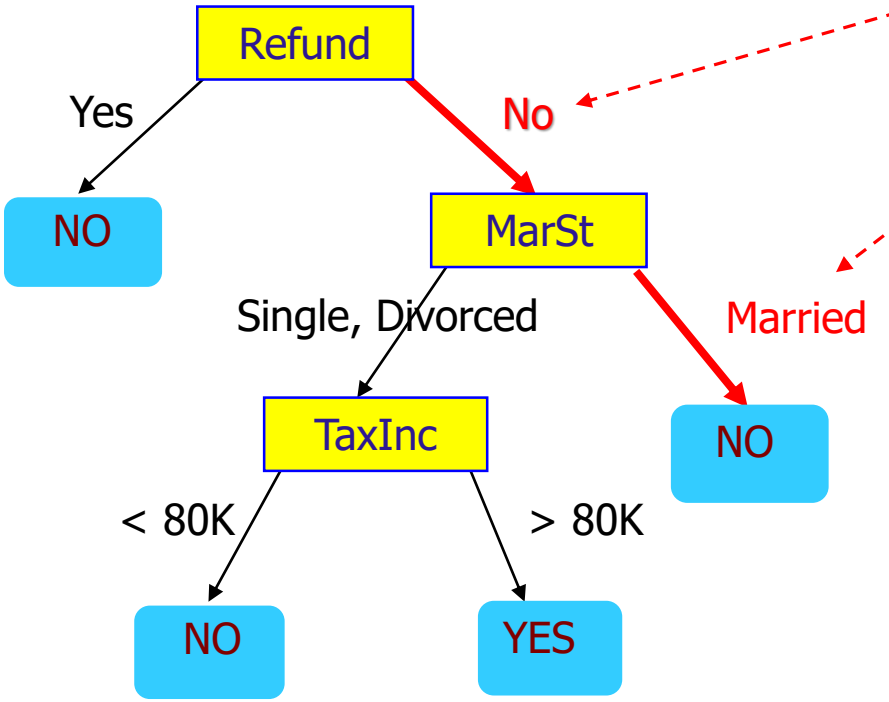
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Εφαρμογή του μοντέλου στο σετ Ελέγχου

Δεδομένα ελέγχου

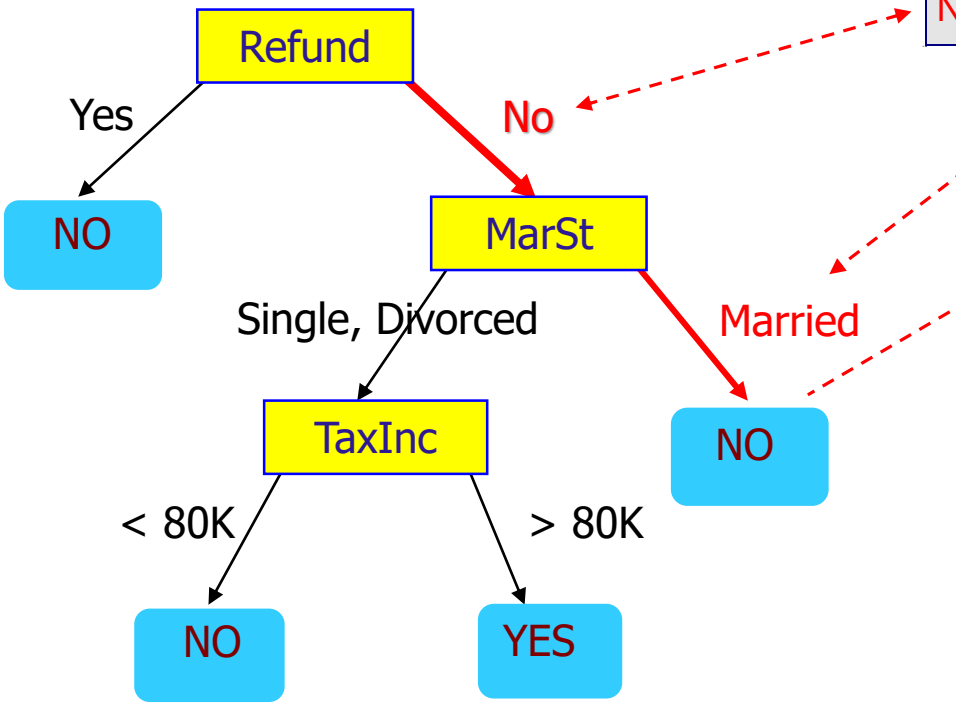
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Εφαρμογή του μοντέλου στο σετ Ελέγχου

