ΕΜΠ ΔΠΜΣ

Εφαρμοσμένες Μαθηματικές Επιστήμες

Αλγόριθμοι Εξόρυξης Πληροφορίας

Διάλεξη 04:

Απεικόνιση Γνώσης, Αξιοπιστία & Αποτίμηση

Η μορφή των εξαγομένων και η σημασία της

Δεδομένα

input

Αλγόριθμοι Εξόρυξης Πληροφορίας

Εξαγόμενα

output

- Οι τεχνικές Μηχανικής Μάθησης παρέχουν διάφορες δομικές περιγραφές των εξαγόμενων
- Καθεμία από αυτές υπαγορεύει το είδος του αλγορίθμου που πρέπει να χρησιμοποιηθεί για το σχηματισμό της από τα δεδομένα
- Η κατανόηση του τρόπου περιγραφής συμβάλλει κατά πολύ στην κατανόηση του τρόπου παραγωγής της

Εξαγόμενα: Απεικόνιση γνώσης

- Πίνακες απόφασης (decision tables)
- Δένδρα απόφασης (decision trees)
- Κανόνες απόφασης (decision rules)
- Κανόνες συσχέτισης (association rules)
- Κανόνες με εξαιρέσεις (rules with exceptions)
- Κανόνες με συσχετίσεις (rules involving relations)
- Γραμμική παλινδρόμηση (linear regression)
- Δένδρα για αριθμητική πρόβλεψη (trees for numeric prediction)
- Απεικόνιση με βάση υποδείγματα (instance-based representation)
- Ομάδες (clusters)

Πίνακες απόφασης (decision tables)

- Ο απλούστερος τρόπος απεικόνισης των εξαγομένων:
  - Χρήση όμοιας περιγραφής με τα δεδομένα!
- Παράδειγμα: πίνακας απόφασης για το πρόβλημα καιρού:

Outlook	Humidity	**Play**
Sunny	High	**No**
Sunny	Normal	**Yes**
Overcast	High	**Yes**
Overcast	Normal	**Yes**
Rainy	High	**No**
Rainy	Normal	**No**

- Κύριο πρόβλημα: επιλογή των κατάλληλων χαρακτηριστικών

Δένδρα απόφασης (decision trees)

- Παράγονται με τη μέθοδο ‘διαίρει και βασίλευε’ (‘divide & conquer’)
- Οι κόμβοι υλοποιούν έλεγχο τιμής ενός χαρακτηριστικού
- Συνήθως, η τιμή του χαρακτηριστικού συγκρίνεται με μία σταθερά
  - Άλλες δυνατότητες: Σύγκριση τιμών δύο χαρακτηριστικών ή χρήση συνάρτησης για ένα ή περισσότερα χαρακτηριστικά
- Τα φύλλα εκχωρούν ταξινόμηση, σύνολο ταξινομήσεων ή κατανομές πιθανότητας στα υποδείγματα
- Ένα άγνωστης τάξεως υπόδειγμα ακολουθεί πορεία από την αρχή ως κάποιο φύλλο του δένδρου για την ταξινόμησή του

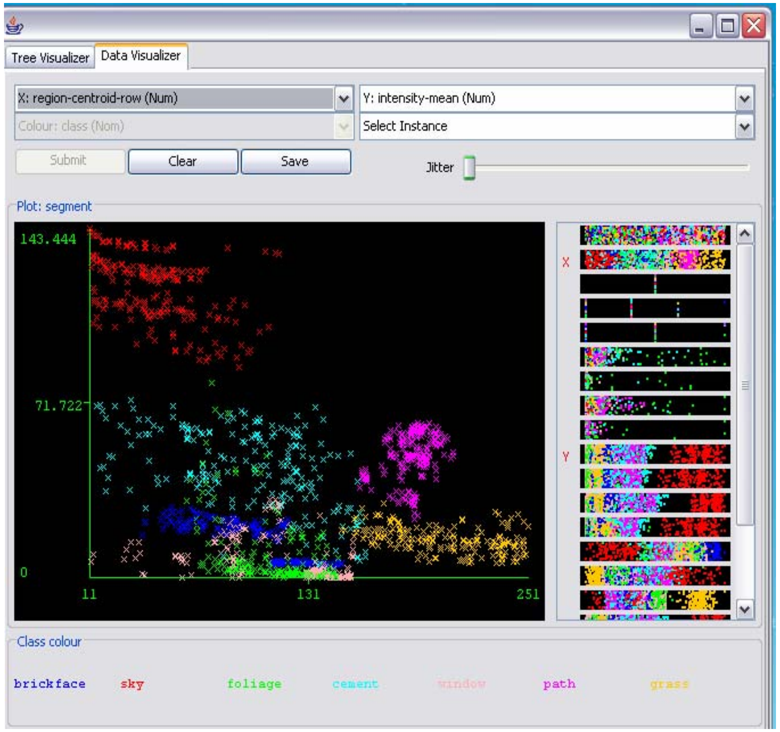
Ονομαστικά & αριθμητικά χαρακτηριστικά

- Ονομαστικά χαρακτηριστικά:
  αριθμός κλάδων μετά από κόμβο συνήθως ίσος προς τον αριθμό των διακριτών τιμών
  ⇒ το χαρακτηριστικό δεν ελέγχεται περισσότερες από μία φορές
  - εναλλακτικά: διαίρεση σε υποσύνολα, συνήθως δύο
- Αριθμητικά χαρακτηριστικά:
  σύγκριση τιμής χαρακτηριστικού με σταθερά (για ακέραιο) ή σύνολο τιμών (για πραγματικό αριθμό)
  ⇒ το χαρακτηριστικό μπορεί να ελεγχθεί περισσότερες από μία φορές
  - εναλλακτικά : διαίρεση σε τρία (ή και περισσότερα) υποσύνολα

Άγνωστες τιμές

- Σε περίπτωση άγνωστης τιμής (missing value) του υπό ελέγχου χαρακτηριστικού, σε ποιο κλάδο εκχωρείται το παράδειγμα;
- Συχνά, η 'άγνωστη' καταχωρείται ως ξεχωριστή τιμή του χαρακτηριστικού
- Αν όχι, τότε απαιτείται ειδική μεταχείριση
  - Λύση Α: εκχώρηση του παραδείγματος στον κλάδο με τη μεγαλύτερη συχνότητα
  - Λύση Β: διαμελισμός του παραδείγματος
    - εκχώρηση τμημάτων του σε κάθε κλάδο, με στάθμιση ανάλογη της συχνότητας κάθε κλάδου στα υποδείγματα εκπαίδευσης
    - άθροιση των καταληκτικών υποδείξεων από κάθε κλάδο με ίδια στάθμιση

"Χειροποίητη" κατασκευή δένδρου απόφασης



X: region-centroid-row (Num)

Y: intensity-mean (Num)

Colour: class (Nom)

Select Instance

- *Open* segment-challenge.arff
- *Classify → choose → trees → UserClassifier*
- *Test options*: supplied test set → segment-test.arff
- *Start → data visualizer*
- X: region-centroid-row(Num) & Y: intensity-mean (Num)
- Rectangle, επιλογή όλων των υποδειγμάτων τάξης 'sky', *submit*
- *Tree visualizer…*
- Συνέχεια…

Κανόνες ταξινόμησης (classification rules)

- Δημοφιλής εναλλακτική στα δένδρα απόφασης
- *Προϋπόθεση κανόνα (antecedent)*: σύνολο ελέγχων (όμοιοι με τους ελέγχους στους κόμβους ενός δένδρου απόφασης)
  - Οι έλεγχοι συνήθως συμπλέκονται με λογική σύζευξη (ΚΑΙ, ωστόσο η χρήση και άλλων λογικών πράξεων είναι εφικτή)
- *Συμπέρασμα κανόνα* (*consequent)*: εκχώρηση ταξινόμησης, συνόλου ταξινομήσεων ή κατανομής πιθανότητας
  - Ανεξάρτητοι κανόνες συμπλέκονται με λογική διάζευξη (Ή)
  - Πρόβλημα: κάποιες φορές οι υποδείξεις των κανόνων είναι διαφορετικές για το ίδιο παράδειγμα

Μετατροπή δένδρου σε σύνολο κανόνων

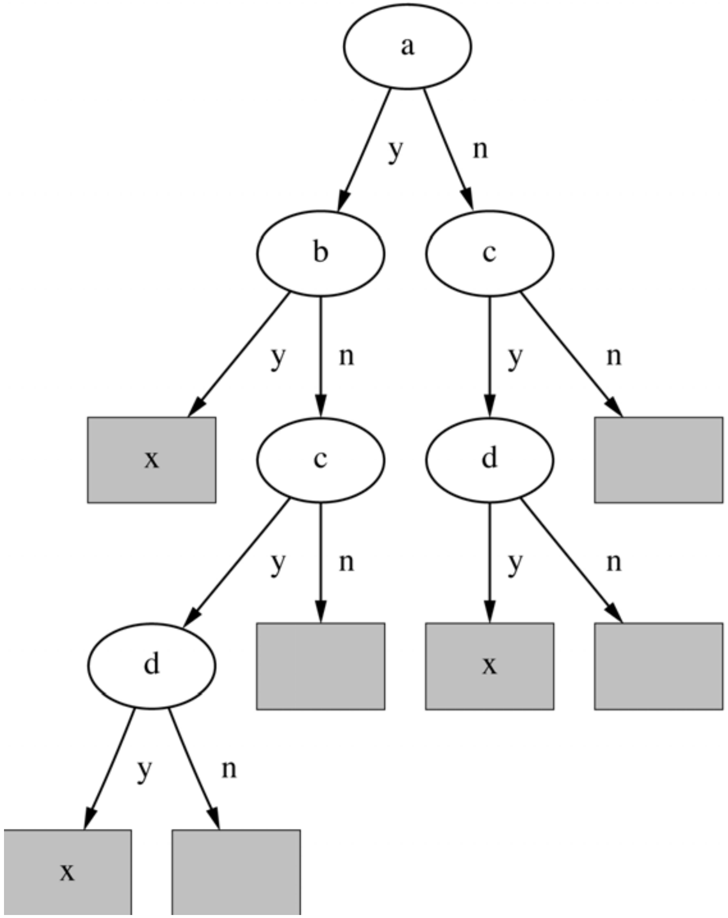
- Υλοποιείται εύκολα:
  - Ένας κανόνας για κάθε φύλλο
    - Η προϋπόθεση περιέχει μία συνθήκη για κάθε κόμβο που συναντάται από τη ‘ρίζα’ ως το φύλλο
    - Ως συμπέρασμα ορίζεται η τάξη εκχώρησης
- Οι παραγόμενοι κανόνες είναι σαφείς και ορίζονται μονοσήμαντα
  - Η σειρά εκτέλεσης δεν επηρεάζει το αποτέλεσμα
- Ωστόσο: οι κανόνες προκύπτουν υπερβολικά περίπλοκοι
  - Απαιτείται ‘κλάδεμα’ για την απομάκρυνση των περιττών ελέγχων και κανόνων

Μετατροπή συνόλου *κανόνων* σε δένδρο

- Συγκριτικά δυσκολότερη
  - Το δένδρο αδυνατεί να εκφράσει εύκολα τη λογική διάζευξη μεταξύ κανόνων
- Παράδειγμα: κανόνες ελέγχου διαφορετικών χαρακτηριστικών

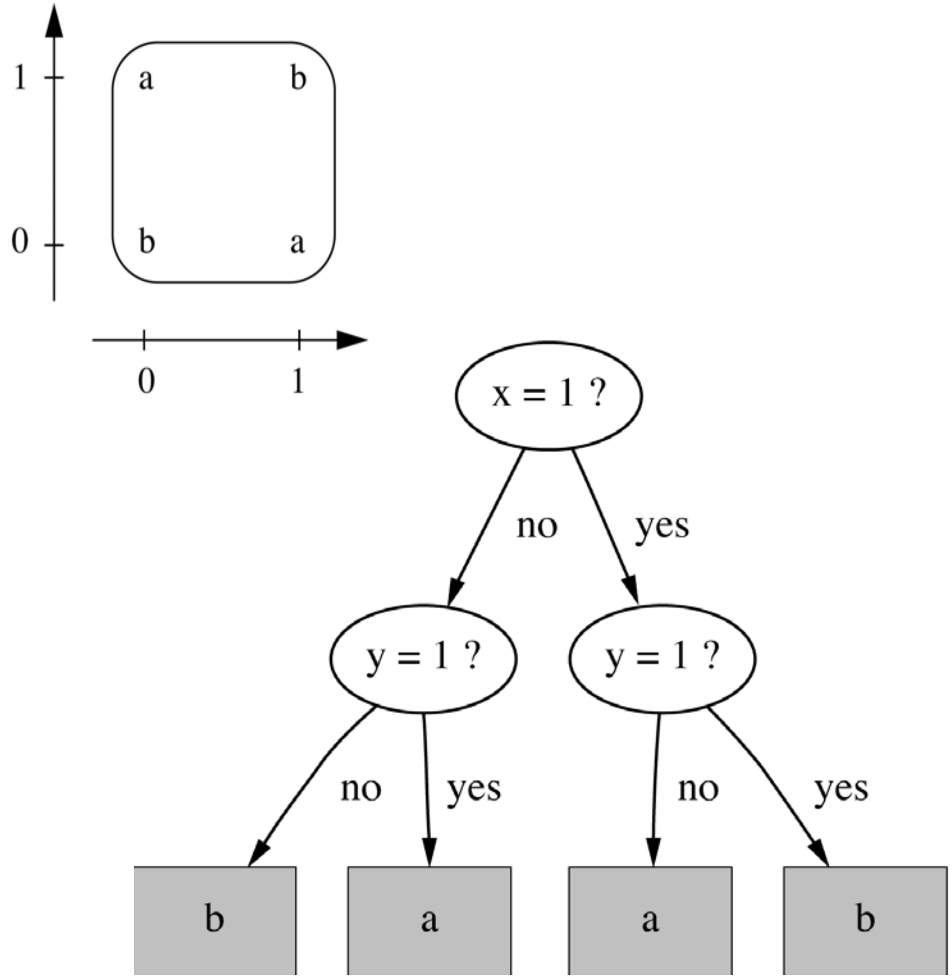
```
If a and b then x
If c and d then x
```

- Το αντίστοιχο δένδρο περιέχει πανομοιότυπα "υποδένδρα" (⇒ "πρόβλημα επαναλαμβανόμενου υποδένδρου", “replicated subtree problem”)



Δένδρο απόφασης απλής λογικής διάζευξης

Το πρόβλημα XOR

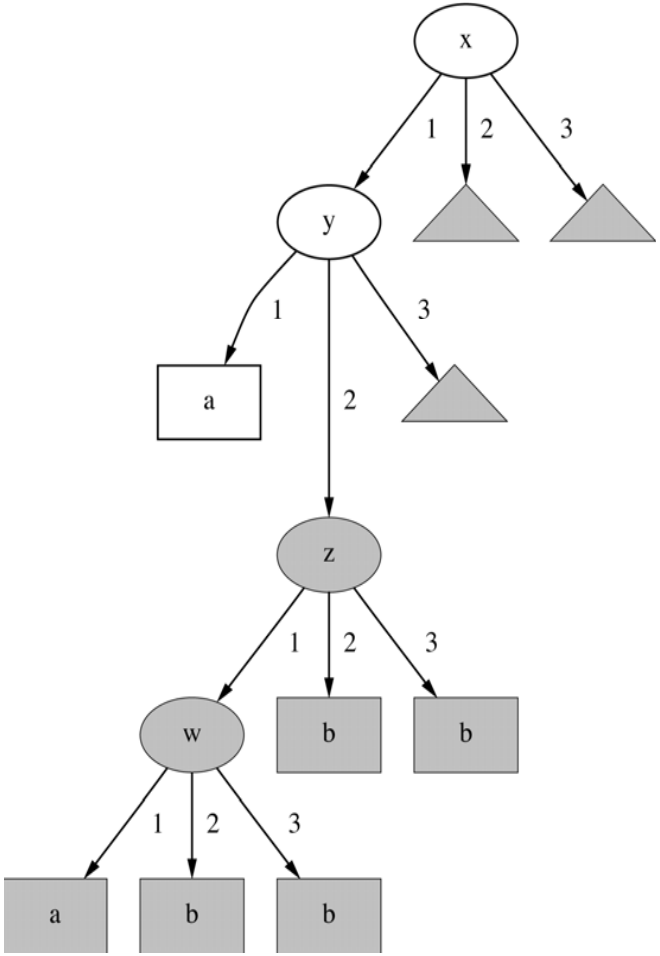


```
If x = 1 and y = 0
   then class = a
If x = 0 and y = 1
   then class = a
If x = 0 and y = 0
   then class = b
If x = 1 and y = 1
   then class = b
```

Στο συγκεκριμένο παράδειγμα, η δομική περιγραφή του συνόλου κανόνων δεν είναι αισθητά περισσότερο συμπαγής από εκείνη του δένδρου απόφασης.

