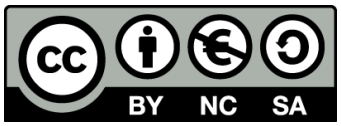
ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Αποθήκες Δεδομένων και Εξόρυξη Δεδομένων

Ενότητα 6: Κατηγοριοποίηση – Μέρος Β΄

Αναστάσιος Γούναρης, Επίκουρος Καθηγητής
Τμήμα Πληροφορικής ΑΠΘ



Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Κατηγοριοποίηση – Μέρος Β΄

Κατηγοριοποιητές, μέθοδοι αποτίμησης ακρίβειας κατηγοριοποιητών

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Περιεχόμενα ενότητας

1. Bayesian κατηγοριοποιητές.
2. Κατηγοριοποιητές πλησιέστερων γειτόνων.
3. Αποτίμηση ακρίβειας κατηγοριοποιητών.

Σκοποί ενότητας

- Παρουσίαση των κατηγοριοποιητών, όπως Bayesian και πλησιέστερων γειτόνων.
- Ανάλυση διαφορετικών μεθόδων αποτίμησης ακρίβειας κατηγοριοποιητών.

Χαρακτηριστικά Δένδρων Απόφασης (1/2)

- Η κατασκευή του βέλτιστου δένδρου απόφασης απαιτεί αποτρεπτικό χρόνο (είναι NP-complete πρόβλημα).
  - Για το λόγο αυτό χρησιμοποιούνται ευρετικοί αλγόριθμοι, οι οποίοι είναι άπληστοι και δεν χρησιμοποιούν οπισθοδρόμηση.
  - Τα ευρετικά μειώνουν κατά πολύ το χρόνο κατασκευής.
  - Το αποτέλεσμα είναι ότι τα δένδρα απόφασης **κλιμακώνονται σε μεγάλους όγκους δεδομένων.**
- **Γρήγορη εφαρμογή.**

Χαρακτηριστικά Δένδρων Απόφασης (2/2)

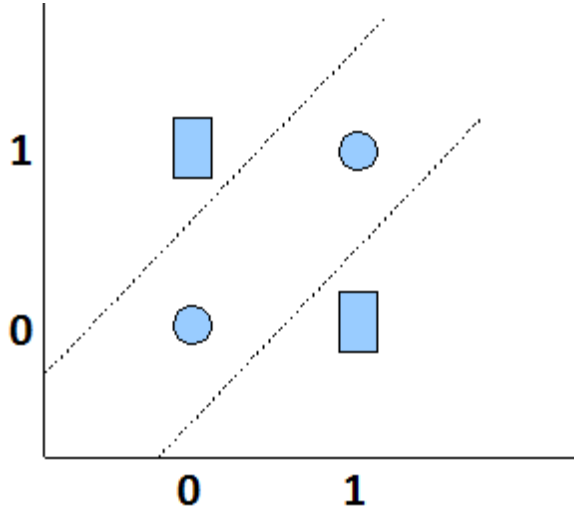
- Η **ακρίβεια πρόβλεψης** των δένδρων απόφασης είναι αποδεκτή για τις περισσότερες περιπτώσεις, συγκρίσιμη με την ακρίβεια άλλων κατηγοριοποιητών.
- Το μοντέλο που προκύπτει είναι πολύ **εύκολο στην κατανόηση**.
- Τα δένδρα απόφασης έχουν καλή **ανοχή στο θόρυβο:**
  - ειδικά όταν εφαρμόζεται ψαλιδισμός.

Επιπλέον

- Τα ΔΑ μπορούν να διαχειριστούν πολυδιάστατα δεδομένα.
  - 1 διάσταση τη φορά χρησιμοποιείται κατά την ανάπτυξη του μοντέλου.
- … και κάθε τύπο μεταβλητών.
  - Συμβολικές, αριθμητικές, κλπ.

Μειονεκτήματα

- Αγνοούν εξαρτήσεις μεταξύ των ιδιοτήτων.
- Προβλήματα όταν λείπουν πολλά δεδομένα.
- Διάσπαση ως προς μία ιδιότητα:
  - αντιστοίχιση με περιοχές, τα όρια των οποίων είναι παράλληλα με τους άξονες.



Άλλοι κατηγοριοποιητές;

- Bayesian κατηγοριοποιητές.
- Κατηγοριοποιητές πλησιέστερων γειτόνων.