Δεδομένα ελέγχου

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δώσε τιμή 'NO' στο Cheat

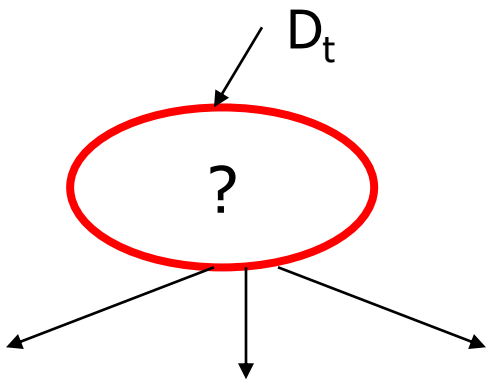
Η επαγωγή του δένδρου απόφασης

- Πολλοί αλγόριθμοι:
  - Ο αλγόριθμος του Hunt (από τους πρώτους)
  - CART
  - ID3, C4.5, C5.0
  - SLIQ,SPRINT

Η γενική δομή του αλγορίθμου του Hunt

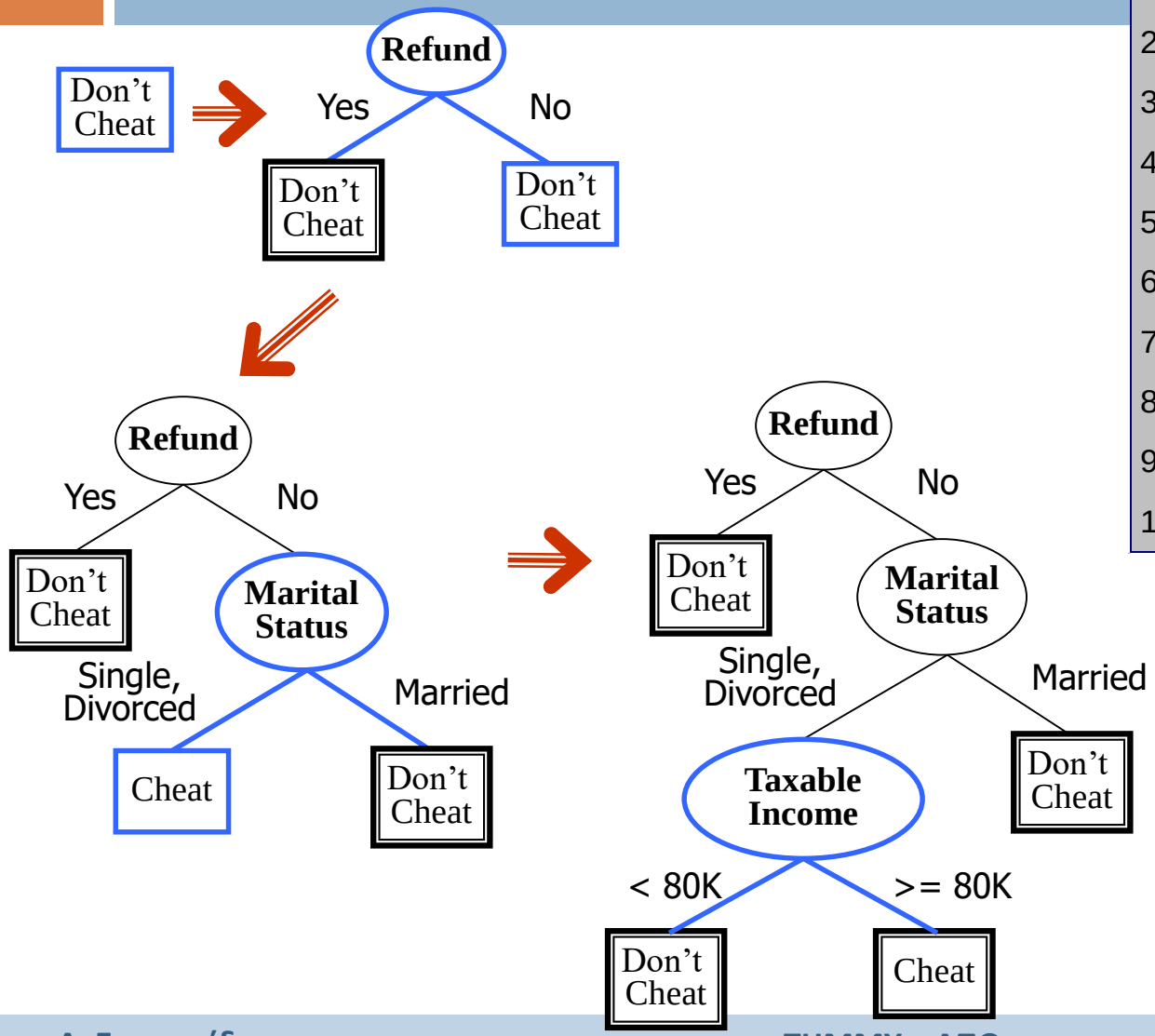
- Έστω D_t το σετ των εγγραφών εκπαίδευσης που φτάνουν στον κόμβο t
- Γενική διαδικασία:
  - Εάν το D_t περιέχει εγγραφές της ίδιας κλάσης y_t, τότε το t είναι φύλλο του δένδρου με ταμπέλα y_t
  - Εάν το D_t είναι κενό σετ, τότε το t είναι φύλλο του δένδρου με ταμπέλα την default κλάση, y_d
  - Ένα το D_t περιέχει εγγραφές που ανήκουν σε περισσότερες από μια κλάσεις, χρησιμοποιήστε μια συνθήκη ελέγχου των χαρακτηριστικών με βάση την οποία να χωρίσετε το σετ σε υποσέτ. Αναδρομικά εφαρμόστε τη διαδικασία σε κάθε υποσέτ.