Δένδρο με επαναλαμβανόμενο υποδένδρο

```
If x = 1 and y = 1
   then class = a
If z = 1 and w = 1
   then class = a
Otherwise class = b
```



Σε αυτό το παράδειγμα, τα γκρι τρίγωνα περιέχουν το σύνολο του γκρι υποδένδρου. Η ύπαρξη γενικού κανόνα (“otherwise…”) δεν μπορεί να περιγραφεί με συμπαγή τρόπο από ένα δένδρο.

Ανεξάρτητα τμήματα πληροφορίας

- Συνιστούν οι κανόνες ανεξάρτητα τμήματα πληροφορίας;
  - Η προσθήκη ενός κανόνα σε ένα σύνολο κανόνων φαντάζει εύκολα υλοποιήσιμη
  - Ωστόσο, μία τέτοια προσθήκη θα αγνοούσε τη διαδοχή εκτέλεσης
- Πιθανές διαδοχές εκτέλεσης ενός συνόλου κανόνων:
  - Διατεταγμένο σύνολο κανόνων (*λίστα απόφασης*, *decision list*)
    - Η διάταξη είναι σημαντική για την ερμηνεία
  - Μη διατεταγμένο σύνολο κανόνων
    - Οι κανόνες δύναται να επικαλύπτονται και να οδηγούν σε διαφορετικά συμπεράσματα για το ίδιο υπόδειγμα

Ερμηνεία κανόνων

- Πρόβλημα:
  - Δύο ή περισσότεροι κανόνες δίδουν αντικρουόμενες υποδείξεις
- Λύση:
  - Αδυναμία εκπόνησης συμπεράσματος;
  - Χρήση κανόνα με μεγαλύτερη συχνότητα στα δεδομένα ελέγχου;
  - …
- Πρόβλημα:
  - Κανένας κανόνας δεν εφαρμόζεται σε υπόδειγμα ελέγχου
- Λύση:
  - Αδυναμία εκπόνησης συμπεράσματος;
  - Εκχώρηση στην τάξη με τη μεγαλύτερη συχνότητα στα δεδομένα ελέγχου;
  - …

Ειδική περίπτωση: Δυαδική τάξη (boolean class)

- Παραδοχή: αν το υπόδειγμα δεν ανήκει στην τάξη “yes”, τότε ανήκει στην τάξη “no” (είδος ‘υπόθεσης κλειστού κόσμου’)
- Τέχνασμα: δημιουργία κανόνων μόνο για την τάξη “yes” και χρήση της τάξης “no” ως προεπιλεγμένης σε κάθε άλλη περίπτωση

```
If x = 1 and y = 1 then class = a
If z = 1 and w = 1 then class = a
Otherwise class = b
```

- Η διαδοχή των κανόνων δεν επηρεάζει τα εξαγόμενα, επίσης δεν παρατηρούνται αντικρουόμενες υποδείξεις
  - κάθε κανόνας συνιστά νέο, ανεξάρτητο τμήμα πληροφορίας
- Ο κανόνας μπορεί να γραφεί σε *διαζευκτική κανονικοποιημένη μορφή* (*disjunctive normal form*, διάζευξη διαδοχικών συζεύξεων)

Κανόνες συσχέτισης (association rules)

- Ειδοποιός (και μόνη ουσιαστική) διαφορά τους από τους κανόνες ταξινόμησης…
  - … δύνανται να προβλέψουν την τιμή κάθε χαρακτηριστικού ή συνδυασμού χαρακτηριστικών και όχι μόνο της τάξης
- Επίσης δεν χρησιμοποιούνται σύνολα κανόνων
- Πρόβλημα: αχανής αριθμός πιθανών συσχετίσεων
  - Αναγκαία η εισαγωγή κριτηρίων επιλογής των κανόνων με μεγαλύτερη συχνότητα εφαρμογής

Outlook	Temp	Humidity	Windy	**Play**
Sunny	Hot	High	False	**No**
Sunny	Hot	High	True	**No**
Overcast	Hot	High	False	**Yes**
Rainy	Mild	High	False	**Yes**
Rainy	Cool	Normal	False	**Yes**
Rainy	Cool	Normal	True	**No**
Overcast	Cool	Normal	True	**Yes**
Sunny	Mild	High	False	**No**
Sunny	Cool	Normal	False	**Yes**
Rainy	Mild	Normal	False	**Yes**
Sunny	Mild	Normal	True	**Yes**
Overcast	Mild	High	True	**Yes**
Overcast	Hot	Normal	False	**Yes**
Rainy	Mild	High	True	**No**

Υποστήριξη & εμπιστοσύνη κανόνα

- *Υποστήριξη* (*support*): η κάλυψη του κανόνα στο σύνολο δεδομένων, ο αριθμός των υποδειγμάτων στα οποία αυτός εφαρμόζεται
- *Εμπιστοσύνη* (*confidence*): η ακρίβεια του κανόνα, ο αριθμός των σωστών υποδείξεών ως ποσοστό της υποστήριξής του
- Παράδειγμα

  `If temperature = cool then humidity = normal`

  - 4 ημέρες με temp ‘cool’ & humidity ‘normal’

  $\Rightarrow$ support = 4, confidence = 100%
- Στις περισσότερες των περιπτώσεων τίθενται ελάχιστα όρια για την υποστήριξη και την εμπιστοσύνη των υπό αναζήτηση κανόνων
  - Παράδειγμα: 58 κανόνες με support $\geq$ 2 και confidence $\geq$ 95% για τα δεδομένα καιρού

Ερμηνεία κανόνων συσχέτισης

- Η ερμηνεία των κανόνων συσχέτισης δεν είναι πάντα προφανής και μονοσήμαντη, για παράδειγμα ο κανόνας

```
If windy = false and play = no
   then outlook = sunny and humidity = high
```

δεν ταυτίζεται με τον

```
If windy = false and play = no
   then outlook = sunny
If windy = false and play = no
   then humidity = high
```

- Ωστόσο, συνεπάγεται για παράδειγμα την ισχύ του:

```
If humidity = high and windy = false and
   play = no
   then outlook = sunny
```

Κανόνες με εξαιρέσεις (rules with exceptions)

- Μία φυσική προέκταση των κανόνων ταξινόμησης, είναι η προσθήκη *εξαιρέσεων*
  - Καθιστούν εφικτή την τροποποίηση ενός συνόλου κανόνων χωρίς να απαιτείται συνολικός εκ θεμελίων ανασχεδιασμός του
- Κύριο όφελος της μεθόδου απεικόνισης αποτελεί η δυνατότητα συγκριτικά εύκολης επέκτασης του συνόλου κανόνων για την συμπερίληψη νέων υποδειγμάτων
- Η δομή *default* . . . *except if* . . . *then* . . . είναι λογικά ισοδύναμη με τη *if* . . . *then* . . . *else* . . .
- Ωστόσο η δομή εκχώρησης προεπιλεγμένης τιμής και ελέγχου περί εξαιρέσεων είναι περισσότερο συμβατή με τον τρόπο αντίληψης ενός ειδικού, επομένως και ευκολότερα κατανοητή

Καταχώρηση νέου υποδείγματος

Sepal length	Sepal width	Petal length	Petal width	**Type**
5.1	3.5	2.6	0.2	**?**

- Το νέο υπόδειγμα ανήκει στην τάξη *Iris Setosa*
- Ωστόσο, οι προηγούμενα δημιουργημένοι κανόνες το καταχωρούν ως *Iris Versicolor*
  - *If petal length* ≥ 2.45 *and petal length* < 4.45 *then Iris versicolor*
  - *If petal length* ≥ 2.45 *and petal length* < 4.95 *and petal width* < 1.55 *then Iris versicolor*
- Εισαγωγή εξαίρεσης
  - *If petal length* ≥ 2.45 *and petal length* < 4.45 *then Iris versicolor EXCEPT if petal width* < 1.0 *then Iris setosa*
  - Είναι δυνατή η ύπαρξη εξαιρέσεων στις εξαιρέσεις και ούτω καθεξής, γεγονός που προσδίδει χαρακτήρα δένδρου στο σύνολο κανόνων

Παράδειγμα συνόλου κανόνων με εξαιρέσεις

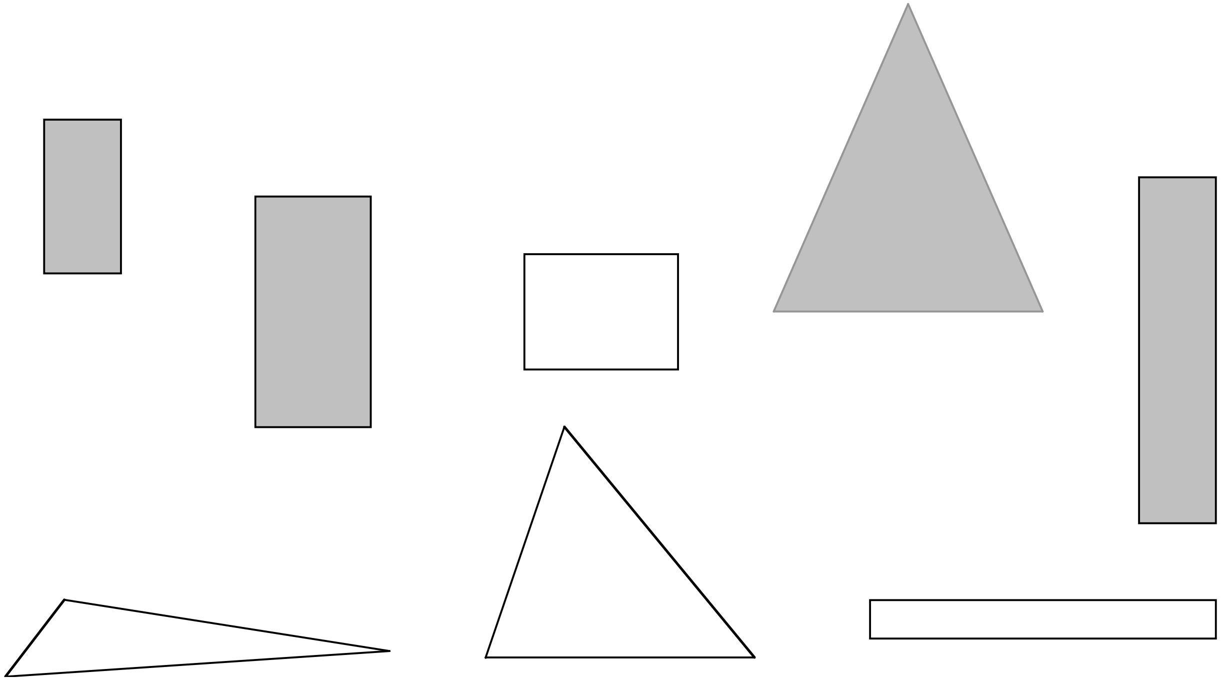
```
Default: Iris-setosa
   except if petal-length ≥ 2.45 and petal-length < 5.355
                                  and petal-width < 1.75
   then Iris-versicolor
        except if petal-length ≥ 4.95 and petal-width < 1.55
               then Iris-virginica
                    else if sepal-length < 4.95
                                and sepal-width ≥ 2.45
                         then Iris-virginica
   else if petal-length ≥ 3.35
        then Iris-virginica
             except if petal-length < 4.85 and sepal-length < 5.95
                    then Iris-versicolor
```

Κανόνες με συσχετίσεις (rules involving relations)

- Ως τώρα έγινε αναφορά σε ‘προτασιακούς’ (‘propositional’) κανόνες
  - Υλοποιούν σύγκριση της τιμής ενός χαρακτηριστικού με μία σταθερά, για παράδειγμα temperature < 45
  - Η ισχύς των εκφράσεών τους είναι αντίστοιχη της προτασιακής λογικής
- Η ισχύς αυτή δεν είναι επαρκής για προβλήματα που περιέχουν συσχετίσεις μεταξύ των υποδειγμάτων, όπως το δένδρο οικογένειας στη διάλεξη02
- Απαιτείται εναλλακτική δομική περιγραφή
  - Εισαγωγή συσχετίσεων σε σύνολα κανόνων

Πρόβλημα θέσης σχήματος

- Αντίληψη - στόχος: *όρθια ή μη θέση σχήματος*
- Υποδείγματα με σκίαση: *όρθια θέση*
  υποδείγματα χωρίς σκίαση: *μη όρθια θέση*



Μία προτασιακή επίλυση

Width	Height	Sides	**Class**
2	4	4	**Standing**
3	6	4	**Standing**
4	3	4	**Lying**
7	8	3	**Standing**
7	6	3	**Lying**
2	9	4	**Standing**
9	1	4	**Lying**
10	2	3	**Lying**

If width ≥ 3.5 and height < 7.0
then lying

If height ≥ 3.5 then standing

Μία συσχετιστική επίλυση

- Σύγκριση χαρακτηριστικών μεταξύ τους

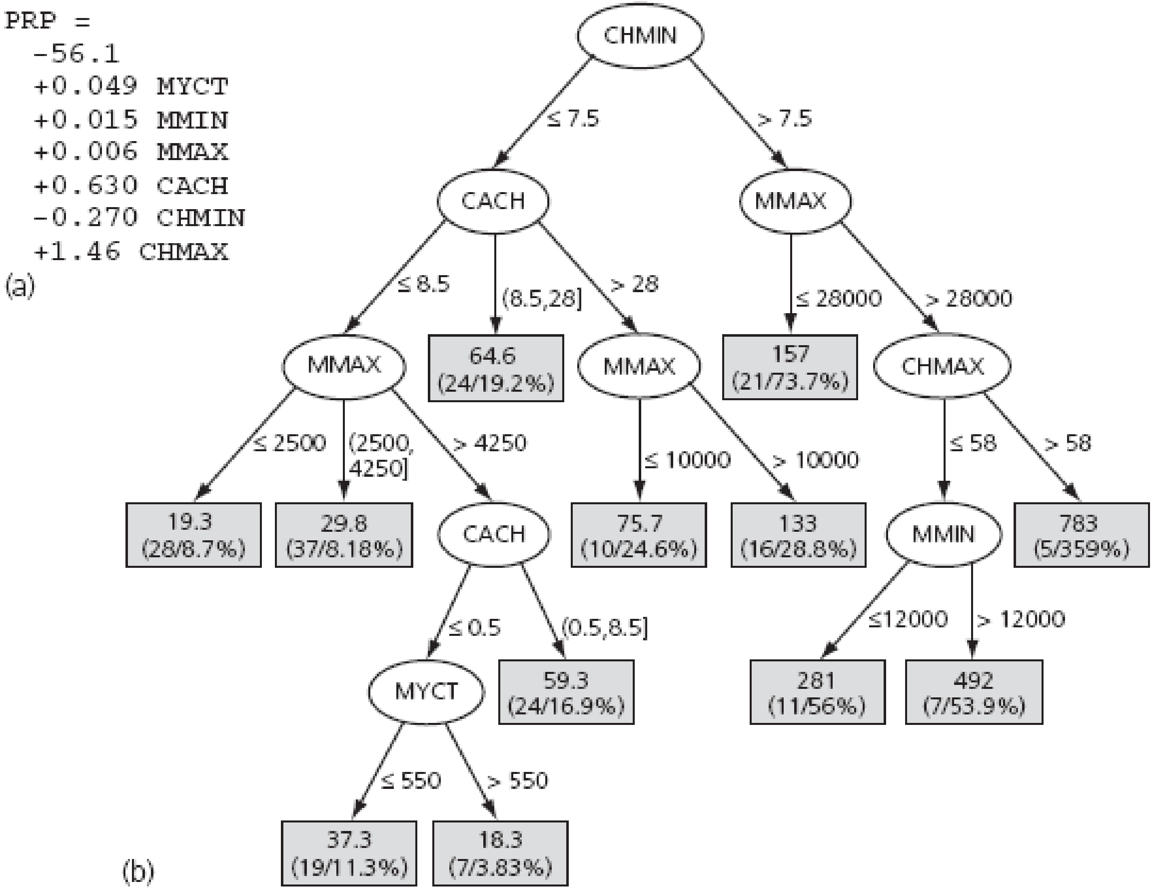
```
If width > height then lying
If height > width then standing
```

