

5

Δίνεται ένα σύνολο κανόνων με 200 θετικά δείγματα και 300 αρνητικά. Για καθένα από τους ακόλουθους κανόνες

R1 : A \rightarrow + (καλύπτει 5 θετικά και 2 αρνητικά δείγματα),

R2 : B \rightarrow + (καλύπτει 40 θετικά και 20 αρνητικά δείγματα),

R3 : C \rightarrow + (καλύπτει 80 θετικά και 80 αρνητικά δείγματα),

υπολογίστε α) την ορθότητά του (accuracy) και β) το κέρδος πληροφορίας (information gain) FOIL του.

(4 Points)

Enter your answer

Θετικά & αρνητικά δείγματα, accuracy, gain (rule based classification) ⑤

R1: 5/7 accuracy, coverage 7/500, $P_1=5, n_1=2$

R2: 40/60 accuracy, coverage 60/500, $P_2=40, n_2=60$

R3: 80/160 " " " 160/500, $P_3=80, n_3=160$

R0: $\{ \} \rightarrow$ positive

coverage $\frac{500}{500}$

accuracy = $\frac{200}{500}$

$P_0=200, n_0=300$

Info Gain Foil R1

$$\text{Gain}(R_0, R_1) = P_1 \times \left[\log_2 \left(\frac{P_1}{P_1 + n_1} \right) - \log_2 \left(\frac{P_0}{P_0 + n_0} \right) \right]$$

$$= 5 \times \left[\log_2 \left(\frac{5}{5+2} \right) - \log_2 \left(\frac{200}{200+300} \right) \right]$$

όμοιος για gains (R_1, R_2)
 (R_2, R_3)

Max Gain \rightarrow κριτήριο επιλογής

Youtube info Gain foil 10 vid.

1ΔΙΟ ΘΕΜΑ ΑΛΛΑ ΝΟΥΜΕΡΑ ΑΝΑΛΥΤΙΚΟΤΕΡΗ ΛΥΣΗ

Rule-based classification example

(a) In rule based classification, consider a training set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules:

$R_1: A \rightarrow +$ (covers 4 positive and 1 negative examples)

$R_2: B \rightarrow +$ (" 30 " " 10 " ")

$R_3: C \rightarrow +$ (" 100 " " 90 " ")

determine which is the best and worst candidate rule according to:

(i) Rule accuracy, (ii) FOIL's information gain.

(b) Explain why rule accuracy and FOIL's information gain rank the rules differently.

ANSW

(a)(i) $R_1: A \rightarrow +$
 ↑ ↑
 antecedent consequent

$$\text{Coverage of rule } R_1 = \frac{\text{\# of records that satisfy antecedent}}{\text{\# of total records/samples}}$$

$$\text{Accuracy of rule } R_1 = \frac{\text{\# of records that satisfy antecedent AND consequent}}{\text{\# of records that satisfy (only) antecedent}}$$

$$R_1: A \rightarrow + : \text{Accuracy}_{R_1} = \frac{4}{4+1} = \frac{4}{5} = 80\%$$

$$R_2: B \rightarrow + : \text{Accuracy}_{R_2} = \frac{30}{30+10} = \frac{3}{4} = 75\%$$

$$R_3: C \rightarrow + : \text{Accuracy}_{R_3} = \frac{100}{100+90} = \frac{10}{19} = 52.6\%$$

→ according to accuracy, the best candidate rule is R_1 and the worst is R_3 .

(ii) FOIL's info gain : $\left. \begin{array}{l} R_0: \{ \} \rightarrow \text{class} \\ R_i: \{ A \} \rightarrow \text{class} \end{array} \right\} \text{Gain}(R_0, R_i) = t \cdot \left(\log_2 \frac{p_i}{p_i + n_i} - \log_2 \frac{p_0}{p_0 + n_0} \right)$

t : number of positive instances covered by both R_0 and R_i .

p_0 : number " " " " by R_0

n_0 : number of negative instances covered by R_0

p_i : " " " " " " R_i

n_i : " " " " " " R_i

αυτό θα είναι ίσο με p_i
 → αυτό προκύπτει είναι το accuracy.

$$R_1: \text{Gain}(R_0, R_1) = 4 \cdot \left[\log_2 \left(\frac{4}{4+1} \right) - \log_2 \left(\frac{100}{100+400} \right) \right] = 4 \cdot \left(\log_2 \frac{4}{5} - \log_2 \frac{1}{5} \right) = 4 \log_2 4 = 8$$

$$R_2: \text{Gain}(R_0, R_2) = 30 \cdot \left[\log_2 \frac{30}{30+10} - \log_2 \frac{100}{500} \right] = 30 \cdot \left(\log_2 \frac{3}{4} - \log_2 \frac{1}{5} \right) \approx 57.2$$

$$R_3: \text{Gain}(R_0, R_3) = 100 \cdot \left[\log_2 \frac{100}{100+90} - \log_2 \frac{100}{500} \right] = 100 \cdot \left(\log_2 \frac{10}{19} - \log_2 \frac{1}{5} \right) \approx 139.6$$

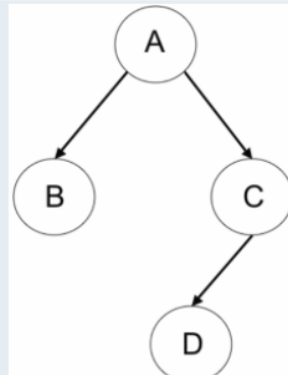
Therefore, based on information gain: R_3 : best, R_1 : worst. which is the opposite of what we found on (i).

(b) Rule accuracy only accounts for the portion/percentage of samples that satisfy the antecedent as well as the consequent. out of the ones that satisfy the antecedent.

However, such a rule can concern only a small portion of the total samples of the dataset and hence it might not be a particularly useful rule for splitting the data.

FOIL's information gain on the other hand has a weight (t) or a multiplier that expresses this dependence on how many samples are satisfying the rule's antecedent, and thus is a better ~~indicator~~ metric for rule ranking especially for the top-levels of a rule-based classifier.

6



Δίνεται το Μπεϋζιανό δίκτυο πεποίθησης του σχήματος και οι εξείς πιθανότητες: $P(A)=0.3$, $P(B|A)=0.7$, $P(B|\sim A)=0.5$, $P(C|A)=0.2$, $P(C|\sim A)=0.6$, $P(D|C)=0.7$, $P(D|\sim C)=0.7$.
Υπολογίστε την πιθανότητα $P(A|B)$

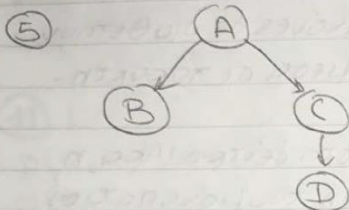
(5 Points)

Αν η εικόνα δεν εμφανίζεται δείτε την στο <https://imgur.com/zgg4eDk>

Enter your answer

DATA MINING

ΟΜΑΔΑ Ε



$$P(A) = 0.3$$

$$P(B|A) = 0.7 \quad P(B|\sim A) = 0.5$$

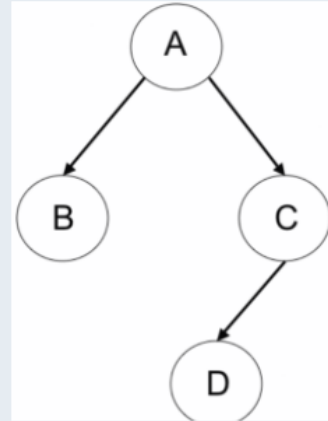
$$P(C|A) = 0.2 \quad P(C|\sim A) = 0.6$$

$$P(D|C) = 0.7 \quad P(D|\sim C) = 0.7$$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{0.7 \cdot 0.3}{P(B)} = \frac{0.21}{P(B)} \quad (1)$$

$$\begin{aligned} P(B) &= P(B|A) \cdot P(A) + P(B|\sim A) \cdot P(\sim A) \\ &= 0.7 \cdot 0.3 + 0.5 \cdot (1 - 0.3) \\ &= 0.21 + 0.35 = 0.56 \quad (2) \end{aligned}$$

$$(1), (2) \Rightarrow P(A|B) = \frac{0.21}{0.56} = 0.375$$



Δίνεται το Μπεϋζιανό δίκτυο πεποίθησης του σχήματος και οι εξείς πιθανότητες: $P(A)=0.3$, $P(B|A)=0.7$, $P(B|\sim A)=0.5$, $P(C|A)=0.2$, $P(C|\sim A)=0.6$, $P(D|C)=0.7$, $P(D|\sim C)=0.7$.