Bayesian κατηγοριοποιητής για 1 ιδιότητα

- Ιδιότητα X (συμβολική).
  - m διακριτές τιμές.
- Ιδιότητα κλάσης C
  - n διακριτές τιμές.
- Θέλουμε να υπολογίσουμε για κάθε *j: 0<j<n+1*:

$$P(C = c_j \mid X = x_i)$$ **Άγνωστη ποσότητα**

Παράδειγμα

Οικογενειακή Κατάσταση	Αγοραστής
Διαζευγμένος	ΝΑΙ
Διαζευγμένος	ΝΑΙ
Έγγαμος	ΟΧΙ
Άγαμος	ΝΑΙ
Άγαμος	ΝΑΙ
Έγγαμος	ΟΧΙ
Διαζευγμένος	ΝΑΙ
Διαζευγμένος	ΝΑΙ
Διαζευγμένος	ΝΑΙ
Άγαμος	ΟΧΙ

- **Αν**
  - **Οικογενειακή κατάσταση: Άγαμος**
  - **Αγοραστής: ναι ή όχι;**
- **P(Ναι | Άγαμος) = ;**
- **P(Όχι | Άγαμος) = ;**

Θεώρημα Bayes

$$P(C = c_j \mid X = x_i)$$ **Άγνωστη ποσότητα**

$$P(C = c_j \mid X = x_i) = \frac{P(X = x_i \mid C = c_j)}{P(X = x_i)} P(C = c_j)$$

$$P(X = x_i \mid C = c_j), P(C = c_j)$$ **Είναι υπολογίσιμα**

$$P(X = x_i)$$ **Είναι ανεξάρτητο της κλάσης**

Άρα αρκεί να βρούμε την κλάση για την οποία μεγιστοποιείται το

$$P(X = x_i \mid C = c_j) P(C = c_j)$$

Στο παράδειγμα…

Ηλικία	Οικογενειακή Κατάσταση	Αγοραστής
20	Διαζευγμένος	ΝΑΙ
30	Διαζευγμένος	ΝΑΙ
25	Έγγαμος	ΟΧΙ
30	Άγαμος	ΝΑΙ
40	Άγαμος	ΝΑΙ
20	Έγγαμος	ΟΧΙ
30	Διαζευγμένος	ΝΑΙ
25	Διαζευγμένος	ΝΑΙ
40	Διαζευγμένος	ΝΑΙ
20	Άγαμος	ΟΧΙ

- P(Ναι | Άγαμος) → P(Άγαμος | Ναι) P(Ναι) = 2/7*7/10 = 0.2
- P(Όχι | Άγαμος) → P(Άγαμος | Όχι) P(Όχι) = 1/3 * 3/10 = 0.1

Τι γίνεται για περισσότερες ιδιότητες;

- Έστω ότι μας δίνεται η τιμή d χαρακτηριστικών.
- Πρέπει να υπολογίσουμε την πιθανότητα:

$$P(X =< x_1, x_2, .., x_d > | C = c_j)$$

- Απλούστευση: ανεξαρτησία των d ιδιοτήτων.

$$P(X =< x_1, x_2, .., x_d > | C = c_j) = \prod_{i=1}^{d} P(X = x_i \ | C = c_j)$$

- **Αφελείς Bayesian κατηγοριοποιητές:**

$$\arg \max_{\forall 1 \le j \le m} \prod_{i=1}^{d} P(X = x_i \ | C = c_j) \ P(C = c_j)$$

Παράδειγμα (1/4)

Ηλικία	Οικογενειακή Κατάσταση	Αγοραστής
20	Διαζευγμένος	ΝΑΙ
30	Διαζευγμένος	ΝΑΙ
25	Έγγαμος	ΟΧΙ
30	Άγαμος	ΝΑΙ
40	Άγαμος	ΝΑΙ
20	Έγγαμος	ΟΧΙ
30	Διαζευγμένος	ΝΑΙ
25	Διαζευγμένος	ΝΑΙ
40	Διαζευγμένος	ΝΑΙ
20	Άγαμος	ΟΧΙ

- Οικογενειακή κατάσταση = Άγαμος, Ηλικία = 35
- Αγοραστής: ναι ή όχι;
- Πρέπει να υπολογιστούν τα P(Ναι | Άγαμος, 35), P(Όχι | Άγαμος, 35)

Παράδειγμα (2/4)

- P(Ναι|Άγαμος, 35) → P(Άγαμος,35|Ναι) * P(Ναι)=;
- P(Όχι|Άγαμος, 35) → P(Άγαμος,35| Όχι) * P(Όχι)=;

- Υπόθεση: Ανεξαρτησία οικογενειακής κατάστασης και ηλικίας.

- P(Ναι|Άγαμος, 35) → P(Άγαμος|Ναι) * P(35 |Ναι) * P(Ναι)=;
- P(Όχι|Άγαμος, 35) → P(Άγαμος| Όχι) * P(35 | Όχι) * P(Όχι)=;

- Από το παράδειγμα μιας ιδιότητας, έχω ήδη υπολογίσει:
  - P(Άγαμος|Ναι) * P(Ναι) = 0.2
  - P(Άγαμος|Όχι) * P(Όχι) = 0.1

Παράδειγμα (3/4)

- P(35|Ναι)=; P(35|Όχι)=;
- Ηλικία: συνεχής μεταβλητή.