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



Ο αλγόριθμος του Hunt

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



Επαγωγή δένδρου

- Άπληστη στρατηγική
  - Διαχωρισμός σε υποσέτ με βάση ένα κριτήριο ελέγχου στις μεταβλητές που βελτιστοποιεί ένα συγκεκριμένο κριτήριο.
- Θέματα προς επίλυση
  - Καθορισμός του κριτηρίου διαχωρισμού
    - Πώς να ορίσεις το καλύτερο κριτήριο ελέγχου
    - Πώς να ορίσεις τον καλύτερο διαχωρισμό
  - Καθορισμός της συνθήκης τερματισμού διαχωρισμού

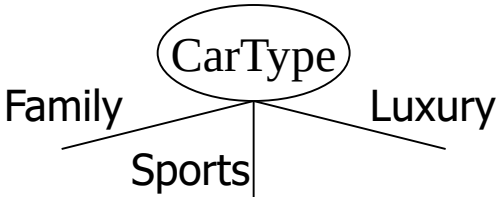
Βέλτιστο κριτήριο ελέγχου

- Εξαρτάται από τους τύπους των μεταβλητών
  - Κατηγορικά (Nominal)
  - Κατάταξης (Ordinal)
  - Συνεχή (Continuous)

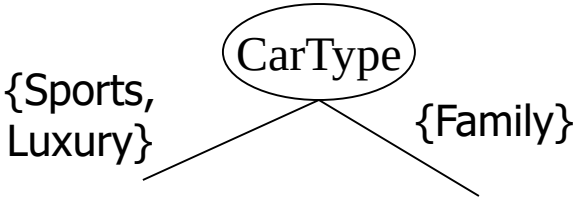
- Εξαρτάται από τον τρόπο διαχωρισμού
  - 2-way split
  - Multi-way split

Διαχωρισμός σε κατηγορικά δεδομένα

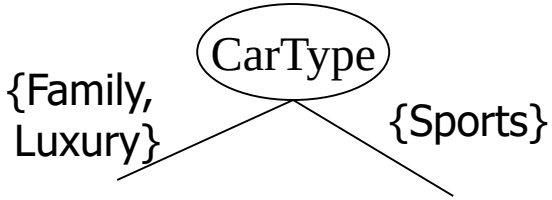
- Multi-way split: Τόσοι διαχωρισμοί, όσες και οι διακριτές τιμές της μεταβλητής



- Δυαδικό split: Διαχωρίζει σε δυο υποσέτ.
  Πρέπει να βρεθεί ο βέλτιστος διαχωρισμός

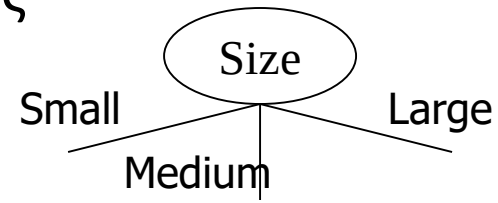


Ή

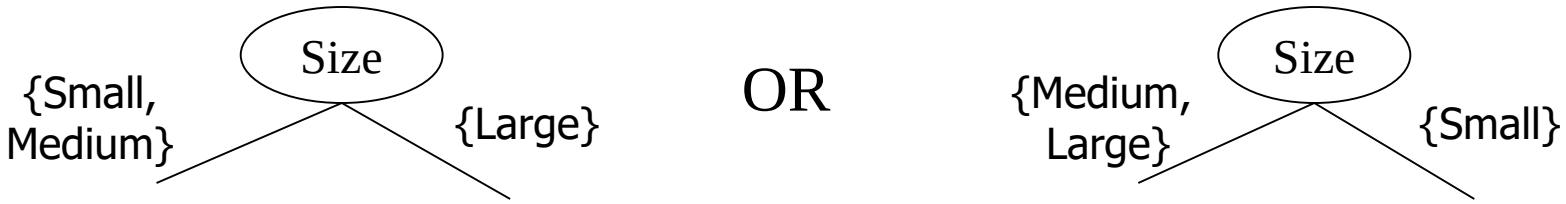


Διαχωρισμός σε δεδομένα κατάταξης

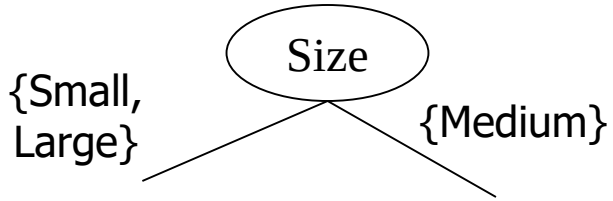
- Multi-way split: Τόσοι διαχωρισμοί, όσες και οι διακριτές τιμές της μεταβλητής



- Δυαδικό split: Διαχωρίζει σε δυο υποσέτ. Πρέπει να βρεθεί ο βέλτιστος διαχωρισμός