Υπολογίστε την πιθανότητα $P(C|B)$

(5 Points)

Αν η εικόνα δεν εμφανίζεται δείτε την στο <https://imgur.com/zag4eDk>

⑦

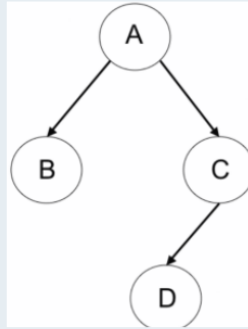
$P(A)=0.3$
 $P(B|A)=0.7$ $P(B|\sim A)=0.5$
 $P(C|A)=0.2$ $P(C|\sim A)=0.6$
 $P(D|C)=0.7$ $P(D|\sim C)=0.7$

$P(A,B)=P(B|A) \cdot P(A) = 0.7 \cdot 0.3 = 0.21$
 $P(A,C)=P(C|A) \cdot P(A) = 0.2 \cdot 0.3 = 0.06$
 $P(B) = P(B|A) \cdot P(A) + P(B|\sim A)P(\sim A)$
 $= 0.21 + 0.5 \cdot 0.7 = 0.56$

Όμως B και C conditionally independent given A (fork), δηλαδή

$P(B,C|A) = P(B|A) \cdot P(C|A) = 0.7 \cdot 0.2 = 0.14$
 $P(B,C|\sim A) = P(B|\sim A) \cdot P(C|\sim A) = 0.5 \cdot 0.6 = 0.3$
 $P(B,C) = P(B,C|A) \cdot P(A) + P(B,C|\sim A)P(\sim A)$
 $= 0.14 \cdot 0.3 + 0.3 \cdot 0.7$
 $= 0.042 + 0.21 = 0.252$

$P(C|B) = \frac{P(B,C)}{P(B)} = \frac{0.252}{0.56} = 0.45$



Δίνεται το Μπεϋζιανό δίκτυο πεποίθησης του σχήματος και οι εξείς πιθανότητες: $P(A)=0.3$, $P(B|A)=0.7$, $P(B|\sim A)=0.5$, $P(C|A)=0.2$, $P(C|\sim A)=0.6$, $P(D|C)=0.7$, $P(D|\sim C)=0.7$.
Υπολογίστε την πιθανότητα $P(B|D)$
(5 Points)

Αν η εικόνα δεν εμφανίζεται δείτε την στο <https://imgur.com/z9g4eDk>

9

$P(A)=0.3$
 $P(B|A)=0.7$ $P(B|\sim A)=0.5$
 $P(C|A)=0.2$ $P(C|\sim A)=0.6$
 $P(D|C)=0.7$ $P(D|\sim C)=0.7$

$P(B|D) = \frac{P(B,D)}{P(D)}$
 (σχέση ότι B και D independent given A (fork)
 ~~$P(A,B,C,D) = P(A) \cdot P(B|A) \cdot P(C|A) \cdot P(D|C)$~~
 A και D independent given C (chain)

$P(A,B,C,D) = \underbrace{P(A) \cdot P(B|A) \cdot P(C|A)}_{P(ABC)} \cdot P(D|C)$
 (child/parent)

$P(B,D) = \sum_A \sum_C P(A,B,C,D)$

$P(D) = \sum_C P(D|C)$

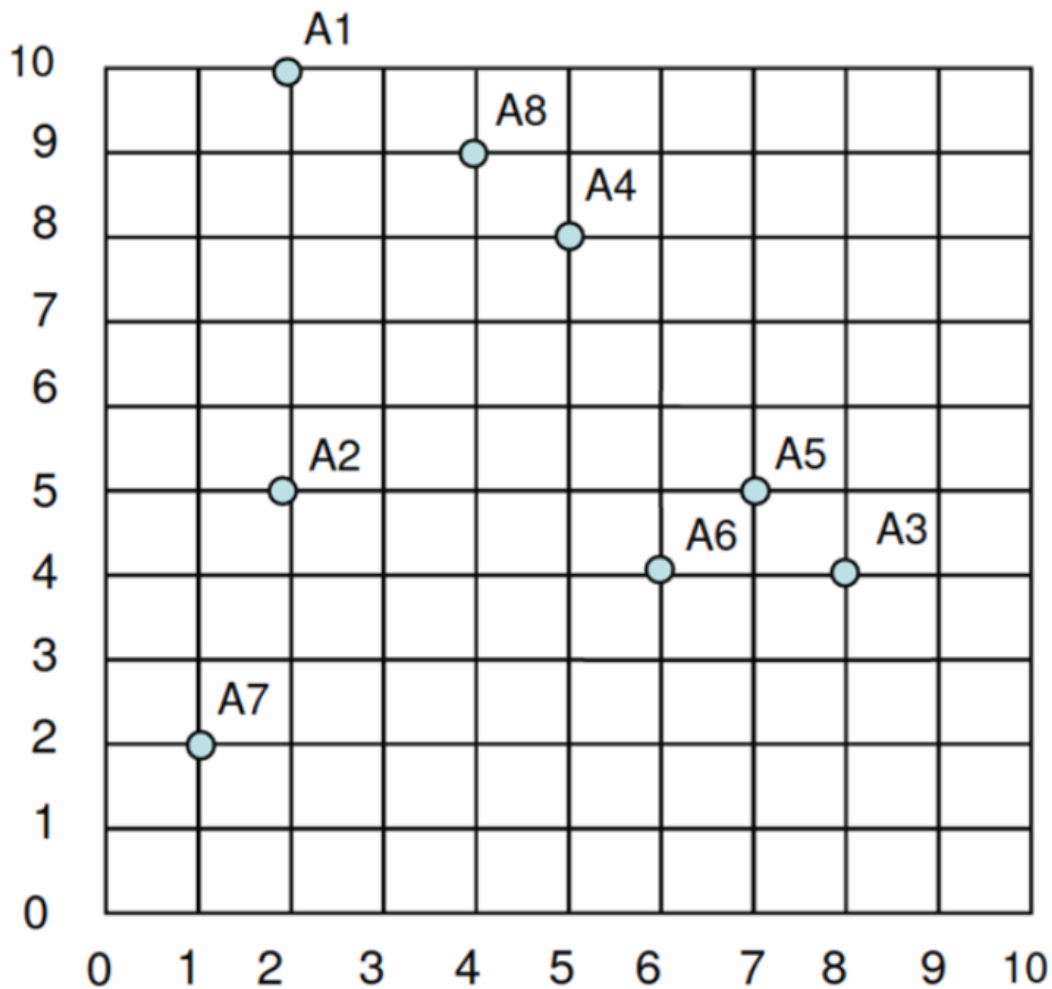
$A=0, C=0 \quad (1-P(A)) \cdot P(B|\sim A) \cdot P(\sim C|\sim A) \cdot P(D|\sim C)$
 $A=0, C=1 \quad \dots$
 $A=1, C=0 \quad \dots$
 $A=1, C=1 \quad \dots$

$P(D) = P(D|C) \cdot P(C) + P(D|\sim C) \cdot P(\sim C)$
 $P(C) = P(C|A) \cdot P(A) + P(C|\sim A) \cdot P(\sim A)$

ΕΛΠ ΕΛΠ

Δίνονται τα 8 σημεία (A1 ως A8) στον δισδιάστατο χώρο, τα οποία τα συσταδοποιούμε με τη χρήση του αλγορίθμου DBSCAN με $\text{MinPts}=3$ και $\text{Eps}=\sqrt{5}$. Να γράψετε ποιες συστάδες σχηματίζονται, αν υπάρχουν οριακά σημεία (border points), ποια είναι και σε ποια συστάδα είναι οριακά καθώς και αν υπάρχουν σημεία θορύβου (noise points) και ποια είναι αυτά. (5 Points)

Αν η εικόνα δεν εμφανίζεται, δείτε την στο <https://imgur.com/b4TgyJy.png>.