- Αρτιότερη γενίκευση σε νέα δεδομένα
- Τυπικές συσχετίσεις: =, <, >
- Μειονέκτημα: η εκμάθηση κανόνων με συσχετίσεις είναι υψηλού υπολογιστικού κόστους
- Απλή λύση: εισαγωγή πρόσθετων χαρακτηριστικών και εφαρμογή προτασιακής επίλυσης
  - Για παράδειγμα, δυαδικό χαρακτηριστικό *is width < height?*

Δένδρα για αριθμητική πρόβλεψη

- Το σύνολο των δένδρων απόφασης και των κανόνων που αναφέρθηκαν ως τώρα αφορούν την πρόβλεψη ονομαστικών παρά αριθμητικών ποσοτήτων
- Η πρόβλεψη αριθμητικών ποσοτήτων δύναται να υλοποιηθεί από τις ίδιες δομικές περιγραφές, με μικρές τροποποιήσεις
  - Οι κόμβοι του δένδρου περιέχουν αριθμητική ποσότητα που αποτελεί τη μέση τιμή όλων των υποδειγμάτων που καταλήγουν στα φύλλα που ακολουθούν
  - Τα συμπεράσματα των κανόνων αναφέρουν τη μέση τιμή των υποδειγμάτων στα οποία εφαρμόζεται ο κανόνας
  - Τα δένδρα με αριθμητικές τιμές στα καταληκτικά φύλλα καλούνται δένδρα παλινδρόμησης (*regression trees*)

Παράδειγμα δένδρου παλινδρόμησης

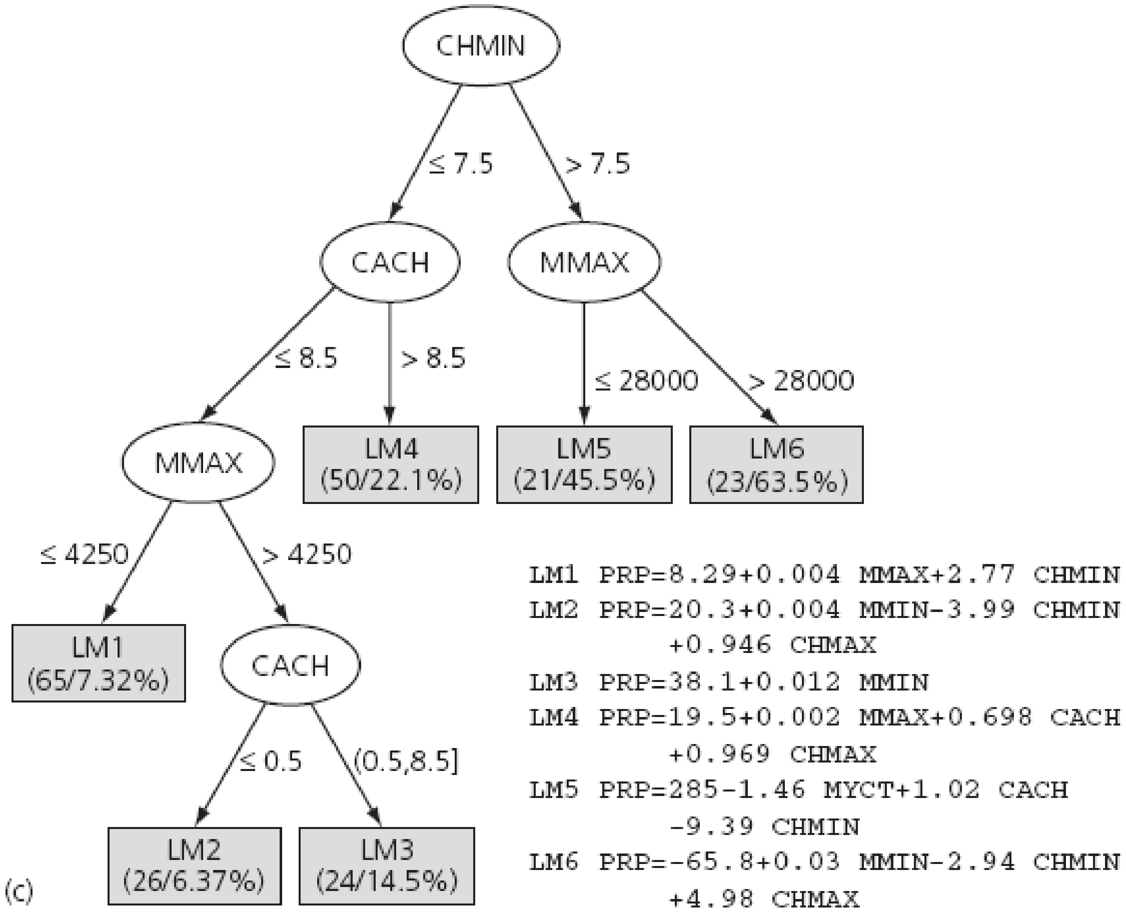


```
PRP =
 -56.1
 +0.049 MYCT
 +0.015 MMIN
 +0.006 MMAX
 +0.630 CACH
 -0.270 CHMIN
 +1.46 CHMAX
```

(a)

- Γραμμική παλινδρόμηση & δένδρο παλινδρόμησης
- Το δένδρο παρέχει μεγαλύτερη ακρίβεια
- Ωστόσο παραμένει ογκώδες και δυσνόητο

Συνδυασμός δένδρου και εξισώσεων παλινδρόμησης



```
LM1 PRP=8.29+0.004 MMAX+2.77 CHMIN
LM2 PRP=20.3+0.004 MMIN-3.99 CHMIN
        +0.946 CHMAX
LM3 PRP=38.1+0.012 MMIN
LM4 PRP=19.5+0.002 MMAX+0.698 CACH
        +0.969 CHMAX
LM5 PRP=285-1.46 MYCT+1.02 CACH
        -9.39 CHMIN
LM6 PRP=-65.8+0.03 MMIN-2.94 CHMIN
        +4.98 CHMAX
```

- Είναι εφικτός ο συνδυασμός ενός δένδρου με μία εξίσωση παλινδρόμησης
- *Model tree:* δένδρο του οποίου τα φύλλα περιέχουν αντί τιμών πρόβλεψης εξισώσεις παλινδρόμησης

Απεικόνιση με βάση υποδείγματα (instance-based representation)

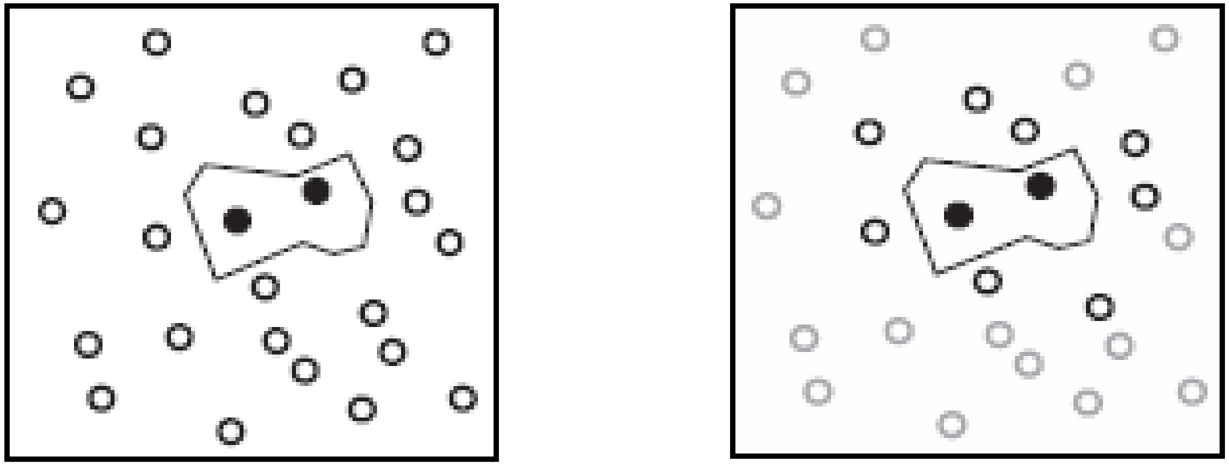
- Απλούστερη μορφή μάθησης: αποστήθιση (*rote learning*)
  - Το σύνολο υποδειγμάτων εκπαίδευσης καταχωρείται στη μνήμη
  - Πραγματοποιείται αντιστοίχιση των νέων παραδειγμάτων με όμοιά τους από το σύνολο εκπαίδευσης υποδείγματα
  - Απομένει μονάχα ο ορισμός του μέτρου ομοιότητας
- *Απεικόνιση γνώσης με βάση τα υποδείγματα*: εντελώς διαφορετική οπτική του προβλήματος εξόρυξης γνώσης
  - Χρήση των υποδειγμάτων ως έχουν για την απεικόνιση της γνώσης αντί για τη συναγωγή μοντέλου από αυτά
  - Ταξινόμηση κάθε νέου παραδείγματος με εκχώρηση τάξης του περισσότερο όμοιου υποδείγματος εκπαίδευσης
- Η θεμελιώδης διαφορά της μεθόδου έγκειται στη στιγμή κατά την οποία λαμβάνει χώρα η εκμάθηση
  - Η μέθοδος είναι αδρανής (‘lazy’), καθώς αναβάλλει τον κύριο όγκο εργασίας μέχρι την έλευση νέων παραδειγμάτων

Ταξινόμηση πλησιέστερων γειτόνων (nearest neighbor classification)

- Κάθε νέο παράδειγμα συγκρίνεται με τα ήδη υπάρχοντα υποδείγματα με χρήση κατάλληλου μέτρου απόστασης
- Το παράδειγμα εκχωρείται στην τάξη του πλησιέστερου υποδείγματος (ταξινόμηση *k πλησιέστερων γειτόνων*, *k-nearest-neighbor* classification)
- Μέτρα απόστασης
  - Αριθμητικά χαρακτηριστικά: ευκλείδεια απόσταση (παραδοχή: τα χαρακτηριστικά είναι κανονικοποιημένα και ίσης βαρύτητας)
  - Ονομαστικά χαρακτηριστικά: *απόσταση = 0* για ταυτόσημες τιμές, διαφορετικά *απόσταση =1*
  - Περισσότερο εξεζητημένα μέτρα απόστασης είναι ίσως επιθυμητά
    - για παράδειγμα διαφορετικές βαρύτητες χαρακτηριστικών

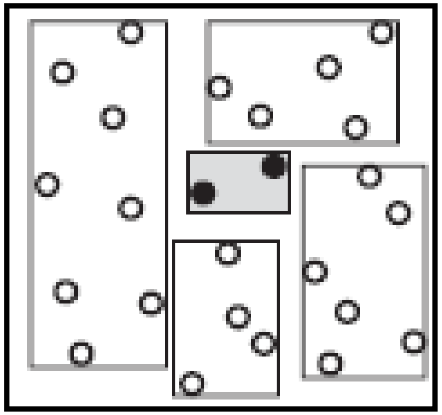
Επιλογή υποσυνόλου υποδειγμάτων

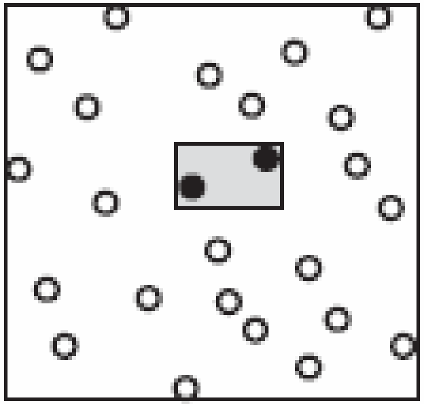
- Αποθήκευση *του συνόλου* των υποδειγμάτων εκπαίδευσης;
  - Απαιτεί υψηλό κόστος αποθήκευσης και εκτέλεσης
- Επιλογή ποιων υποδειγμάτων;
  - Σταθερές (ως προς την τάξη) περιοχές: απαιτούνται λίγα υποδείγματα
- Εμφανές μειονέκτημα: οι εξαγόμενες δομικές περιγραφές δεν είναι ρητές
  - Η δομική περιγραφή δεν είναι περιγραφική των προτύπων στα δεδομένα
  - Παραβιάζει την έννοια της γνώσης, όπως αυτή ορίστηκε στη διάλεξη01



Γενίκευση στην απεικόνιση με βάση υποδείγματα

- Γενίκευση των υποδειγμάτων: δημιουργία ορθογώνιων περιοχών που περιέχουν υποδείγματα της ίδιας τάξης
  - Νέα παραδείγματα που εμπίπτουν σε τέτοιες περιοχές ταξινομούνται αντίστοιχα
  - Σε περίπτωση που δεν εμπίπτουν σε καμία από αυτές, υλοποιείται ο κλασσικός κανόνας πλησιέστερου γείτονα
- Περισσότερο πολύπλοκη επιλογή: ένθετα ορθογώνια (nesting, ανάλογο των εξαιρέσεων στους κανόνες)
- Οπτικοποίηση της τεχνικής: προβληματική για μεγάλο αριθμό διαστάσεων και ονομαστικά χαρακτηριστικά

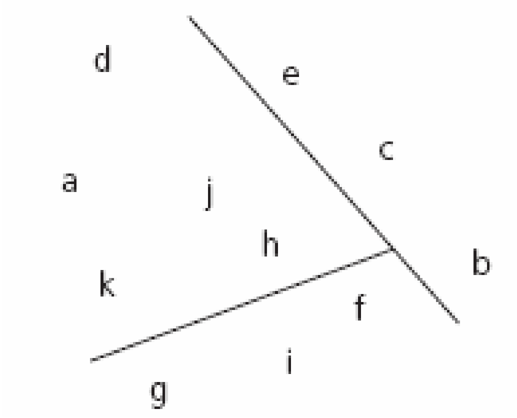


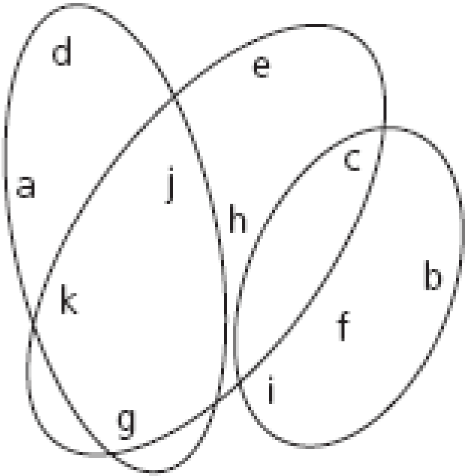


Ομάδες (clusters)

