Τεχνητή Νοημοσύνη

7ο φροντιστήριο (2023-24)

Επιμέλεια: Σοφία Ελευθερίου,
Φοίβος Χαραλαμπάκος

Άσκηση 15.1

Έστω ένα αντικείμενο προς κατάταξη το οποίο παριστάνεται με το διάνυσμα $< X_1, X_2, X_3 > = < 0{,}1{,}0 >$. 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.

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

Το παράδειγμα 1 ($< 1{,}0{,}0 >$) απέχει απόσταση 2 από το αντικείμενο προς εξέταση ($< 0{,}1{,}0 >$) αφού διαφέρουν σε δύο ψηφία.
Όμοια βρίσκουμε και τις αποστάσεις με τα υπόλοιπα παραδείγματα.
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 $C = 0$.

Άσκηση 15.2

Αν δεν υπάρχουν ασυνεπή διανύσματα εκπαίδευσης και αξιολογήσουμε έναν ταξινομητή k κοντινότερων γειτόνων στο ίδιο σύνολο διανυσμάτων στο οποίο τον εκπαιδεύσαμε, τι εκτίμηση μπορούμε να κάνουμε για το ποσοστό ορθότητας όταν $k = 1$ και όταν $\kappa > 1$;

- Αν $k = 1$,

Αφού αξιολογούμε τον ταξινομητή στο ίδιο σύνολο διανυσμάτων στο οποίο τον εκπαιδεύσαμε, κάθε διάνυσμα προς εξέταση θα υπάρχει στα παραδείγματα εκπαίδευσης (απόσταση μηδέν) τουλάχιστον μια φορά με την ίδια κατάταξη αφού δεν υπάρχουν ασυνεπή διανύσματα εκπαίδευσης. Άρα θα κατατάσσεται σωστά.

- Αν $k > 1$,

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%.

Εντροπία

- **Εντροπία** της τυχαίας μεταβλητής *C*.
  - Δείχνει **πόσο αβέβαιοι** είμαστε για την **τιμή της *C***.
  - Πόση είναι η **ελάχιστη ποσότητα πληροφορίας** που πρέπει να μας δοθεί για να γνωρίζουμε με βεβαιότητα την τιμή της *C*.
  - Ποιος είναι στην **καλύτερη περίπτωση** (με την ιδανική κωδικοποίηση) ο **αναμενόμενος αριθμός δυφίων** που πρέπει να μας μεταδοθεί για να καθοριστεί η τιμή της *C*;

$$H(C) = -\sum_{c} P(C=c)\log_2 P(C=c)$$

Κέρδος Πληροφορίας

$$I(C,X) = H(C) - \sum_{x} P(X=x) H(C|X=x)$$

Άσκηση 15.3

α) Βάσει των δεδομένων εκπαίδευσης του διπλανού πίνακα, ποια είναι η εντροπία της κατηγορίας C;

X	Y	Z	C
0	1	0	Θετικό
0	1	1	Θετικό
0	1	0	Θετικό
1	0	1	Θετικό
1	1	0	Αρνητικό
0	0	1	Αρνητικό
0	0	0	Αρνητικό
0	0	1	Αρνητικό

Η πιθανότητα η C να κατατάσσεται στο “θετικό” είναι

$$P(C = \text{"}\theta\varepsilon\tau\iota\kappa\acute{o}\text{"}) = \frac{4}{8} = \frac{1}{2}$$

Όμοια, η πιθανότητα η C να κατατάσσεται στο “αρνητικό” είναι

$$P(C = \text{"}\alpha\rho\nu\eta\tau\iota\kappa\acute{o}\text{"}) = \frac{4}{8} = \frac{1}{2}$$

Επομένως η εντροπία είναι

$$H(C) = -\sum_{c} P(C = c)\log_2 P(C = c) =$$

$$-P(C = \text{"}\theta\varepsilon\tau\iota\kappa\acute{o}\text{"})\log_2\big(P(C = \text{"}\theta\varepsilon\tau\iota\kappa\acute{o}\text{"})\big) - P(C = \text{"}\alpha\rho\nu\eta\tau\iota\kappa\acute{o}\text{"})\log_2 P(C = \text{"}\alpha\rho\nu\eta\tau\iota\kappa\acute{o}\text{"})$$

$$= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = -2\frac{1}{2}(\log_2 1 - \log_2 2) = 1.$$