1) $A \rightarrow 60\%$ κέρδη

$B \rightarrow 50\%$ κέρδη

$A \cap B \rightarrow 30\%$

$A \cup B \rightarrow A + B - A \cap B$

$\rightarrow 60 + 50 - 30 \rightarrow 80\%$

①

2) $\overline{E_{OTW}}$ or $\leq \sqrt{5}$ for outlier

Example

$C1 = (A1, A3, A4)$

• $\geq \text{Min pts} = \text{Core pts}$

• $\text{border} = (\text{points } \leq \text{Min pts})$

border

Core (Min=3)

border (Min=2)

• All the noise

$C2 = (A6, A5, A7)$

all core, none border

$A2, A7$ outliers

Δες youtube dbscan example 2ο βήμα

Καταζω ομαδικά τα A_{xi}

• Αν κύκλος με κέντρο A_{xi} και ακτίνα ϵ έχει $\geq \text{Min pts}$ τότε Core

Άλλως ni border ni noise

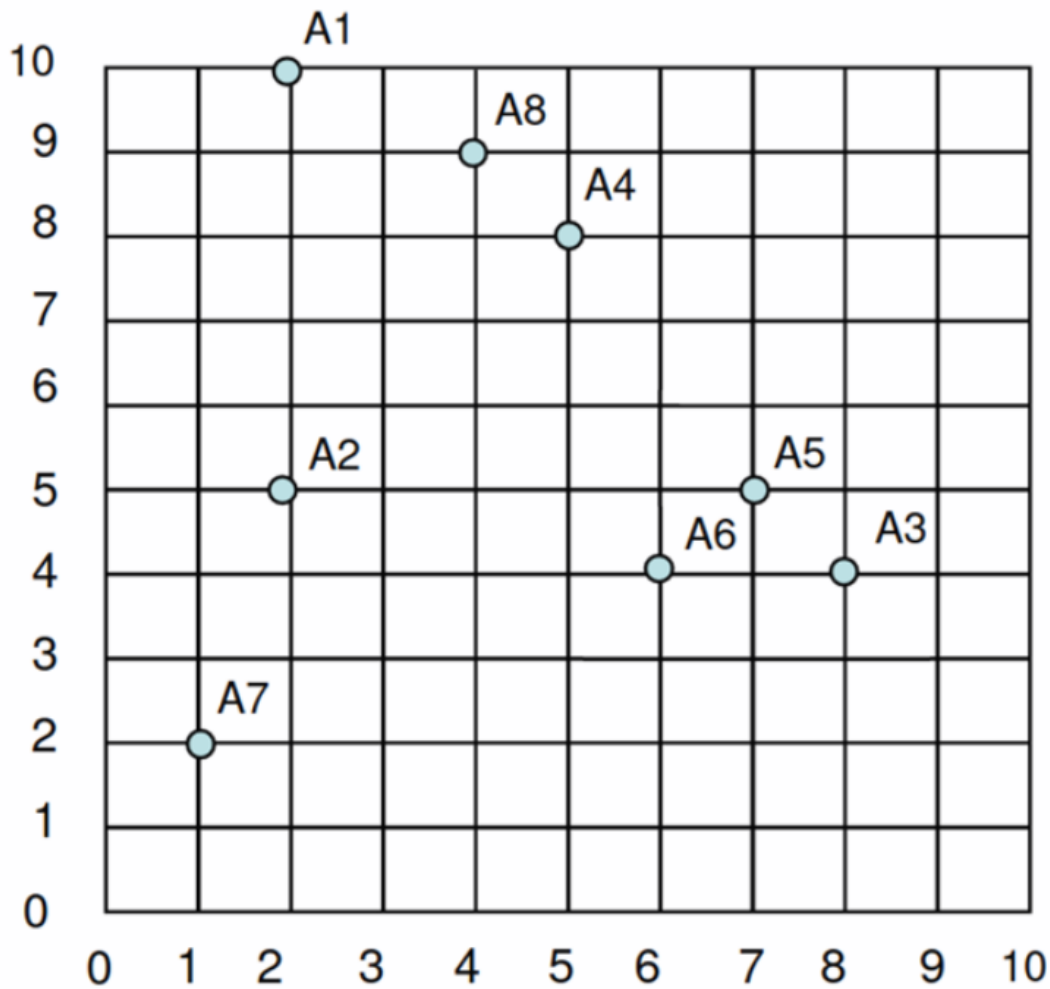
Μόλις βρω όλα τα cores καταζω

• αν τα (border/noise) ανήκουν σε κύκλο με Core point \Rightarrow τότε ανήκουν border

• αν όχι, τότε noise \neq outlier

Δίνονται τα 8 σημεία (A1 ως A8) στον διδιάστατο χώρο, τα οποία τα συσταδοποιούμε με τη χρήση του αλγορίθμου DBSCAN με $\text{MinPts}=2$ και $\text{Eps}=\sqrt{10}$. Να γράψετε ποιες συστάδες σχηματίζονται, αν υπάρχουν οριακά σημεία (border points), ποια είναι και σε ποια συστάδα είναι οριακά καθώς και αν υπάρχουν σημεία θορύβου (noise points) και ποια είναι αυτά. (5 Points)

Αν η εικόνα δεν εμφανίζεται, δείτε την στο <https://imgur.com/b4TqyJy.png>



Enter your answer

$A_1: 2 \text{ pts} = \text{Min pts} \rightarrow \text{Core}$
 $A_8: 3 \text{ pts} \geq \text{Min pts} \rightarrow \text{Core}$
 $A_4: 2 \text{ pts} = \text{Min pts} \rightarrow \text{Core}$

$\notin \emptyset \text{ border} = C_1(A_1, A_8, A_4)$

$A_2, A_7 = \text{Min pts}$ και τα δύο, άρα και τα δύο Core.
 $\notin \emptyset \text{ border}, C_2(A_2, A_7)$

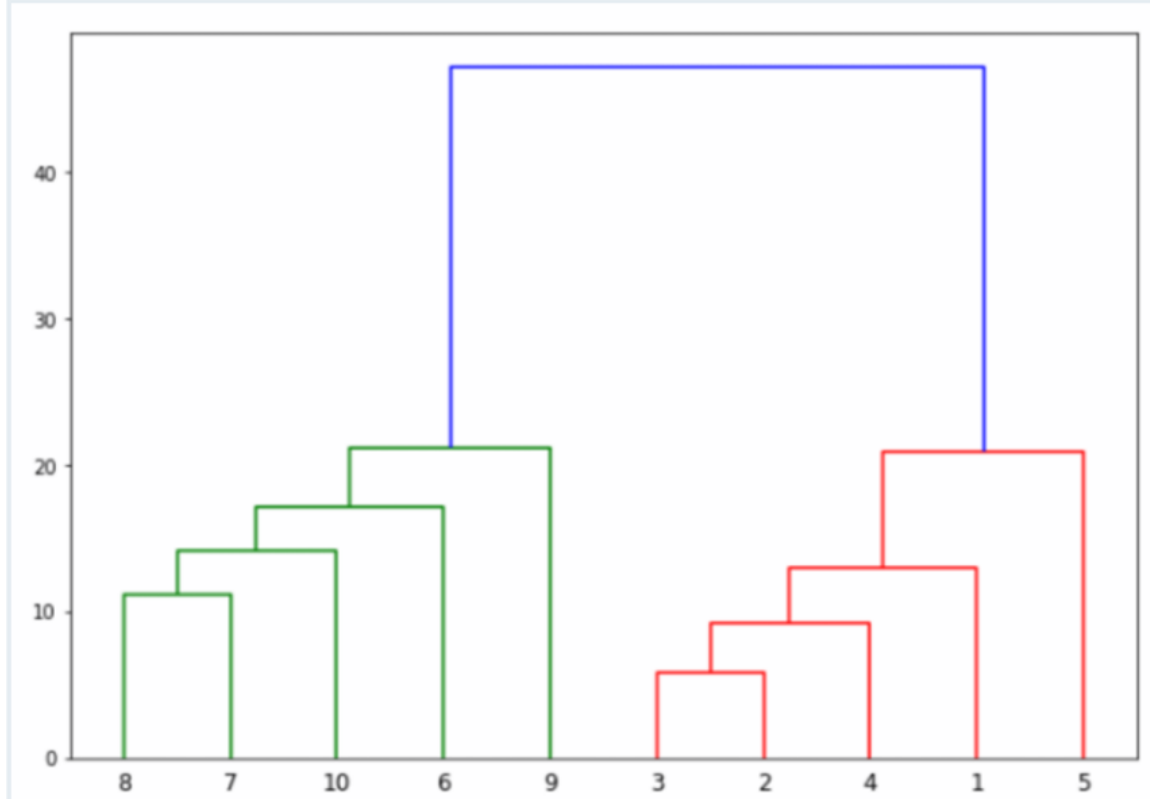
$A_5, A_6, A_3 \geq \text{Min pts}$ άρα $\emptyset \text{ border}$
 $\text{Άρα } C_3(A_3, A_5, A_6)$