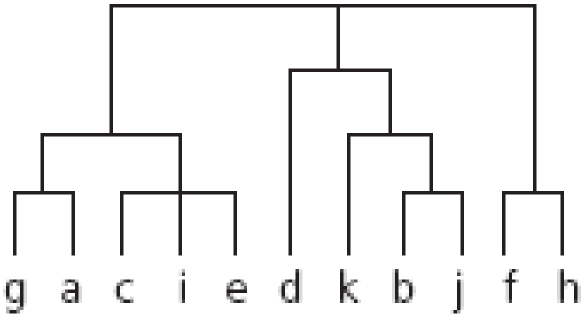
- Όταν ζητούμενο είναι η ομαδοποίηση και όχι η ταξινόμηση, ως εξαγόμενο προκύπτει η απεικόνιση του τρόπου εκχώρησης των υποδειγμάτων σε ομάδες
- Απλούστερη περίπτωση:
  - Συσχέτιση αριθμού ομάδας με κάθε υπόδειγμα, επίσης απεικόνιση με διαμελισμό του χώρου στις ομάδες και υπέρθεση των υποδειγμάτων
- Διάγραμμα Venn:
  - Επιτρέπει σε ένα υπόδειγμα να ανήκει σε περισσότερες της μίας τάξεων, επομένως και την επικάλυψη ομάδων
- Πιθανοκρατική συσχέτιση υποδειγμάτων με ομάδες:
  - Για κάθε υπόδειγμα δίδεται η πιθανότητα (degree of membership) να αποτελεί μέλος της κάθε ομάδας
- Δενδρόγραμμα
  - Ιεραρχική δομή ομάδων με διαδοχικούς διαμελισμούς του χώρου των υποδειγμάτων

Παραδείγματα ομαδοποιήσεων





	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1



Αξιοπιστία & Αποτίμηση

Αξιοπιστία & αποτίμηση

- Ζητήματα: εκπαίδευση, έλεγχος, ρύθμιση
- Πρόγνωση απόδοσης: όρια εμπιστοσύνης
- Παρακράτηση (holdout), διασταυρωμένη επικύρωση (cross-validation), μέθοδος bootstrap
- Σύγκριση σχημάτων: t-test
- Πρόγνωση πιθανοτήτων: συναρτήσεις απωλειών (loss functions)
- Κριτήρια εξαρτώμενα από κόστος (cost-sensitive measures)
- Αποτίμηση αριθμητικής πρόβλεψης
- Η αρχή ελαχίστου μήκους περιγραφής (the Minimum Description Length principle)

Αποτίμηση: το κλειδί της επιτυχίας

- Ποια η προβλεπτική ικανότητα του μοντέλου που προέκυψε από τον αλγόριθμο εκμάθησης;
    - Πλήθος μεθόδων εξαγωγής δομικών περιγραφών από δεδομένα
    - Ποιες από αυτές είναι οι βέλτιστες για συγκεκριμένο πρόβλημα;
    - Απαιτείται συστηματική προσέγγιση για την αποτίμηση και τη σύγκριση αποδοτικότητας των περιγραφών αυτών
- Αναφέρεται ρητά ότι το σφάλμα στα δεδομένα εκπαίδευσης **δεν** αποτελεί αξιόπιστο δείκτη απόδοσης σε μελλοντικά δεδομένα
- Λύση: διάσπαση των δεδομένων σε *σύνολο εκπαίδευσης* (*training set*) και *σύνολο ελέγχου* (*test set*)
    - Απαιτείται μεγάλος όγκος (ταξινομημένων) δεδομένων
- Ωστόσο: ο όγκος των δεδομένων είναι συνήθως περιορισμένος
    - Αναγκαία η χρήση περισσότερο εκλεπτυσμένων τεχνικών

Εκπαίδευση & έλεγχος

- Μέτρο απόδοσης σε προβλήματα ταξινόμησης: *τιμή σφάλματος* (*error rate)*
  - *Επιτυχία* (*success)*: σωστή πρόβλεψη της τάξης του υποδείγματος
  - *Σφάλμα* (*error)*: λανθασμένη πρόβλεψη της τάξης του υποδείγματος
  - Τιμή σφάλματος (error rate): αναλογία σφαλμάτων στο σύνολο των υποδειγμάτων
- Ωστόσο, ενδιαφέρει η πιθανή μελλοντική απόδοση σε νέα παραδείγματα και όχι η απόδοση στα ήδη δεδομένα υποδείγματα εκπαίδευσης…
- Αποτελεί η τιμή σφάλματος σε ήδη γνωστά δεδομένα αξιόπιστη ένδειξη της τιμής σφάλματος σε νέα δεδομένα;

Εκπαίδευση & έλεγχος

- Η απάντηση συνιστά ένα ηχηρό ΌΧΙ – όχι στην περίπτωση που τα παλαιά αυτά δεδομένα έχουν χρησιμοποιηθεί στη διαμόρφωση του μοντέλου κατά τη διαδικασία εκπαίδευσης
  - *Σφάλμα επαναληπτικής αντικατάστασης* (*resubstitution error*):
    - η τιμή σφάλματος που λαμβάνεται από τα δεδομένα εκπαίδευσης
    - υπολογίζεται με επανατροφοδότηση ενός μοντέλου ταξινόμησης με τα υποδείγματα που χρησιμοποιήθηκαν για την κατασκευή του
- Καθώς η δομική περιγραφή έχει σχεδιαστεί με τρόπο ώστε να ελαχιστοποιεί το συγκεκριμένο σφάλμα, οποιαδήποτε εκτίμηση της αποδοτικότητάς της βασισμένη στα υποδείγματα αυτά αποτελεί αισιόδοξη, αν όχι ανέλπιστα αισιόδοξη εκτίμηση

Εκπαίδευση & έλεγχος

- *Σύνολο ελέγχου* (*test set*): ανεξάρτητα υποδείγματα που δεν συμμετείχαν με κανένα τρόπο στη διαδικασία εκπαίδευσης
  - Παραδοχή: αμφότερα τα δεδομένα εκπαίδευσης και ελέγχου συνιστούν αντιπροσωπευτικά δείγματα του υποκείμενου προβλήματος
- Ωστόσο, τα σύνολα μπορούν να διαφέρουν στη φύση τους
  - Παράδειγμα: κατασκευή μοντέλων ταξινόμησης με χρήση δεδομένων πελατών από δύο διαφορετικές πόλεις A & B
    - Για την εκτίμηση της απόδοσης του μοντέλου της πόλης A σε νέα πόλη, έλεγχος στα δεδομένα της πόλης B

Περί ρύθμισης των παραμέτρων

- Επανάληψη: ιδιαίτερα σημαντική είναι η **μη** συμμετοχή με **οποιοδήποτε** τρόπο των δεδομένων ελέγχου στη δημιουργία του μοντέλου
- Κάποιοι αλγόριθμοι εκμάθησης απαρτίζονται από δύο στάδια:
  - Βήμα 1: κατασκευή της βασικής δομής
  - Βήμα 2: βελτιστοποίηση των παραμέτρων της
- Τα δεδομένα ελέγχου δεν πρέπει να χρησιμοποιηθούν ούτε για τη ρύθμιση των παραμέτρων!
- Σε αυτή την περίπτωση τα δεδομένα διαχωρίζονται σε *τρία* σύνολα:
  - *Δεδομένα εκπαίδευσης* (*training data*): για τη δημιουργία των μοντέλων ταξινόμησης
  - *Δεδομένα επικύρωσης* (*validation data*): για τη βελτιστοποίηση των παραμέτρων κάθε μοντέλου ή και για την επιλογή του βέλτιστου από αυτά
  - *Δεδομένα ελέγχου* (*test data*): για τον υπολογισμό της τιμή σφάλματος της τελικά επιλεγμένης και βελτιστοποιημένης μεθόδου

Βέλτιστη εκμετάλλευση των δεδομένων

- Καθώς η επικύρωση έχει ολοκληρωθεί, το *σύνολο* των δεδομένων μπορεί να χρησιμοποιηθεί για την κατασκευή ή και παραμετροποίηση του τελικού μοντέλου
- Γενικά, η ποιότητα του μοντέλου είναι ανάλογη του όγκου των διαθέσιμων δεδομένων
  - Αν και συνήθως βαίνει μειούμενη όταν ό όγκος του συνόλου εκπαίδευσης υπερβεί κάποιο όριο
- Η αξιοπιστία της εκτίμησης του σφάλματος είναι επίσης ανάλογη του όγκου των δεδομένων ελέγχου
- Διαδικασία *παρακράτησης* (*holdout*): διαχωρισμός του αρχικού συνόλου δεδομένων σε σύνολο εκπαίδευσης και ελέγχου
  - Δίλημμα: ιδανικά, αμφότερα τα σύνολα πρέπει να είναι μεγάλα!

Πρόβλεψη απόδοσης

- Έστω ότι το εκτιμώμενο σφάλμα ανέρχεται σε 25%. Πόσο κοντά είναι αυτή η εκτίμηση στο πραγματικό σφάλμα;
  - Εξαρτάται από τον όγκο των δεδομένων ελέγχου
- Η διαδικασία μπορεί να προσομοιωθεί με τη ρίψη ενός (έντονα μεροληπτικού!) κέρματος
  - Η ‘κεφαλή’ ισοδυναμεί με ‘επιτυχία’, τα ‘γράμματα’ με ‘σφάλμα’
- Μία διαδοχή τέτοιων ανεξάρτητων γεγονότων καλείται *διαδοχή Bernoulli*
  - Η στατιστική θεωρία παρέχει τα ζητούμενα επίπεδα εμπιστοσύνης

Επίπεδο εμπιστοσύνης

- Η μεταβλητή p ανήκει σε ορισμένο διάστημα τιμών με συγκεκριμένο βαθμό εμπιστοσύνης
- Παράδειγμα: $S=750$ επιτυχίες σε $N=1000$ δοκιμές
  - Εκτιμώμενος βαθμός επιτυχίας: 75%
  - Πόσο κοντά είναι η εκτίμηση αυτή στην πραγματική τιμή p;
    - απάντηση: με βαθμό εμπιστοσύνης 80%, $p \in [73.2, 76.7]$
- Άλλο παράδειγμα: $S=75$ και $N=100$
  - Εκτιμώμενος βαθμός επιτυχίας: 75%
  - με βαθμό εμπιστοσύνης 80%, $p \in [69.1, 80.1]$

Μέση τιμή και τυπική απόκλιση

- Μέση τιμή και τυπική απόκλιση για μία δοκιμή Bernoulli: p, $p\,(1-p)$
- Εκτιμώμενος βαθμός επιτυχίας $f=S/N$
- Μέση τιμή και διασπορά για τον βαθμό επιτυχίας f: p, $p\,(1-p)/N$
- Για αρκετά μεγάλο N, ο f ακολουθεί κανονική κατανομή
- c% διάστημα εμπιστοσύνης $[-z \le X \le z]$ για τυχαία μεταβλητή με μέση τιμή 0:

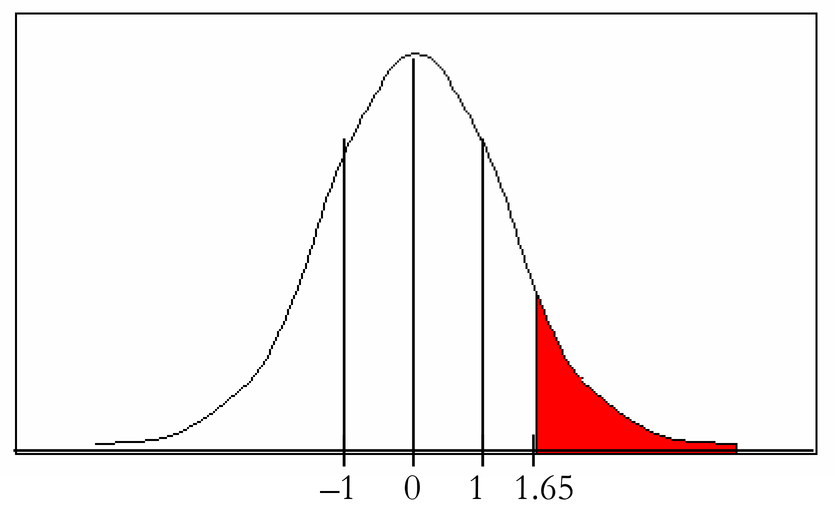
$$\Pr[-z \le X \le z] = c$$

- Για συμμετρική κατανομή:

$$\Pr[-z \le X \le z] = 1 - 2 \times \Pr[X \ge z]$$

Διαστήματα εμπιστοσύνης

- Διαστήματα εμπιστοσύνης για κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση 1:



$\Pr[X \geq z]$	z
0.1%	3.09
0.5%	2.58
1%	2.33
5%	1.65
10%	1.28
20%	0.84
40%	0.25

- Επομένως:

$$\Pr[-1.65 \leq X \leq 1.65] = 90\%$$

- Για να είναι δόκιμη η χρήση αυτών πρέπει να μετασχηματιστεί η τυχαία μεταβλητή f ώστε να έχει μέση τιμή 0 και τυπική απόκλιση 1

Μετασχηματισμός *f*

- Μετασχηματισμένη τιμή της *f* : $\dfrac{f-p}{\sqrt{p(1-p)/N}}$

  (δηλαδή αφαίρεση του μέσου & διαίρεση με την τυπική απόκλιση)

- Προκύπτει η εξίσωση: $\Pr\left[-z \le \dfrac{f-p}{\sqrt{p(1-p)/N}} \le z\right] = c$

- Επίλυση ως προς *p* :

$$p = \left(f + \frac{z^2}{2N} \pm z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) \Bigg/ \left(1 + \frac{z^2}{N} \right)$$

Παραδείγματα

- $f = 75\%$, $N = 1000$, $c = 80\%$ (έτσι ώστε $z = 1.28$):

$$p \in [0.732, 0.767]$$

- $f = 75\%$, $N = 100$, $c = 80\%$ (έτσι ώστε $z = 1.28$):

$$p \in [0.691, 0.801]$$

- Σημείωση: η παραδοχή περί κανονικής κατανομής ισχύει μονάχα για μεγάλα N ($N > 100$)
- $f = 75\%$, $N = 10$, $c = 80\%$ (έτσι ώστε $z = 1.28$):

$$p \in [0.549, 0.881]$$

Εκτίμηση με παρακράτηση (holdout estimation)

- Στην περίπτωση που ο όγκος δεδομένων είναι περιορισμένος;
- Η μέθοδος *παρακράτησης* (*holdout*) διατηρεί ένα συγκεκριμένο ποσοστό των δεδομένων προς έλεγχο και χρησιμοποιεί τα υπόλοιπα προς εκπαίδευση
  - Συνήθως: 1/3 για έλεγχο, 2/3 για εκπαίδευση
  - Επίσης παρακρατεί τμήμα αυτών προς επικύρωση, αν απαιτείται
- Πρόβλημα: τα δείγματα ίσως δεν είναι αντιπροσωπευτικά
  - Παράδειγμα: μία τάξη μπορεί να απουσιάζει στα δεδομένα ελέγχου
- Μία εξελιγμένη μέθοδος χρησιμοποιεί *διαστρωμάτωση* (*stratification, διαστρωματοποιημένη παρακράτηση, stratified holdout)*
  - Διαφυλάσσει την ύπαρξη κάθε τάξης με περίπου ίσες αναλογίες σε όλα τα υποσύνολα
    - Στοιχειώδης εξασφάλιση από ανομοιόμορφες κατανομές στα σύνολα εκπαίδευσης & ελέγχου

Μέθοδος επαναλαμβανόμενης παρακράτησης

- Η εκτίμηση με παρακράτηση μπορεί να βελτιωθεί, ως προς την αξιοπιστία της, με την επανάληψη της μεθόδου με χρήση διαφορετικών υποσυνόλων
    - Σε κάθε επανάληψη, ένα συγκεκριμένο ποσοστό επιλέγεται τυχαία για την εκπαίδευση (ίσως και με διαστρωμάτωση)
    - Ο μέσος όρος των τιμών σφάλματος κάθε επανάληψης υποδεικνύει περισσότερο αξιόπιστα τη ζητούμενη εκτίμηση
- Η μέθοδος καλείται *επαναλαμβανόμενης παρακράτησης* (*repeated holdout*)
- Ωστόσο, ακόμα μη βέλτιστη: τα διαφορετικά υποσύνολα πιθανά επικαλύπτονται
    - Πώς μπορεί να αποφευχθεί η επικάλυψη;

Διασταυρωμένη επικύρωση (cross-validation, CV)

