ΕΜΠ ΔΠΜΣ

Εφαρμοσμένες Μαθηματικές Επιστήμες

Αλγόριθμοι Εξόρυξης Πληροφορίας

Διάλεξη 06:

Αλγόριθμοι εκμάθησης ΜέροςΒ

Bayes, ΚανόνεςΣυσχέτισης, ΑδρανήςΕκμάθηση & Ομαδοποίηση

Αλγόριθμοι

- Η παρούσα διάλεξη αποτελεί το δεύτερο μέρος της εστίασης στον πυρήνα της διαδικασίας εξόρυξης πληροφορίας
- Θα μελετηθούν η στατιστική μοντελοποίηση, οι κανόνες συσχέτισης, η μάθηση με βάση υποδείγματα και η ομαδοποίηση

Στατιστική μοντελοποίηση

- Αντιδιαμετρική οπτική του 1R: χρήση όλων των χαρακτηριστικών
- Υποθέσεις: τα χαρακτηριστικά είναι
  - *Εξ' ίσου σημαντικά*
  - *Στατιστικά ανεξάρτητα* (given the class value)
    - Δεδομένης της τάξεως, η τιμή ενός χαρακτηριστικού δεν φανερώνει οποιαδήποτε πληροφορία για την τιμή ενός άλλου χαρακτηριστικού
- Η υπόθεση ανεξαρτησίας δεν επαληθεύεται σε καμία περίπτωση!
- Ωστόσο… οδηγεί σε ένα απλό σχήμα εκμάθησης που και πάλι λειτουργεί –πέραν των προσδοκιών– ικανοποιητικά στην πράξη

Δεδομένα καιρού Πιθανότητες

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Ένα νέο παράδειγμα

Outlook	**sunny**
Temp.	**cool**
Humidity	**high**
Windy	**true**
Play	**?**

- Πιθανότητα εμφάνισης κάθε τάξης
  - "yes" = $2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$
  - "no" = $3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0206$
- Μετατροπή σε % πιθανότητα μέσω κανονικοποίησης:
  - P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205
  - P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795

Ο κανόνας του Bayes
(περί πιθανότητας υπό συνθήκη)

- Πιθανότητα του γεγονότος H δεδομένου του E :

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

- *Εκ των προτέρων / A priori* πιθανότητα του H : $\Pr[H]$
  - Πιθανότητα του γεγονότος *προτού* γίνει γνωστό το δεδομένο
- *Εκ των υστέρων / A posteriori* πιθανότητα του H : $\Pr[H \mid E]$
  - Πιθανότητα του γεγονότος *αφότου* γίνει γνωστό το δεδομένο

Απλοϊκός (naïve) Bayes για ταξινόμηση

- Μάθηση ταξινόμησης: ποια η πιθανότητα της τάξης, δεδομένου του υποδείγματος;
  - Δεδομένο E = υπόδειγμα
  - Γεγονός H = τιμή τάξης για το υπόδειγμα
- Απλοϊκή υπόθεση: το δεδομένο διαχωρίζεται σε τμήματα (βλέπε χαρακτηριστικά) τα οποία είναι *ανεξάρτητα*

$$\Pr[H \mid E] = \frac{\Pr[E_1 \mid H]\Pr[E_2 \mid H]\ldots\Pr[E_n \mid H]\Pr[H]}{\Pr[E]}$$

Δεδομένα καιρού Παράδειγμα

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← ***Δεδομένο E***

$$\Pr[yes \mid E] = \Pr[Outlook = Sunny \mid yes]$$
$$\times \Pr[Temperature = Cool \mid yes]$$
$$\times \Pr[Humidity = High \mid yes]$$
$$\times \Pr[Windy = True \mid yes]$$
$$\times \frac{\Pr[yes]}{\Pr[E]}$$
$$= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{\Pr[E]}$$

Πιθανότητα του γεγονότος "yes"

Ο παρανομαστής απαλείφεται κατά την κανονικοποίηση

Το πρόβλημα μηδενικής συχνότητας

- Στην περίπτωση που ένα χαρακτηριστικό δεν προκύπτει για μία τιμή της τάξης; (για παράδειγμα "Humidity = high" για την τάξη "yes")
  - Η πιθανότητα μηδενίζεται $\Pr[Humidity = High \mid yes] = 0$
  - Το ίδιο και η *εκ των υστέρων* πιθανότητα $\Pr[yes \mid E] = 0$
- Θεραπεία: προσθήκη μονάδας στην απαρίθμηση κάθε συνδυασμού τιμής χαρακτηριστικού-τάξης (*εκτιμητής Laplace)*
- Αποτέλεσμα: έκλειψη φαινομένου μηδενικής πιθανότητας
  - Παράπλευρη συνέπεια: σταθεροποίηση εκτιμήσεων πιθανότητας

Τροποποιημένες εκτιμήσεις πιθανότητας

- Σε κάποιες περιπτώσεις, η πρόσθεση σταθεράς διαφορετικής της μονάδας εμφανίζει μεγαλύτερη καταλληλότητα
- Παράδειγμα: χαρακτηριστικό *outlook* για την τάξη *yes*

$$\frac{2+\mu/3}{9+\mu} \qquad \frac{4+\mu/3}{9+\mu} \qquad \frac{3+\mu/3}{9+\mu}$$

Sunny *Overcast* *Rainy*

$$\frac{2+\mu p_1}{9+\mu} \qquad \frac{4+\mu p_2}{9+\mu} \qquad \frac{3+\mu p_3}{9+\mu}$$

- Δεν είναι αναγκαία η ισότητα των βαρών
  - ωστόσο το άθροισμά τους πρέπει να είναι ίσο της μονάδας

Άγνωστες τιμές

- Εκπαίδευση: το υπόδειγμα δεν περιλαμβάνεται στην καταμέτρηση συχνότητας των συνδυασμών τιμής χαρακτηριστικού-τάξης
- Ταξινόμηση: το χαρακτηριστικό παραλείπεται κατά τον υπολογισμό
- Παράδειγμα:

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

Πιθανότητα "yes" = 3/9 × 3/9 × 3/9 × 9/14 = 0.0238

Πιθανότητα "no" = 1/5 × 4/5 × 3/5 × 5/14 = 0.0343

P("yes") = 0.0238 / (0.0238 + 0.0343) = 41%

P("no") = 0.0343 / (0.0238 + 0.0343) = 59%

Αριθμητικά χαρακτηριστικά

- Συνήθης υπόθεση: τα χαρακτηριστικά ακολουθούν *κανονική* κατανομή (δεδομένης της τάξης)
- Η *συνάρτηση πυκνότητας πιθανότητας* της κανονικής κατανομής καθορίζεται από δύο παραμέτρους:
  - *Μέσο* μ του δείγματος

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - *Τυπική απόκλιση* σ

$$\sigma = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2$$

Η συνάρτηση πυκνότητας *f(x)* τότε είναι

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Στατιστικά δεδομένα καιρού

Outlook			Temperature		Humidity		Windy			Play	
	Yes	*No*	*Yes*	*No*	*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	64, 68,	65, 71,	65, 70,	70, 85,	False	6	2	9	5
Overcast	4	0	69, 70,	72, 80,	70, 75,	90, 91,	True	3	3		
Rainy	3	2	72, …	85, …	80, …	95, …					
Sunny	2/9	3/5	μ =73	μ =75	μ =79	μ =86	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	σ =6.2	σ =7.9	σ =10.2	σ =9.7	True	3/9	3/5		
Rainy	3/9	2/5									

- Παράδειγμα τιμής πυκνότητας:

$$f(temperature = 66 \mid yes) = \frac{1}{\sqrt{2\pi}\,6.2} e^{-\frac{(66-73)^2}{2*6.2^2}} = 0.0340$$

Ταξινόμηση νέου παραδείγματος

- Νέο παράδειγμα:

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Πιθανότητα "yes" = 2/9 × 0.0340 × 0.0221 × 3/9 × 9/14 = 0.000036

Πιθανότητα "no" = 3/5 × 0.0291 × 0.0380 × 3/5 × 5/14 = 0.000136

P("yes") = 0.000036 / (0.000036 + 0. 000136) = 20.9%

P("no") = 0.000136 / (0.000036 + 0. 000136) = 79.1%

- Οι άγνωστες τιμές δεν λαμβάνονται υπ' όψιν κατά τη διάρκεια της εκπαίδευσης για τον υπολογισμό του μέσου και της τυπικής απόκλισης

Γενικά περί απλοϊκού Bayes

- Αποδίδει ικανοποιητικά, πέραν του αναμενόμενου (ακόμα και αν η υπόθεση περί ανεξαρτησίας παραβιάζεται ξεκάθαρα)
  - Πάγια τακτική: αρχικός πειραματισμός με απλοϊκά σχήματα
- Αίτιο; Η ταξινόμηση δεν απαιτεί ακριβείς εκτιμήσεις πιθανότητας *με την προϋπόθεση ότι η μέγιστη πιθανότητα εκχωρείται στη σωστή τάξη*
- Ωστόσο: η προσθήκη μεγάλου αριθμού περιττών χαρακτηριστικών προκαλεί προβλήματα (για παράδειγμα πανομοιότυπα χαρακτηριστικά)
- Επίσης: συχνά τα αριθμητικά χαρακτηριστικά δεν ακολουθούν κανονική κατανομή (→ *εκτιμητές πυκνότητας πυρήνα (kernel density estimators) ή διακριτοποίηση*)