$A_9 \text{ noise}$ ε ότι $= \sqrt{10} \leadsto$ ~~δεν ανήκει~~ το σημείο στο κελό

τότε $A_2, A_7 \rightarrow \text{noise}$

Δίνεται το δενδρόγραμμα ενός συσσωρευτικού ιεραρχικού αλγορίθμου συσταδοποίησης (agglomerative hierarchical clustering) δέκα σημείων. Ποια από τις παρακάτω συστάδες σχηματίζεται τελευταία;
(5 Points)

Αν δεν εμφανίζεται η εικόνα, δείτε την στο <https://i.imgur.com/YfoqshN.png>



 Καμία από τις υπόλοιπες απαντήσεις

- ☐ Η συστάδα των σημείων 8,7
- ☐ Η συστάδα των σημείων 3,2,4,1
- ☐ Η συστάδα των σημείων 3,2,4
- ☐ Η συστάδα των σημείων 8,7,10,6
- ☐ Δεν απαντώ

ΚΑΝΟΝΙΚΑ ΟΙ ΤΕΛΕΥΤΑΙΕΣ ΣΥΣΤΑΔΕΣ ΠΟΥ ΣΧΗΜΑΤΙΖΟΝΤΑΙ ΕΙΝΑΙ ΟΙ 8,7,10,6,9 ΚΑΙ 3,2,4,1,5. ΕΠΕΙΔΗ ΣΧΗΜΑΤΙΖΟΝΤΑΙ ΤΑΥΤΟΧΡΟΝΑ Η ΣΩΣΤΗ ΑΠΑΝΤΗΣΗ ΕΙΝΑΙ ΤΟ "ΚΑΜΙΑ ΑΠΟ ΤΙΣ ΥΠΟΛΟΙΠΕΣ".

ΑΝ ΔΕΝ ΥΠΗΡΧΕ ΑΥΤΗ Η ΕΠΙΛΟΓΗ ΚΑΙ ΣΟΝΙ ΚΑΙ ΝΤΕ ΘΕΛΑΜΕ ΝΑ ΒΡΟΥΜΕ ΤΗ ΣΥΣΤΑΔΑ ΠΟΥ ΣΧΗΜΑΤΙΖΕΤΑΙ ΤΕΛΕΥΤΑΙΑ ΑΠΟ ΤΙΣ ΠΑΡΑΚΑΤΩ ΕΠΙΛΟΓΕΣ, ΑΡΚΕΙ ΝΑ ΔΟΥΜΕ ΤΟ ΥΨΟΣ ΤΗΣ ΣΤΟΝ ΑΞΟΝΑ Υ. ΟΣΟ ΠΙΟ ΨΗΛΑ ΒΡΙΣΚΕΤΑΙ, ΤΟΣΟ ΠΙΟ ΜΕΤΕΠΕΙΤΑ ΔΗΜΙΟΥΡΓΗΘΗΚΕ ΑΡΑ ΘΑ ΑΠΑΝΤΟΥΣΑΜΕ 8,7,10,6

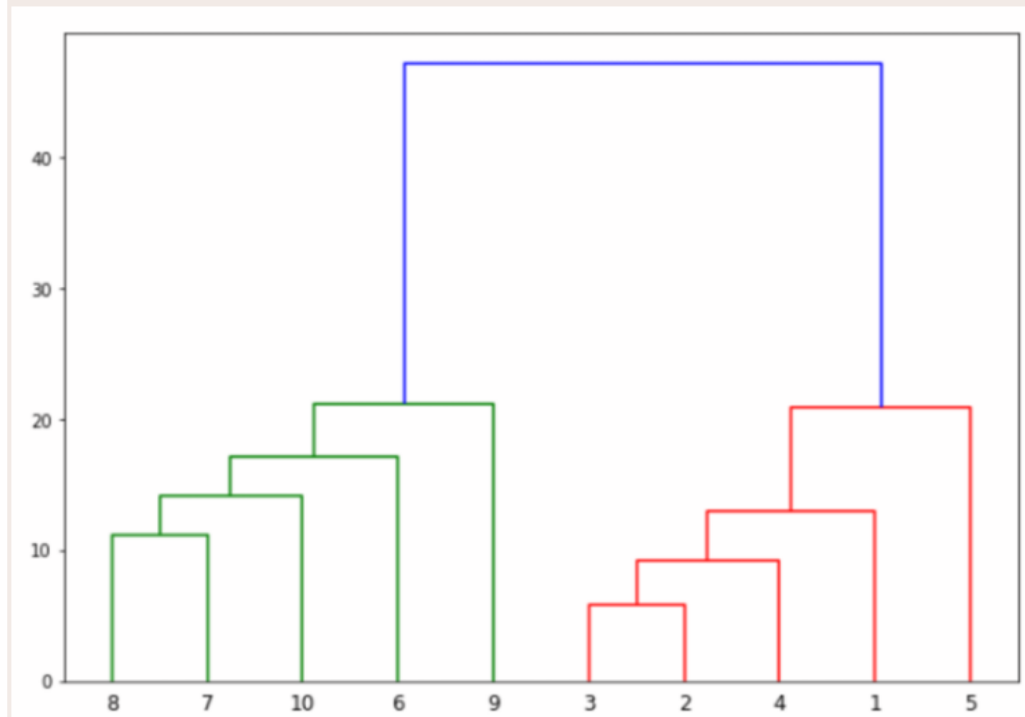
?

14

Δίνεται το δενδρόγραμμα ενός συσσωρευτικού ιεραρχικού αλγορίθμου συσταδοποίησης (agglomerative hierarchical clustering) δέκα σημείων. Ποια από τις παρακάτω συστάδες σχηματίζεται πρώτη;

(5 Points)

Αν δεν εμφανίζεται η εικόνα, δείτε την στο <https://i.imgur.com/YfoqshN.png>.