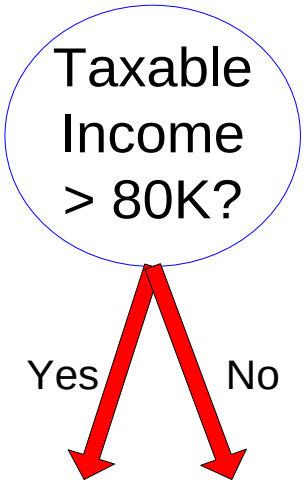
- Αυτή η περίπτωση επιτρέπεται;



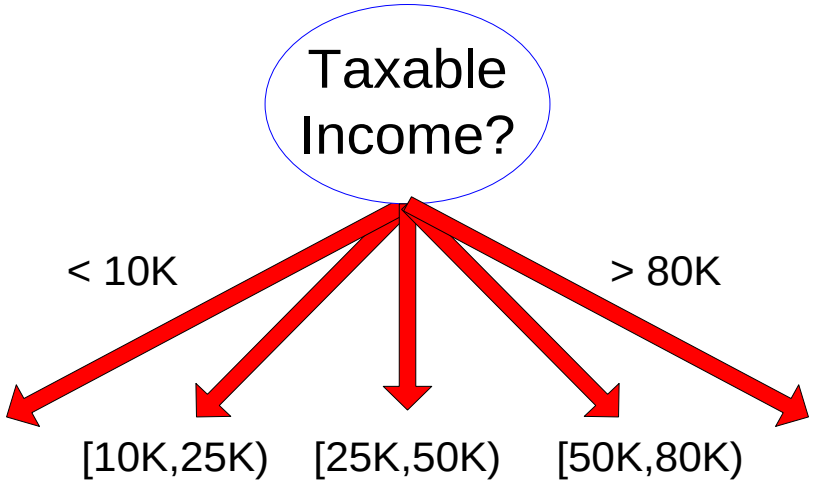
Διαχωρισμός σε συνεχή δεδομένα

- Τρόποι διαχείρισης συνεχών μεταβλητών
  - Διακριτοποίηση ώστε να μετατραπούν σε μεταβλητές κατάταξης
    - Στατικά – διακριτοποίηση με βάση συγκεκριμένες τιμές που ορίζει ο χρήστης
    - Δυναμικά – το εύρος τιμών μπορεί να βρεθεί με διακριτοποίηση σε ίσα διαστήματα, ίσες συχνότητες διαστημάτων (percentiles), ή ομαδοποίηση.
  - Δυαδική απόφαση: $(A < v)$ ή $(A \geq v)$
    - Θεωρήστε όλους τους πιθανούς διαχωρισμούς και βρείτε το βέλτιστο σημείο τομής
    - Υπολογιστικά πολυέξοδο

Διαχωρισμός σε συνεχή δεδομένα



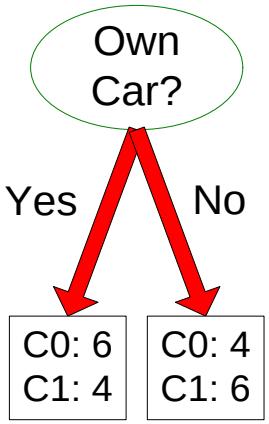
(i) Binary split

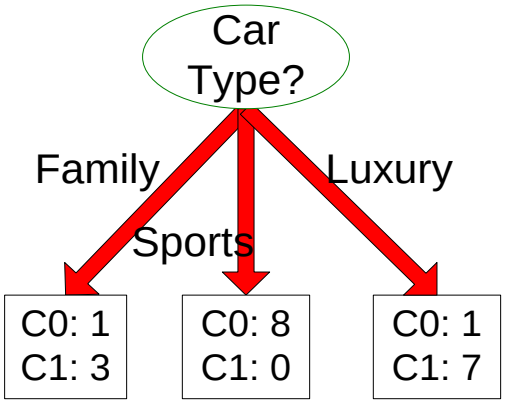


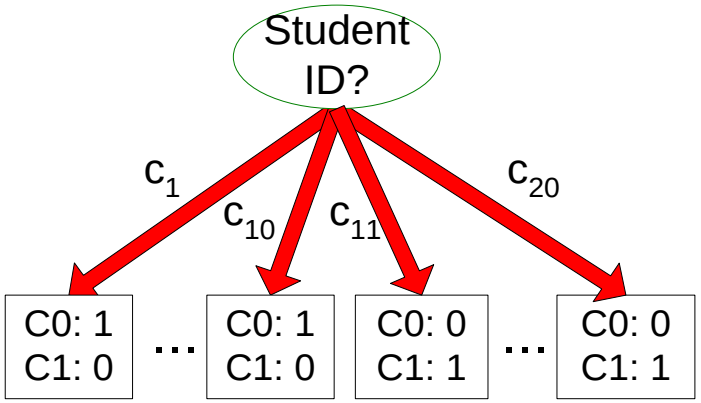
(ii) Multi-way split

Βέλτιστος διαχωρισμός

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







Ποιος είναι ο βέλτιστος διαχωρισμός;