@weka

- *AODE* Averaged, one-dependence estimators
- *BayesNet* Learn Bayesian nets
  - Για περισσότερα http://weka.sourceforge.net/manuals/weka.bn.pdf
- *ComplementNaiveBayes* Build a Complement Naïve Bayes classifier
- *NaiveBayes* Standard probabilistic Naïve Bayes classifier
- *NaiveBayesMultinomial* Multinomial version of Naïve Bayes
- *NaiveBayesSimple* Simple implementation of Naïve Bayes
- *NaiveBayesUpdateable* Incremental Naïve Bayes classifier that learns one instance at a time

Κανόνες συσχέτισης (association rules)

- Οι κανόνες συσχέτισης
  - μπορούν να προβλέψουν οποιοδήποτε χαρακτηριστικό ή συνδυασμό χαρακτηριστικών και όχι μόνο την τάξη
  - δεν χρησιμοποιούνται υποχρεωτικά ως σύνολο
- Πρόβλημα: αχανής αριθμός πιθανών συσχετίσεων
  - Απαιτείται η θέσπιση περιορισμών ώστε να αναδειχθούν μόνο οι σημαντικότερες των προβλεπτικών συσχετίσεων

    $\Rightarrow$ μόνο εκείνες με υψηλή *υποστήριξη (support)* και υψηλή *εμπιστοσύνη (confidence)*

Υποστήριξη & εμπιστοσύνη κανόνα

- Υποστήριξη (support): αριθμός υποδειγμάτων επιτυχούς πρόβλεψης
- Εμπιστοσύνη (confidence): αριθμός επιτυχών προβλέψεων, ως ποσοστό του συνόλου των υποδειγμάτων στα οποία εφαρμόζεται ο κανόνας
- Παράδειγμα: 4 ημέρες με temperature 'cool' και humidity 'normal'

  `If temperature = cool then humidity = normal`

  ⇒Support = 4, confidence = 100%
- Συνήθως προκαθορίζονται ελάχιστες τιμές υποστήριξης και εμπιστοσύνης
  - για παράδειγμα 58 κανόνες με υποστήριξη≥ 2 και εμπιστοσύνη≥ 95% για τα δεδομένα καιρού

Εξόρυξη κανόνων συσχέτισης

- Απλοϊκή μέθοδος εύρεσης κανόνων συσχέτισης:
  - Χρήση μεθόδου διαίρει & βασίλευε (separate-and-conquer)
  - Χειρισμός κάθε δυνατού συνδυασμού τιμών των χαρακτηριστικών ως ξεχωριστή τάξη
- Προβλήματα:
  - Υπολογιστική πολυπλοκότητα
  - Προκύπτει ογκώδες σύνολο κανόνων (απαιτείται ο περιορισμός του με βάση τα μεγέθη της εμπιστοσύνης και της υποστήριξης)
- Ωστόσο: η απευθείας αναζήτηση κανόνων με υψηλή υποστήριξη είναι εφικτή!

Σύνολα στοιχείων (item sets)

- Υποστήριξη (support): αριθμός υποδειγμάτων που καλύπτονται επιτυχώς από τον κανόνα συσχέτισης
  - Όμοιος του αριθμού υποδειγμάτων που καλύπτονται από *το σύνολο* των ελέγχων ενός κανόνα
- *Στοιχείο (item)*: ένα ζεύγος ελέγχου / χαρακτηριστικού-τιμής one test/attribute-value pair
- *Σύνολο στοιχείων (item set)*: το σύνολο των items που προκύπτουν σε ένα κανόνα
- Στόχος: το σύνολο των κανόνων που υπερβαίνουν προκαθορισμένη υποστήριξη

  $\Rightarrow$ Υλοποίηση: εύρεση όλων των συνόλων στοιχείων δεδομένης ελάχιστης υποστήριξης και παραγωγή κανόνων από αυτά

Σύνολα στοιχείων για τα δεδομένα καιρού

One-item sets	Two-item sets	Three-item sets	Four-item sets
Outlook = Sunny (5)	Outlook = Sunny Temperature = Hot (2)	Outlook = Sunny Temperature = Hot Humidity = High (2)	Outlook = Sunny Temperature = Hot Humidity = High Play = No (2)
Temperature = Cool (4)	Outlook = Sunny Humidity = High (3)	Outlook = Sunny Humidity = High Windy = False (2)	Outlook = Rainy Temperature = Mild Windy = False Play = Yes (2)
…	…	…	…

- Τελικά: 12 σύνολα ενός στοιχείου, 47 σύνολα δύο στοιχείων, 39 σύνολα τριών στοιχείων, 6 σύνολα τεσσάρων στοιχείων και 0 σύνολα πέντε στοιχείων (με ελάχιστη υποστήριξη ίση με δύο)

Παραγωγή κανόνων από σύνολο στοιχείων

- Καθώς έχουν παραχθεί όλα τα σύνολα στοιχείων δεδομένης ελάχιστης υποστήριξης, η μετατροπή τους σε κανόνες είναι πλέον εφικτή
- Παράδειγμα συνόλου στοιχείων:

  Humidity = Normal, Windy = False, Play = Yes (4)

- Επτά (2^N-1) πιθανοί κανόνες:

If *Humidity = Normal* **and** *Windy = False* **then** *Play = Yes* **4/4**

If *Humidity = Normal* **and** *Play = Yes* **then** *Windy = False* **4/6**

If *Windy = False* **and** *Play = Yes* **then** *Humidity = Normal* **4/6**

If *Humidity = Normal* **then** *Windy = False* **and** *Play = Yes* **4/7**

If *Windy = False* **then** *Humidity = Normal* **and** *Play = Yes* **4/8**

If *Play = Yes* **then** *Humidity = Normal* **and** *Windy = False* **4/9**

If - **then** *Humidity = Normal* **and** *Windy = False*
and *Play = Yes* **4/12**

Κανόνες για τα δεδομένα καιρού

- Κανόνες με υποστήριξη > 1 και εμπιστοσύνη = 100%:

	Association rule		Sup.	Conf.
1	Humidity=Normal Windy=False	⇒ Play=Yes	4	100%
2	Temp=Cool	⇒ Humidity=Normal	4	100%
3	Outlook=Overcast	⇒ Play=Yes	4	100%
4	Temp=Cold Play=Yes	⇒ Humidity=Normal	3	100%
	...	...	...	...
58	Outlook=Sunny Temp=Hot	⇒ Humidity=High	2	100%

- Συνολικά:
  - 3 κανόνες με υποστήριξη 4
  - 5 με υποστήριξη 3
  - 50 με υποστήριξη 2

Παράδειγμα κανόνων από το ίδιο σύνολο

- Σύνολο στοιχείων:

Temperature = Cool, Humidity = Normal, Windy = False, Play = Yes (2)

- Κανόνες που προκύπτουν (100% εμπιστοσύνη):

Temperature = Cool, Windy = False ⇒ Humidity = Normal, Play = Yes
Temperature = Cool, Windy = False, Humidity = Normal ⇒ Play = Yes
Temperature = Cool, Windy = False, Play = Yes ⇒ Humidity = Normal

με βάση τα ακόλουθα "συχνά" σύνολα στοιχείων:

Temperature = Cool, Windy = False (2)
Temperature = Cool, Humidity = Normal, Windy = False (2)
Temperature = Cool, Windy = False, Play = Yes (2)

Αποτελεσματική δημιουργία συνόλου στοιχείων

- Αναζητείται αποτελεσματικός τρόπος εύρεσης όλων των συνόλων στοιχείων μεγάλης συχνότητας
- Η εύρεση των συνόλων ενός στοιχείου είναι εύκολη
- Ιδέα: χρήση συνόλων ενός στοιχείου για την παραγωγή συνόλων δύο στοιχείων, συνόλων δύο στοιχείων για την παραγωγή συνόλων τριών στοιχείων, …
  - Αν (A B) είναι σύνολο στοιχείων υψηλής συχνότητας, τότε τα (A) και (B) πρέπει να είναι επίσης συχνά σύνολα στοιχείων!
  - Γενίκευση: αν X ένα σύνολο k-στοιχείων υψηλής συχνότητας, τότε τα υποσύνολα (k-1)-στοιχείων του X είναι επίσης υψηλής συχνότητας

  $\Rightarrow$ Υπολογισμός συνόλου k-στοιχείων με συγχώνευση των συνόλων (k-1)-στοιχείων

Παράδειγμα

- Δεδομένα: 5 σύνολα 3-στοιχείων

  (A B C), (A B D), (A C D), (A C E), (B C D)

- Σε 'αλφαβητική' σειρά!
- Υποψήφια σύνολα 4 στοιχείων:

  (A B C D) OK λόγω του (B C D)

  (A C D E) Not OK λόγω του (C D E)

