

Ασκήσεις μελέτης της 8^{ης} διάλεξης

8.1. Έστω ένα αντικείμενο προς κατάταξη το οποίο παριστάνεται με το διάνυσμα $\langle X_1, X_2, X_3 \rangle = \langle 0, 1, 0 \rangle$. Σε ποια από τις δύο κατηγορίες ($C = 0$ ή $C = 1$) θα κατατάξει το αντικείμενο ο αλγόριθμος των k κοντινότερων γειτόνων με $k = 3$ και μέτρο απόστασης δύο διανυσμάτων τον αριθμό των ιδιοτήτων στις οποίες έχουν διαφορετικές τιμές; Δείξτε αναλυτικά τους υπολογισμούς σας.

d	0	1	0	C?
2	1	0	0	0
1	0	1	1	0
3	1	0	1	1
2	1	1	1	1

Απάντηση: Η πρώτη στήλη του διπλανού πίνακα δείχνει τις αποστάσεις (d) του διανύσματος του προς κατάταξη αντικειμένου από τα διανύσματα των παραδειγμάτων εκπαίδευσης. Οι $k = 3$ κοντινότεροι γείτονες είναι εκείνοι που βρίσκονται σε αποστάσεις 1 και 2. Μεταξύ αυτών πλειοψηφεί η $C = 0$. Επομένως το αντικείμενο θα καταταγεί στη $C = 0$.

8.2. Αν δεν υπάρχουν ασυνεπή διανύσματα εκπαίδευσης και αξιολογήσουμε έναν ταξινομητή k κοντινότερων γειτόνων στο ίδιο σύνολο διανυσμάτων στο οποίο τον εκπαιδεύσαμε, το ποσοστό ορθότητάς του θα βρεθεί να είναι:

___ σίγουρα 100%

___ ☒ σίγουρα 100% αν $k = 1$, αλλά όχι σίγουρα 100% αν $k \neq 1$

___ τίποτα από τα παραπάνω (το αποτέλεσμα εξαρτάται από τις τυχαιότητες των δεδομένων).

Απάντηση: Αν $k = 1$, κάθε διάνυσμα αξιολόγησης θα κατατάσσεται στην κατηγορία του κοντινότερου διανύσματος εκπαίδευσης, που θα είναι ο εαυτός του (ή ένα αντίγραφο του εαυτού του, της ίδιας κατηγορίας, αφού δεν υπάρχουν ασυνεπή διανύσματα εκπαίδευσης), αφού τα διανύσματα αξιολόγησης είναι τα ίδια με τα διανύσματα εκπαίδευσης, οπότε θα κατατάσσεται σωστά.

Αν $k \neq 1$, κάθε διάνυσμα αξιολόγησης θα κατατάσσεται στην κατηγορία που πλειοψηφεί μεταξύ των k πιο παρόμοιων διανυσμάτων εκπαίδευσης (και αξιολόγησης) και μπορεί η πλειοψηφούσα κατηγορία να είναι διαφορετική από τη σωστή κατηγορία του διανύσματος αξιολόγησης. Άρα στην περίπτωση αυτή το ποσοστό ορθότητας δεν θα είναι σίγουρα 100%.

8.3. (α) Έστω ένα αντικείμενο προς κατάταξη το οποίο παριστάνεται με το διάνυσμα $\langle X_1, X_2, X_3 \rangle = \langle 0, 1, 0 \rangle$. Σε ποια από τις δύο κατηγορίες ($C = 0$ ή $C = 1$) θα το κατατάξει ένας αφελής ταξινομητής Bayes (της πολυ-μεταβλητής μορφής Bernoulli που συναντήσαμε στο μάθημα) που έχει στη διάθεσή του τα δεδομένα εκπαίδευσης του πίνακα; Γράψτε αναλυτικά τους υπολογισμούς σας. Χρησιμοποιήστε εκτιμητήρια Laplace κατά τις εκτιμήσεις των πιθανοτήτων $P(X_i|C)$. Όλες οι ιδιότητες X_i είναι δυαδικές (έχουν ως πεδίο τιμών το $\{0, 1\}$).