Βέλτιστος διαχωρισμός

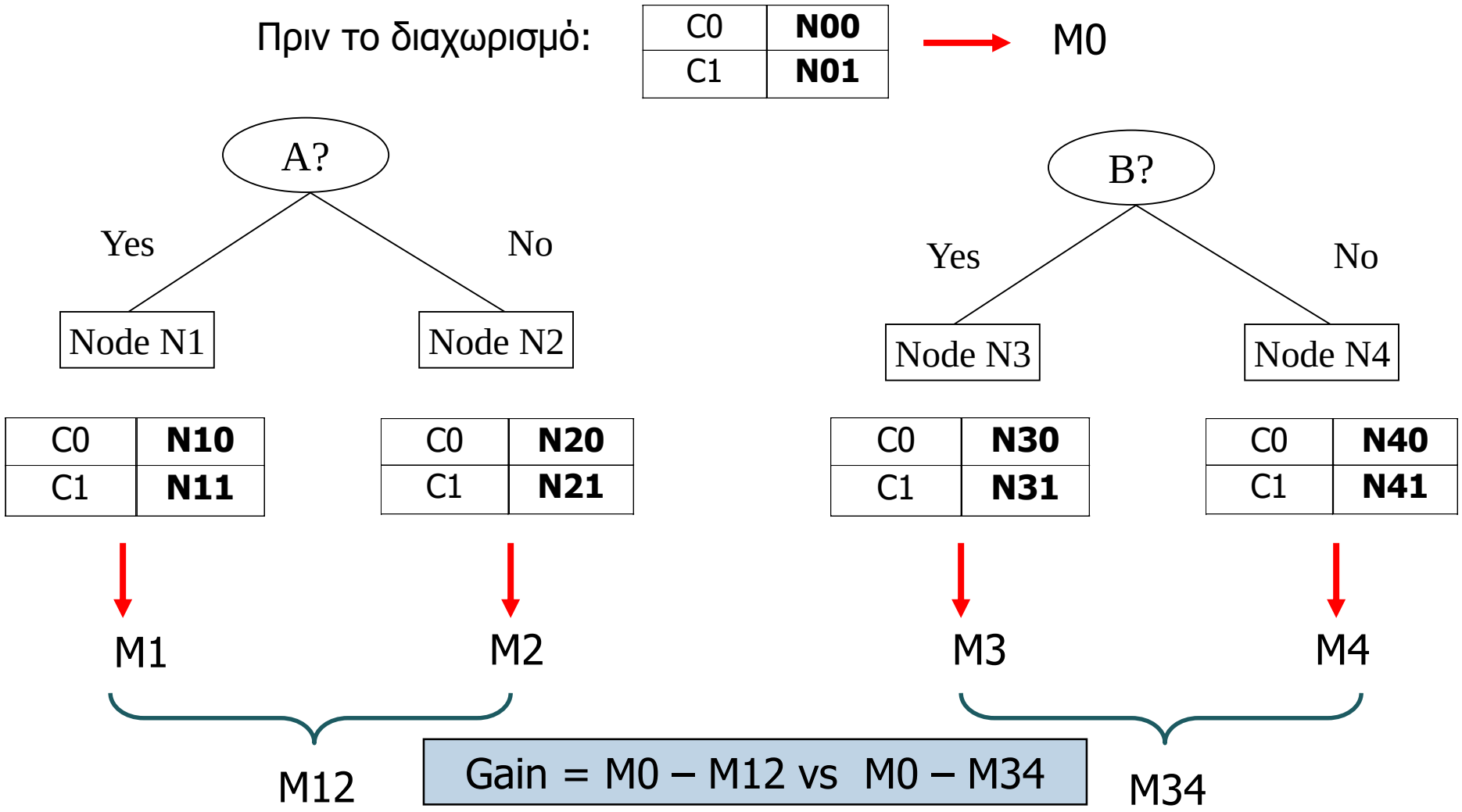
- Άπληστη προσέγγιση:
  - Προτιμώνται κόμβοι με ομογενή κατανομή
- Χρειάζεται μια μετρική «καθαρότητας»:

C0: 5 C1: 5	C0: 9 C1: 1
Μη ομογενής κατανομή, Χαμηλός βαθμός καθαρότητας	Ομογενής κατανομή, Υψηλός βαθμός καθαρότητας

Μετρικές καθαρότητας

- Ο δείκτης Gini (Gini Index)
- Εντροπία
- Το σφάλμα εσφαλμένης ταξινόμησης (Misclassification error)

Βέλτιστος διαχωρισμός



Μετρική καθαρότητας: GINI

- Το Gini Index για δεδομένο κόμβο t:

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

(Σημείωση: $p(j \mid t)$ είναι η σχετική συχνότητα της κλάσης j στον κόμβο t)

- Μέγιστο (1 - 1/n_c), όταν οι εγγραφές έχουν ίδια κατανομή από όλες τις κλάσεις. Συνεπάγεται την ύπαρξη ελάχιστης χρήσιμης πληροφορίας
- Ελάχιστο (0.0), όταν οι εγγραφές ανήκουν σε μια κλάση. Συνεπάγεται την ύπαρξη απόλυτα χρήσιμης πληροφορίας.