- Τελικός έλεγχος με καταμέτρηση υποδειγμάτων στο σύνολο δεδομένων
- Τα σύνολα (k –1)-στοιχείων αποθηκεύονται σε *πίνακα κατακερματισμού (hash table)*

Αποδοτική δημιουργία κανόνων

- Αναζήτηση του συνόλου των κανόνων με υψηλή εμπιστοσύνη
  - Η υποστήριξη των προγόνων λαμβάνεται από τον πίνακα κατακερματισμού
  - Η εξαντλητική αναζήτηση υποδεικνύει 2^N-1 κανόνες
- Βελτιωμένη έκδοση: κατασκευή κανόνα με (c + 1)-επακόλουθα από κανόνες με c-επακόλουθα
  - Παρατήρηση: ο κανόνας με (c + 1)-επακόλουθα έχει ισχύ μόνο εάν το σύνολο των κανόνων με c-επακόλουθα ισχύουν επίσης
- Προκύπτει αλγόριθμος όμοιος εκείνου περί κατασκευής εκτενούς συνόλου στοιχείων

Παράδειγμα

- Κανόνες με 1-επακόλουθο:

```
If Outlook = Sunny and Windy = False and Play = No
   then Humidity = High (2/2)

If Humidity = High and Windy = False and Play = No
   then Outlook = Sunny (2/2)
```

- Αντίστοιχος κανόνας 2-επακόλουθων:

```
If Windy = False and Play = No
   then Outlook = Sunny and Humidity = High (2/2)
```

- Τελικός έλεγχος συμβατότητας προγόνων με τον πίνακα κατακερματισμού

Γενικά περί κανόνων συσχέτισης

- Η προηγούμενη μέθοδος υλοποιεί μία σάρωση των δεδομένων για κάθε διαφορετικό μέγεθος συνόλου στοιχείων
  - Εναλλακτικά: δημιουργία συνόλων (k+2)-στοιχείων αμέσως μετά από τη δημιουργία συνόλων (k+1)-στοιχείων
  - Προκύπτουν περισσότερα των αναγκαίων σύνολα (k+2)-στοιχείων, αλλά λιγότερες σαρώσεις των δεδομένων
  - Χρήσιμη μέθοδος στην περίπτωση που ο όγκος των δεδομένων είναι μεγάλος για τη διαθέσιμη μνήμη
- Πρακτικό ζήτημα: παραγωγή συγκεκριμένου αριθμού κανόνων (για παράδειγμα μέσω βηματικής μείωσης της ελάχιστης επιτρεπόμενης υποστήριξης)

Άλλα ζητήματα

- Το τυπικό format .arff είναι μη αποδοτικό για *δεδομένα καλαθιού αγοράς* (*market basket data)*
  - Τα χαρακτηριστικά απεικονίζουν τα αντικείμενα ενός καλαθιού και συνήθως τα περισσότερα προϊόντα απουσιάζουν
  - Απαιτείται τρόπος απεικόνισης αραιών (sparse) δεδομένων
- Τα υποδείγματα καλούνται επίσης *συναλλαγές (transactions)*
- Η εμπιστοσύνη δεν είναι απαραίτητα το βέλτιστο μέτρο
  - Παράδειγμα: το προϊόν γάλα προκύπτει σε κάθε σχεδόν συναλλαγή ενός supermarket
  - Άλλα μέτρα έχουν επινοηθεί, όπως για παράδειγμα 'lift'

@weka

- *Apriori* Find association rules using the Apriori algorithm
- *PredictiveApriori* Find association rules sorted by predictive accuracy
- *Tertius* Confirmation-guided discovery of association or classification rules

Απεικόνιση με βάση υποδείγματα (instance-based representation)

- Απλούστερη μορφή μάθησης: αποστήθιση (*rote learning*)
  - Τα υποδείγματα εκπαίδευσης ερευνώνται για την εύρεση του περισσότερο όμοιου προς το νέο παράδειγμα υποδείγματος
  - Τα υποδείγματα αυτά καθαυτά αναπαριστούν τη γνώση
  - Η μάθηση καλείται επίσης ως *βασισμένη στα υποδείγματα*
- Η συνάρτηση ομοιότητας καθορίζει το αποτέλεσμα της μαθησιακής διαδικασίας
- Η βασισμένη στα υποδείγματα μάθηση είναι *αδρανής* (*lazy)*
- Μέθοδοι:
  - *Πλησιέστερος γείτονας (nearest-neighbor, NN)*
  - *k-πλησιέστεροι γείτονες (k-nearest-neighbor, k-NN)*
  - …

Συνάρτηση απόστασης

- Απλούστερη περίπτωση: ένα αριθμητικό χαρακτηριστικό
  - Η απόσταση ορίζεται ως η διαφορά των δύο τιμών του χαρακτηριστικού (ή μία συνάρτηση αυτών)
- Περισσότερα αριθμητικά χαρακτηριστικά: συνήθως χρησιμοποιείται η Ευκλείδεια απόσταση, όπου οι τιμές των χαρακτηριστικών έχουν κανονικοποιηθεί
- Ονομαστικά χαρακτηριστικά: απόσταση = 1 για διαφορετικές τιμές, απόσταση = 0 για ίδιες τιμές του χαρακτηριστικού
- Είναι όλα τα χαρακτηριστικά εξ ίσου σημαντικά;
  - Η εκχώρηση βαρύτητας στα χαρακτηριστικά είναι ίσως αναγκαία

Μάθηση βασισμένη στα υποδείγματα

- Η συνάρτηση απόσταση καθορίζει το αποτέλεσμα της μάθησης
- Τα περισσότερα βασισμένα σε υποδείγματα σχήματα χρησιμοποιούν την *Ευκλείδεια απόσταση*:

$$\sqrt{(a_1^{(1)} - a_1^{(2)})^2 + (a_2^{(1)} - a_2^{(2)})^2 + ... + (a_k^{(1)} - a_k^{(2)})^2}$$

  $\mathbf{a}^{(1)}$ και $\mathbf{a}^{(2)}$: δύο υποδείγματα με k χαρακτηριστικά

- Η τετραγωνική ρίζα δεν είναι αναγκαία όταν υλοποιείται σύγκριση αποστάσεων
- Άλλο σύνηθες μέτρο: *μέτρο οικοδομικού τετραγώνου (city-block metric)*
  - Πρόσθεση απόλυτων διαφορών και όχι των τετραγώνων τους

Κανονικοποίηση και άλλα ζητήματα

- Τα διάφορα χαρακτηριστικά μετρώνται σε διαφορετικές κλίμακες $\Rightarrow$ απαιτείται *κανονικοποίηση* αυτών:

$$a_i = \frac{v_i - \min v_i}{\max v_i - \min v_i}$$

v_i: η πραγματική τιμή του χαρακτηριστικού i

- Ονομαστικά χαρακτηριστικά: απόσταση ίση με 0 ή 1
- Συνήθης πολιτική για απούσες τιμές: θεώρηση μέγιστης απόστασης (για κανονικοποιημένα χαρακτηριστικά)

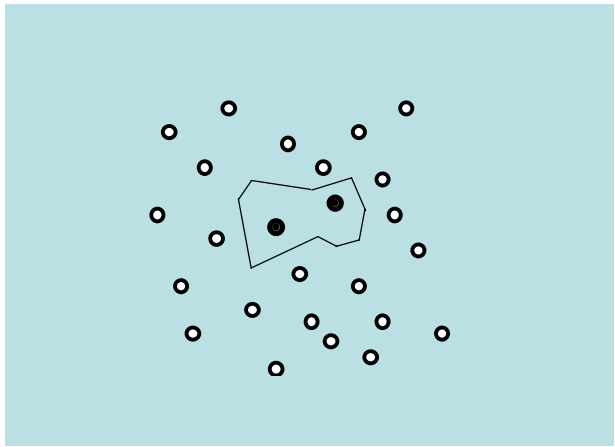
Γενικά περί αλγορίθμου 1-NN

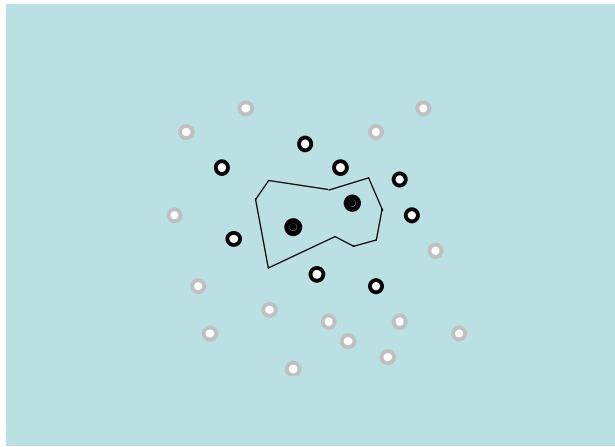
- Συχνά πολύ ακριβής αλγόριθμος
- Ωστόσο αργός…
  - Η απλοϊκή έκδοση σαρώνει όλα τα δεδομένα εκπαίδευσης για την εκπόνηση μιας πρόβλεψης
- Υποθέτει ισοδύναμη σημαντικότητα όλων των χαρακτηριστικών
  - Θεραπεία: επιλογή χαρακτηριστικών είτε εκχώρηση βαρύτητας
- Πρόβλημα υποδειγμάτων με υψηλό θόρυβο:
  - Εκχώρηση ψήφου πλειοψηφίας των k πλησιέστερων γειτόνων
  - Απομάκρυνση των υποδειγμάτων αυτών από το σύνολο των δεδομένων (δύσκολη!)
- Η μέθοδος χρησιμοποιείται στη στατιστική ήδη από το 1950
  - Αν $n \to \infty$ και $k/n \to 0$, το σφάλμα προσεγγίζει την ελάχιστη τιμή του
    - n ο αριθμός των υποδειγμάτων