☐ Δεν απαντώ

☒ Η συστάδα των σημείων 3,2,4

☐ Καμία από τις υπόλοιπες απαντήσεις

☐ Η συστάδα των σημείων 8,7

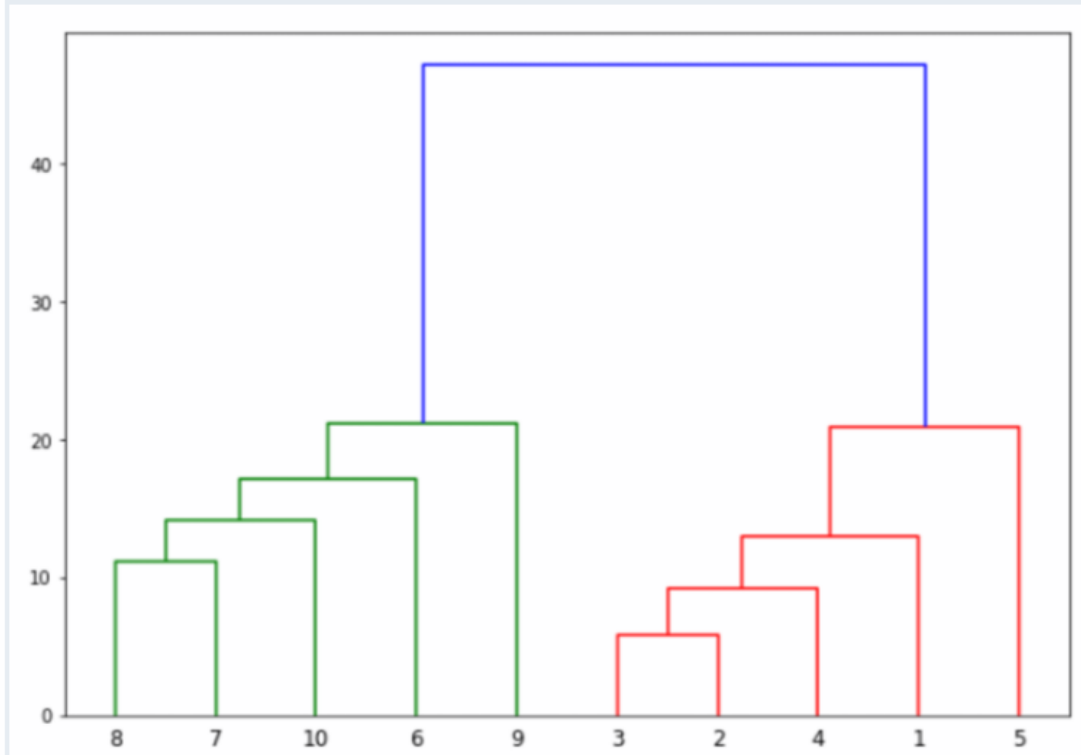
☐ Η συστάδα των σημείων 3,2,4,1

☐ Η συστάδα των σημείων 8,7,10,6

Δίνεται το δενδρόγραμμα ενός διαχωριστικού ιεραρχικού αλγορίθμου συσταδοποίησης (divisive hierarchical clustering) δέκα σημείων. Ποια από τις παρακάτω συστάδες σχηματίζεται πρώτη;
(5 Points)

SUS - top/down

Αν δεν εμφανίζεται η εικόνα, δείτε την στο <https://i.imgur.com/YfoqshN.png>



- ☐ Καμία από τις υπόλοιπες απαντήσεις
- ☐ Δεν απαντώ
- ☒ Η συστάδα των σημείων 8,7,10,6
- ☐ Η συστάδα των σημείων 8,7
- ☐ Η συστάδα των σημείων 3,2,4
- ☐ Η συστάδα των σημείων 3,2,4,1

9

Εκτελούμε τον αλγόριθμο Expectation-Maximization για ένα μοντέλο μίξης δύο γκαουσιανών κατανομών με ίδια βάρη για ένα σύνολο 3 δειγμάτων δεδομένων και σε κάποιο βήμα της επανάληψης έχουμε $p(x_1 = -2|\theta_1) = A$, $p(x_1 = -2|\theta_2) = B$, $p(x_2 = 1|\theta_1) = C$, $p(x_2 = 1|\theta_2) = D$, $p(x_3 = 3|\theta_1) = E$, $p(x_3 = 3|\theta_2) = F$. Έστω $A=0.3$, $B=0.7$, $C=0.45$, $D=0.55$, $E=0.2$, και $F=0.8$. Να υπολογίσετε την πιθανότητα το κάθε δείγμα δεδομένων να προέρχεται από κάθε μια από τις δύο κατανομές καθώς και τις μέσες τιμές των δύο κατανομών μετά την ολοκλήρωση αυτού του βήματος (στο διάστημα $[0, 1]$, με ακρίβεια 2 δεκαδικών ψηφίων)
(5 Points)

Enter your answer

ΜΙΛΑΕΙ ΓΙΑ ΙΔΙΑ ΒΑΡΗ, ΑΡΑ ΙΣΟΠΙΘΑΝΑ ΑΡΑ ΠΑΝΤΟΥ $P(\theta_1) = P(\theta_2) = 0.5$

ΓΙΑ ΤΗΝ 1Η ΚΑΤΑΝΟΜΗ ΨΑΧΝΩ 3 ΠΙΘΑΝΟΤΗΤΕΣ

$P(x_1, \theta_1) = P(x_1|\theta_1) * (P(x_1, \theta_1) + P(\sim x_1, \theta_1)) = 0.3 * P(\theta_1)$, ΑΦΟΥ

$P(x) = P(x|\theta_1) * P(\theta_1) + P(x|\theta_2) * P(\theta_2)$, ΑΡΑ ΖΗΤΟΥΜΕΝΟ $P = 0.3 * 0.5$

ΟΜΟΙΩΣ ΓΙΑ ΤΙΣ ΑΛΛΕΣ 2 ΠΙΘΑΝΟΤΗΤΕΣ

ΟΜΟΙΩΣ ΓΙΑ 3 ΠΙΘΑΝΟΤΗΤΕΣ 2ΗΣ ΚΑΤΑΝΟΜΗΣ

(ΠΑΝΤΟΥ $P(\theta_1) = P(\theta_2) = 0.5$)

11) ΕΜ

$p(x_1 = -2|\theta_1) = A = 0.3$
 $p(x_1 = -2|\theta_2) = B = 0.7$
 $p(x_2 = 1|\theta_1) = C = 0.45$
 $p(x_2 = 1|\theta_2) = D = 0.55$
 $p(x_3 = 3|\theta_1) = E = 0.2$
 $p(x_3 = 3|\theta_2) = F = 0.8$

ΜΙΛΑΕΙ ΓΙΑ ΙΔΙΑ ΒΑΡΗ, ΑΡΑ ΙΣΟΠΙΘΑΝΑ ΑΡΑ ΠΑΝΤΟΥ $P(\theta_1) = P(\theta_2) = 0.5$
 ΓΙΑ ΤΗΝ 1Η ΚΑΤΑΝΟΜΗ ΨΑΧΝΩ 3 ΠΙΘΑΝΟΤΗΤΕΣ
 $P(x_1, \theta_1) = P(x_1|\theta_1) * (P(x_1, \theta_1) + P(\sim x_1, \theta_1)) = 0.3 * P(\theta_1)$, ΑΦΟΥ
 $P(x) = P(x|\theta_1) * P(\theta_1) + P(x|\theta_2) * P(\theta_2)$, ΑΡΑ ΖΗΤΟΥΜΕΝΟ $P = 0.3 * 0.5$

ΟΜΟΙΩΣ ΓΙΑ ΤΙΣ ΑΛΛΕΣ 2 ΠΙΘΑΝΟΤΗΤΕΣ

ΟΜΟΙΩΣ ΓΙΑ 3 ΠΙΘΑΝΟΤΗΤΕΣ 2ΗΣ ΚΑΤΑΝΟΜΗΣ
 (ΠΑΝΤΟΥ $P(\theta_1) = P(\theta_2) = 0.5$)

$p(x_1|\theta_1) = \frac{p(x_1, \theta_1)}{p(x_1, \theta_1) + p(\sim x_1, \theta_1)} = 0.3$ (E-step)
 Ομοίως για $p(x_2, \theta_1)$, $p(x_3, \theta_1)$ για των 1η κατανομή
 $\mu_1 = \frac{\sum_{j=1}^3 x_j \frac{p(j|x_i, \theta)}{\sum_{j=1}^3 p(j|x_i, \theta)}}{\sum_{j=1}^3 p(j|x_i, \theta)}$ ← E-step
 $= \frac{-2 \cdot 0.3}{0.3 + 0.45 + 0.2} + \frac{1 \cdot 0.45}{0.3 + 0.45 + 0.2} + \frac{3 \cdot 0.2}{0.3 + 0.45 + 0.2}$
 $= \frac{-0.6 + 0.45 + 0.6}{0.95} = 0.474$
 $\mu_2 = \frac{-2 \cdot 0.7}{0.7 + 0.55 + 0.8} + \frac{1 \cdot 0.55}{0.7 + 0.55 + 0.8} + \frac{3 \cdot 0.8}{0.7 + 0.55 + 0.8}$
 $= \frac{-1.4 + 0.55 + 2.4}{2.05} = 0.756$

10

Έστω ένα δέντρο αποφάσεων. Έχετε δύο επιλογές: α) να μετατρέψετε το δέντρο σε κανόνες και να κλαδέψετε τους κανόνες, β) να κλαδέψετε το δέντρο και στη συνέχεια να μετατρέψετε το δέντρο σε κανόνες. Ποια είναι τα πλεονεκτήματα και μειονεκτήματα της κάθε μεθόδου (εν συντομία);
(4 Points)

Enter your answer

Once a decision tree has been constructed, it is a simple matter to convert it into an equivalent set of rules.