- Η μέθοδος CV αποφεύγει το πρόβλημα επικάλυψης των συνόλων εκπαίδευσης
  - Πρώτο βήμα: διάσπαση των δεδομένων σε *k* ισομεγέθη υποσύνολα
  - Δεύτερο βήμα: χρήση κάθε υποσυνόλου διαδοχικά για έλεγχο, τα υπόλοιπα δεδομένα για εκπαίδευση
  - *Διασταυρωμένη επικύρωση k-πτυχών (k-fold CV)*
- Συχνά πραγματοποιείται διαστρωμάτωση πριν την εφαρμογή της διασταυρωμένης επικύρωσης
- Ο μέσος όρος των *k* εκτιμήσεων για το σφάλμα αποτελεί μία αξιόπιστη εκτίμηση

Περισσότερα περί διασταυρωμένης επικύρωσης

- Τυπική μέθοδος αποτίμησης: *διασταυρωμένη επικύρωση 10-πτυχών με διαστρωμάτωση* (*stratified ten-fold CV*)
- & γιατί 10;
  - Αναλυτικά πειράματα υποδεικνύουν τον αριθμό ως τη βέλτιστη επιλογή για τη λήψη αξιόπιστης εκτίμησης
  - Επίσης υπάρχουν θεωρητικές μαρτυρίες επ’ αυτού
- Η διαστρωμάτωση μειώνει τη διακύμανση των εκτιμήσεων
- Ακόμη περισσότερο αξιόπιστη επιλογή: *επαναλαμβανόμενη διασταυρωμένη επικύρωση με διαστρωμάτωση* (*repeated stratified CV*)
  - Για παράδειγμα, διασταυρωμένη επικύρωση 10-πτυχών επαναλαμβάνεται 10 φορές και υπολογίζεται ο μέσος όρος των αποτελεσμάτων
- Η λήψη αξιόπιστου μέτρου της απόδοσης αποτελεί εγχείρημα υψηλού υπολογιστικού κόστους

Διασταυρωμένη επικύρωση με αποκλεισμό μιας τιμής

- *Leave-One-Out CV*: μία ξεχωριστή μορφή διασταυρωμένης επικύρωσης
  - Αριθμός πτυχών ίσος με αριθμό υποδειγμάτων εκπαίδευσης
  - Για παράδειγμα, για *n* υποδείγματα εκπαίδευσης, κατασκευή μοντέλου ταξινόμησης *n*-φορές
- Πραγματοποιεί βέλτιστη αξιοποίηση των δεδομένων
  - Χρησιμοποιεί το μέγιστο δυνατό υποσύνολο προς εκπαίδευση
- Δεν περιλαμβάνει τυχαία δειγματοληψία
  - Η διαδικασία είναι ντετερμινιστική
- Υπερβολικά δαπανηρή σε υπολογιστικό κόστος

… & διαστρωμάτωση;

- Μειονέκτημα της διασταυρωμένης επικύρωσης με αποκλεισμό μιας τιμής:
  - μη εφικτή διαστρωμάτωση
  - για την ακρίβεια, εγγύηση ανομοιόμορφου υποδείγματος (κάθε σύνολο ελέγχου περιέχει ένα και μόνο υπόδειγμα)
- Ακραίο παράδειγμα: τυχαίο υποσύνολο χωρίζεται ισοπίθανα σε δύο τάξεις
  - Το βέλτιστο μοντέλο προβλέπει την πλειοψηφούσα –στο σύνολο εκπαίδευσης- τάξη
  - Ακρίβεια 50% σε νέα δεδομένα
  - Ωστόσο η (…) μέθοδος εκτιμά 100% σφάλμα!

Η μέθοδος bootstrap

- (bootstrap = θηλιά για το τράβηγμα μπότας προς το γόνατο)
- Η διασταυρωμένη επικύρωση χρησιμοποιεί δειγματοληψία *χωρίς επανατοποθέτηση*
  - Ένα υπόδειγμα, όταν επιλεγεί, δεν μπορεί να επιλεγεί ξανά για το συγκεκριμένο ζεύγος συνόλων εκπαίδευσης & ελέγχου
- Η μέθοδος *bootstrap* χρησιμοποιεί δειγματοληψία *με επανατοποθέτηση* για τη δημιουργία του συνόλου εκπαίδευσης
  - Για τη διαμόρφωση ενός συνόλου εκπαίδευσης *n* υποδειγμάτων, δειγματοληψία *n*-φορές με επανατοποθέτηση
  - Χρήση των υποδειγμάτων από το αρχικό σύνολο τα οποία δεν συναντώνται στο σύνολο εκπαίδευσης για έλεγχο

Η μέθοδος 0.632 bootstrap

- Ένα συγκεκριμένο υπόδειγμα έχει πιθανότητα *μη* επιλογής ίση με 1–1/*n*
- Επομένως, η πιθανότητα να καταλήξει στο σύνολο ελέγχου (να μην επιλεγεί σε κάποια δειγματοληψία με επανατοποθέτηση) είναι ίση προς:

$$\left(1-\frac{1}{n}\right)^n \approx e^{-1} = 0.368$$

- Κατά συνέπεια, τα δεδομένα εκπαίδευσης θα περιέχουν περίπου το 63.2% των υποδειγμάτων

Υπολογισμός σφάλματος με τη μέθοδο bootstrap

- Η βασισμένη στα δεδομένα ελέγχου εκτίμηση του σφάλματος είναι πεσιμιστική
  - Η εκπαίδευση πραγματοποιήθηκε μόνο στο 63% των υποδειγμάτων (ανεξάρτητα από το μέγεθός της *n*)
- Για το λόγο αυτό, εισάγεται ο συνδυασμός του με το σφάλμα επαναληπτικής αντικατάστασης (resubstitution error):

$$err = 0.632 \cdot e_{\text{test instances}} + 0.368 \cdot e_{\text{training instances}}$$

- Το σφάλμα επαναληπτικής αντικατάστασης λαμβάνει μικρότερη βαρύτητα από το σφάλμα των δεδομένων ελέγχου
- Υλοποιείται διαδοχική επανάληψη της διαδικασίας με διαφορετικά δείγματα, στη συνέχεια υπολογίζεται το μέσο των αποτελεσμάτων

Περισσότερα περί της bootstrap

- Ίσως ο πλέον βέλτιστος τρόπος για τον υπολογισμό της απόδοσης για πολύ μικρά σύνολα δεδομένων
- Ωστόσο, τα προβλήματα δεν είναι απόντα
  - Έστω το τυχαίο σύνολο δεδομένων που αναφέρθηκε προηγούμενα
  - Το καλύτερο δυνατό μοντέλο θα επιτύχει
    0% σφάλμα επαναληπτικής αντικατάστασης και
    ~50% σφάλμα στα δεδομένα ελέγχου
  - Η εκτίμηση της μεθόδου για αυτό το μοντέλο είναι:

$$err = 0.632 \cdot 50\% + 0.368 \cdot 0\% = 31.6\%$$

  - Παραμένει πεσιμιστική, το πραγματικά αναμενόμενο σφάλμα είναι ίσο με 50%

Σύγκριση διαφορετικών σχημάτων απεικόνισης γνώσης

- Ποιο από τα διάφορα σχήματα μάθησης εμφανίζει καλύτερη απόδοση;
  - Σημείωση: η απάντηση δεν είναι οικουμενική και εξαρτάται από το πρόβλημα
  - Προφανής λύση: σύγκριση εκτιμήσεων διασταυρωμένης επικύρωσης 10-πτυχών → ικανοποιητική στις περισσότερες των εφαρμογών
- Πρόβλημα: διακύμανση των εκτιμήσεων
  - Μπορεί να μειωθεί με επαναλαμβανόμενες διασταυρωμένες επικυρώσεις
- Ωστόσο, η αξιοπιστία των αποτελεσμάτων παραμένει κάποιες φορές αμφιλεγόμενη
  - Η παρατηρούμενη διαφορά μπορεί να οφείλεται σε σφάλμα κατά τη διαδικασία εκτίμησης
  - Απαιτείται τεκμηρίωση περί του αντιθέτου

Έλεγχος σημαντικότητας

- Οι έλεγχοι σημαντικότητας αναδεικνύουν το βαθμό εμπιστοσύνης στις υποδείξεις περί διαφοράς απόδοσης των συγκρινόμενων σχημάτων
  - Αν η διαφορά προκύψει σημαντική, εξασφαλίζεται ότι αυτή δεν οφείλεται σε τυχαίους παράγοντες (για παράδειγμα στο επιλεγμένο υποσύνολο του πειράματος)
  - Επειδή η απόδοση είναι συνάρτηση του μεγέθους των δεδομένων εκπαίδευσης, όλα τα σύνολα εκπαίδευσης πρέπει να είναι ισομεγέθη
  - Πράγματι, το πείραμα επαναλαμβάνεται με διαφορετικά μεγέθη ώστε να προκύψει μία καμπύλη μάθησης

Έλεγχος σημαντικότητας

- Υπόθεση
  - *Μηδενική υπόθεση*: η διαφορά δεν είναι στατιστικά σημαντική
  - *Εναλλακτική υπόθεση*: η διαφορά είναι στατιστικά σημαντική
  - Ο έλεγχος σημαντικότητας καταμετρά τις μαρτυρίες περί αποδοχής ή απόρριψης της μηδενικής υπόθεσης
  - Έστω ότι χρησιμοποιείται διασταυρωμένη επικύρωση 10-πτυχών
  - Ερώτηση: οι μέσοι όροι των εκτιμήσεων διαφέρουν σημαντικά μεταξύ των σχημάτων μάθησης;

Ζευγαρωτό (paired) t-test

- Το *t-test* του *Student* ελέγχει αν οι μέσες τιμές δύο δειγμάτων διαφέρουν σημαντικά
- Επιλογή ανεξάρτητων δειγμάτων με χρήση διασταυρωμένης επικύρωσης
- Χρήση του ζευγαρωτού t-test καθώς τα ανεξάρτητα δείγματα αποτελούν ζεύγος
  - Η ίδια διασταυρωμένη επικύρωση εφαρμόζεται δύο φορές

Κατανομή των μέσων τιμών

- Έστω $x_1\ x_2 \ldots x_k$ και $y_1\ y_2 \ldots y_k$ τα $2k$ δείγματα που προέκυψαν από την διασταυρωμένη επικύρωση k-πτυχών
- m_x και m_y οι μέσες τιμές
- Αν τα δείγματα είναι αρκετά, οι μέσες τιμές του συνόλου των ανεξάρτητων τμημάτων ακολουθούν κανονική κατανομή

- Εκτιμώμενες διακυμάνσεις των μέσων: σ_x^2/k και σ_y^2/k

- Αν μ_x και μ_y οι πραγματικές μέσες τιμές, τότε $\dfrac{m_x-\mu_x}{\sqrt{\sigma_x^2/k}}$ $\dfrac{m_y-\mu_y}{\sqrt{\sigma_y^2/k}}$ ανήκουν *κατά προσέγγιση* σε κανονική κατανομή με μέσο 0 και τυπική απόκλιση 1

Κατανομή Student

- Αν τα δείγματα είναι μικρά ($k < 100$) οι μέσες τιμές ακολουθούν κατανομή *Student με* k–1 *βαθμούς ελευθερίας*
- Όρια εμπιστοσύνης:

9 βαθμοί ελευθερίας

$\Pr[X \geq z]$	z
0.1%	4.30
0.5%	3.25
1%	2.82
5%	1.83
10%	1.38
20%	0.88

κανονική κατανομή

$\Pr[X \geq z]$	z
0.1%	3.09
0.5%	2.58
1%	2.33
5%	1.65
10%	1.28
20%	0.84

Κατανομή της διαφοράς των μέσων

- Έστω $m_d = m_x - m_y$
- Η διαφορά των μέσων m_d ακολουθεί επίσης κατανομή Student με k–1 βαθμούς ελευθερίας
- Έστω επίσης σ_d^2 η διακύμανση της διαφοράς
- Η τυποποιημένη έκφραση του m_d καλείται ως στατιστικό t:

$$t = \frac{m_d}{\sqrt{\sigma_d^2 / k}}$$

- Η τιμή του t χρησιμοποιείται για το t-test

Εκτέλεση του ελέγχου

- Επιλογή επιπέδου σημαντικότητας α
  - Αν η διαφορά είναι στατιστικά σημαντική στο επίπεδο α%, τότε υπάρχει πιθανότητα (100-α)% η διαφορά αυτή να μην οφείλεται σε τυχαίους παράγοντες
- Διαίρεση του επιπέδου σημαντικότητας σε δύο μέρη, μιας και το τεστ είναι διμερές (two-tailed)
  - Η διαφορά πρέπει να είναι μεγαλύτερη ή μικρότερη από α/2
- Εύρεση της τιμής του z που αντιστοιχεί σε πιθανότητα α/2
- Αν $t \le -z$ ή $t \ge z$ τότε η διαφορά είναι σημαντική και η μηδενική υπόθεση μπορεί να απορριφθεί

Ασύζευκτες (unpaired) παρατηρήσεις

- Σε περίπτωση της CV, οι εκτιμήσεις προέρχονται από διαφορετικά τυχαία δείγματα και μοιραία δεν αποτελούν ζεύγος
  - Ίδια κατάληξη έχει και η χρήση *k* -fold CV για το ένα σχήμα και η χρήση *j* -fold CV για το άλλο
- Τότε απαιτείται η χρήση ενός *un*paired t-test με min(*k* , *j*) – 1 βαθμούς ελευθερίας
- Το *t*-statistic μετατρέπεται σε:

$$t = \frac{m_d}{\sqrt{\sigma_d^2 / k}} \quad \Longrightarrow \quad t = \frac{m_x - m_y}{\sqrt{\dfrac{\sigma_x^2}{k} + \dfrac{\sigma_y^2}{j}}}$$

Ερμηνεία του αποτελέσματος

- Το σύνολο των εκτιμήσεων της CV βασίζεσαι στο ίδιο σύνολο δεδομένων
- Τα δείγματα δεν είναι ανεξάρτητα
- Πρέπει να χρησιμοποιηθεί δείγμα από διαφορετικό υποσύνολο τιμών για καθεμία από τις *k* εκτιμήσεις που χρησιμοποιούνται για την κρίση της απόδοσης σε διάφορα σύνολα εκπαίδευσης
- Εναλλακτικά, χρήση ευρετικού ελέγχου, για παράδειγμα *corrected resampled t-test*

$$t = \frac{\bar{d}}{\sqrt{\left(\frac{1}{k} + \frac{n_2}{n_1}\right)\sigma_d^2}}$$

Πρόβλεψη πιθανότητας

- Μέτρο απόδοσης έως τώρα: *Βαθμός επιτυχίας (success rate)*
- Επίσης καλείται ως *συνάρτηση απωλειών 0-1 (loss function)*:

$$\sum_i \begin{cases} 0 & \text{αν πρόβλεψη : σωστή} \\ 1 & \text{αν πρόβλεψη : λάθος} \end{cases}$$

- Ωστόσο, οι περισσότεροι ταξινομητές υποδεικνύουν πιθανότητα ανά τάξη
- Ανάλογα της εφαρμογής, πιθανόν να είναι επιθυμητός ο έλεγχος της ακρίβειας των εκτιμήσεων πιθανότητας
- Η συνάρτηση απωλειών 0-1 είναι ακατάλληλη

Δευτεροβάθμια συνάρτηση απωλειών (Quadratic loss function)

- $p_1 \ldots p_k$: εκτιμήσεις πιθανότητας για συγκεκριμένο παράδειγμα
- c : δείκτης πραγματικής τάξης παραδείγματος
- $a_1 \ldots a_k = 0$, εκτός από a_c όπου ίσο με1
- *Quadratic loss*: $\sum_j (p_j - a_j)^2 = \sum_{j \neq c} p_j^2 + (1 - p_c)^2$
- Η ελαχιστοποίηση του $E\left[\sum_j (p_j - a_j)^2\right]$

προκύπτει όταν $p_j = p_j^*$, οι αληθινές πιθανότητες

Συνάρτηση απωλειών πληροφορίας (Informational loss function)