Μάθηση βασισμένη στα υποδείγματα

- Πρακτικά προβλήματα του σχήματος 1-NN:
  - Αργό (ωστόσο υπάρχουν διαθέσιμες ταχύτερες προσεγγίσεις βασισμένες σε δενδρική δομή)
    - Θεραπεία: απομάκρυνση μη συσχετιζόμενων δεδομένων
  - Θόρυβος (ωστόσο ο αλγόριθμος k -NN αντιμετωπίζει επιτυχώς το θόρυβο)
    - Θεραπεία: απομάκρυνση υποδειγμάτων με υψηλό ποσοστό θορύβου
  - Η βαρύτητα όλων των χαρακτηριστικών θεωρείται σταθερή
    - Θεραπεία: εκχώρηση βαρύτητας στα χαρακτηριστικά (ή απλά επιλογή)
  - Δεν υλοποιείται γενίκευση
    - θεραπεία: πλησιέστεροι γείτονες βασισμένοι σε κανόνες

Εκμάθηση προτοτύπων





- Αναγκαία η αποθήκευση των υποδειγμάτων εκείνων που συνδιαμορφώνουν αποφάσεις και μόνο αυτών
- Αποκοπή υποδειγμάτων με υψηλό θόρυβο
- Ιδέα: χρήση των *πρωτότυπων (prototypical)* παραδειγμάτων και μόνο αυτών

Επιτάχυνση και αντιμετώπιση θορύβου

- IB2 (Instance-Based Learner version 2): μικρότερες απαιτήσεις μνήμης, επιτάχυνση ταξινόμησης
  - Λειτουργεί βηματικά
  - Ενσωματώνει μονάχα τα υποδείγματα που ταξινομούνται ανεπιτυχώς
  - Πρόβλημα: σε αυτά συγκαταλέγονται δεδομένα με θόρυβο
- IB3: χειρισμός θορύβου
  - Μέθοδος 1: k-πλησιέστεροι γείτονες
    - Εκχώρηση της πλειοψηφούσας τάξης στο άγνωστο παράδειγμα
  - Μέθοδος 2: απόρριψη υποδειγμάτων με χαμηλή απόδοση
  - Υπολογισμός επιπέδου εμπιστοσύνης για
    1. Το βαθμό επιτυχίας κάθε υποδείγματος
    2. Προκαθορισμένη ακρίβεια της τάξης του
  - Αποδοχή / απόρριψη υποδειγμάτων
    - Αποδοχή εάν το κατώτερο όριο του 1 υπερβαίνει το ανώτερο όριο του 2
    - Απόρριψη εάν το ανώτερο όριο του 1 χαμηλότερο του κατωτέρου ορίου του 2

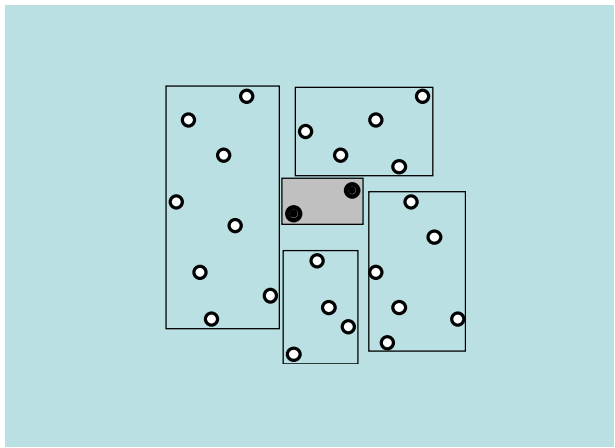
Βαρύτητα χαρακτηριστικών

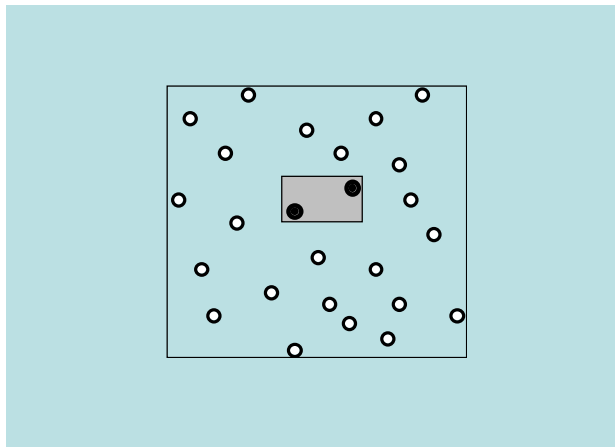
- IB5: εκχώρηση βαρύτητας σε κάθε χαρακτηριστικό
  - οι βαρύτητες δύναται να εξαρτώνται και από την τάξη
- Σταθμισμένη Ευκλείδεια απόσταση:

$$\sqrt{w_1^2(x_1-y_1)^2+\ldots+w_n^2(x_n-y_n)^2}$$

- Ενημέρωση βαρών με βάση τον πλησιέστερο γείτονα
  - Πρόβλεψη τάξης αληθής: αύξηση βαρύτητας
  - Πρόβλεψη τάξης ψευδής: μείωση βαρύτητας
  - Το ποσό μεταβολής του *i-οστού* χαρακτηριστικού εξαρτάται από την απόλυτη διαφορά $|x_i - y_i|$

Ορθογώνια γενίκευσης



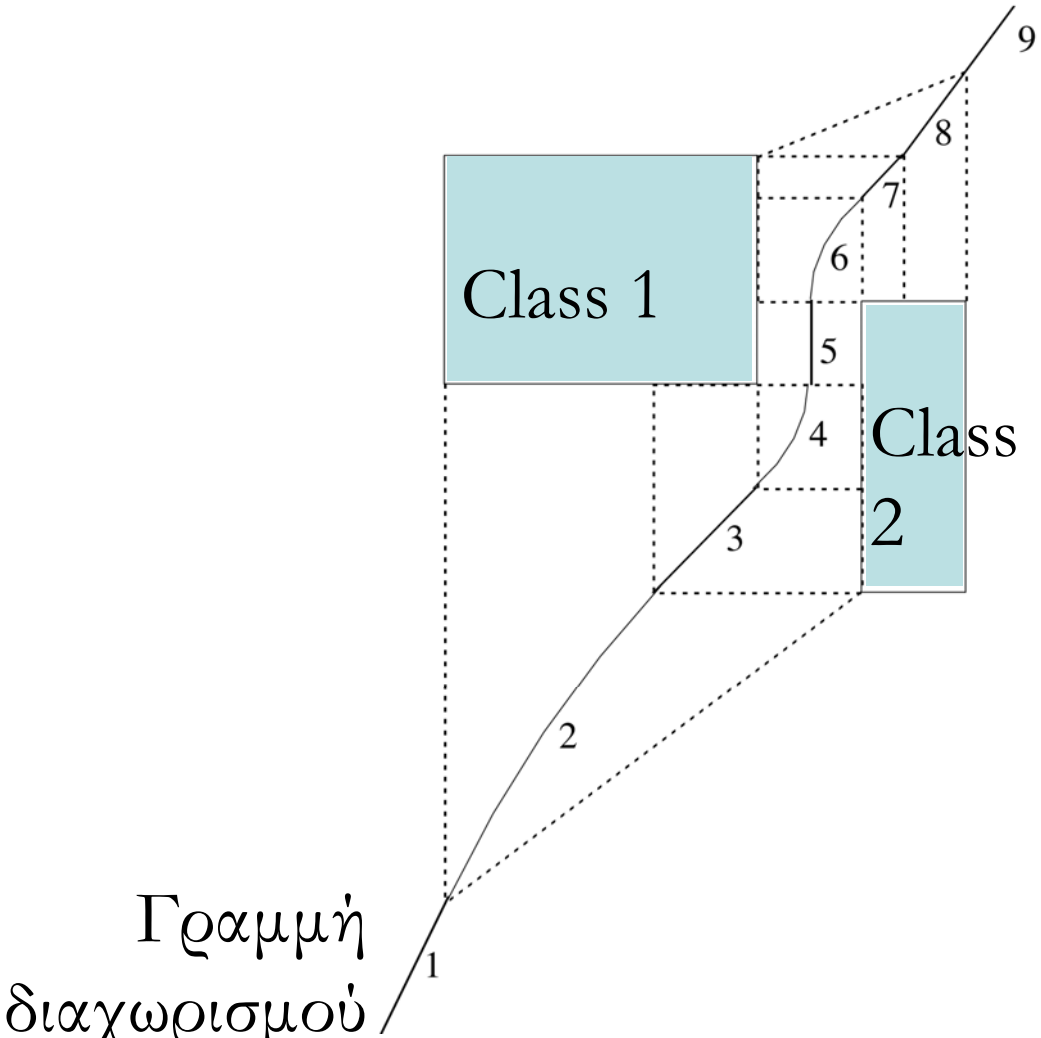


- Ο κανόνας NN χρησιμοποιείται εκτός των ορθογωνίων
- Τα ορθογώνια ουσιαστικά συνιστούν κανόνες
  - Ωστόσο είναι περισσότερο συντηρητικά από τους συνήθεις κανόνες
- Τα ένθετα ορθογώνια αποτελούν κανόνες με εξαιρέσεις