Converting a decision tree to rules before pruning has three main advantages:

Converting to rules allows distinguishing among the different contexts in which a decision node is used. Since each distinct path through the decision tree node produces a distinct rule, the pruning decision regarding that attribute test can be made differently for each path.

In contrast, if the tree itself were pruned, the only two choices would be:

Remove the decision node completely, or

Retain it in its original form.

Converting to rules removes the distinction between attribute tests that occur near the root of the tree and those that occur near the leaves.

We thus avoid messy bookkeeping issues such as how to reorganize the tree if the root node is pruned while retaining part of the subtree below this test.

Converting to rules improves readability.

Rules are often easier for people to understand.

To generate rules, trace each path in the decision tree, from root node to leaf node, recording the test outcomes as antecedents and the leaf-node classification as the consequent.

ΕΝΑΛΛΑΚΤΙΚΗ ΑΠΑΝΤΗΣΗ ΣΤΑ ΕΛΛΗΝΙΚΑ

Η αναγωγή δέντρου απόφασης σε κανόνες πριν το pruning έχει τα ακόλουθα πλεονεκτήματα:
- Οι κανόνες είναι πιο εύκολα αναγνώσιμοι από τους αυθριώδεις, ιδιαίτερα στην περίπτωση που το

δέντρο απόφασης είναι μεγάλο

- Το pruning στο δέντρο μπορεί να απαιτεί επανακατασκευή του δέντρου εάν αποκοπεί η ρίζα, ενώ οι κανόνες δε διαθέτουν τέτοιου είδους διαχωριστική ικανότητα ανάμεσα σε ποσοτική-φυσικά κομμάτια στη ρίζα ή στα φύλλα

- Κάθε κανόνας αντιστοιχεί σε ένα μονοπάτι στο δέντρο. Άρα, η αποκοπή ενός κανόνα ισοδυναμεί με διαγραφή του μονοπατιού (ακμών) και όχι με την αποκοπή ενός στόκληρου κόμβου, όπως στην περίπτωση του tree pruning.

Αναφορικά με την επιλογή tree pruning, πολλοί αποδοτικοί αλγόριθμοι έχουν προταθεί, οδηγώντας σε ένα αποτελεσματικό δέντρο, από το οποίο θα προκύψουν απλούστεροι και πιο ευανάγνωστοι κανόνες.

11

Ένας επενδυτής εξετάζει δύο μετοχές, την A και την B. Οι αναλυτές λένε ότι η πιθανότητα να φέρει κέρδη η A σε ένα τρίμηνο είναι 60% και η B 50%. Εκτιμούν επίσης ότι η πιθανότητα το να φέρουν κέρδη και οι δύο σε ένα τρίμηνο είναι 30%. Ποια είναι η πιθανότητα ο επενδυτής να έχει κέρδη στο τρίμηνο από την A, την B ή και τις δύο μετοχές

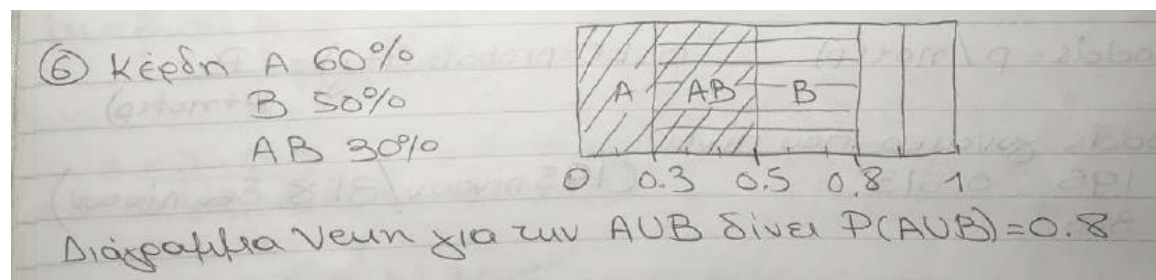
(3 Points)

A → 60%

B → 50%

Γ → 30%

$A \cup B \rightarrow A + B - A \cap B = 60 + 50 - 30 = 80\%$



Σε εκλογές, N υποψήφιοι ανταγωνίζονται ο ένας τον άλλον. Οι ψηφοφόροι ψηφίζουν για έναν από τους υποψηφίους και δεν επικοινωνούν μεταξύ τους την ώρα που ψηφίζουν. Ποια από τις παρακάτω μεθόδους συνόλων λειτουργεί παρόμοια με αυτή την εκλογική διαδικασία;

(3 Points)

- ☐ Boosting
- ☐ Δεν απαντώ
- ☐ Καμία από τις υπόλοιπες απαντήσεις
- ☐ Bagging και Boosting
- ☒ Bagging

⑩ N υποψήφιοι ανταγωνίζονται, οι ψηφοφόροι δεν επικοινωνούν.
Bagging αφού ψηφίζουν ανεξάρτητα και στο τέλος A W συμφωνούνται, άρα E .

Έχουμε έξι δείγματα δύο χαρακτηριστικών: [4 1], [6 6], [9 5], [1 2], [7 3], [5 4]. Οι ετικέτες τους είναι [1 0 1 0 1 0]. Για την ταξινόμηση φτιάχνουμε ένα δέντρο βάθους δύο. Οι διχοτομήσεις προκύπτουν με κανόνες που θέτουν ένα χαρακτηριστικό μεγαλύτερο ίσο από ένα κατώφλι. α) στη ρίζα του δέντρου, ποιοι κανόνες δίνουν μηδενικό κέρδος πληροφορίας; β) ποιος είναι ο κανόνας της πρώτης διχοτόμησης;

(5 Points)

13) 6 δείγματα, 2 χαρακτηριστικών

4 1	1	1 2	0	Διχοτομήσεις
6 6	0	4 1	1	α) Πίθα, κανόνες με μηδενικό IG
9 5	1	5 4	0	β) Κανόνες της διχοτόμησης
1 2	0	6 6	0	
7 3	1	7 3	1	
5 4	0	9 5	1	ταξινόμηση ως προς x_1

α) Μηδενικό information gain οι κανόνες που χωρίζουν το σύνολο σε 2 υποσύνολα όπου τα labels κάθε κατηγορίας είναι ισομοιρασμένα. Πχ τέτοιος κανόνας είναι ο $x_1 < 5$ (παράγει [0 1] [0 0 1 1]) ή $x_2 < 5$ (παράγει [1 0 1 0] [0 1]) ή $x_2 < 3$ (παράγει [1 0] [1 0 1 0])

Για το 1^ο ζητούμενο θέλω πρακτικά να βρω έναν κανόνα ώστε $E(\text{children}) = 1$, αφού θέλουμε $\text{Gain} = 0$, όπου

$\text{IG}(Y, X) = E(Y)_{\text{parent}} - E(Y|X)_{\text{children}}$. Άρα πρακτικά θέλουμε έναν κανόνα που χωρίζει τα δεδομένα μας σε 2 κλαδιά, όπου σε κάθε κλαδί βρίσκονται μισά στοιχεία της κλάσης 0 και μισά της 1 (κάθε κλαδί ισομοιρασμένα στοιχεία).

β) Ο κανόνας της διχοτόμησης καθό είναι να αδειχεί σε αμοιρασμένα υποσύνολα (δηλ. όσα 0 στο ένα, όσα 1 στο άλλο). Ο διαχωρισμός γίνεται με βάση κατώφλι (διαφορετικά ο κανόνας $x_2 \bmod 2 = 0$ θα έδινε την τέλεια διχοτόμηση), άρα: $x_1 \geq 7$ δίνει: [0 1 0 0] [1 1]

Έχουμε έξι δείγματα δύο χαρακτηριστικών το καθένα: [4 1], [6 6], [9 5], [1 2], [7 3], [5 4]. Οι ετικέτες τους είναι [1 0 1 0 1 0]. Για την ταξινόμηση φτιάχνουμε ένα δέντρο αποφάσεων βάθους δύο. Οι διχοτομήσεις προκύπτουν με κανόνες που θέτουν ένα χαρακτηριστικό μεγαλύτερο ίσο από ένα κατώφλι.