X_1	X_2	X_3	C
1	0	0	0
0	1	1	0
1	0	1	1
1	1	1	1

Απάντηση:

$$P(C = 1|X_1 = 0, X_2 = 1, X_3 = 0) \cong$$

$$\begin{aligned} & \frac{P(C = 1)}{P(X_1 = 0, X_2 = 1, X_3 = 0)} \cdot P(X_1 = 0|C = 1) \cdot P(X_2 = 1|C = 1) \cdot P(X_3 = 0|C = 1) \\ & \cong \frac{1/2}{P(X_1 = 0, X_2 = 1, X_3 = 0)} \cdot \frac{0+1}{2+2} \cdot \frac{1+1}{2+2} \cdot \frac{0+1}{2+2} \\ & = \frac{1/2}{P(X_1 = 0, X_2 = 1, X_3 = 0)} \cdot \frac{1}{4} \cdot \frac{2}{4} \cdot \frac{1}{4} \end{aligned}$$

$$\begin{aligned} P(C = 0|X_1 = 0, X_2 = 1, X_3 = 0) & \cong \frac{P(C = 0)}{P(X_1 = 0, X_2 = 1, X_3 = 0)} \cdot P(X_1 = 0|C = 0) \cdot P(X_2 = 1|C = 0) \cdot P(X_3 = 0|C = 0) \\ & \cong \frac{1/2}{P(X_1 = 0, X_2 = 1, X_3 = 0)} \cdot \frac{1+1}{2+2} \cdot \frac{1+1}{2+2} \cdot \frac{1+1}{2+2} \\ & = \frac{1/2}{P(X_1 = 0, X_2 = 1, X_3 = 0)} \cdot \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{2}{4} \end{aligned}$$

Επομένως θα το κατατάξει στην $C = 0$.

8.4. Έστω ένα αντικείμενο με διάνυσμα $\langle X_1, X_2, X_3, X_4 \rangle = \langle b, d, b, a \rangle$. Σε ποια από τις τρεις κατηγορίες θα κατατάξει το αντικείμενο ένας αφελής ταξινομητής Bayes (της πολυμεταβλητής μορφής Bernoulli που συναντήσαμε στο μάθημα) ο οποίος διαθέτει τα δεδομένα εκπαίδευσης του πίνακα; Γράψτε αναλυτικά τους υπολογισμούς σας. Χρησιμοποιήστε εκτιμητήρια Laplace κατά τις εκτιμήσεις των πιθανοτήτων $P(X_i|C)$. Θεωρήστε ότι κάθε τυχαία μεταβλητή X_i έχει τέσσερις δυνατές τιμές: a, b, c, d.

X_1	X_2	X_3	X_4	C
a	b	b	a	1
b	a	a	b	1
b	b	a	b	1
a	a	b	b	2
a	b	b	b	2
b	a	b	a	2
c	d	d	c	3
d	c	c	d	3
d	d	c	d	3

Απάντηση:

Έχουμε: $P(C = 1) = P(C = 2) = P(C = 3) = 3/9$. Επομένως, οι a priori πιθανότητες δεν επηρεάζουν την απόφαση.

Έχουμε επίσης:

$$\begin{aligned} P(X_1 = b|C = 1) &= \frac{2+1}{3+4} = \frac{3}{7} & P(X_2 = d|C = 1) &= \frac{0+1}{3+4} = \frac{1}{7} \\ P(X_3 = b|C = 1) &= \frac{1+1}{3+4} = \frac{2}{7} & P(X_4 = a|C = 1) &= \frac{1+1}{3+4} = \frac{2}{7} \end{aligned}$$

και:

$$\begin{aligned} P(X_1 = b|C = 2) &= \frac{1+1}{3+4} = \frac{2}{7} & P(X_2 = d|C = 2) &= \frac{0+1}{3+4} = \frac{1}{7} \\ P(X_3 = b|C = 2) &= \frac{3+1}{3+4} = \frac{4}{7} & P(X_4 = a|C = 2) &= \frac{1+1}{3+4} = \frac{2}{7} \end{aligned}$$

και:

$$\begin{aligned} P(X_1 = b|C = 3) &= \frac{0+1}{3+4} = \frac{1}{7} & P(X_2 = d|C = 3) &= \frac{2+1}{3+4} = \frac{3}{7} \\ P(X_3 = b|C = 3) &= \frac{0+1}{3+4} = \frac{1}{7} & P(X_4 = a|C = 3) &= \frac{0+1}{3+4} = \frac{1}{7} \end{aligned}$$