- Συνάρτηση απωλειών πληροφορίας: $\mathbf{-\log_2(p_c)}$ όπου c δείκτης της πραγματικής τάξης παραδείγματος
  - Εκφράζει τον αριθμό των bits που απαιτούνται για την περιγραφή της πραγματικής τάξης
- Έστω $p_1^* \ldots p_k^*$ οι αληθείς πιθανότητες κάθε τάξης
- Αναμενόμενη τιμή συνάρτησης απωλειών:

$$-p_1^* \log_2 p_1 - \ldots - p_k^* \log_2 p_k$$

- Τεκμηρίωση: ελάχιστη όταν $p_j = p_j^*$
- Μειονέκτημα: *Πρόβλημα μηδενικής συχνότητας (zero-frequency problem)*
  - Αν εκχωρηθεί μηδενική πιθανότητα σε τάξη που τελικά προκύψει, η τιμή της συνάρτησης τείνει στο μείον άπειρο

Επιλογή κριτηρίου

- Ποια η βέλτιστη συνάρτηση απωλειών προς χρήση;
  - Επιλογή μάλλον υποκειμενική και κατά περίπτωση
    - Αμφότερες οι συναρτήσεις επιτελούν επιτυχώς το ρόλο τους
  - Η πρώτη λαμβάνει υπ’ όψιν τις εκτιμήσεις πιθανότητας όλων των τάξεων για κάθε υπόδειγμα
  - Η δεύτερη εστιάζει στην εκτίμηση πιθανότητας της πραγματικής τάξης και μόνο
  - Η δευτεροβάθμια έχει όριο τιμών
    - Δεν μπορεί να υπερβεί την τιμή 2
  - Η απώλεια πληροφορίας μπορεί να είναι άπειρη
- Η συνάρτηση απωλειών πληροφορίας συσχετίζεται με την *Αρχή Ελαχίστου Μήκους Περιγραφής (MDL)* [βλέπε προηγούμενη διάλεξη]

Καταμέτρηση κόστους

- Στην πράξη, διαφορετικοί τύποι σφάλματος ταξινόμησης επιφέρουν διαφορετικό κόστος
    - Οι ως τώρα αποτιμήσεις δεν λαμβάνουν υπ’ όψιν το κόστος λήψης λανθασμένης απόφασης από λανθασμένη ταξινόμηση
    - Η ελαχιστοποίηση του σφάλματος χωρίς παράλληλη αποτίμηση του κόστους οδηγεί συχνά σε παράξενα αποτελέσματα
- Παράδειγμα:
    - Πρόγνωση σεισμού
        - “όχι σεισμός” σωστή πρόβλεψη στο 99.99% των περιπτώσεων
        - Αλλά ποιο το κόστος μη ανίχνευσης ενός σεισμού;
- Στην πραγματικότητα είναι δύσκολο να εντοπίσει κανείς εφαρμογή στην οποία το κόστος διαφορετικών σφαλμάτων είναι σταθερό

Καταμέτρηση κόστους

- *Πίνακας σύγχυσης (confusion matrix)*:

		Predicted class	
		Yes	No
Actual class	Yes	True Positive (TP)	False Negative (FN)
	No	False Positive (FP)	True Negative (TN)

- Success_rate = (TP+TN)/(TP+FN+FP+TN)
- Διάφοροι τύποι κόστους!
  - Για παράδειγμα, κόστος συλλογής δεδομένων εκπαίδευσης

Ταξινόμηση εξαρτώμενη από κόστος (cost sensitive classification)

- Τα διάφορα σφάλματα έχουν πιθανά διαφορετικό κόστος, όμοια οι επιτυχείς ταξινομήσεις φέρουν διαφορετικό κέρδος: *πίνακας κόστους (cost matrix)*
- Ο βαθμός επιτυχίας μπορεί να αντικατασταθεί πλέον από το μέσο κόστος ανά απόφαση
- Είναι εφικτός ο υπολογισμός του κόστους ενός συγκεκριμένου μοντέλου εκμάθησης σε δεδομένου σύνολο ελέγχου
- Αν το μοντέλο παρέχει πιθανότητα για κάθε πρόβλεψη, μπορεί να ρυθμιστεί ώστε να ελαχιστοποιεί το αναμενόμενο κόστος των προβλέψεων

		Predicted class		
		a	b	c
Actual class	a	0	1	1
	b	1	0	1
	c	1	1	0

Μάθηση εξαρτώμενη από κόστος

- Τα περισσότερα σχήματα εκμάθησης δεν υλοποιούν μάθηση εξαρτώμενη από κόστος
  - Παράγουν την ίδια ταξινόμηση ανεξάρτητα από το κόστος που εκχωρείται σε κάθε τάξη
    - Παράδειγμα: τυπικό δένδρο απόφασης
- Απλές μέθοδοι υλοποίησης μάθησης εξαρτώμενης από κόστος:
  - Επαναληπτική δειγματοληψία των υποδειγμάτων ανάλογα του κόστους
  - Εκχώρηση βαρύτητας στα υποδείγματα ανάλογα με το κόστος
- Κάποια σχήματα εκμάθησης δύνανται να λάβουν υπ’ όψιν το κόστος τροποποιώντας την τιμή μιας παραμέτρου
    - για παράδειγμα naïve Bayes

Σωρευτικά διαγράμματα (lift charts)

- Στην πράξη, τα διάφορα κόστη είναι σπανίως γνωστά
- Οι αποφάσεις λαμβάνονται με σύγκριση διαφόρων δυνατών σεναρίων
- Παράδειγμα: προώθηση προϊόντος με αποστολή επιστολής σε 1.000.000 νοικοκυριά
  - Αποστολή σε όλα, ποσοστό απόκρισης 0.1% (1000 νοικοκυριά)
  - Εργαλείο εξόρυξης δεδόμένων υποδεικνύει υποσύνολο των 100.000 περισσότερο πιθανών με απόκριση 0.4% (400)
    40% των ενδιαφερόμενων πελατών για το 10% του κόστους;
  - Ανάδειξη υποσυνόλου 400,000 περισσότερο πιθανών, απόκριση 0.2% (800)
- Ένα σωρευτικό διάγραμμα επιτρέπει την οπτική αντιπαραβολή

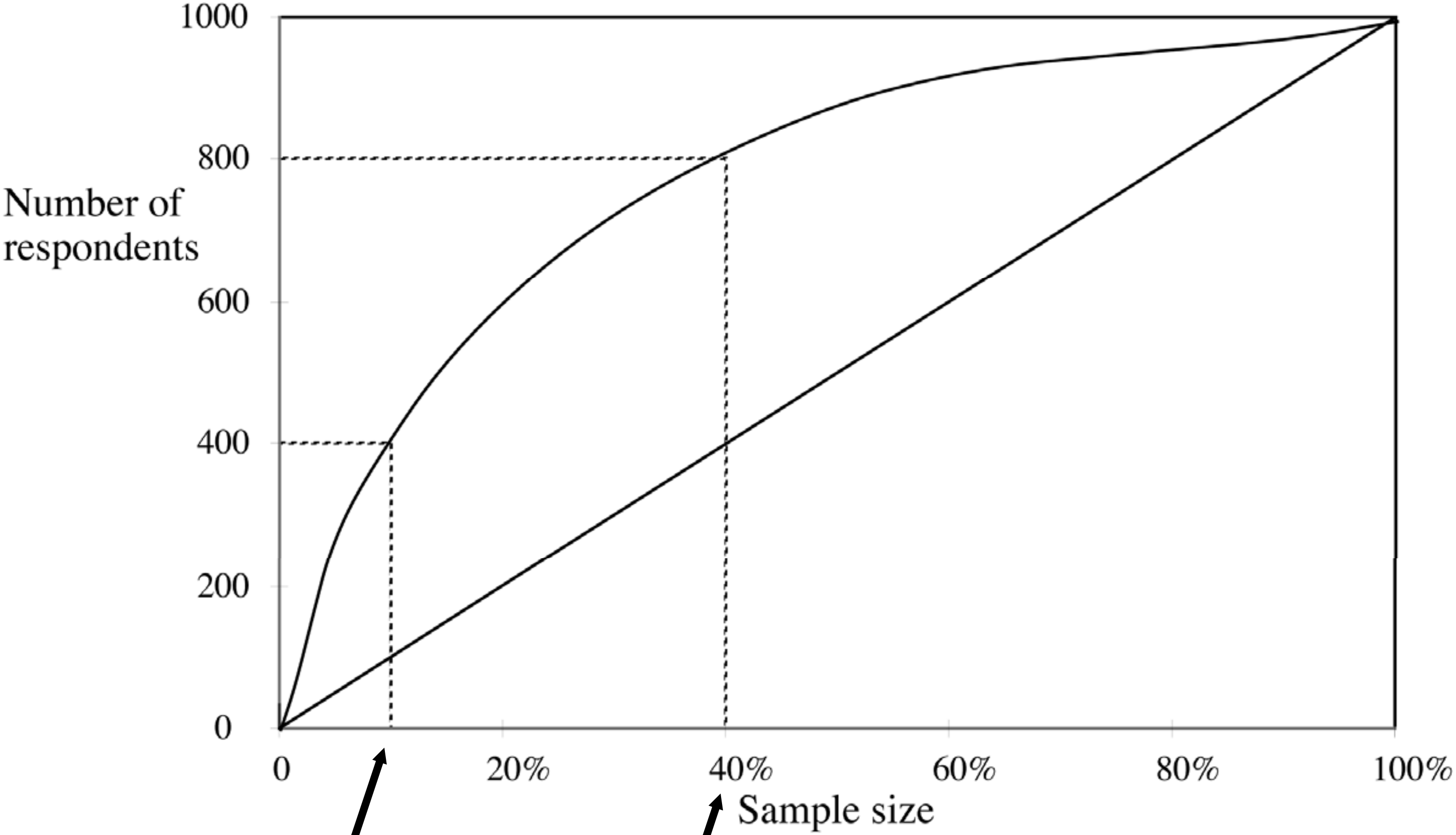
Δημιουργία σωρευτικού διαγράμματος

- Ταξινόμηση υποδειγμάτων συνόλου ελέγχου σε φθίνουσα σειρά προβλεπόμενης πιθανότητας να ανήκουν στη ζητούμενη τάξη:

	Προβλεπόμενη πιθανότητα	**Πραγματική τάξη**
1	0.95	Yes
2	0.93	Yes
3	0.93	No
4	0.88	Yes
…	…	…

- *Άξονας x:* μέγεθος δείγματος
  Άξονας y: αριθμός αληθών θετικών υποδειγμάτων

Ένα υποθετικό σωρευτικό διάγραμμα



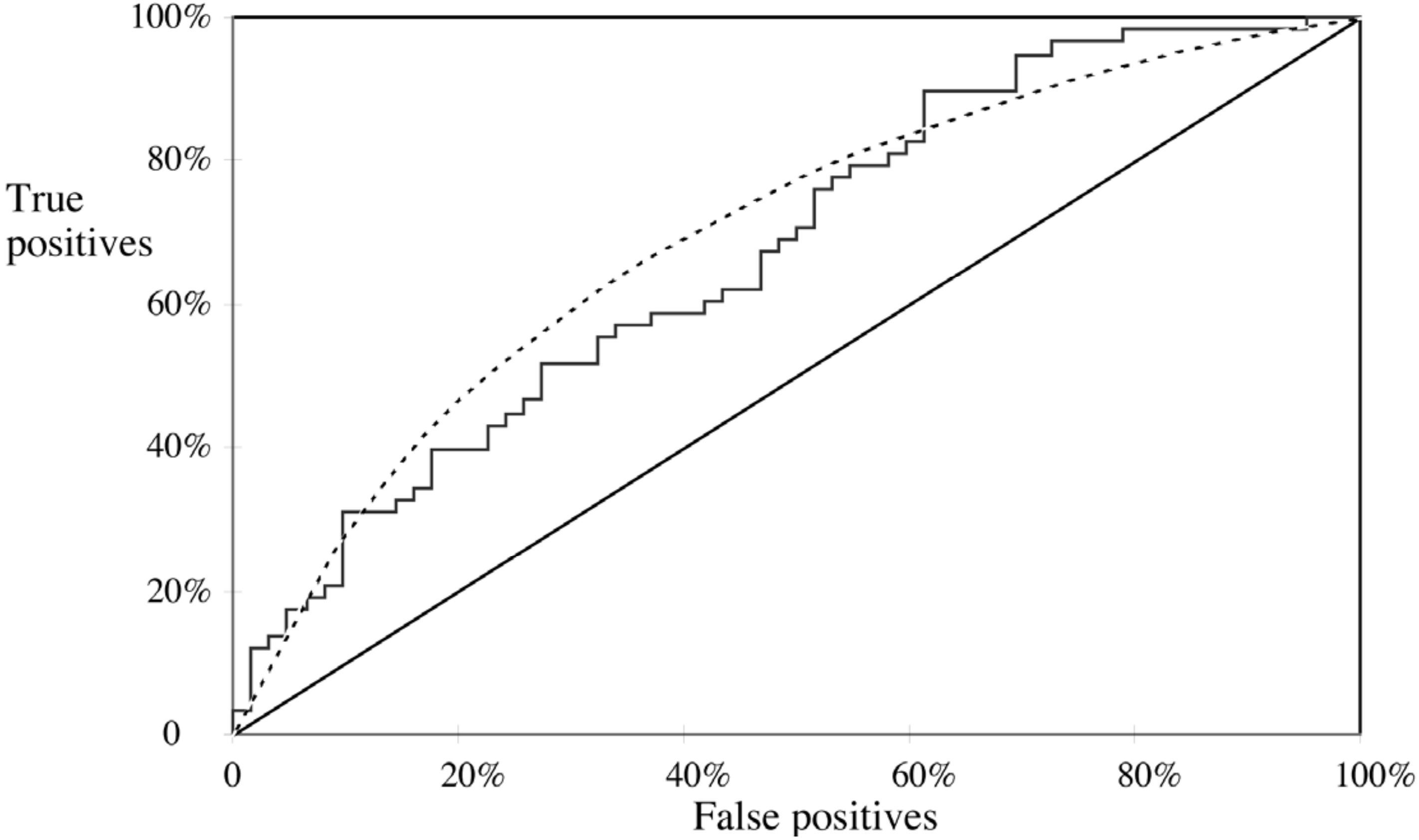
40% των αποκρίσεων
για 10% του κόστους

80% των αποκρίσεων
για 40% του κόστους

Καμπύλες ROC

- Παρόμοιες των σωρευτικών διαγραμμάτων
  - ROC: Receiver Operating Characteristic
    - Προέλευση: ανίχνευση σήματος, δείκτης της εξισσορόπησης μεταξύ του βαθμού επιτυχίας και των λανθασμένων συναγερμών σε ένα κανάλι με θόρυβο
- Απεικονίζουν την απόδοση ενός ταξινομητή, ανεξάρτητα της κατανομής της τάξης ή του κόστους των σφαλμάτων
- Διαφοροποίηση από το σωρευτικό διάγραμμα:
  - Άξονας *y*: ποσοστό αληθών θετικών στο δείγμα *(αντί απόλυτου αριθμού)*
  - Άξονας *x* : ποσοστό ψευδών θετικών στο δείγμα *(αντί μεγέθους δείγματος)*
- Επιθημητή είναι η παρουσία στο βορειοδυτικό άκρο του διαγράμματος, όπως και στα σωρευτικά διαγράμματα