β) Βάσει των δεδομένων του πίνακα του σκέλους (α), ποια μεταβλητή έχει το υψηλότερο κέρδος πληροφορίας;

X	Y	Z	C
0	1	0	Θετικό
0	1	1	Θετικό
0	1	0	Θετικό
1	0	1	Θετικό
1	1	0	Αρνητικό
0	0	1	Αρνητικό
0	0	0	Αρνητικό
0	0	1	Αρνητικό

Η εντροπία της C αν μάθουμε ότι X = 0 είναι

$$\mathrm{H}(\mathrm{C}|\mathrm{X}=0) = -\sum_c P(C=c|X=0)\log_2 P(C=c|X=0) =$$

$$-\left(\frac{1}{2}\cdot\log_2\frac{1}{2}\right)\cdot 2 = 1$$

Η εντροπία της C αν μάθουμε ότι X = 1 είναι

$$\mathrm{H}(\mathrm{C}|\mathrm{X}=1) = -\sum_c P(C=c|X=1)\log_2 P(C=c|X=1) =$$

$$-\left(\frac{1}{2}\cdot\log_2\frac{1}{2}\right)\cdot 2 = 1$$

Οπότε το κέρδος πληροφορίας του X είναι

$$IG(C,X) = \mathrm{H}(C) - \sum_x P(X=x)H(C|X=x)$$

$$1 - \frac{6}{8}\cdot 1 - \frac{2}{8}\cdot 1 = 1 - 1 = 0$$

Αναμενόμενο αφού είτε μάθουμε ότι $X = 1$ είτε ότι $X = 0$, η πιθανότητα να έχουμε C = θετικό παραμένει $\frac{1}{2}$ και ίση με την πιθανότητα να έχουμε C = αρνητικό.

Η εντροπία της C αν μάθουμε ότι $\mathrm{Y} = 0$ είναι

$$\mathrm{H}(\mathrm{C}|\mathrm{Y}=0) = -\sum_{c} P(C=c|Y=0)\log_2 P(C=c|Y=0) =$$

$$-\frac{1}{4}\cdot\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.811$$

Η εντροπία της C αν μάθουμε ότι $\mathrm{Y} = 1$ είναι

$$\mathrm{H}(\mathrm{C}|\mathrm{Y}=1) = -\sum_{c} P(C=c|Y=1)\log_2 P(C=c|Y=1) =$$

$$-\frac{1}{4}\cdot\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.811$$

Οπότε το κέρδος πληροφορίας του Y είναι

$$IG(C,Y) = \mathrm{H}(C) - \sum_{x} P(Y=y)H(C|Y=y)$$

$$1 - \frac{4}{8}\cdot 0.811 - \frac{4}{8}\cdot 0.811 = 1 - 0.811 = 0.189$$

Είναι αναμενόμενο να έχουμε θετικό κέρδος πληροφορίας για το Y αφού αν για παράδειγμα μάθουμε ότι $Y = 1$ τότε η πιθανότητα να είναι C = θετικό αυξάνει από $\frac{1}{2}$ σε $\frac{3}{4}$

X	Y	Z	C
0	1	0	Θετικό
0	1	1	Θετικό
0	1	0	Θετικό
1	0	1	Θετικό
1	1	0	Αρνητικό
0	0	1	Αρνητικό
0	0	0	Αρνητικό
0	0	1	Αρνητικό