1. Κβάντωση
2. Υπόθεση συνεχούς κανονικής κατανομής:

( )

( )

( ) $\frac{1}{\sqrt{\quad}} \quad \frac{\quad}{\quad} \quad \frac{1}{\sqrt{\quad}} \quad \frac{(\quad)}{\quad}$

( ) $\frac{1}{\sqrt{\quad}} \quad \frac{\quad}{\quad} \quad \frac{1}{\sqrt{\quad}} \quad \frac{(\quad)}{\quad}$

Παράδειγμα (4/4)

- P(Ναι|Άγαμος, 35) →

P(Άγαμος|Ναι) P(35 |Ναι) * P(Ναι)= 0.2 * 0.11ε = 0.022 ε

- P(Όχι|Άγαμος, 35) →

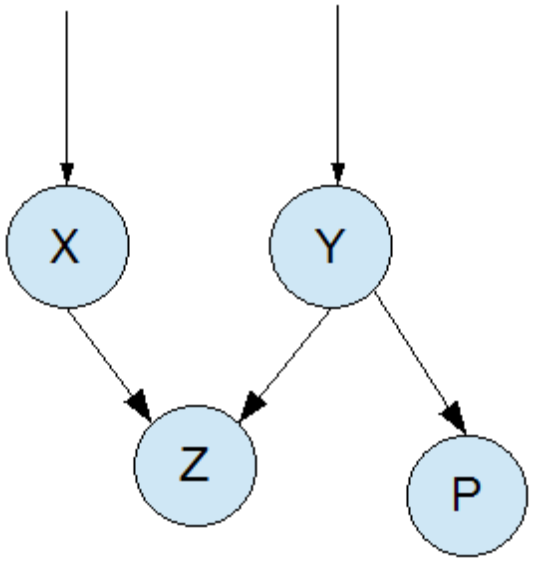
P(Άγαμος| Όχι) P(35 | Όχι) * P(Όχι)= 0.1 * 10^{-14} ε= 10^{-15} ε

- Άρα, αγοραστής: ΝΑΙ

Χαρακτηριστικά Αφελών Bayesian

- Η **ακρίβεια πρόβλεψης** των αφελών Bayesian κατηγοριοποιητών **επηρεάζεται αρνητικά** από το γεγονός ότι σε πραγματικά δεδομένα σχεδόν πάντοτε υπάρχουν εξαρτήσεις μεταξύ των μεταβλητών.
- Το μοντέλο που προκύπτει είναι απλά και σχετικά **εύκολο στην κατανόηση**.
- Η κατασκευή των ιστογραμμάτων για τους υπολογισμούς των πιθανοτήτων, απαιτεί μόνο μία ανάγνωση του συνόλου δεδομένων. Επομένως, οι Bayesian κατηγοριοποιητές **κλιμακώνονται** σε μεγάλους όγκους δεδομένων.
- Οι Bayesian κατηγοριοποιητές έχουν **καλή ανοχή στο θόρυβο**, επειδή οι θορυβώδεις τιμές εξομαλύνονται από τις υπόλοιπες κατά τους υπολογισμούς των εν μέρει πιθανοτήτων.
- Οι Bayesian κατηγοριοποιητές **δεν επηρεάζονται από τις ελλιπείς τιμές**, επειδή μπορούν να αγνοηθούν.