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	

GINI: παραδείγματα υπολογισμού

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

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

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

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

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

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

Gini = 1 – (1/6)^2 – (5/6)^2 = 0.278

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

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

Gini = 1 – (2/6)^2 – (4/6)^2 = 0.444

Διαχωρισμός βάσει GINI

- Χρησιμοποιείται στους αλγορίθμους CART, SLIQ, SPRINT.
- Όταν ο κόμβος p διαχωρίζεται σε k τμήματα (παιδιά), η ποιότητα του διαχωρισμού υπολογίζεται ως:

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

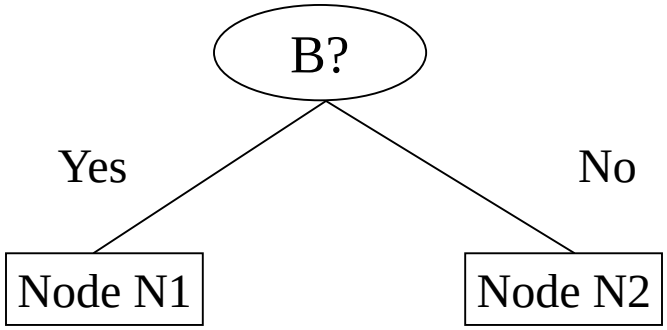
όπου, n_i = ο αριθμός των εγγραφών στον κομβο i,

n = ο αριθμός των εγγραφών στον κομβο p.

2-way διαχωρισμός: Υπολογισμός GINI Index

- Υπολογίζεται τη καθαρότητα των τμημάτων με τη χρήση βαρών
- Τα μεγαλύτερα και πιο καθαρά τμήματα προτιμώνται

	Parent
C1	**6**
C2	**6**
Gini = 0.500	



Gini(N1)
$= 1 - (5/7)^2 - (2/7)^2$
$= 0.408$

Gini(N2)
$= 1 - (1/5)^2 - (4/5)^2$
$= 0.32$

	N1	N2
C1	**5**	**1**
C2	**2**	**4**
Gini=0.375		

Gini(Children)
$= 7/12 * 0.408 +$
$5/12 * 0.32$
$= 0.375$

Multi-way διαχωρισμός: Υπολογισμός GINI Index

- Για κάθε διακριτή τιμή, δημιουργία ενός πίνακα απαρίθμησης των εγγραφών για κάθε κλάση στο σετ
- Χρήση του πίνακα απαρίθμησης για τη λήψη απόφασης

Multi-way split

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

Two-way split

(εύρεση του καλύτερου διαχωρισμού)

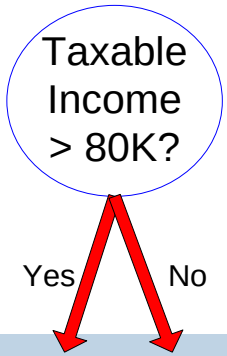
	CarType	
	{Sports, Luxury}	{Family}
C1	3	1
C2	2	4
Gini	0.400	

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

Διαχωρισμός σε συνεχή μεταβλητή: Υπολογισμός GINI Index

- Χρήση δυαδικής απόφασης σε μια τιμή διαχωρισμού
- Πολλές επιλογές για την τιμή διαχωρισμού
  - Αριθμός των πιθανών τιμών διαχωρισμού = Αριθμός των διακριτών τιμών
- Κάθε τιμή διαχωρισμού έχει έναν πίνακα απαρίθμηση συνδεδεμένο με αυτή
  - Απαρίθμηση των στιγμιοτύπων που ανήκουν σε κάθε κλάση για τα υποσέτ διαχωρισμού, $A < v$ και $A \geq v$
- Απλή μέθοδος επιλογής της βέλτιστης τιμής v
  - Για κάθε v, σάρωσε το σετ για να φτιάξεις τον πίνακα απαρίθμησης και να υπολογίσεις το Gini index
  - Υπολογιστικά πολυέξοδο! Επαναλαμβανόμενη δουλειά.

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



Διαχωρισμός σε συνεχή μεταβλητή: Υπολογισμός GINI Index

- Υπολογιστικά ανεκτή υλοποίηση: για κάθε συνεχή μεταβλητή,
  - Ταξινόμησε τις εγγραφές με βάση την αύξουσα τιμή της μεταβλητής
  - Όρισε τιμές διαχωρισμού
  - Σάρωσε γραμμικά τις τιμές, υπολογίζοντας κάθε φορά τον πίνακα απαρίθμησης και το gini index
  - Επέλεξε το σημείο διαχωρισμού με το μικρότερο gini index

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	

Μετρική καθαρότητας: Εντροπία (Κέρδος πληροφορίας)

- Η εντροπία σε έναν κόμβο t:

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

(Σημείωση: $p(j \mid t)$ είναι η σχετική συχνότητα της κλάσης j στον κόμβο t)

- Μετρά την ομογένεια ενός κόμβου
  - Μέγιστο ($\log n_c$), όταν οι εγγραφές στον κόμβο έχουν ίδια κατανομή από όλες τις κλάσεις. Συνεπάγεται την ύπαρξη ελάχιστης χρήσιμης πληροφορίας
  - Ελάχιστο (0.0), όταν οι εγγραφές ανήκουν σε μια κλάση. Συνεπάγεται την ύπαρξη απόλυτα χρήσιμης πληροφορίας.
- Οι υπολογισμοί της εντροπίας για τους διάφορους τύπους μεταβλητών είναι ίδιοι με αυτούς του GINI index

Εντροπία: παραδείγματα υπολογισμού

$$Entropy(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

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

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

Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

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

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

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

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

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

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

Διαχωρισμός βάσει κέρδους πληροφορίας

- Κέρδος Πληροφορίας

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

- Ο κόμβος p διαχωρίζεται σε k τμήματα
- n_i είναι ο αριθμός των εγγραφών στο τμήμα i

- Μετρά τη μείωση της εντροπίας λόγω ενός διαχωρισμού.
- Επιλογή του διαχωρισμού που επιτυγχάνει τη μέγιστη μείωση (μεγιστοποιεί το GAIN)
- Χρησιμοποιείται στους ID3 και C4.5
- Μειονέκτημα: Προτιμά διαχωρισμούς που οδηγούν σε πολλά, μικρά, αλλά καθαρά τμήματα.

Διαχωρισμός βάσει κέρδους πληροφορίας

- GainRatio:

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

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

- Ο κόμβος p διαχωρίζεται σε k τμήματα
- n_i είναι ο αριθμός των εγγραφών στο τμήμα i

  - Κανονικοποιεί το κέρδος πληροφορίας ως προς την εντροπία του διαχωρισμού (SplitINFO).
  - Η υψηλότερη εντροπία διαχωρισμού (μεγάλος αριθμός μικρών τμημάτων) τιμωρείται!
  - Χρησιμοποιείται στον C4.5
  - Εξαλείφει το μειονέκτημα του κέρδους πληροφορίας

Μετρική καθαρότητας: Σφάλμα λανθασμένης ταξινόμησης

- Σφάλμα λανθασμένης ταξινόμησης (Classification error) στον κόμβο t:

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

- Υπολογίζει το σφάλμα λανθασμένης ταξινόμησης ενός κόμβου:
  - Μέγιστο (1 - 1/n_c), όταν οι εγγραφές στον κόμβο έχουν ίδια κατανομή από όλες τις κλάσεις. Συνεπάγεται την ύπαρξη ελάχιστης χρήσιμης πληροφορίας
  - Ελάχιστο (0.0), όταν οι εγγραφές ανήκουν σε μια κλάση. Συνεπάγεται την ύπαρξη απόλυτα χρήσιμης πληροφορίας.