Η εντροπία της C αν μάθουμε ότι Z = 0 είναι

$$\mathrm{H}(\mathrm{C}|\mathrm{Z}=0) = -\sum_{c} P(C=c|Z=0)\log_2 P(C=c|Z=0) =$$

$$-\left(\frac{1}{2}\cdot\log_2\frac{1}{2}\right)\cdot 2 = 1$$

Η εντροπία της C αν μάθουμε ότι Z = 1 είναι

$$\mathrm{H}(\mathrm{C}|\mathrm{Z}=1) = -\sum_{c} P(C=c|Z=1)\log_2 P(C=c|Z=1) =$$

$$-\left(\frac{1}{2}\cdot\log_2\frac{1}{2}\right)\cdot 2 = 1$$

Οπότε το κέρδος πληροφορίας του Z είναι

$$IG(C,Z) = \mathrm{H}(C) - \sum_{x} P(Z=z)H(C|Z=z)$$

$$1 - \frac{4}{8}\cdot 1 - \frac{4}{8}\cdot 1 = 1 - 1 = 0$$

Αναμενόμενο αφού είτε μάθουμε ότι $Z = 1$ είτε ότι $Z = 0$, η πιθανότητα να έχουμε C = θετικό παραμένει $\frac{1}{2}$ και ίση με την πιθανότητα να έχουμε C = αρνητικό.

X	Z	Z	C
0	1	0	Θετικό
0	1	1	Θετικό
0	1	0	Θετικό
1	0	1	Θετικό
1	1	0	Αρνητικό
0	0	1	Αρνητικό
0	0	0	Αρνητικό
0	0	1	Αρνητικό

Άσκηση 15.5

α) Εξηγήστε γιατί οι αλγόριθμοι επιβλεπόμενης μάθησης πρέπει να αξιολογούνται σε διαφορετικά δεδομένα από αυτά με τα οποία εκπαιδεύτηκαν

Αν αξιολογούνταν στα δεδομένα εκπαίδευσης, ένας αλγόριθμος που θα απομνημόνευε τις σωστές απαντήσεις των παραδειγμάτων εκπαίδευσης και θα απαντούσε τυχαία σε κάθε άλλη περίπτωση θα πετύχαινε ακρίβεια 100% στην αξιολόγηση, χωρίς αυτή η επίδοση να είναι ενδεικτική του πόσο καλά θα τα πήγαινε σε διαφορετικά δεδομένα αξιολόγησης, αφού τότε θα απαντούσε τυχαία.

Γενικότερα, υπάρχει ο κίνδυνος το μοντέλο απόφασης που παράγεται να είναι υπερ-εξειδικευμένο στα δεδομένα εκπαίδευσης, με αποτέλεσμα να επιτυγχάνεται υψηλή ακρίβεια στα δεδομένα εκπαίδευσης, που δεν είναι ενδεικτική της ακρίβειας που επιτυγχάνεται σε διαφορετικά δεδομένα αξιολόγησης.

β) Ένας ερευνητής υπέβαλε σε ένα επιστημονικό συνέδριο εργασία του στην οποία περιέγραφε ένα σύστημα αναγνώρισης ονομάτων οντοτήτων που χρησιμοποιούσε επιβλεπόμενη μηχανική μάθηση. Στην εργασία περιέγραφε, μεταξύ άλλων, πειράματα στα οποία δοκίμασε πολλά διαφορετικά σύνολα ιδιοτήτων. Για κάθε σύνολο ιδιοτήτων, είχε εκπαιδεύσει το σύστημα σε ένα σύνολο κειμένων εκπαίδευσης (το ίδιο για όλα τα σύνολα ιδιοτήτων) και τον είχε αξιολογήσει σε ένα εντελώς διαφορετικό σύνολο κειμένων αξιολόγησης (το ίδιο για όλα τα σύνολα ιδιοτήτων). Στην εργασία παρέθετε τα αποτελέσματα της αξιολόγησης για κάθε διαφορετικό σύνολο ιδιοτήτων, από τα οποία προέκυπτε το καλύτερο σύνολο ιδιοτήτων και η αντίστοιχη καλύτερη επίδοση του συστήματος στο σύνολο αξιολόγησης. Ωστόσο, οι κριτές του συνεδρίου απέρριψαν την εργασία λέγοντας ότι μέρος της εκπαίδευσης είχε γίνει στα δεδομένα αξιολόγησης. Είχαν δίκιο οι κριτές; Εξηγήστε γιατί. Αν πιστεύετε ότι είχαν δίκιο, εξηγήστε επίσης τι θα έπρεπε να κάνει ο ερευνητής για να αντιμετωπίσει το πρόβλημα.