Άρα:

$$P(X_1 = b|C = 1) \cdot P(X_2 = d|C = 1) \cdot P(X_3 = b|C = 1) \cdot P(X_4 = a|C = 1) = \frac{3 \cdot 1 \cdot 2 \cdot 2}{7 \cdot 7 \cdot 7 \cdot 7}$$

$$P(X_1 = b|C = 2) \cdot P(X_2 = d|C = 2) \cdot P(X_3 = b|C = 2) \cdot P(X_4 = a|C = 2) = \frac{2 \cdot 1 \cdot 4 \cdot 2}{7 \cdot 7 \cdot 7 \cdot 7}$$

$$P(X_1 = b|C = 3) \cdot P(X_2 = d|C = 3) \cdot P(X_3 = b|C = 4) \cdot P(X_4 = a|C = 3) = \frac{1 \cdot 3 \cdot 1 \cdot 1}{7 \cdot 7 \cdot 7 \cdot 7}$$

και επομένως το αντικείμενο κατατάσσεται στην κατηγορία $C = 2$.

8.5. Χρησιμοποιούμε μια παραλλαγή του αφελούς ταξινομητή Bayes που συναντήσαμε στο μάθημα (δηλαδή μια παραλλαγή της πολυ-μεταβλητής μορφής Bernoulli), με δύο κατηγορίες ($C = 0$ και $C = 1$) και m δυαδικές ιδιότητες $X_1, \dots, X_m$, η οποία κατατάσσει στη $C = 1$ αν:

$$P(C = 1) \cdot \prod_{i=1}^m P(X_i = x_i|C = 1) \geq K$$

όπου K μια σταθερά, ενώ διαφορετικά κατατάσσει στη $C = 0$. Αποδείξτε ότι ο ταξινομητής αυτός είναι ένας γραμμικός διαχωριστής. Δείξτε αναλυτικά τους υπολογισμούς σας. Υπόδειξη: Αν παραστήσουμε με t_i το ενδεχόμενο που παριστάνεται με $X_i = 1$ (π.χ. την εμφάνιση μιας συγκεκριμένης λέξης), τότε:

$$P(X_i = x_i|C = 1) = P(t_i|C = 1)^{x_i} \cdot [1 - P(t_i|C = 1)]^{1-x_i}$$

Υπενθυμίζεται, επίσης, ότι $\log(a \cdot b) = \log a + \log b$ και $\log a^b = b \cdot \log a$.

Απάντηση: Ο ταξινομητής κατατάσσει στην κατηγορία $C = 1$ αν και μόνο αν (ανν):

$$\log[P(C = 1) \cdot \prod_{i=1}^m P(X_i = x_i|C = 1)] \geq \log K$$

ισοδύναμα:

$$\log P(C = 1) + \sum_{i=1}^m \log P(X_i = x_i|C = 1) \geq \log K$$

ισοδύναμα:

$$\log P(C = 1) + \sum_{i=1}^m \log\{P(t_i|C = 1)^{x_i} \cdot [1 - P(t_i|C = 1)]^{1-x_i}\} \geq \log K$$

ισοδύναμα:

$$\log P(C = 1) + \sum_{i=1}^m x_i \cdot \log P(t_i|C = 1) + (1 - x_i) \cdot \log[1 - P(t_i|C = 1)] \geq \log K$$

Θέτοντας $K_1 = \log P(C = 1)$, $K_{2,i} = \log P(t_i|C = 1)$, $K_{3,i} = \log[1 - P(t_i|C = 1)]$, $K_4 = \log K$, η προηγούμενη σχέση γίνεται:

$$K_1 + \sum_{i=1}^m [x_i \cdot K_{2,i} + (1 - x_i) \cdot K_{3,i}] \geq K_4$$

ισοδύναμα:

$$K_1 + \sum_{i=1}^m K_{3,i} - K_4 + \sum_{i=1}^m x_i \cdot (K_{2,i} - K_{3,i}) \geq 0$$

Θέτοντας $w_0 = K_1 + \sum_{i=1}^m K_{3,i} - K_4$ και $w_i = (K_{2,i} - K_{3,i})$, η προηγούμενη σχέση γίνεται:

$$w_0 + \sum_{i=1}^m w_i \cdot x_i \geq 0$$

Επομένως πρόκειται για γραμμικό ταξινομητή.