Γενικευμένα υποδείγματα

- Γενίκευση υποδειγμάτων σε *υπερορθογώνια (hyperrectangles)*
  - Online: βηματική τροποποίηση ορθογωνίων
  - Offline version: αναζήτηση μικρού συνόλου ορθογωνίων που καλύπτουν τα υποδείγματα
- Σημαντικές αποφάσεις κατά το σχεδιασμό:
  - Επιτρεπόμενη η επικάλυψη των ορθογωνίων;
    - Απαιτεί απόφαση περί αντικρουόμενων υποδείξεων
  - Επιτρεπόμενη η ένθεση των ορθογωνίων;
  - Σχετικά με τα μη καλυπτόμενα υποδείγματα;

Διαχωρισμός κατά τη γενίκευση υποδειγμάτων



@weka

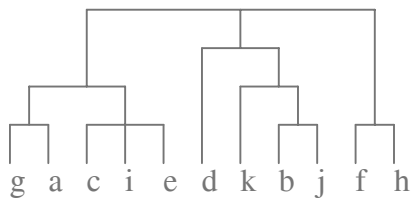
- *IB1* Basic nearest-neighbor instance-based learner
- *IBk* *k*-nearest-neighbor classifier
- *KStar* Nearest neighbor with generalized distance function
- *LBR* Lazy Bayesian Rules classifier
- *LWL* General algorithm for locally weighted learning

Ομαδοποίηση (clustering)

- Μάθηση *χωρίς επίβλεψη (unsupervised)*: δεν υπάρχει διαθέσιμη τιμή-στόχος της πρόβλεψης
- Διαφοροποιήσεις μεταξύ μοντέλων / αλγορίθμων:
  - Αποκλειστικά ή επικαλυπτόμενα
  - Ντετερμινιστικά ή πιθανοκρατικά
  - Ιεραρχικής ή επίπεδης δομής
  - Μάθηση βηματική ή σε παρτίδες (batch)
- Πρόβλημα:
  Αποτίμηση; —συνήθως με επιθεώρηση
- Ωστόσο:
  αν θεωρηθεί ως πρόβλημα εκτίμησης πυκνότητας,
  οι ομάδες μπορούν να αποτιμηθούν στα δεδομένα ελέγχου

Ιεραρχική ομαδοποίηση

- Από κάτω προς τα πάνω
  - Εκκίνηση με ομάδες ενός υποδείγματος
  - Σε κάθε βήμα, ένωση των δύο πλησιέστερων ομάδων
  - Απόφαση με βάση την απόσταση μεταξύ των ομάδων
    - Για παράδειγμα δύο πλησιέστερα υποδείγματα ή απόσταση μεταξύ μέσων ομάδων
- Από πάνω προς τα κάτω
  - Εκκίνηση με μία καθολική ομάδα
  - Εύρεση δύο υποομάδων
  - Επανάληψη διάσπασης σε κάθε υποσύνολο
  - Μπορεί να αποδειχθεί ιδιαίτερα γρήγορη
- Αμφότερες οι μέθοδοι παράγουν *δενδρόγραμμα*



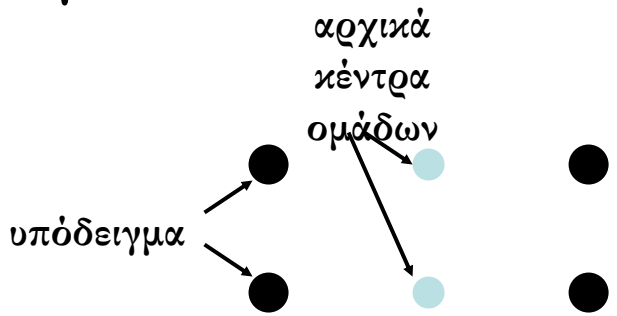
Ο αλγόριθμος *k*-μέσων

Για την ομαδοποίηση δεδομένων σε *k* ομάδες:
(*k*: προκαθορισμένο)

1. Επιλογή *k* κέντρων ομάδων
   - Για παράδειγμα με τυχαίο τρόπο
2. Εκχώρηση υποδειγμάτων σε ομάδες
   - Με βάση την απόσταση από το κέντρο των ομάδων
3. Υπολογισμός *κεντροειδών* (*centroids*) των ομάδων
4. Επανάληψη του βήματος 1
   - Μέχρι τη σύγκλιση

Γενικά περί ιεραρχικής ομαδοποίησης

- Το αποτέλεσμα μπορεί να ποικίλει σημαντικά
  - Ανάλογα της αρχικής επιλογής των κάντρων
- Ο αλγόριθμος συχνά παγιδεύεται σε τοπικό ακρότατο
  - Παράδειγμα:



- Για την αύξηση της πιθανότητας εύρεσης ολικού βέλτιστου: επανάληψη της διαδικασίας με διαφορετικές τυχαίες αρχικές επιλογές

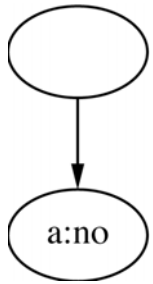
Βηματική ομαδοποίηση

- Ευρετική μέθοδος (Cobweb/Classit)
- Βηματική διαμόρφωση ιεραρχίας ομάδων
- Εκκίνηση:
  - Το δένδρο αποτελείται από τον κενό κόμβο της ρίζας
- Στη συνέχεια:
  - Προσθήκη υποδειγμάτων ένα προς ένα
  - Ενημέρωση του δένδρου με κατάλληλο τρόπο σε κάθε βήμα
    - Μέσω εύρεσης κατάλληλου φύλλου για κάθε υπόδειγμα
  - Ίσως περιλαμβάνει αναδόμηση του δένδρου
- Οι αποφάσεις προσθήκης και ενημέρωσης βασίζονται στο μέγεθος *ωφέλειας τάξης (category utility*, βλέπε παρακάτω)

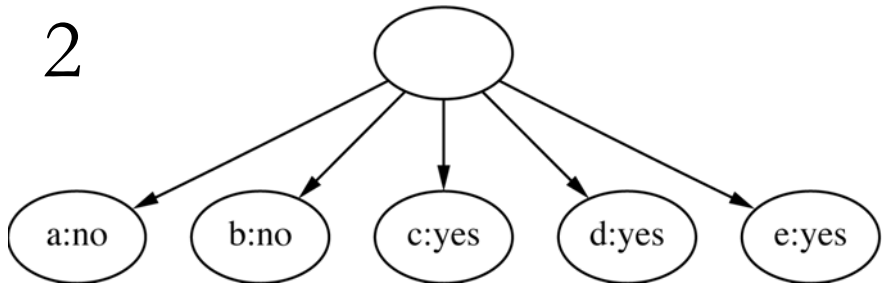
Ομαδοποίηση δεδομένων καιρού

ID	Outlook	Temp.	Humidity	Windy
A	Sunny	Hot	High	False
B	Sunny	Hot	High	True
C	Overcast	Hot	High	False
D	Rainy	Mild	High	False
E	Rainy	Cool	Normal	False
F	Rainy	Cool	Normal	True
G	Overcast	Cool	Normal	True
H	Sunny	Mild	High	False
I	Sunny	Cool	Normal	False
J	Rainy	Mild	Normal	False
K	Sunny	Mild	Normal	True
L	Overcast	Mild	High	True
M	Overcast	Hot	Normal	False
N	Rainy	Mild	High	True