Είχαν δίκιο, γιατί ο ερευνητής επέλεξε τις ιδιότητες που οδηγούσαν στα καλύτερα αποτελέσματα αξιολόγησης. Ουσιαστικά, δηλαδή, χρησιμοποίησε τα δεδομένα αξιολόγησης για την επιλογή ιδιοτήτων, η οποία αποτελεί μέρος της εκπαίδευσης. Υπάρχει ο κίνδυνος να επέλεξε έτσι ιδιότητες που οδηγούν σε καλά αποτελέσματα στο συγκεκριμένο σύνολο αξιολόγησης και μόνο (πρόβλημα υπερ-εφαρμογής).
Θα έπρεπε να είχε χρησιμοποιήσει ένα ξεχωριστό σύνολο δεδομένων επικύρωσης.
Για κάθε σύνολο ιδιοτήτων, θα έπρεπε να είχε εκπαιδεύσει το σύστημα με τα δεδομένα εκπαίδευσης και να είχε μετρήσει την επίδοσή του στα δεδομένα επικύρωσης, ώστε να επιλέξει το σύνολο ιδιοτήτων με την καλύτερη επίδοση στα δεδομένα επικύρωσης. Κατόπιν, θα έπρεπε να είχε αξιολογήσει το σύστημα με το επιλεγμένο σύνολο ιδιοτήτων στα δεδομένα αξιολόγησης.
Αν είχα ένα σύνολο ιδιοτήτων, τότε η αξιολόγησή του θα ήταν σωστή. Αν όμως επιλέξω το καλύτερο σύμφωνα με την απόδοσή τους, το καλύτερο πρέπει να αξιολογηθεί σε άλλα δεδομένα.

Άσκηση 16.1

Έστω ένα αντικείμενο προς κατάταξη το οποίο παριστάνεται με το διάνυσμα $< X_1, X_2, X_3 > = < 0,1,0 >$. 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 εκτιμήτρια 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

Άσκηση 16.1

Έστω ένα αντικείμενο προς κατάταξη το οποίο παριστάνεται με το διάνυσμα $< X_1, X_2, X_3 > = < 0,1,0 >$. 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 εκτιμήτρια Laplace κατά τις εκτιμήσεις των πιθανοτήτων $P(X_i|C)$. Όλες οι ιδιότητες X_i είναι δυαδικές (έχουν ως πεδίο τιμών το {0, 1}).

$$P(X_1 = 0|C = 1) = \frac{P(X_1 = 0 \cap C = 1)}{P(C = 1)} = \frac{0+1}{2+2} = \frac{\mathbf{1}}{\mathbf{4}}$$

$$P(X_2 = 1|C = 1) = \frac{P(X_2 = 1 \cap C = 1)}{P(C = 1)} = \frac{1+1}{2+2} = \frac{\mathbf{2}}{\mathbf{4}}$$

$$P(X_3 = 0|C = 1) = \frac{P(X_3 = 1 \cap C = 1)}{P(C = 1)} = \frac{0+1}{2+2} = \frac{\mathbf{1}}{\mathbf{4}}$$

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

Επιπλέον $P(C = 1) = \frac{1}{2}$. Άρα

$$P\left(C = 1 \mid \vec{X} = \text{<}0,1,0\text{>}\right) = \frac{1}{P\left(\vec{X} = \text{<}0,1,0\text{>}\right)} P(C = 1) \prod_{i=1}^{m} P(X_i = x_i \mid C = 1) =$$

$$= \frac{1}{P\left(\vec{X} = \text{<}0,1,0\text{>}\right)} \cdot \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{2}{4} \cdot \frac{1}{4} = \frac{1}{P\left(\vec{X} = \text{<}0,1,0\text{>}\right)} \cdot \mathbf{\frac{1}{4^3}}$$