Misclassification Error: παραδείγματα υπολογισμού

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

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

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

Error = 1 – max (0, 1) = 1 – 1 = 0

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

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

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

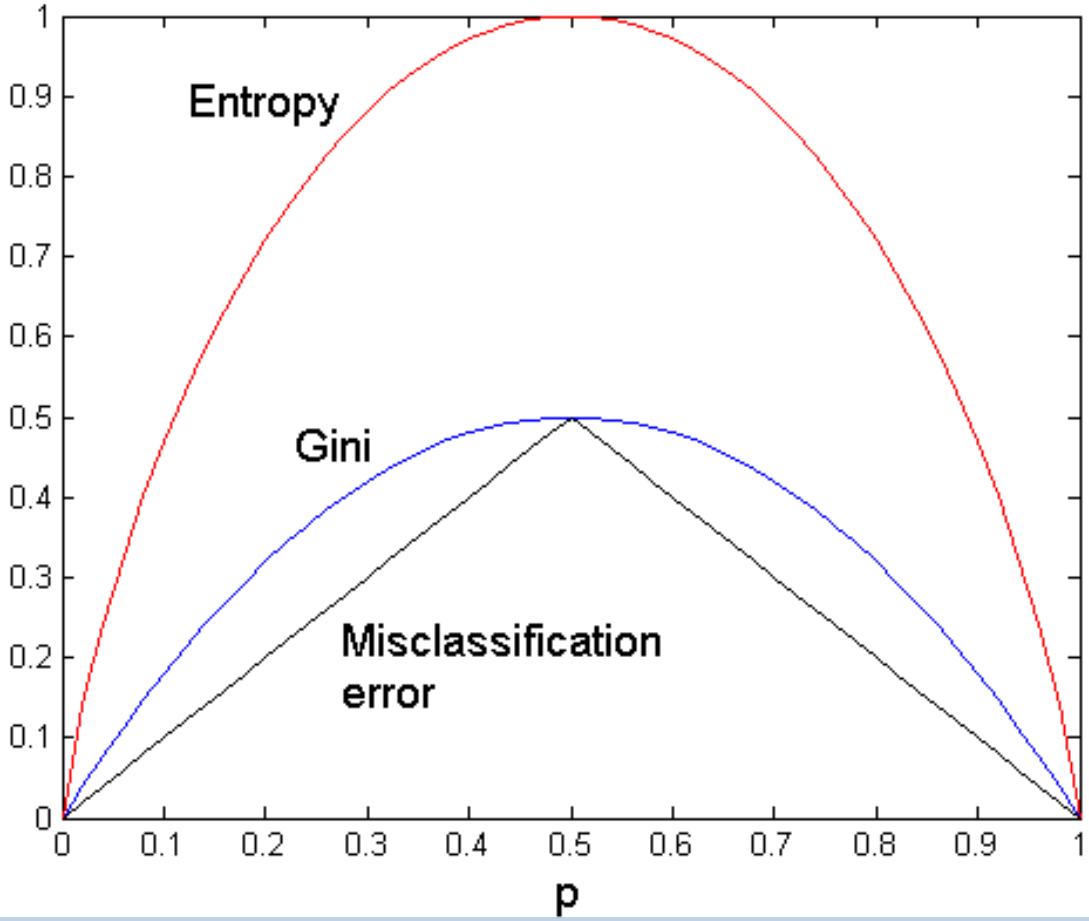
C1	**2**
C2	**4**

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

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

Σύγκριση των κριτηρίων καθαρότητας

Για ένα πρόβλημα 2 κλάσεων:



Παύση επέκτασης του δένδρου

- Παύση διαχωρισμού ενός κόμβου όταν όλες οι εγγραφές ανήκουν στην ίδια κλάση
- Παύση διαχωρισμού ενός κόμβου όταν όλες οι εγγραφές έχουν παρόμοιες τιμές
- Πρόωρη παύση

Πλεονεκτήματα Δένδρων απόφασης

- Υπολογιστικά εύκολα στην κατασκευή
- Πολύ γρήγορα στην ταξινόμηση νέων εγγραφών
- Κατανοητά για δένδρα μικρού μεγέθους
- Ακρίβεια συγκρίσιμη με άλλες τεχνικές σε πολλά σετ δεδομένων

Αξιολόγηση Μοντέλων

- Μετρικές για την αξιολόγηση του μοντέλου
- Καθορισμός αξιόπιστων εκτιμήσεων
- Σύγκριση μοντέλων (σχετική επίδοση)

Μετρικές για την αξιολόγηση μοντέλων

- Στόχος η ικανότητα πρόβλεψης του μοντέλου (όχι η ταχύτητα κατασκευής μοντέλου ή ταξινόμησης, η κλιμάκωση κτλ)
- Confusion Matrix (Πίνακας Σύγχυσης):

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

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Μετρικές για την αξιολόγηση μοντέλων

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

- Η πιο συνήθης μετρική – **Ακρίβεια**:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Περιορισμοί του Accuracy

- Θεωρήστε ένα πρόβλημα 2 κλάσεων:
  - Στιγμιότυπα που ανήκουν στην Class 0 = 9990
  - Στιγμιότυπα που ανήκουν στην Class 1 = 10
- Εάν το μοντέλο προβλέψει ότι όλες οι εγγραφές πρέπει να ανήκουν στην κλάση 0, τότε:

$$\text{Accuracy} = 9990/10000 = 99.9\ \%$$

- Πολύ καλή τιμή για Accuracy, πολύ κακό μοντέλο!!!

Πίνακας κόστους (Cost Matrix)

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

C(i|j): Το κόστος λάθος ταξινόμησης μιας εγγραφής της κλάσης j ως κλάση i

Υπολογισμός του κόστους Ταξινόμησης

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

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

Cost vs 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

Το Accuracy είναι ανάλογο του κόστους εάν:

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

$$N = a + b + c + d$$
$$\text{Accuracy} = (a + d)/N$$

$$\begin{aligned}\text{Cost} &= p\,(a + d) + q\,(b + c)\\ &= p\,(a + d) + q\,(N - a - d)\\ &= q\,N - (q - p)(a + d)\\ &= N\,[q - (q-p) \times \text{Accuracy}]\end{aligned}$$

Μετρικές ευαίσθητες στο κόστος

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F - measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

- Το Precision είναι πολωμένο προς τα: C(Yes|Yes) & C(Yes|No)
- Το Recall είναι πολωμένο προς τα: C(Yes|Yes) & C(No|Yes)
- Το F-measure είναι πολωμένο προς όλα εκτός από τα: C(No|No)

$$\text{Weighted Accuracy} = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

Ανακεφαλαίωση

- Βασικές έννοιες ταξινόμησης
- Δένδρα απόφασης
  - Κριτήρια διαχωρισμού
  - Κριτήρια τερματισμού
  - Πλεονεκτήματα/Μειονεκτήματα
- Πηγές:
  - Introduction to Data Mining, Tan, Steinbach, Kumar.
  - Pattern recognition, Theodoridis & Koutroumbas.