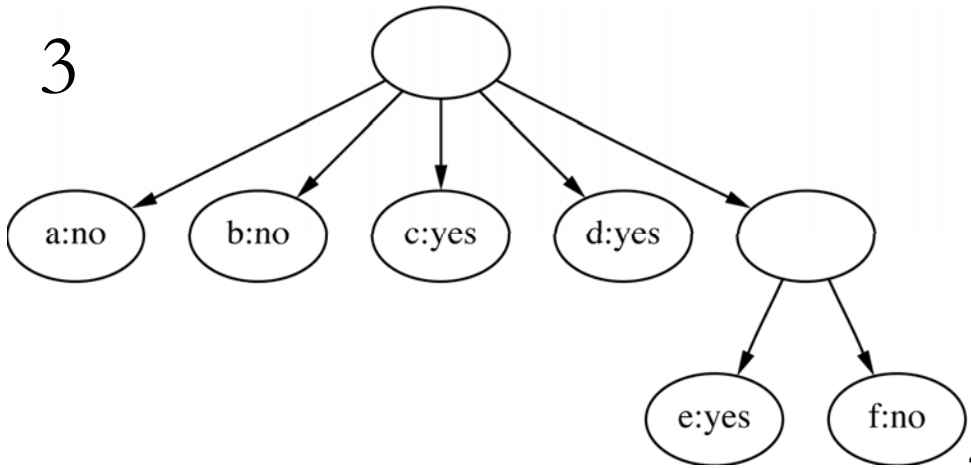
1



2



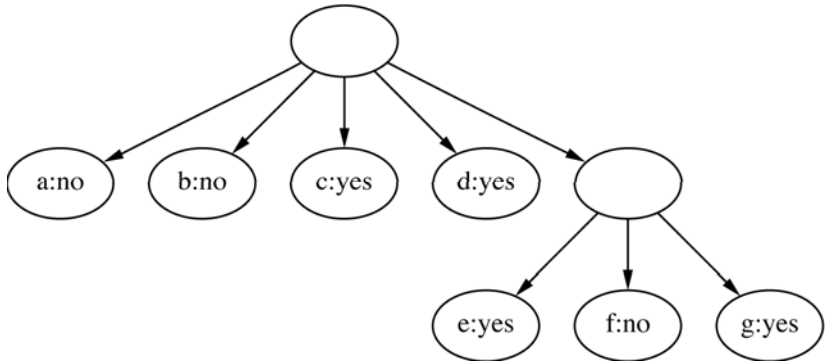
3



Ομαδοποίηση δεδομένων καιρού

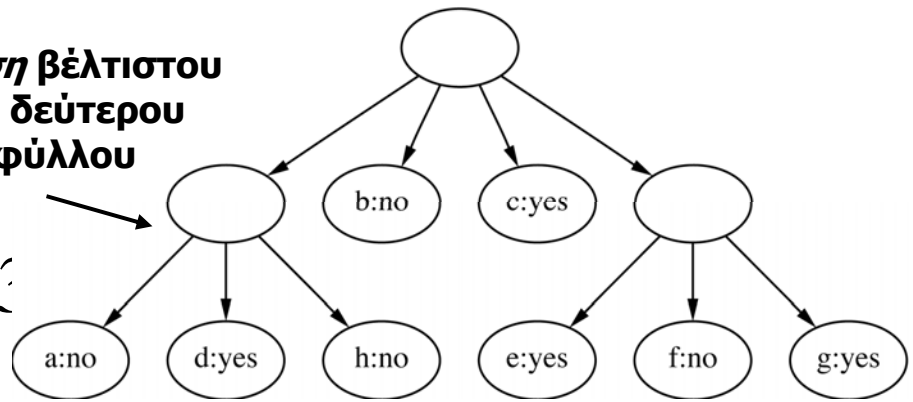
ID	Outlook	Temp.	Humidity	Windy
A	Sunny	Hot	High	False
B	Sunny	Hot	High	True
C	Overcast	Hot	High	False
D	Rainy	Mild	High	False
E	Rainy	Cool	Normal	False
F	Rainy	Cool	Normal	True
G	Overcast	Cool	Normal	True
H	Sunny	Mild	High	False
I	Sunny	Cool	Normal	False
J	Rainy	Mild	Normal	False
K	Sunny	Mild	Normal	True
L	Overcast	Mild	High	True
M	Overcast	Hot	Normal	False
N	Rainy	Mild	High	True

4



5

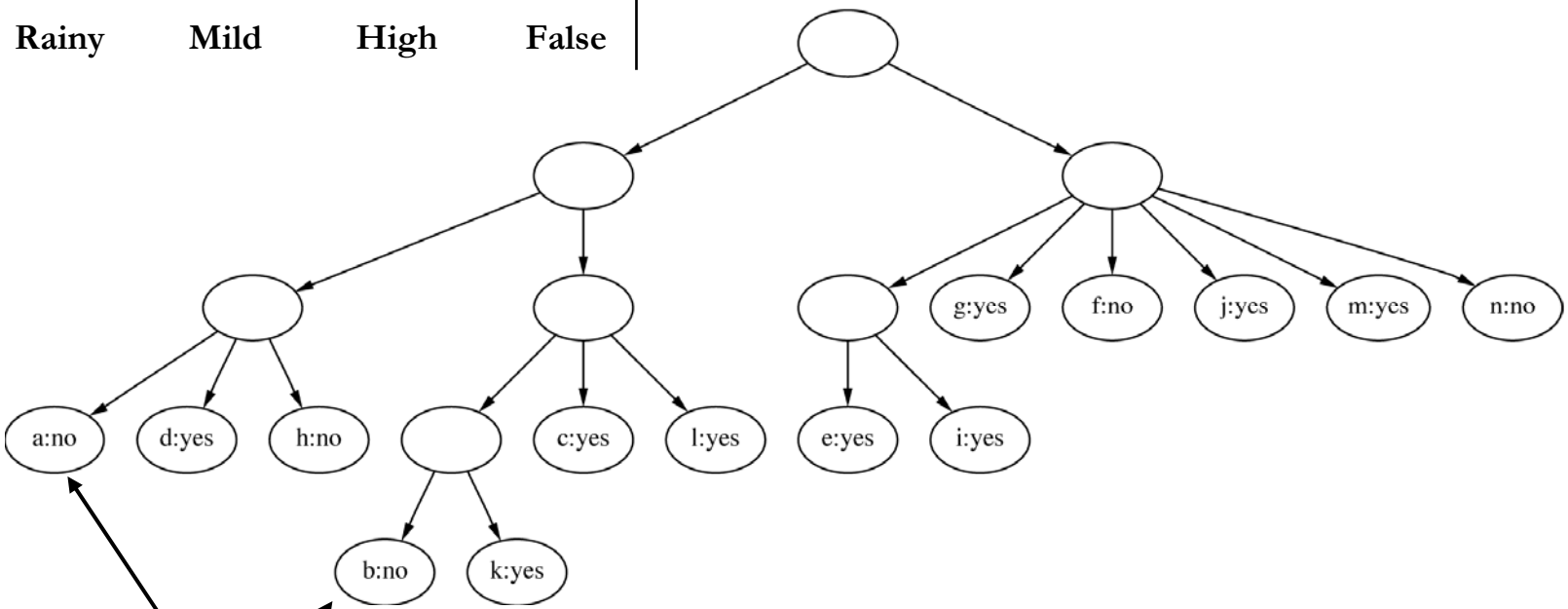
ένωση **βέλτιστου και δεύτερου φύλλου**



Διάσπαση **βέλτιστου φύλλου εάν η *ένωση* δεν ωφελεί**

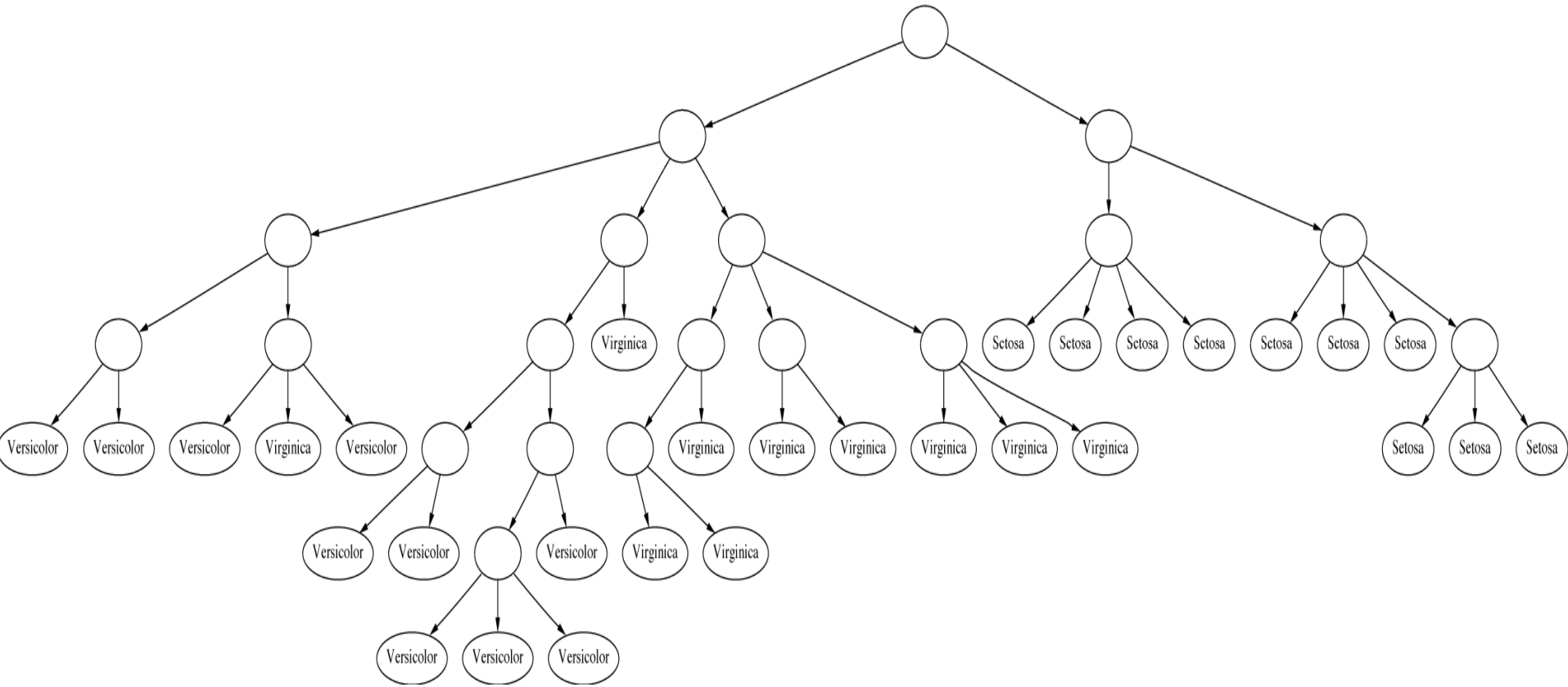
Τελική ιεραρχία

ID	Outlook	Temp.	Humidity	Windy
A	Sunny	Hot	High	False
B	Sunny	Hot	High	True
C	Overcast	Hot	High	False
D	Rainy	Mild	High	False