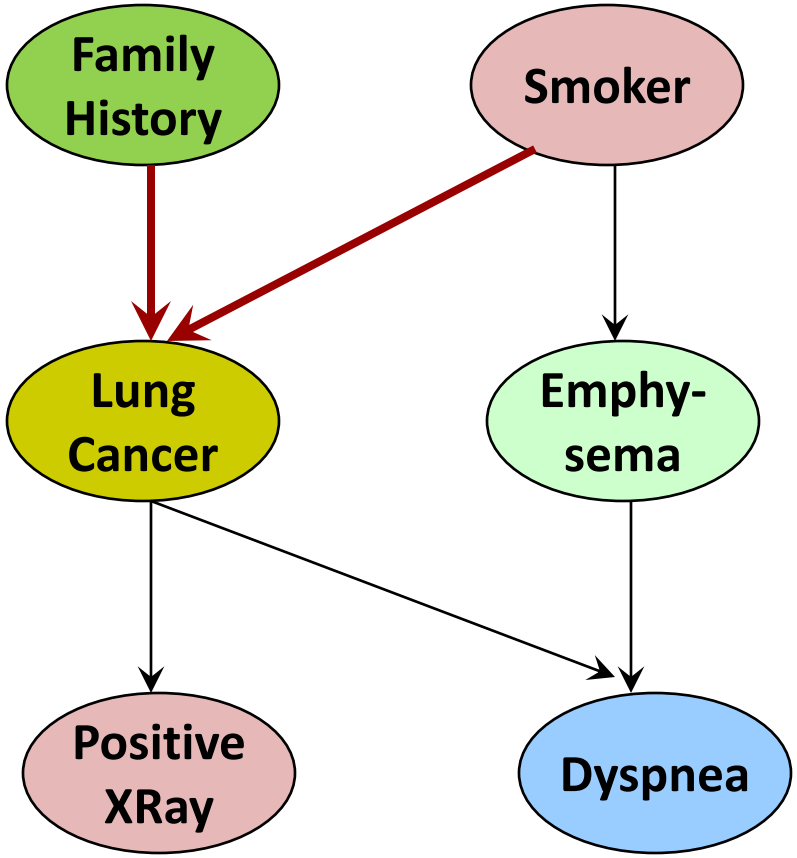
Bayesian Belief Networks



- Κόμβοι: χαρακτηριστικά.
- Συνδέσεις: εξαρτήσεις.
- Τα X και Y είναι οι γονείς του Z
- και το Y είναι γονέας του P.
- Τα Z και P είναι ανεξάρτητα.
- Δεν υπάρχουν κύκλοι.

- Μοντελοποίηση εξαρτήσεων μεταξύ των χαρακτηριστικών.
- Γραφικό μοντέλο.
- Ορίζει την κοινή κατανομή πιθανότητας.

Παράδειγμα



Bayesian Belief Networks

Πίνακας με υπο-συνθήκη πιθανότητες για Lung Cancer:

	(FH, S)	(FH, ~S)	(~FH, S)	(~FH, ~S)
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

Δείχνει την υπο-συνθήκη πιθανότητα για κάθε συνδυασμό γονέων.

$$P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid A(x_i))$$

A(x): κόμβοι γονείς του x

Κατηγοριοποιητής *k* πλησιέστερων γειτόνων

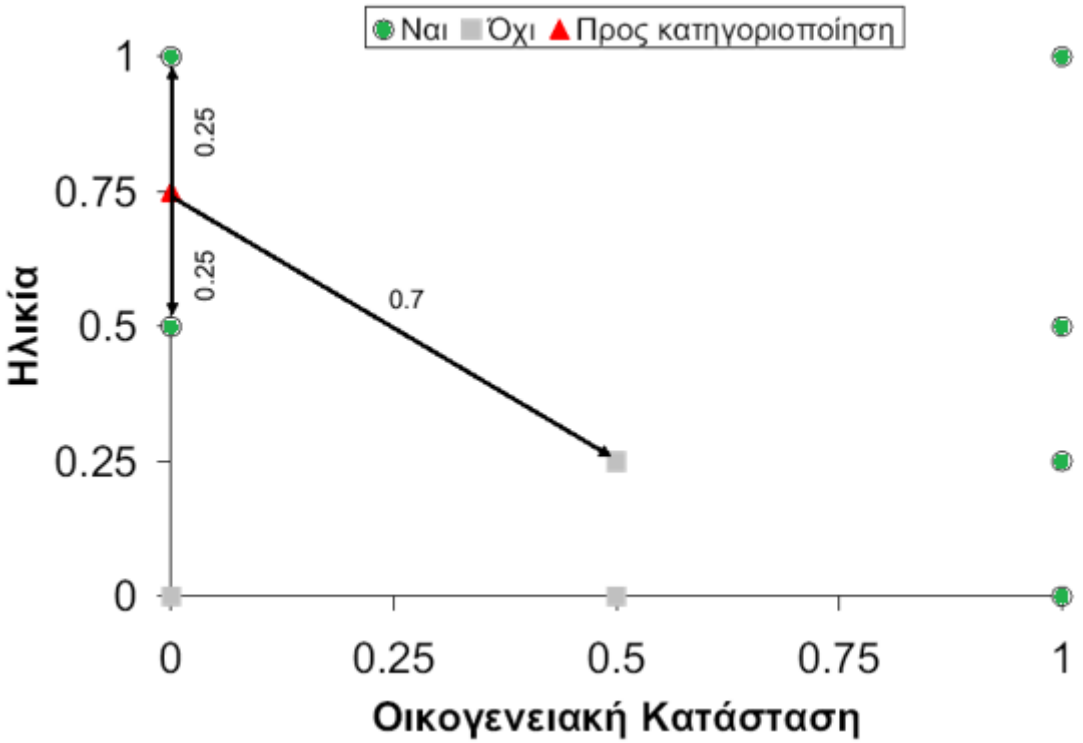
- Κατηγοριοποιεί ένα αντικείμενο στην κλάση στην οποία ανήκει η πλειοψηφία των *k* πλησιέστερων σε αυτό αντικειμένων.
- Απαιτείται ορισμός μέτρου ομοιότητας
  - (ή απόστασης).

Παράδειγμα (1/2)

- {Άγαμος, Έγγαμος, Διαζευγμένος} → {0, 0.5, 1}
- Ηλικία x → (x-20)/(40-20)
- Ευκλείδειος χώρος [0,1] x [0,1]
- Ευκλείδεια απόσταση.