α) ποια είναι η εντροπία της ρίζας του δέντρου; β) ποιος είναι ο κανόνας της δεύτερης διχοτόμησης;

(5 Points)

⑤

4	1	1
6	6	0
9	5	1
1	2	0
7	3	1
5	4	0

α) Εντροπία ρίζας δέντρου

$3 \rightarrow 1 \quad 3 \rightarrow 0 \Rightarrow p^+ = 0.5 = p^-$

$$\text{Entropy}(\text{root}) = -p^+ \log_2 p^+ - p^- \log_2 p^-$$

$$= -0.5 \log_2 0.5 - 0.5 \log_2 0.5$$

$$= -\log_2 0.5 = 1 \text{ (Μέγιστη Εντροπία)}$$

β) Κανόνας 2ης διχοτόμησης

1η διχοτόμηση (σε 2 όσο το δυνατό πιο ομοιόμορφα σύνολα)

1	2	0
4	1	1
5	4	0
6	6	0
7	3	1
9	5	1

$x_1 \geq 7 \rightarrow [0 \ 1 \ 0 \ 0] [1 \ 1]$

$S_1 \quad S_2$

Το S_2 δε χρειάζεται περαιτέρω διχοτόμηση

Μένουν:

1	2	0
4	1	1
5	4	0
6	6	0

$x_2 \geq 2 \rightarrow [1] [0 \ 0 \ 0]$

Σωστό αυτό που λέει το screenshot, άρα όντως έχω μέγιστη εντροπία.

Έστω ένα μοντέλο λογιστικής παλινδρόμησης και ένα γραμμικά διαχωρίσιμο πρόβλημα. Αν έχουμε μια απεικόνιση $g(x)$ των χαρακτηριστικών x , και βάρη w , εξηγήστε ποιοτικά πώς το μοντέλο μέγιστης πιθανότητας προκύπτει αν βρούμε την ευθεία με $w^T g(x) = 0$ και στη συνέχεια θέσουμε τη νόρμα του w στο άπειρο.

(5 Points)

Στα πλαίσια της λογιστικής παλινδρόμησης για ένα σύνολο δεδομένων $\{\phi_n, t_n\}$ με $t_n \in \{0, 1\}$ (δυαδική ταξινόμηση), $\phi_n = \phi(x_n)$ και $n = 1, \dots, N$, η a-posteriori πιθανότητα της μιας κατηγορίας, έστω της \mathcal{C}_1 , ορίζεται ως

$$p(\mathcal{C}_1|\phi_n) = \sigma(w^T \phi_n) \equiv y_n, \quad (9.1)$$

επομένως $p(\mathcal{C}_2|\phi_n) = 1 - y_n$.

9.1 Σε ένα γραμμικά διαχωρίσιμο σύνολο δεδομένων, τα σημεία n για τα οποία ισχύει $p(\mathcal{C}_1|\phi_n) > p(\mathcal{C}_2|\phi_n)$ θα ταξινομούνται στην κατηγορία \mathcal{C}_1 , ενώ τα σημεία για τα οποία ισχύει $p(\mathcal{C}_1|\phi_n) < p(\mathcal{C}_2|\phi_n)$ θα ταξινομούνται στην κατηγορία \mathcal{C}_2 . Έτσι, η επιφάνεια απόφασης θα ορίζεται βάσει των

$$\begin{aligned} p(\mathcal{C}_1|\phi_n) = p(\mathcal{C}_2|\phi_n) &\Leftrightarrow y_n = 1 - y_n \Leftrightarrow \sigma(w^T \phi_n) = 0.5 \Leftrightarrow \sigma(w^T \phi_n) = \sigma(0) \\ &\Leftrightarrow w^T \phi_n = 0, \end{aligned} \quad (9.2)$$

όπου στην τελευταία ισοδυναμία αξιοποιείται το γεγονός πως η $\sigma(x)$ είναι 1-1, ενώ η ύπαρξη ενός τέτοιου w διασφαλίζεται από τη γραμμική διαχωρισιμότητα του συνόλου. Η συνάρτηση πιθανοφάνειας για ένα τέτοιο πρόβλημα μπορεί να γραφεί ως [3]

$$\ell(w) = \prod_{n=1}^N y_n^{t_n} (1 - y_n)^{1-t_n} \quad (9.3)$$

και επειδή ο φυσικός λογάριθμος είναι γνησίως αύξουσα συνάρτηση του ορίσματος της, η μεγιστοποίηση της πιθανοφάνειας ισοδυναμεί με τη μεγιστοποίηση του λογαρίθμου της, δηλαδή

$$\begin{aligned} \ln \ell(w) &= \ln \prod_{n=1}^N y_n^{t_n} (1 - y_n)^{1-t_n} = \sum_{n=1}^N \ln [y_n^{t_n} (1 - y_n)^{1-t_n}] = \sum_{n=1}^N [\ln y_n^{t_n} + \ln (1 - y_n)^{1-t_n}] \\ &= \sum_{n=1}^N [t_n \ln y_n + (1 - t_n) \ln (1 - y_n)] = -E(w), \end{aligned} \quad (9.4)$$

όπου η $E(w)$ είναι η δοθείσα συνάρτηση σφάλματος (cross-entropy). Προκύπτει, λοιπόν, πως στην προκειμένη περίπτωση η μεγιστοποίηση της πιθανοφάνειας ισοδυναμεί με την ελαχιστοποίηση της συνάρτησης σφάλματος. Για το σκοπό αυτό, γράφει κανείς

$$\begin{aligned} -\nabla E(w) &= \sum_{n=1}^N \nabla [t_n \ln y_n + (1 - t_n) \ln (1 - y_n)] \\ &= \sum_{n=1}^N \left(\frac{t_n}{y_n} - \frac{1 - t_n}{1 - y_n} \right) \nabla y_n = \sum_{n=1}^N \left(\frac{t_n}{y_n} - \frac{1 - t_n}{1 - y_n} \right) y_n (1 - y_n) \phi_n \\ &= \sum_{n=1}^N \frac{t_n - y_n}{y_n (1 - y_n)} y_n (1 - y_n) \phi_n = \sum_{n=1}^N (t_n - y_n) \phi_n, \end{aligned} \quad (9.5)$$

όπου στην τρίτη ισότητα αξιοποιήθηκε η ιδιότητα

$$\frac{d\sigma(x)}{dx} = \sigma(x) [1 - \sigma(x)] \quad (9.6)$$

της σιγμοειδούς συνάρτησης. Έτσι, η μεγιστοποίηση της πιθανοφάνειας, η οποία ισοδυναμεί με την απαίτηση $-\nabla E(w) = 0$, βάσει της Σχέσης (9.5) ανάγεται στην εύρεση ενός w τέτοιου, ώστε

$$\sigma(w^T \phi_n) \rightarrow \begin{cases} 1, & \text{εάν } t_n = 1 \\ 0, & \text{εάν } t_n = 0 \end{cases}, \quad \forall n = 1, \dots, N \quad (9.7)$$

και επειδή η σιγμοειδής συνάρτηση είναι 1-1,

$$w^T \phi_n \rightarrow \begin{cases} \infty, & \text{εάν } t_n = 1 \\ -\infty, & \text{εάν } t_n = 0 \end{cases}, \quad \forall n = 1, \dots, N. \quad (9.8)$$

Συμπεραίνει κανείς πως, προκειμένου η απαίτηση (9.8) να ικανοποιείται για κάθε n , θα πρέπει με βεβαιότητα για το w να ισχύει

$$\|w\|_2 \rightarrow \infty. \quad (9.9)$$

Αξίζει να σημειωθεί στο σημείο αυτό πως η περίπτωση αυτή ισοδυναμεί με τη χρήση της βηματικής συνάρτησης Heaviside για την ταξινόμηση σημείων, με αποτέλεσμα η a-posteriori πιθανότητα για κάθε σημείο εκπαίδευσης να ισούται με τη μονάδα και συνεπώς το μοντέλο να κινδυνεύει σημαντικά από υπερπροσαρμογή (overfitting).