Υπόδειγμα καμπύλης ROC

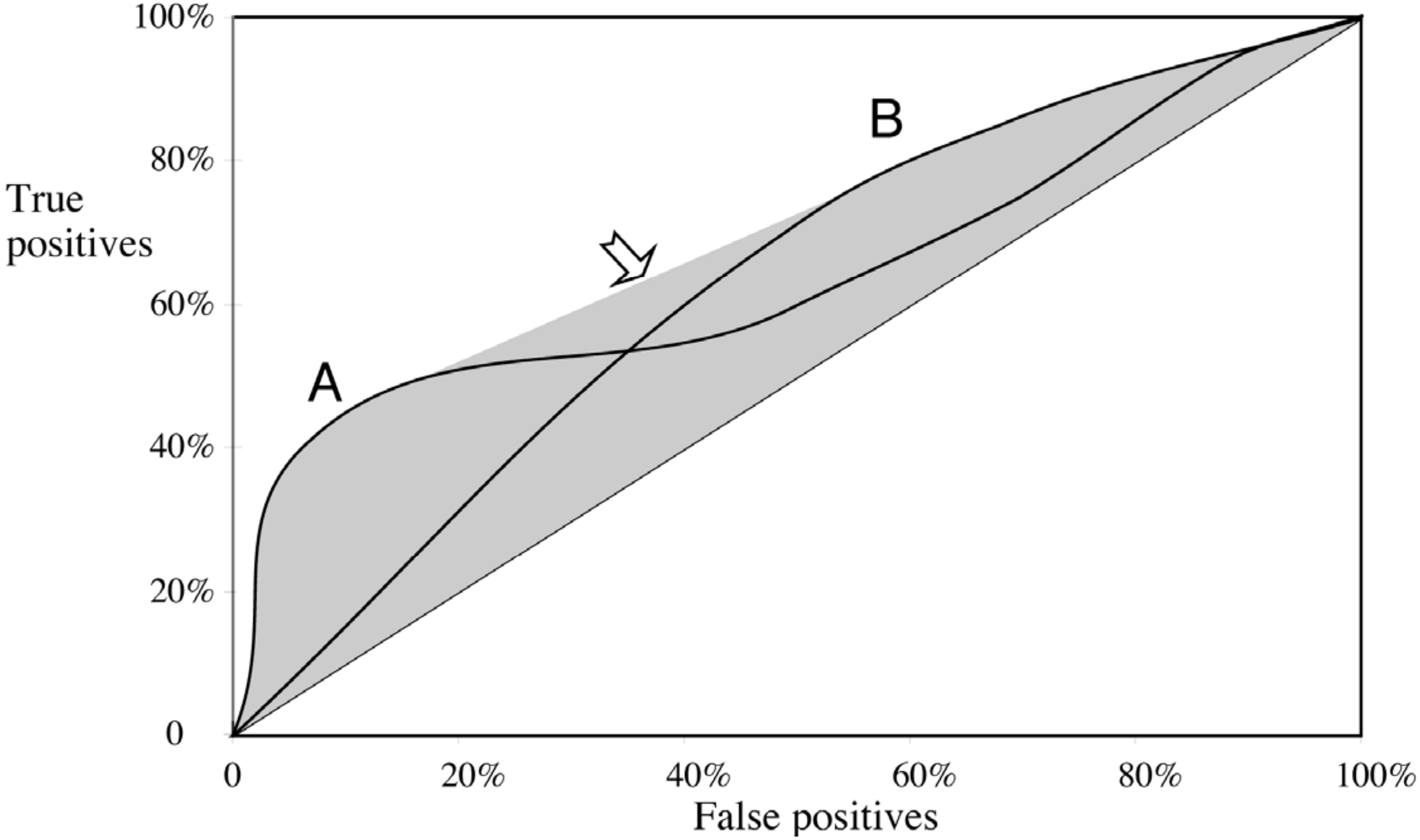


- Πριονωτή καμπύλη – ένα σύνολο δεδομένων ελέγχου
- Λεία καμπύλη – χρήση διασταυρωμένης επικύρωσης

Διασταυρωμένη επικύρωση και καμπύλες ROC

- Απλή μέθοδος δημιουργίας καμπύλης ROC με χρήση διασταυρωμένης επικύρωσης:
  - Συλλογή πιθανότητας υποδειγμάτων από τις πτυχές ελέγχου (test folds)
  - Ταξινόμηση υποδειγμάτων σε φθίνουσα σειρά πιθανότητας
  - *Αυτή η μέθοδος υλοποιείται στο WEKA*
- Ωστόσο, υπάρχουν και άλλες δυνατότητες
  - Για παράδειγμα, δημιουργία καμπύλης ROC για κάθε πτυχή και εύρεση μέσου όρου

Σύγκριση σχημάτων εκμάθησης με χρήση καμπύλων ROC



- Για ένα μικρό δείγμα με υψηλή εστίαση επέλεξε τη μέθοδο Α
- Για ένα μεγαλύτερο, επέλεξε τη μέθοδο Β
- Για ενδιάμεσα μεγέθη, επέλεξη μεταξύ Α & Β με κατάλληλες πιθανότητες

Το κυρτό κέλυφος
(convex hull)

- Για τα δεδομένα δύο σχήματα εκμάθησης της προηγούμενης διαφάνειας, μπορεί να επιτευχθεί οποιοδήποτε σημείο μέσα στο κυρτό κέλυφος (σημεία με σκίαση)!
- TP & FP rates για το σχήμα 1: t_1 & f_1
- TP & FP rates για το σχήμα 2: t_2 & f_2
- Αν το σχήμα 1 χρησιμοποιείται για την πρόβλεψη 100×q % των περιπτώσεων και το σχήμα 2 για τις υπόλοιπες, τότε
  - TP rate για το συνδυασμό τους:
    $q \times t_1+(1-q) \times t_2$
  - FP rate για το συνδυασμό τους:
    $q \times f_2+(1-q) \times f_2$
- Ένα σχήμα πρέπει να χρησιμοποιείται στο 100% των περιπτώσεων μόνο όταν, για το ζητούμενο ποσοστό, παράγει σημείο που ανήκει στο κυρτό κέλυφος

Καμπύλες ανάκλησης–ακρίβειας (recall–precision)

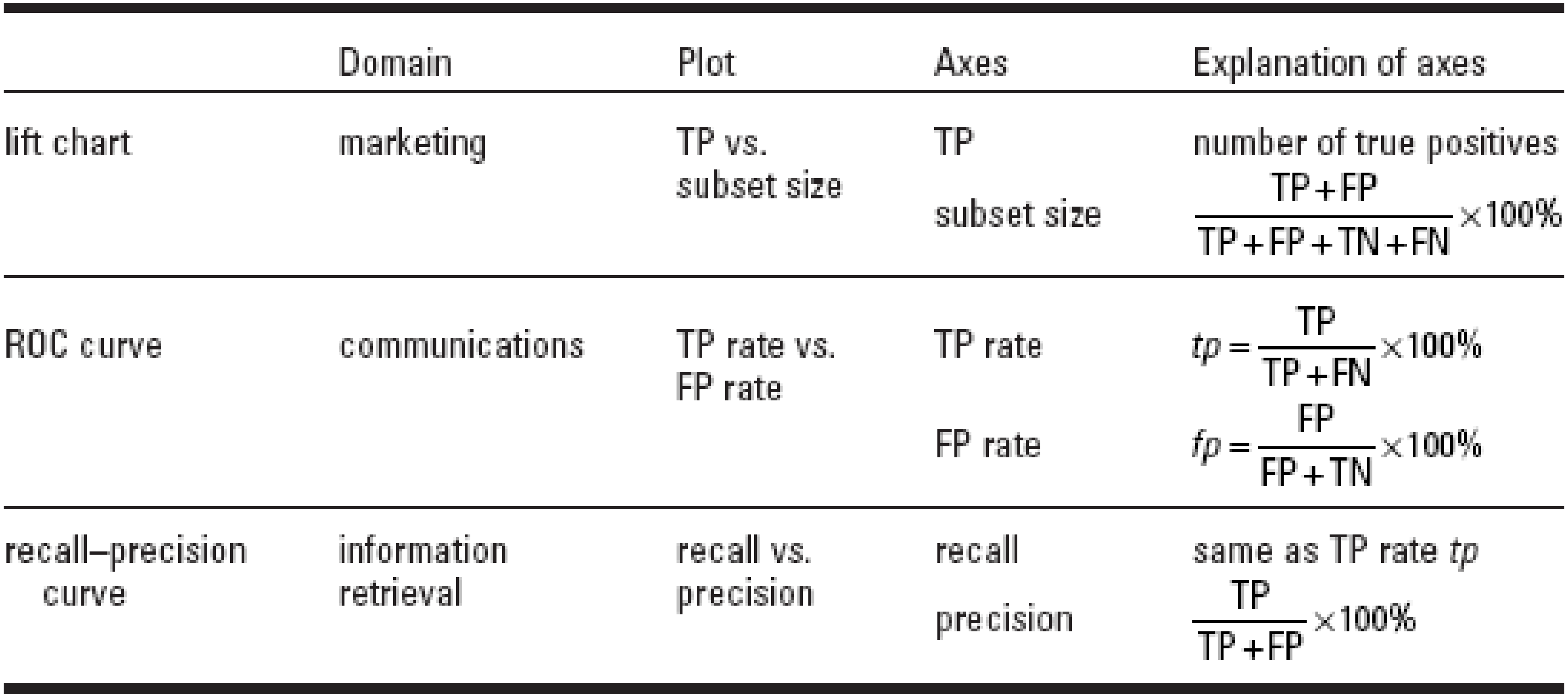
- Για κάθε υποβαλλόμενο ερώτημα, μία μηχανή διαδικτυακής αναζήτησης παράγει μία λίστα συνδέσεων υποτιθέμενα συναφών με το ερώτημα

$$\text{recall} = \frac{\text{number of documents retrieved that are relevant}}{\text{total number of documents that are relevant}}$$

$$\text{precision} = \frac{\text{number of documents retrieved that are relevant}}{\text{total number of documents that are retrieved}}.$$

  - *recall* =TP/(TP+FN)
  - *precision*=TP/(TP+FP)
- Σε προβλήματα ανάκτησης πληροφορίας, χρησιμοποιούνται καμπύλες ανάκλησης-ακρίβειας ακριβώς όπως οι καμπύλες ROC curves και τα σωρευτικά διαγράμματα

Σύνοψη κριτηρίων εξισορρόπησης μεταξύ FP & FN



	Domain	Plot	Axes	Explanation of axes
lift chart	marketing	TP vs. subset size	TP	number of true positives
			subset size	$\frac{TP+FP}{TP+FP+TN+FN} \times 100\%$
ROC curve	communications	TP rate vs. FP rate	TP rate	$tp = \frac{TP}{TP+FN} \times 100\%$
			FP rate	$fp = \frac{FP}{FP+TN} \times 100\%$
recall–precision curve	information retrieval	recall vs. precision	recall	same as TP rate *tp*
			precision	$\frac{TP}{TP+FP} \times 100\%$

Κριτήρια εξισορρόπησης (trade-off measures)

- *Μέση ακρίβεια 3-σημείων (3-point average recall)*
  - Μέση ακρίβεια για τιμές ανάκλησης 20%, 50% & 80%
- *Μέση ακρίβεια 11-σημείων*
  - 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% & 100%
- *Ευαισθησία & ιδιαιτερότητα (sensitivity & specificity)*
  - Ευαισθησία: *tp*
    - Για παράδειγμα, ποσοστό ατόμων με θετικό αποτέλεσμα εξέτασης που πράγματι φέρουν την ασθένεια
  - Ιδιαιτερότητα: 1 - *fp*
    - Για παράδειγμα, ποσοστό ατόμων που δε φέρουν την ασθένεια και έχουν αρνητικό αποτέλεσμα
  - Το γινόμενό τους χρησιμοποιείται ως γενικό μέτρο

    ευαισθησία x ιδιαιτερότητα $= tp \times (1-fp) = (TP \times TN) / (TP+FN) \times (FP+TN)$

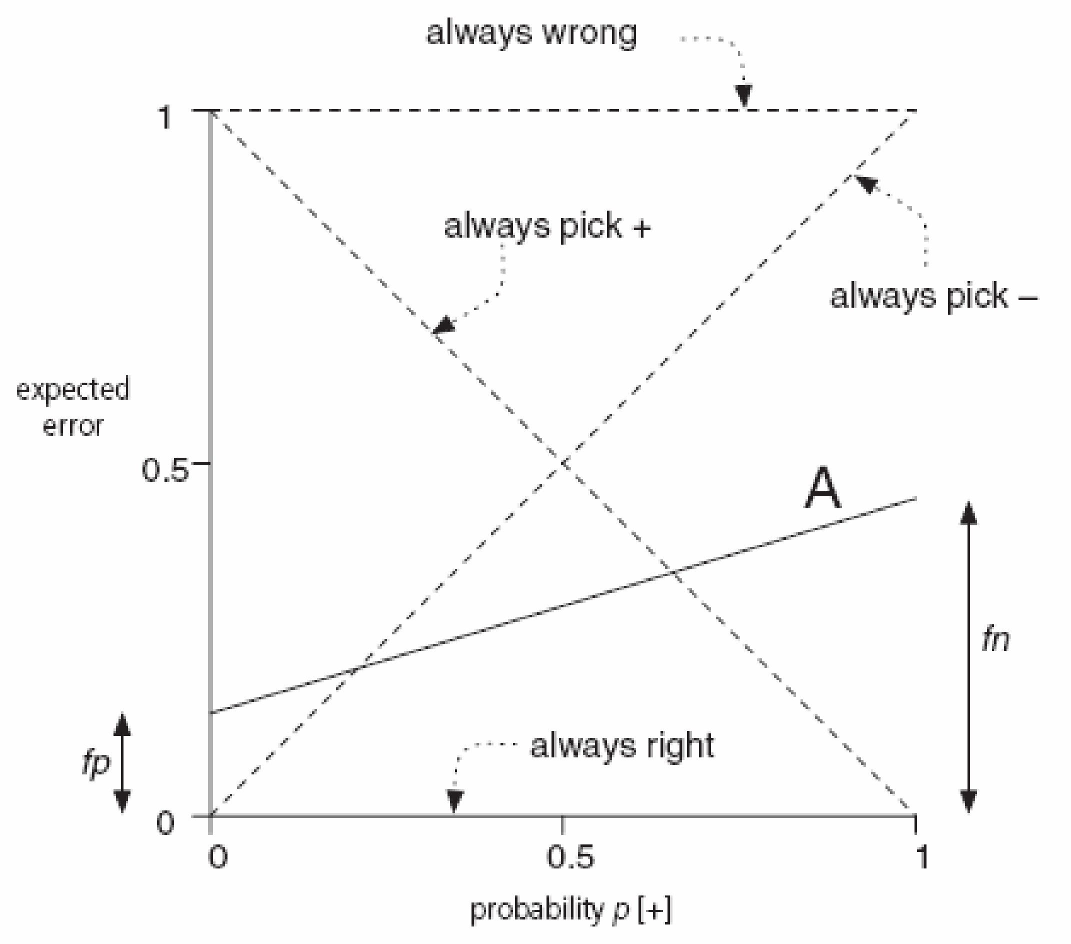
Κριτήρια εξισορρόπησης

- *Βαθμός επιτυχίας*: (TP+TN)/(TP+FP+TN+FN)
- *F-measure = (2* x *ανάκληση* x *ακρίβεια) / (ανάκληση + ακρίβεια) =*
  = 2TP / (2TP+FP+FN)
- Εμβαδόν περιοχής υπό της καμπύλης ROC
  - Όσο μεγαλύτερη η περιοχή, τόσο καλύτερο το μοντέλο
  - Ερμηνεία: η πιθανότητα το μοντέλο να αποδώσει σε ένα τυχαία επιλεγμένο θετικό παράδειγμα μεγαλύτερη τιμή προβλεπόμενης πιθανότητας από ότι σε ένα τυχαία επιλεγμένο αρνητικό
- Γενικά, κανένα μοναδιαίο μέτρο δε δύναται να απεικονίσει τον όγκο πληροφορίας ενός δισδιάστατου διαγράμματος

Καμπύλες κόστους

- Οι προηγούμενες καμπύλες είναι πολύ χρήσιμες για την εξερεύνηση των ισορροπιών μεταξύ διαφόρων ταξινομητών και του εύρους του συνολικού κόστους
  - Ωστόσο, δεν είναι κατάλληλες για την αποτίμηση των μοντέλων όταν τα κόστη των διάφορων σφαλμάτων είναι γνωστά
- Οι *καμπύλες κόστους (cost curves)* συνιστούν διαφορετική μορφή απεικόνισης
  - Μία ευθεία γραμμή περιγράφει τη μεταβολή της απόδοσης ενός μοντέλου καθώς αλλάζει η κατανομή της τάξης