Παράδειγμα (2/2)

- k = 3, προς κατηγοριοποίηση Άγαμος, 35



Επιλογή *k*

- Η τιμή του *k* μπορεί επηρεάζει το αποτέλεσμα.
- Μικρές τιμές του *k* εξετάζουν μόνο την άμεση γειτονιά, επομένως είναι επιρρεπείς στο θόρυβο.
- Μεγάλες τιμές του *k* αγνοούν την αρχή της τοπικότητας, και είναι επιρρεπείς στην πλειοψηφούσα κλάση σε όλο το σύνολο δεδομένων.
- Συχνά χρησιμοποιούμενη τιμή είναι $k = \text{sqrt}(n)$, όπου n είναι ο αριθμός των αντικειμένων στο σύνολο εκμάθησης.
- Σε εμπορικά συστήματα η default τιμή είναι $k = 10$.

Χαρακτηριστικά κατηγοριοποιητών *k* πλησιέστερων γειτόνων

- Η **ακρίβεια πρόβλεψης** των κατηγοριοποιητών k πλησιέστερων γειτόνων είναι **ευαίσθητη** στην τιμή του k.
- Παρά ταύτα, οι κατηγοριοποιητές k πλησιέστερων γειτόνων αξιοποιούν την τοπικότητα και εξετάζουν **μη γραμμικές περιοχές** (αντίθετα από τα δένδρα απόφασης), κάτι που σε αρκετές περιπτώσεις αποτελεί πλεονέκτημα.
- Το αποτέλεσμα της κατηγοριοποίησης **δεν γίνεται πολύ εύκολα κατανοητό**. Η αρχή της τοπικότητας είναι η μόνη αιτιολόγηση του αποτελέσματος, αλλά είναι πολύ γενική.
- Ο χρόνος εύρεσης απόστασης είναι γραμμικός ως προς τα σημεία, κάτι που **περιορίζει την κλιμάκωσή** (γιατί;) των κατηγοριοποιητών k πλησιέστερων γειτόνων. Μπορούν, όμως, να χρησιμοποιηθούν δομές καταλόγου (π.χ., kd-tree) για την επιτάχυνση της εύρεσης πλησιέστερων γειτόνων.
- Οι κατηγοριοποιητές k πλησιέστερων γειτόνων **δεν έχουν καλή ανοχή στο θόρυβο**, ιδιαίτερα για μικρές τιμές του k.

Χαρακτηριστικά Δένδρων Απόφασης

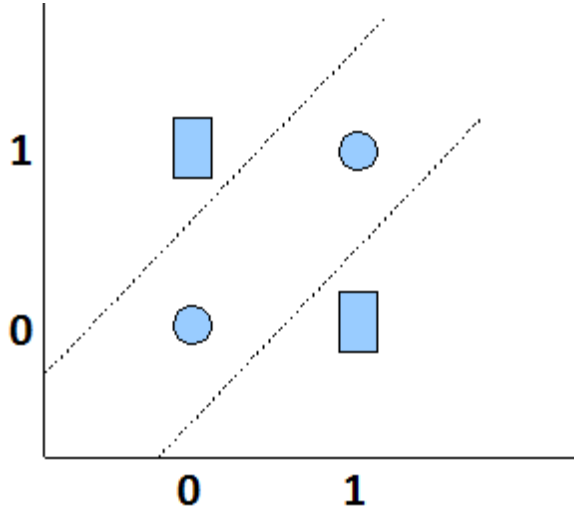
- Η κατασκευή του βέλτιστου δένδρου απόφασης απαιτεί αποτρεπτικό χρόνο (είναι NP-complete πρόβλημα).
  - Για το λόγο αυτό χρησιμοποιούνται ευρετικοί αλγόριθμοι, οι οποίοι είναι άπληστοι και δεν χρησιμοποιούν οπισθοδρόμηση.
  - Τα ευρετικά μειώνουν κατά πολύ το χρόνο κατασκευής.
  - Το αποτέλεσμα είναι ότι τα δένδρα απόφασης **κλιμακώνονται σε μεγάλους όγκους δεδομένων.**
- **Γρήγορη εφαρμογή.**
- Η **ακρίβεια πρόβλεψης** των δένδρων απόφασης είναι αποδεκτή για τις περισσότερες περιπτώσεις, συγκρίσιμη με την ακρίβεια άλλων κατηγοριοποιητών.
- Το μοντέλο που προκύπτει είναι πολύ **εύκολο στην κατανόηση**.
- Τα δένδρα απόφασης έχουν καλή **ανοχή στο θόρυβο.**
  - Ειδικά όταν εφαρμόζεται ψαλιδισμός.