15

964 άνθρωποι ερωτήθηκαν αν πίνουν αλκοόλ. Ο πίνακας συγκεντρώνει τις απαντήσεις τους

Φύλο Ναι Όχι Σύνολο

Άνδρες 152 299 451

Γυναίκες 195 318 513

Σύνολο 347 617 964

Ποιες είναι οι πιθανότητες (odds) για μια γυναίκα να πίνει (ως προς το να μην πίνει) αλκοόλ;

Ποιο είναι ο σχετικός κίνδυνος να πίνουν οι άνδρες σε σχέση με τις γυναίκες;

(4 Points)

Άνδρες: 152 ναι, 299 όχι: $RISK_A = 152/451$, $ODDS_A = 152/299$

Γυναίκες: 195 ναι, 318 όχι: $RISK_F = 195/513$, $ODDS_F = 195/318$

Risk Ratio = $RISK_A / RISK_F = 0,88$

ΑΝ ΖΗΤΗΘΕΙ ΓΥΝΑΙΚΕΣ ΣΕ ΣΧΕΣΗ ΜΕ ΑΝΤΡΕΣ ΤΟΤΕ

$RiskRatio' = 1/RiskRatio$

9) 964 άνθρωποι ερωτήθηκαν αν πίνουν αλκοόλ.

$$\text{odds} = p / \text{not}(p) \quad \text{risk} (= \text{probability}) = \frac{p}{p + \text{not}(p)}$$

odds γυναίκα που πίνει:

$$\frac{195}{318} = 0.613 \quad (195 \text{ πίνουν} / 318 \text{ δεν πίνουν})$$

Σχετικός κίνδυνος να πίνουν οι άντρες σε σχέση με γυναίκες

$$\text{risk for men} = 152 / 451 = 0.337$$

↑ πίνουν ↑ όλοι οι άντρες

$$\text{risk for women} = 195 / 513 = 0.38$$

↑ πίνουν ↑ όλες οι γυναίκες

$$\text{Relative risk} = \frac{\text{risk for men}}{\text{risk for women}} = \frac{0.337}{0.38} = 0.887$$

Ένα λαχείο σύμφωνα με τους εκδότες του έχει πιθανότητα να κερδίζει 0.3. Ποια είναι η πιθανότητα να χρειαστεί κάποιος το λιγότερο τρία λαχεία για να κερδίσει;
(3 Points)

11) $p(\text{λαχείο} = \text{κέρδος}) = 0.3$

Γεωμετρική κατανομή και αριθμός προσπαθειών μέχρι των 1η επιτυχία:

$$p(X=x) = (1-p)^{x-1} \cdot p$$

$x=3$ για 3 προσπαθειές, δίνει πιθανότητα επιτυχίας στην 3η (μετά από 2 αποτυχίες)

$$p(X=3) = (1-0.3)^2 \cdot 0.3 = 0.7^2 \cdot 0.3 = 0.147$$

6

Έχετε ένα πρόβλημα δυαδικής ταξινόμησης. Κάθε δείγμα x_i ανήκει είτε στην κλάση $c=1$ είτε στην κλάση $c=0$. Η διαδικασία συλλογής δεδομένων εκπαίδευσης είναι ωστόσο ατελής και κάποια παραδείγματα δεν λαμβάνουν σωστές ετικέτες. Για κάθε δείγμα x_i αντί να έχουμε την αληθινή ετικέτα του, c_i , έχουμε την πιθανότητα π_i που αντιστοιχεί στο να είναι $c_i=1$. Πώς εκφράζεται η λογαριθμική πιθανότητα (log likelihood) για ένα πιθανοτικό μοντέλο $p(c=1|\theta)$ αν αντί του c_i βάλουμε το π_i ;

(5 Points)

The log-likelihood for a probabilistic model for binary classification is

$$\sum_{i=1}^n y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i)),$$

where $p(x_i)$ is the model predicted probability that the i -th observation is a 1, and y_i is the i -th observation for the response.

In summary, sum up the logs of the predicted probabilities where the actual response was one, and add this to the sum of the logs of (1 - the probabilities) whenever the actual response was zero.

13

Χρησιμοποιήστε τις μεθόδους (a) κανονικοποίηση min-max θέτοντας min=0 και max=1 και (b) κανονικοποίηση z-score για να ομαλοποιήσετε τις ακόλουθες τιμές δεδομένων: 200, 300, 400, 600, 1000

(3 Points)

(a) <https://t4tutorials.com/min-max-normalization-of-data-in-data-mining/>

$v' = (v - \min) / (\max - \min) * (\text{new_max} - \text{new_min}) + \text{new_min}$ όπου $\text{new_min} = 0$ και $\text{new_max} = 1$, και για τα δοθέντα δεδομένα $\min=200$, $\max=1000$ άρα $v' = (v - 200) / 800$. Οπότε:

$v = 200: v' = 0$

$v = 300: v' = 1/8 = 0.125$

$v = 400: v' = 1/4 = 0.25$

$v = 600: v' = 1/2 = 0.5$

$v = 1000: v' = 1$

(b) <https://t4tutorials.com/z-score-normalization-data-mining/>

for each data point: z score = $(v - \mu) / \sigma$, where v is the value

We first need to calculate the mean μ and standard deviation σ of the data:

$\mu = (200 + 300 + 400 + 600 + 1000) / N = 2500 / 5 = 500$, ($N=5$ the number of data points)

$\sigma^2 = 1/N * \sum [(v - \mu)^2] = [(200 - 500)^2 + (300 - 500)^2 + (400 - 500)^2 + (600 - 500)^2 + (1000 - 500)^2] / 5$

$= [9 * 10^4 + 4 * 10^4 + 10^4 + 10^4 + 25 * 10^4] / 5 = [9 + 4 + 1 + 1 + 25] * (10^4) / 5 = 40 * 2000 = 80,000$

$\sigma = \sqrt{\sigma^2} = 282.84$

Therefore:

$v = 200: z = (200 - 500) / 282.84 = -1.06$

$v = 300: z = (300 - 500) / 282.84 = -0.71$

$v = 400: z = (400 - 500) / 282.84 = -0.35$

$v = 600: z = (600 - 500) / 282.84 = 0.35$

$v = 1000: z = (1000 - 500) / 282.84 = 1.77$

(c) (BONUS) decimal scaling

Decimal Scaling: $\text{norm_value} = \text{value} / (10^j)$, j = είναι ο μικρότερος ακέραιος έτσι ώστε $\max(|\text{norm_value}|) < 1$ πρακτικά είναι j = όσα τα ψηφία του μεγαλύτερου ακεραίου, δηλαδή εδώ που 1,000 ο μεγαλύτερος, $j = 4$ άρα αυτό που κάνω είναι διαιρώ όλες τις τιμές με $10^4 = 10,000$ και περιμένω ότι όλες οι normalized τιμές με decimal scaling θα έχουν απόλυτη τιμή μικρότερη της μονάδας.

$v = 200: v' = 200 / 10,000 = 0.02$

$v = 300: v' = 300 / 10,000 = 0.03$

$v = 400: v' = 400 / 10,000 = 0.04$

$v = 600: v' = 600 / 10,000 = 0.06$

5

Ποιες από τις επόμενες ιδιότητες είναι χαρακτηριστικές των δέντρων αποφάσεων;
(4 Points)

- ☐ Υψηλή πόλωση
- ☒ Μη φραγμένο σύνολο παραμέτρων ?
- ☒ Υψηλή διακύμανση
- ☐ Καμία από τις υπόλοιπες απαντήσεις
- ☒ Μη ομαλότητα στις επιφάνειες απόφασης

7

Ποιο ή ποιά από τα ακόλουθα είναι παραδείγματα εξαγωγής χαρακτηριστικών;
(3 Points)

- ☐ Καμία από τις υπόλοιπες απαντήσεις
- ☒ Εφαρμογή PCA σε δεδομένα υψηλής διαστατικότητας
- ☒ Κατασκευή bag of words vector από ένα email
- ☐ Αφαίρεση stopwords σε μια πρόταση ?

9

Ποιος είναι ο σκοπός της διασταυρούμενης επικύρωσης;
(3 Points)

- ☐ Η εκτίμηση της απόδοσης του μοντέλου εκτός του συνόλου δεδομένων εκπαίδευσης
- ☐ Καμία από τις υπόλοιπες απαντήσεις
- ☐ Η εκτίμηση της απόδοσης του μοντέλου στα σύνολα εκπαίδευσης και επικύρωσης
- ☒ Η εκτίμηση της προγνωστικής απόδοσης των μοντέλων