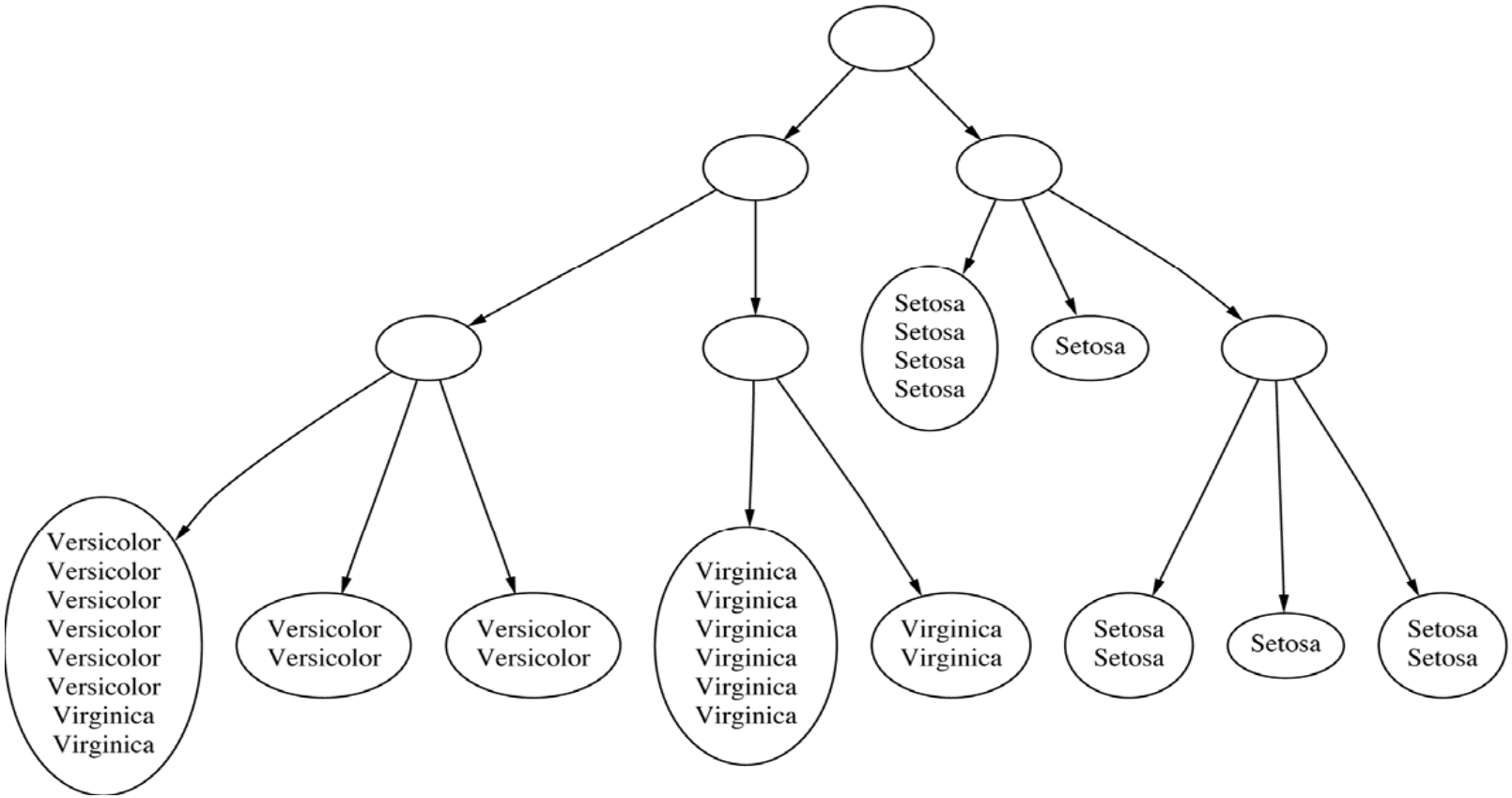


Τα a & b είναι στην πράξη παρόμοια, όμως περιλήφθηκαν σε άλλη ομάδα

Παράδειγμα: δεδομένα ίρις (υποσύνολο)



Ομαδοποίηση με αποκοπή



- Παράμετρος *αποκοπής* (*cutoff*): η αύξηση της ωφέλειας τάξης από την εισαγωγή νέου κλάδου πρέπει να υπερβαίνει την τιμή της παραμέτρου

Ωφέλεια τάξης (category utility)

- Ωφέλεια τάξης (category utility, CU): εκφράζει τη συνολική ποιότητα του διαχωρισμού των υποδειγμάτων σε ομάδες
  - Συνάρτηση απωλειών δευτέρου βαθμού με βάση τις πιθανότητες υπό συνθήκη:

$$CU(C_1, C_2, ..., C_k) = \frac{1}{k} \sum_l \Pr[C_l] \sum_i \sum_j (\Pr[a_i = v_{ij} \mid C_l]^2 - \Pr[a_i = v_{ij}]^2)$$

- Η $\mathbf{Pr[a_i = v_{ij} \mid C_l]}$ αποτελεί καλύτερη εκτίμηση της πιθανότητας το χαρακτηριστικό a_i να έχει τιμή v_{ij}, για ένα υπόδειγμα που ανήκει στην ομάδα C_l, από ότι η $\mathbf{Pr[a_i = v_{ij}]}$.
  - Αν αυτή η πληροφορία δεν έχει αξία, η ομαδοποίηση είναι ανεπιτυχής!
  - Επομένως, το μέτρο εν τέλει αποτιμά το κέρδος πληροφορίας από την εισαγωγή των ομάδων

Αριθμητικά χαρακτηριστικά

- Υπόθεση κανονικής κατανομής: $f(a) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(a-\mu)^2}{2\sigma^2}}$
- Τότε $\sum_j \Pr[a_i = v_{ij}]^2 \Leftrightarrow \int f(a_i)^2 da_i = \frac{1}{2\sqrt{\pi}\sigma_i}$
- Επομένως $CU = \frac{1}{k}\sum_l \Pr[C_l] \frac{1}{2\sqrt{\pi}} \sum_i \left(\frac{1}{\sigma_{il}} - \frac{1}{\sigma_i} \right)$
- Προκαθορισμένη ελάχιστη διακύμανση
  - Εισαγωγή παραμέτρου *οξύτητας (acuity)*
  - Αντιμετωπίζει το πρόβλημα variance = 0 (CU → ∞)
- @weka: Cobweb

Πιθανοκρατική ομαδοποίηση

- Προβλήματα ευρετικής προσέγγισης:
  - Προκαθορισμένος αριθμός ομάδων *k*
  - Προκαθορισμένες / ανά περίπτωση τιμές παραμέτρων αποκοπής και οξύτητας
  - Κατάταξη υποδειγμάτων
  - Αποδοτικότητα λειτουργιών απόδοσης δένδρου
  - Κατάληξη σε τοπικό ακρότατο της συνάρτησης ωφέλειας τάξης
- Πιθανοκρατική θεώρηση $\Rightarrow$
  αναζήτηση των *πλέον πιθανών* ομάδων για τα διαθέσιμα δεδομένα
- Επίσης, κάθε υπόδειγμα ανήκει με *συγκεκριμένη* πιθανότητα σε συγκεκριμένη ομάδα

Πεπερασμένες αναμείξεις (finite mixtures)