Επιπλέον

- Τα ΔΑ μπορούν να διαχειριστούν πολυδιάστατα δεδομένα.
  - 1 διάσταση τη φορά χρησιμοποιείται κατά την ανάπτυξη του μοντέλου.
- … και κάθε τύπο μεταβλητών.
  - Συμβολικές, αριθμητικές, κλπ.

Μειονεκτήματα

- Αγνοούν εξαρτήσεις μεταξύ των ιδιοτήτων.
- Προβλήματα όταν λείπουν πολλά δεδομένα
- Διάσπαση ως προς μία ιδιότητα:
  - αντιστοίχιση με περιοχές, τα όρια των οποίων είναι παράλληλα με τους άξονες.



Αποτίμηση ακρίβειας

- Γνωρίζουμε 3 κατηγοριοποιητές.
- Πως συγκρίνουμε την επίδοσή τους ως προς την ακρίβεια;
- Πως μπορούμε να είμαστε σίγουροι για την ακρίβεια που θα έχει το μοντέλο μας;
  - Μέτρηση με αντικειμενικό τρόπο, που να αποκλείει προκατάληψη: 4 μέθοδοι.
  - Υπολογισμός στατιστικής σημαντικότητας.

Μέθοδοι μέτρησης ακρίβειας: Hold-out

- Χωρίζουμε το σύνολο δεδομένων σε δύο τμήματα:
  - το σύνολο εκμάθησης (π.χ., τα 2/3 πρώτα αντικείμενα)
  - και το σύνολο ελέγχου (π.χ. τα επόμενα 1/3).
- Δημιουργούμε μοντέλο σύμφωνα με το σύνολο εκμάθησης.
- Κατατάσσεται κάθε αντικείμενο του συνόλου ελέγχου.
- *X* είναι ο αριθμός που κατατάσσονται σωστά.
- *N* είναι ο συνολικός αριθμός των αντικειμένων στο σύνολο ελέγχου.
- Ακρίβεια: $A = \frac{X}{N}$
- Εξάρτηση από τη διάταξη των αντικειμένων.

Μέθοδοι μέτρησης ακρίβειας:
Τυχαία Υποδειγματοληψία-Random subsampling

- Αποφυγή εξάρτησης από διάταξη.
- Εφαρμόζουμε τυχαία δειγματοληψία *χωρίς* επανατοποθέτηση.
- Επιλέγουμε Ν αντικείμενα, τα οποία θέτουμε στο σύνολο ελέγχου.
- Τα εναπομείναντα αντικείμενα τα θέτουμε στο σύνολο εκμάθησης.
- Επανάληψη της διαδικασίας αυτής *k* φορές.
- Αν X_i είναι ο αριθμός των σωστά καταταγμένων αντικειμένων στην *i*-οστή επανάληψη:

$$A = \frac{1}{k} \sum_{i=1}^{k} \frac{X_i}{N}$$

Μέθοδοι μέτρησης ακρίβειας: Δια-εγκυροποίηση - Cross validation

- Για να μην λαμβάνουμε τυχαία τα δείγματα:
- *M* ο αριθμός των αντικειμένων στο σύνολο δεδομένων.
- Θέλουμε *k* επαναλήψεις.
- Χωρίζουμε το σύνολο σε *k* τμήματα με *M*/*k* (διαδοχικά) αντικείμενα το κάθε ένα.
- Στην *i*-οστή επανάληψη, το *i*-οστό τμήμα λειτουργεί ως σύνολο ελέγχου, ενώ τα υπόλοιπα *k*-1 τμήματα απαρτίζουν το σύνολο εκμάθησης.
- Μία συχνά χρησιμοποιούμενη τιμή του *k* είναι το 10
- Η μέθοδος 10-fold cross-validation θεωρείται ως μία από τις πιο αξιόπιστες για την αποτίμηση της ακρίβειας κατηγοριοποιητών.
- Στην ειδική περίπτωση που *k=M*, τότε η μέθοδος ονομάζεται leave-one-out
  - χρησιμοποιείται μόνο για μικρά σύνολα δεδομένων.

Μέθοδοι μέτρησης ακρίβειας: bootstrap (1/2)

- Δειγματοληψία *με επανατοποθέτηση.*
- *M* αντικείμενα → 0.632*M* στο δείγμα.
- Δειγματοληπτούμε *M* φορές.
- Σε κάθε προσπάθεια: P(επιλογή *x*) = 1/*M.*
- Σε *M* προσπάθειες:

$$\left(\left(\left(\right)\right)\right)$$

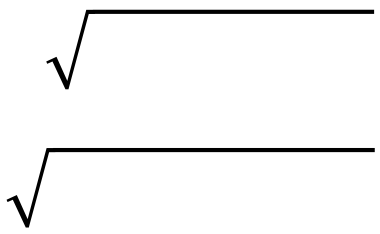
Μέθοδοι μέτρησης ακρίβειας: bootstrap (2/2)