Η καμπύλη σφάλματος



- Άξονας x : αναμενόμενο ποσοστό σφάλματος
- Άξονας y : πιθανότητα εμφάνισης τάξης [+]
- Επιλογή μοντέλου Α μόνο όταν

  $0.2 < p[+] < 0.65$
- Δεν λαμβάνει υπ’ όψιν το κόστος

Η καμπύλη κόστους

- $C[+|-]$: κόστος πρόβλεψης + όταν το παράδειγμα είναι στην πραγματικότητα –
- $C[-|+]$: το αντίστροφο
- Κανονικοποιημένο αναμενόμενο κόστος

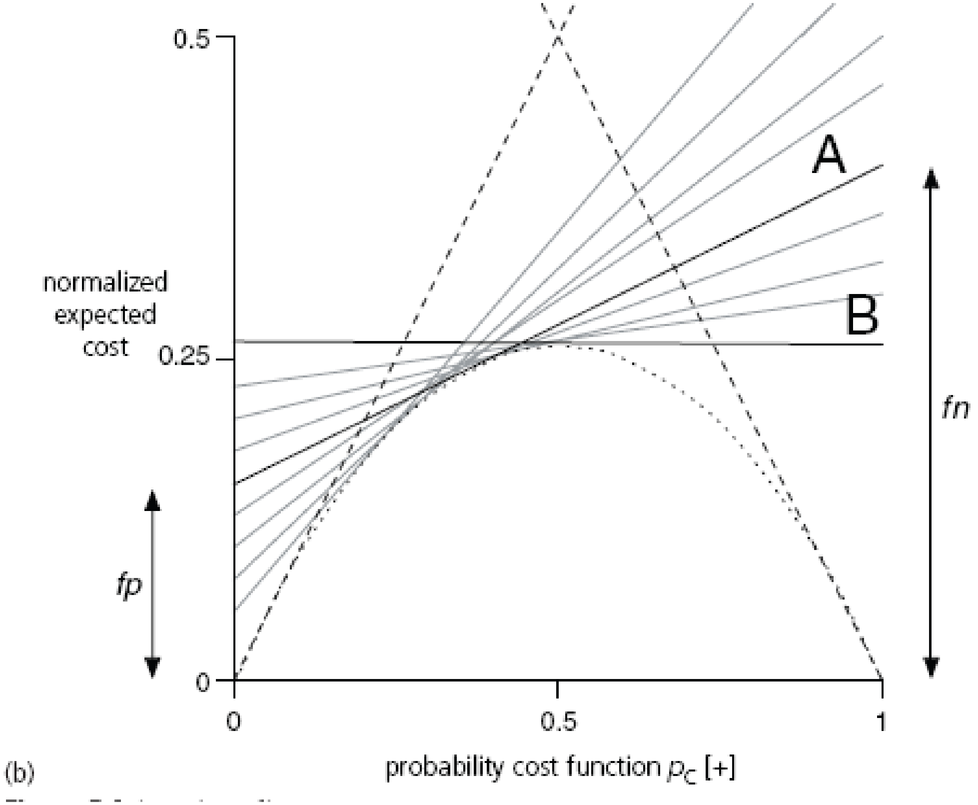
$$fn \times p_C[+] + fp \times (1 - p_C[+])$$

- Συνάρτηση πιθανότητας με βάση το κόστος

$$p_c[+] = \frac{p[+]C[+|-]}{p[+]C[+|-] + p[-]C[-|+]}$$

  - Θεωρήθηκε ότι $C[+|+] = C[-|-] = 0$

Η καμπύλη κόστους



- Άξονας x : κανονικοποιημένο αναμενόμενο κόστος
- Άξονας y : συνάρτηση πιθανότητας με βάση το κόστος
- Μοντέλο B:

  $C[+|-] = C[-|+]$
- Επιλέγεται αντί του A όταν $p_c[+] > 0.45$

Αποτίμηση αριθμητικής πρόβλεψης

- Τα μέτρα αποτίμησης που έχουν αναφερθεί μέχρι στιγμής αφορούν την ταξινόμηση
  - Τι συμβαίνει στην περίπτωση αριθμητικής πρόβλεψης;
- Ίδιες τεχνικές: ανεξάρτητο σύνολο ελέγχου, διασταυρωμένη επικύρωση, έλεγχοι σημαντικότητας, κτλ.
- Διαφορές: κριτήρια σφάλματος
  - Η τιμή σφάλματος (error rate) δεν μπορεί να εφαρμοστεί
  - Το σφάλμα δεν μπορεί να περιγραφεί από δυαδικό δεδομένο (υπάρχει / δεν υπάρχει) αλλά από αριθμητικό (διαφέρει ως προς το μέγεθος)

Κριτήρια σφάλματος

- Mean-Squared Error

$$\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}$$

- Root Mean Squared Error

$$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}$$

- Mean Absolute Error

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{n}$$

- Relative Squared Error

$$\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{(a_1 - \overline{a})^2 + \ldots + (a_n - \overline{a})^2}, \text{όπου } \overline{a} = \frac{1}{n}\sum_j a_j$$

- Root Relative Squared Error

$$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{(a_1 - \overline{a})^2 + \ldots + (a_n - \overline{a})^2}}$$

- Relative Absolute Error

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|a_1 - \overline{a}| + \ldots + |a_n - \overline{a}|}$$

- Correlation Coefficient

$$\frac{S_{PA}}{\sqrt{S_P S_A}}, \text{ όπου } S_{PA} = \frac{\sum_i (p_i - \overline{p})(a_i - \overline{a})}{n-1}$$

$$S_P = \frac{\sum_i (p_i - \overline{p})^2}{n-1}, \text{ και } S_A = \frac{\sum_i (a_i - \overline{a})^2}{n-1}$$

a : πραγματικές τιμές & p: πρόγνωση

Κριτήρια σφάλματος

- MSE: το πλέον δημοφιλές μέτρο σφάλματος
- RMSE: ίδιες διαστάσεις με την προβλεπόμενη τιμή
- MAE: λιγότερο ευαίσθητο σε τιμές προς εξαίρεση (outliers) από το MSE
- R..: τιμές σφάλματος σε σύγκριση με το σφάλμα που θα προέκυπτε αν κάθε πρόβλεψη ήταν ίση με τη μέση τιμή
- CC: αποτίμηση της συσχέτισης μεταξύ των πραγματικών τιμών *a* και των προβλέψεων *p*
  - Κυμαίνεται από 1 για απολύτων συσχετιζόμενα αποτελέσματα, μέχρι 0 όταν δεν υπάρχει καμία συσχέτιση, και -1 όταν τα αποτελέσματα έχουν πλήρως αρνητική συσχέτιση
  - Το μοναδικό κριτήριο με τιμές ανεξάρτητες της κλίμακας των δεδομένων
  - Επιθυμητή η μεγιστοποίησή του, για όλα τα άλλα κριτήρια επιθυμητή η ελαχιστοποίησή τους

Επιλογή κριτηρίου

- Προτείνεται η εξέταση όλων
- Στις περισσότερες των περιπτώσεων οι συστάσεις τους ταυτίζονται
- Παράδειγμα:

	A	**B**	**C**	**D**
Root mean-squared error	67.8	91.7	63.3	57.4
Mean absolute error	41.3	38.5	33.4	29.2
Root rel squared error	42.2%	57.2%	39.4%	35.8%
Relative absolute error	43.1%	40.1%	34.8%	30.4%
Correlation coefficient	0.88	0.88	0.89	0.91

- D: βέλτιστο μοντέλο
- C: δεύτερο καλύτερο
- A, B: αμφισβητήσιμη επιλογή

Αρχή Ελαχίστου Μήκους Περιγραφής (MDL Principle)

- Το μήκος περιγραφής ορίζεται ως:

απαιτούμενος χώρος για την περιγραφή της θεωρίας

\+

απαιτούμενος χώρος για την περιγραφή των λαθών της θεωρίας

- Στη συγκεκριμένη περίπτωση
  - Θεωρία: μοντέλο ταξινόμησης
  - Λάθη: σφάλματα ταξινόμησης στο σύνολο εκπαίδευσης
- Στόχος: αναζήτηση μοντέλου με ελάχιστο μήκος περιγραφής
- Η αρχή MDL συνιστά *κριτήριο απλότητας/κομψότητας* για την *επιλογή μοντέλου*

Κριτήρια επιλογής μοντέλου

- Αποπειρώνται να επιτύχουν θετικό συμβιβασμό μεταξύ:
  - Της πολυπλοκότητας του μοντέλου
  - Της ακρίβειας των προβλέψεών του στα δεδομένα εκπαίδευσης
- Συλλογιστική: **άρτιο θεωρείται το μοντέλο που αφενός είναι απλό και αφετέρου επιτυγχάνει υψηλή ακρίβεια στα διαθέσιμα δεδομένα**
  - Επίσης γνωστή και ως *Ξυράφι του Όκκαμ (Occam's Razor)*: **βέλτιστη είναι η μικρότερη θεωρία που περιγράφει όλες τις πληροφορίες**
- Κατά τον Αλβέρτο Αϊνστάιν: "**Everything should be made as simple as possible, but no simpler**."

Κομψότητα vs. σφάλμα

- Θεωρία Α: πολύ απλή, κομψή θεωρία που περιγράφει τα δεδομένα με σχεδόν απόλυτο τρόπο
- Θεωρία Β: σημαντικά περισσότερο πολύπλοκη θεωρία που περιγράφει τα δεδομένα χωρίς λάθη
- Η Θεωρία Α είναι προφανώς προτιμώμενη
  - Κλασσικό παράδειγμα: οι τρεις νόμοι του Κέπλερ για την κίνηση των πλανητών
    - Λιγότερο ακριβείς από την τελευταία βελτίωση του Κοπέρνικου στη θεωρία των επικύκλων του Πτολεμαίου

MDL και συμπίεση

- Η αρχή MDL σχετίζεται με τη συμπίεση των δεδομένων
  - Βέλτιστη είναι η θεωρία που επιτυγχάνει τη μέγιστη συμπίεση των δεδομένων
  - Για παράδειγμα, για να συμπιέσουμε ένα σύνολο δεδομένων δημιουργούμε ένα μοντέλο και μετά αποθηκεύουμε αυτό και τα λάθη του
- Απαιτείται ο υπολογισμός:
(α) του μεγέθους του μοντέλου και
(β) του απαιτούμενου χώρου για την κωδικοποίηση των σφαλμάτων
- (β) εφικτός με χρήση της συνάρτησης απώλειας πληροφορίας
- (α) χρειάζεται μέθοδος κωδικοποίησης του μοντέλου

MDL και θεώρημα Bayes

- L[T]=μήκος (length) της θεωρίας
- L[E|T]=κωδικοποίηση του συνόλου εκπαίδευσης E δεδομένης της θεωρίας
- Μήκος περιγραφής= L[T] + L[E|T]
- Το θεώρημα του Bayes δίνει κατ' επαγωγή (*a posteriori)* την πιθανότητα της θεωρίας με γνωστό το σύνολο των δεδομένων:

$$\Pr[T \mid E] = \frac{\Pr[E \mid T]\Pr[T]}{\Pr[E]}$$

- Ισοδυναμεί με: (η αρχή MDL!!)

$$-\log \Pr[T \mid E] = -\log \Pr[E \mid T] - \log \Pr[T] + \underbrace{\log \Pr[E]}_{\boldsymbol{\sigma\tau\alpha\theta\varepsilon\rho\acute{o}}}$$

MDL & MAP

- MAP: *μέγιστη κατ' επαγωγή πιθανότητα (maximum a posteriori probability)*
- Η εύρεση της βέλτιστης (περισσότερο πιθανής) θεωρίας με MAP είναι σε συμφωνία με την εύρεση της βέλτιστης (λιτής) θεωρίας με MDL
- Δύσκολο τμήμα στην εφαρμογή της αρχής MAP:
  - καθορισμός της πρότερης πιθανότητας Pr[T] της θεωρίας
- Αντιστοιχεί στο δύσκολο κομμάτι εφαρμογής της αρχής MDL: κωδικοποίηση σχήματος της θεωρίας
  - Για παράδειγμα, εάν γνωρίζουμε εκ των προτέρων ότι μία θεωρία είναι περισσότερο πιθανή, χρειαζόμαστε λιγότερα bits για την κωδικοποίησή της

Περί της αρχής MDL

- Πλεονέκτημα
  - πλήρης χρήση των δεδομένων εκπαίδευσης για την επιλογή μοντέλου
- Μειονεκτήματα
  - κρίσιμη σημασία του κατάλληλου σχήματος κωδικοποίησης / πρότερων πιθανοτήτων
  - δεν εξασφαλίζεται ότι η θεωρία με MDL είναι εκείνη που ελαχιστοποιεί το εκτιμώμενο σφάλμα
- Σημείωση: το ξυράφι του Όκκαμ είναι ένα αξίωμα!
  - Άλλα αξιώματα είναι επίσης διαθέσιμα, για παράδειγμα η *αρχή των πολλαπλών εξηγήσεων* του Επίκουρου: «διατήρησε το σύνολο των συμβατών με τα δεδομένα θεωριών»

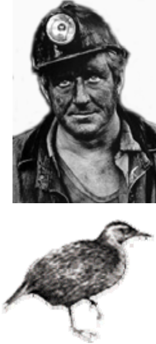
Εύρεση μέσου μοντέλου κατά Bayes

- Εκφράζει την αρχή του Επίκουρου: όλες οι θεωρίες χρησιμοποιούνται για την πρόβλεψη, με βαρύτητα ανάλογη της P[T|E]
- Έστω I ένα νέο παράδειγμα του οποίου ζητείται η τάξη
- Έστω C η τυχαία μεταβλητή που υποδηλώνει την τάξη αυτή
- Τότε η μέθοδος δίνει την πιθανότητα της C με βάση
  - I
  - Τα δεδομένα εκπαίδευσης E
  - Τις πιθανές θεωρίες T_j

$$\Pr[C \mid I, E] = \sum_j \Pr[C \mid I, T_j] \Pr[T_j \mid E]$$

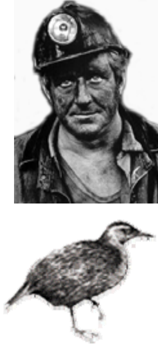
MDL και ομαδοποίηση

- Ομαδοποίηση: εκ φύσεως δύσκολα αποτιμώμενη
  - Δεν υπάρχει αντικειμενικό κριτήριο επιτυχίας
- Μήκος περιγραφής της θεωρίας:
  απαιτούμενα bits για την κωδικοποίηση των ομάδων
  - Για παράδειγμα, κέντρα ομάδων
- Μήκος περιγραφής των δεδομένων με γνωστή τη θεωρία:
  για κάθε παράδειγμα, κωδικοποίηση της ομάδας στην οποία ανήκει και της σχετικής του θέσης
  - Για παράδειγμα, απόσταση από το κέντρο ομάδας
- Αν τα δεδομένα παρουσιάζουν ισχυρή ομαδοποίηση, το μήκος περιγραφής θα είναι αισθητά μικρότερο από την απλή μετάδοση των δεδομένων
  - Ωστόσο, σε αντίθετη περίπτωση, το μήκος μάλλον θα αυξηθεί



Εφαρμογή στο weka

- Test options / Cluster mode / Attribute selection mode
  - *Use training set*
  - *Supplied test set*
  - *Cross-validation*
  - *Percentage split*
  - *Cost-sensitive evaluation*



Εφαρμογή στο weka

- Classifiers / meta
  - *CostSensitiveClassifier*: Make its base classifier cost sensitive
  - *CVParameterSelection*: Perform parameter selection by cross-validation
  - *MetaCost*: Make a classifier cost-sensitive
  - *MultiScheme*: Use cross-validation to select a classifier from several candidates
  - *ThresholdSelector*: Optimize the F-measure for a probabilistic classifier
  - *Vote*: Combine classifiers using average of probability estimates
  - or numeric predictions
  - onDemandDirectory

Τέλος

Επόμενη διάλεξη:

Αλγόριθμοι εκμάθησης