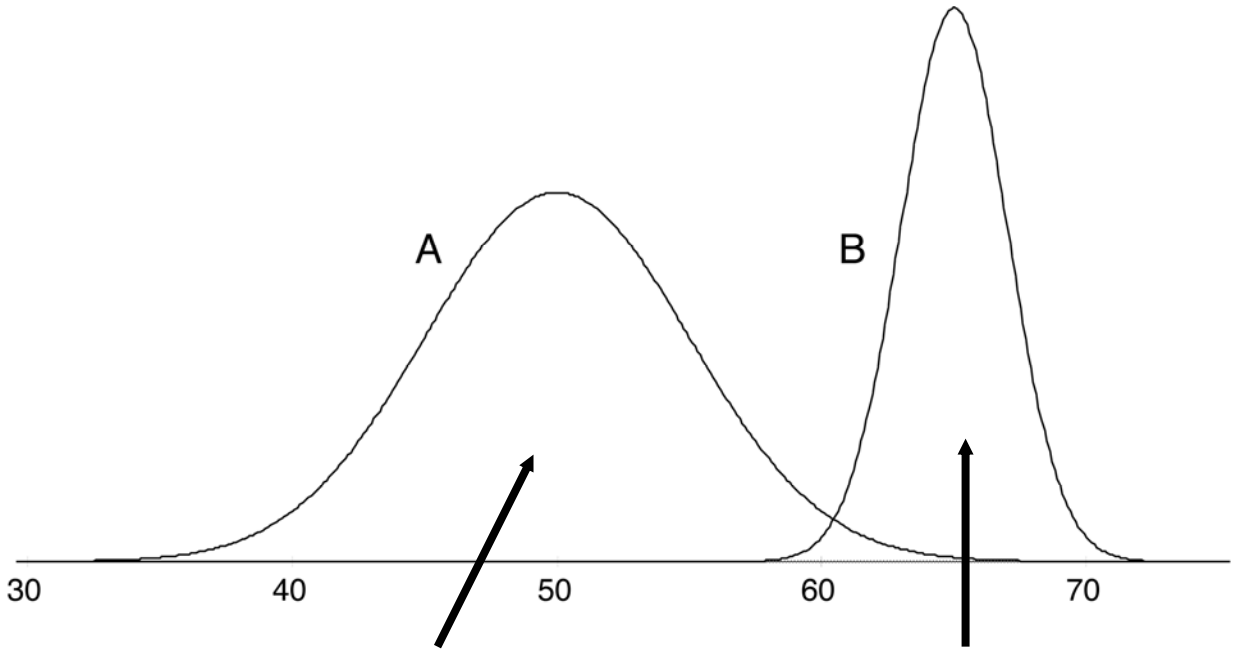
- Μοντελοποίηση των δεδομένων με χρήση ενός *μείγματος* κατανομών
- Μία ομάδα, μία κατανομή
  - Καθορίζει τις πιθανότητες των τιμών των χαρακτηριστικών σε αυτή την ομάδα
- *Πεπερασμένες αναμείξεις (finite mixtures)*: πεπερασμένος αριθμός ομάδων
- Η κατανομή κάθε ομάδας ορίζεται συνήθως ως κανονική
- Συνδυασμός κατανομών με βάση τη βαρύτητα κάθε ομάδας

Μοντέλο ανάμειξης δύο τάξεων

δεδομένα

A	51	B	62	B	64	A	48	A	39	A	51
A	43	A	47	A	51	B	64	B	62	A	48
B	62	A	52	A	52	A	51	B	64	B	64
B	64	B	64	B	62	B	63	A	52	A	42
A	45	A	51	A	49	A	43	B	63	A	48
A	42	B	65	A	48	B	65	B	64	A	41
A	46	A	48	B	62	B	66	A	48		
A	45	A	49	A	43	B	65	B	64		
A	45	A	46	A	40	A	46	A	48		

μοντέλο



μ_A=50, σ_A =5, p_A=0.6 μ_B=65, σ_B =2, p_B=0.4

Χρήση μοντέλου ανάμειξης

- Πιθανότητα το υπόδειγμα x να ανήκει στην ομάδα A:

$$\Pr[A \mid x] = \frac{\Pr[x \mid A]\Pr[A]}{\Pr[x]} = \frac{f(x; \mu_A, \sigma_A) p_A}{\Pr[x]}$$

με $$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

- Πιθανότητα ενός υποδείγματος, δεδομένων των ομάδων

$$\Pr[x \mid \text{the distributions}] = \sum_i \Pr[x \mid \text{cluster}_i]\Pr[\text{cluster}_i]$$

Εκμάθηση ομάδων

- Υπόθεση:
  - *k* ομάδες
- Εκμάθηση ομάδων ⇒
  - Προσδιορισμός των παραμέτρων τους
    - Δηλαδή μέσων τιμών και τυπικών αποκλίσεων
- Κριτήριο απόδοσης:
  - *Πιθανότητα των δεδομένων εκπαίδευσης με γνωστές τις ομάδες*
- Αλγόριθμος ΕΜ
  - Εύρεση τοπικού μεγίστου της πιθανότητας

Αλγόριθμος EM

- EM = Expectation-Maximization (Μεγιστοποίησης Εκτίμησης
  - Γενίκευση των *k*-μέσων σε πιθανοκρατικό πλαίσιο
- Επαναληπτική διαδικασία:
  - Βήμα Ε "expectation":
    Υπολογισμός πιθανότητας ομάδας για κάθε υπόδειγμα
  - Βήμα Μ "maximization" :
    εκτίμηση παραμέτρων κατανομών από τις πιθανότητες των ομάδων
- Αποθήκευση πιθανοτήτων ομάδων ως βαρύτητες υποδειγμάτων
- Διακοπή όταν η βελτίωση είναι αμελητέα

Περισσότερα περί του EM

- Εκτίμηση παραμέτρων από υποδείγματα με βαρύτητες

$$\mu_A = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{w_1 + w_2 + \ldots + w_n}$$

$$\sigma_A^{\ 2} = \frac{w_1 (x_1 - \mu)^2 + w_2 (x_2 - \mu)^2 + \ldots + w_n (x_n - \mu)^2}{w_1 + w_2 + \ldots + w_n}$$

- Διακοπή όταν κορεστεί ο λογάριθμος της πιθανότητας

  – Log-likelihood: $\sum_i \log(p_A \Pr[x_i \mid A] + p_B \Pr[x_i \mid B])$

Επέκταση του μοντέλου ανάμειξης

- Περισσότερες των δύο κατανομών: εύκολη
- Πολλαπλά χαρακτηριστικά: εύκολη – με την υπόθεση ανεξαρτησίας
- Συσχετιζόμενα χαρακτηριστικά: δύσκολη
  - Μοντέλο σύνδεσης: κανονική κατανομή δύο μεταβλητών με (συμμετρικό) πίνακα συνδιακύμανσης (covariance)
  - n χαρακτηριστικά: απαιτείται η εκτίμηση $n + n(n+1)/2$ παραμέτρων

Περισσότερες επεκτάσεις

- Ονομαστικά χαρακτηριστικά: εύκολη αν είναι ανεξάρτητα
- Συσχετιζόμενα ονομαστικά χαρακτηριστικά: δύσκολη
  - Δύο συσχετιζόμενα χαρακτηριστικά $\Rightarrow \nu_1 \nu_2$ παράμετροι
- Άγνωστες τιμές: εύκολη
- Εφικτή η χρήση κατανομών διαφορετικών της κανονικής:
  - "log-normal" για δεδομένο και προκαθορισμένο ελάχιστο
  - "log-odds" για άνω και κάτω όρια
  - Poisson για χαρακτηριστικά καταμέτρησης ακεραίων
- Χρήση διασταυρωμένης επικύρωσης για τον υπολογισμό του k !
- @weka: *EM*

Ομαδοποίηση κατά Bayes

- Πρόβλημα: πληθώρα παραμέτρων ⇒ υπερπροσαρμογή ΕΜ
- *Προσέγγιση Bayes*: εκ των προτέρων εκχώρηση κατανομής πιθανότητας σε κάθε παράμετρο
  - Ενσωμάτωσή της στον υπολογισμό της ολικής πιθανότητας
  - Τιμωρεί την εισαγωγή νέων παραμέτρων
- Παράδειγμα: Εκτιμητής Laplace για ονομαστικά χαρακτηριστικά
- Εφαρμογή της μεθόδου και για τον αριθμό ομάδων
- Υλοποίηση: AUTOCLASS (δημιουργός: NASA)

Γενικά περί Bayes

- Εφικτή η ερμηνεία ομάδων με χρήση εκμάθησης με επίβλεψη
  - Βήμα μετα-επεξεργασίας
- Μείωση της εξάρτησης μεταξύ των χαρακτηριστικών;
  - Βήμα προ-επεξεργασίας
  - Για παράδειγμα, χρήση *principal component analysis*
- Μπορεί να χρησιμοποιηθεί και για την εξάλειψη των άγνωστων τιμών
- Κύριο πλεονέκτημα της πιθανοκρατικής ομαδοποίησης:
  - Δύναται να υπολογίσει την πιθανότητα των δεδομένων
  - Χρήση της μεθόδου για την αντικειμενική σύγκριση διαφορετικών μοντέλων

@weka

- *EM* Cluster using expectation maximization
- *Cobweb* Implements the Cobweb and Classit clustering algorithms
- *FarthestFirst* Cluster using the farthest first traversal algorithm
- *MakeDensityBasedClusterer* Wrap a clusterer to make it return distribution and density
- *SimpleKMeans* Cluster using the *k*-means method

Τέλος

Επόμενη διάλεξη:

Αλγόριθμοι εκμάθησης, μέρος Γ