- Τα 0.632M αντικείμενα αποτελούν το σύνολο εκμάθησης.
- Επανάληψη k φορές της διαδικασίας.
- Αν α_i είναι η ακρίβεια στην i-οστή επανάληψη και α η ακρίβεια όταν σύνολο ελέγχου = σύνολο εκμάθησης = σύνολο:

$$\bar{\alpha} =$$

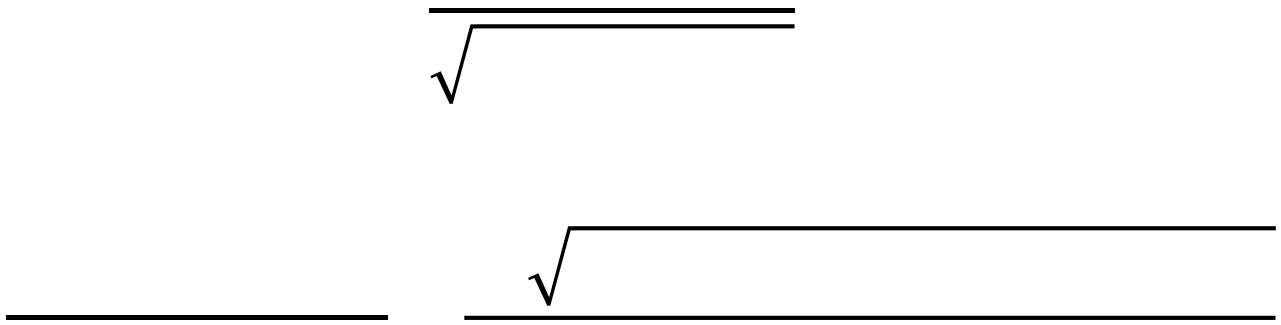
Διαστήματα εμπιστοσύνης για την ακρίβεια πρόβλεψης (1/2)

- Χ από Ν αντικείμενα κατατάχθηκαν σωστά.
- Χ τυχαία μεταβλητή με διωνυμική κατανομή.
- p η πραγματική ακρίβεια πρόβλεψης.
- A = X/N τυχαία μεταβλητή (διωνυμική).



Διαστήματα εμπιστοσύνης για την ακρίβεια πρόβλεψης (2/2)

- Για $N \rightarrow \infty$ $A = X/N$ ακολουθεί κανονική κατανομή (νόμος μεγάλων αριθμών).
- Σε επίπεδο εμπιστοσύνης α:



Ενδεικτικές τιμές

- Έστω ένα μοντέλο με ακρίβεια 80% όταν χρησιμοποιούνται 100 δείγματα για έλεγχο:
  - N=100, A = 0.8
  - α = 0.95 (διάστημα εμπιστοσύνης 95%)
  - Από τον διπλανό πίνακα, Z_{α}=1.96

α	Z
0.99	2.58
0.98	2.33
0.95	1.96
0.90	1.65

N	**50**	**100**	**500**	**1000**	**5000**
p(lower)	0.670	0.711	0.763	0.774	0.789
p(upper)	0.888	0.866	0.833	0.824	0.811

Παράδειγμα

- Έστω ένα σύνολο ελέγχου με N=50 αντικείμενα. Αν η εκτιμώμενη ακρίβεια αποτιμήθηκε ίση με A=85%, να βρεθεί το διάστημα εμπιστοσύνης σε επίπεδο a=0.95 και a=0.90.
- Για a=0.95 ισχύει z_a=1.96. Από την Εξίσωση:

$$\frac{\qquad}{\qquad} \quad \frac{\sqrt{\qquad\qquad\qquad\qquad}}{\qquad\qquad\qquad\qquad}$$

  - προκύπτει ότι p=0.825 ± 0.099. Άρα, αναμένουμε η πραγματική ακρίβεια p να κυμαίνεται μεταξύ 0.726 και 0.924.
- Για a=0.90 ισχύει z_a=1.65. Από την ίδια εξίσωση προκύπτει ότι p=0.832 ± 0.083. Άρα, αναμένουμε η πραγματική ακρίβεια p να κυμαίνεται μεταξύ 0.749 και 0.915.

Βελτίωση της ακρίβειας

- Ξέρουμε πώς να εκτιμούμε σωστά την ακρίβεια.
- Μπορούμε να βελτιώσουμε την ακρίβεια χρησιμοποιώντας διαφορετικά τους γνωστούς μας κατηγοριοποιητές;
  - Κλάδεμα.
  - Σύνολα κατηγοριοποιητών.

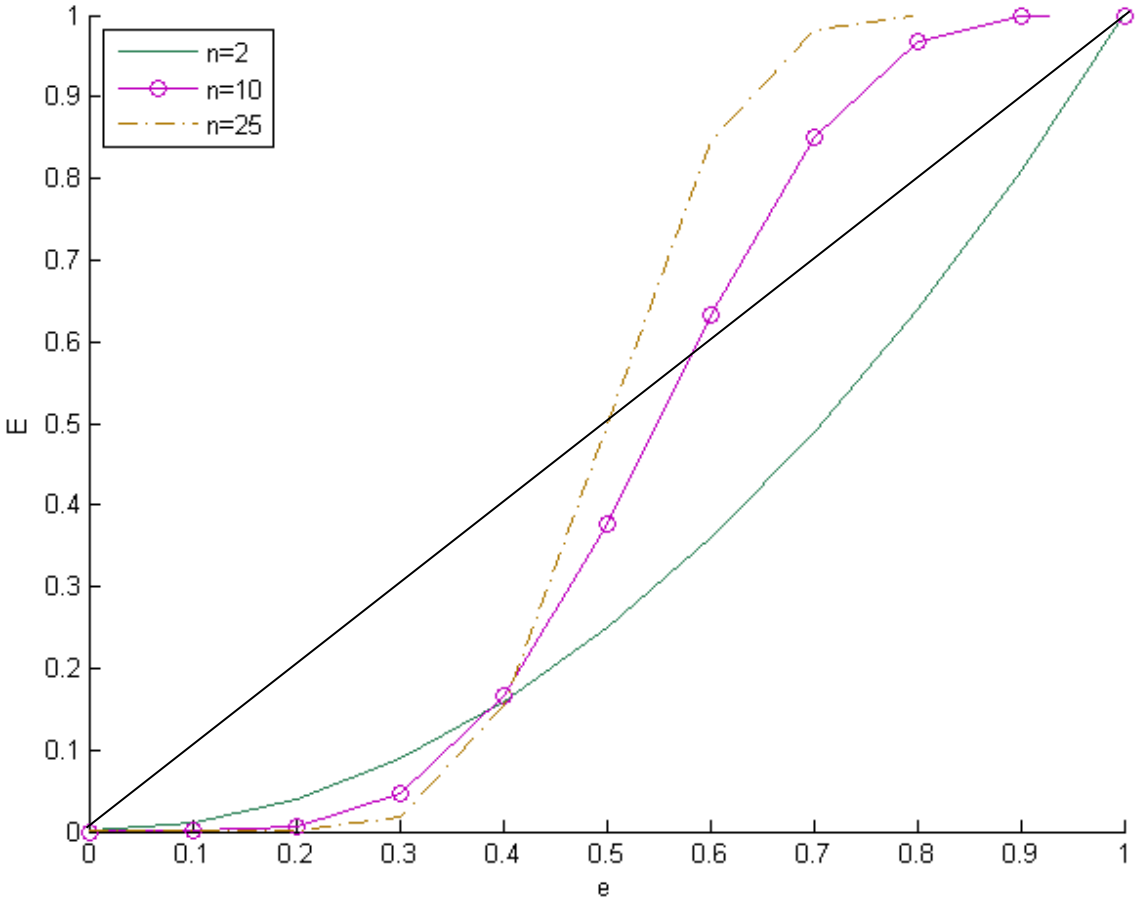
Σύνολα κατηγοριοποιητών

- *n* δυαδικοί (2 κλάσεις) ανεξάρτητοι κατηγοριοποιητές.
- Κάθε ένας έχει πιθανότητα λάθους *e.*
- Αποφασίζουμε την κλάση που λέει η πλειοψηφία των κατηγοριοποιητών.
- Για να γίνει λάθος, περισσότεροι από *n*/2 να κάνουν λάθος.
- Εκτιμώμενο λάθος για το σύνολο:

$$\lfloor - \rfloor$$

Σύνολα κατηγοριοποιητών

- n = 10
- e = 0.2
- E = 0.0064



Bagging (Bootstrap AGGragatING)

- *k* δείγματα *με επανατοποθέτηση* (διαδικασία bootstrap).
  - σύνολα δεδομένων για *k* κατηγοριοποιητές (ίδιος αλγόριθμος κατασκευής).
  - αναμενόμενος αριθμός διακριτών αντικειμένων στο κάθε δείγμα: 63.2% του αρχικού.
- Ένα νέο αντικείμενο κατατάσσεται με καθέναν από *k* κατηγοριοποιητές.
  - Το αναθέτουμε κλάση που πλειοψηφεί.
- Η μέθοδος bagging βελτιώνει την ακρίβεια, όταν υπάρχει διακύμανση στην ακρίβεια των *k* κατηγοριοποιητών.

Σημείωμα Αναφοράς

Διαθέσιμο από τη δικτυακή διεύθυνση:http://eclass.auth.gr/courses/OCRS182/

Σημείωμα Αδειοδότησης

Τέλος ενότητας

Επεξεργασία: Ανδρέας Κοσματόπουλος

Θεσσαλονίκη, Χειμερινό Εξάμηνο 2013-2014

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημειώματα

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.