Όμοια βρίσκουμε ότι

$$P\left(C = 0 \mid \vec{X} = \text{<}0,1,0\text{>}\right) = \frac{1}{P\left(\vec{X} = \text{<}0,1,0\text{>}\right)} P(C = 0) \prod_{i=1}^{m} P(X_i = x_i \mid C = 0) =$$

$$= \frac{1}{P\left(\vec{X} = \text{<}0,1,0\text{>}\right)} \cdot \frac{1}{2} \cdot \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{2}{4} = \frac{1}{P\left(\vec{X} = \text{<}0,1,0\text{>}\right)} \cdot \mathbf{\frac{1}{4^2}}$$

Επομένως θα το κατατάξει στην C = 0.

Άσκηση 16.3

Χρησιμοποιούμε μια παραλλαγή του αφελούς ταξινομητή Bayes (της πολυ-μεταβλητής μορφής Bernoulli που συναντήσαμε στο μάθημα), 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 $\log a^b = b \cdot \log a$.

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

$$P(C=1) \cdot \prod_{i=1}^{m} \left(P(t_i | C = 1)^{x_i} \cdot [1 - P(t_i | C = 1)]^{1 - x_i}\right) \geq K$$

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

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

$$\boxed{\log(a \cdot b) = \log a + \log b}$$

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

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

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

$$\boxed{\log a^b = b \cdot \log a}$$

$$\log P(C=1) + \sum_{i=1}^{m} \log[1 - P(t_i|C=1)] - \log K + \left(\sum_{i=1}^{m} \log P(t_i|C=1) - \sum_{i=1}^{m} \log[1 - P(t_i|C=1)] \right) x_i \geq 0$$

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

Επομένως πρόκειται για γραμμικό ταξινομητή.

Άσκηση 16.4

α) Βάσει των παραδειγμάτων εκπαίδευσης του πίνακα, πόση είναι η εντροπία H(C) της κατηγορίας C και γιατί;

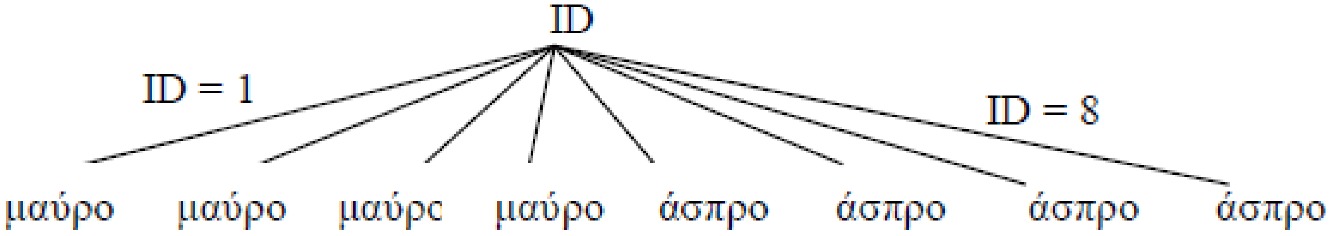
P(C=μαύρο) = P(C=άσπρο) $= \frac{1}{2}$. Έχουμε δύο ισοπίθανα ενδεχόμενα C = μαύρο και C = άσπρο και άρα μέγιστη εντροπία (αβεβαιότητα), που για δύο ενδεχόμενα είναι ίση με 1. Το ίδιο συμπέρασμα προκύπτει από τον ορισμό της εντροπίας, με αριθμητικούς υπολογισμούς.

ID	X	Y	Z	C
1	0	0	1	Μαύρο
2	1	0	1	Μαύρο
3	0	0	1	Μαύρο
4	1	1	0	Μαύρο
5	0	1	1	Άσπρο
6	1	0	0	Άσπρο
7	0	0	0	Άσπρο
8	1	0	0	Άσπρο

β) Σχεδιάστε το δέντρο απόφασης που θα κατασκευάσει ο ID3 (Iterative Dichotomiser 3), αν του επιτρέψουμε να χρησιμοποιήσει ως ιδιότητες τις ID, X, Y, Z (το δέντρο προβλέπει την κατηγορία C). Δεν χρειάζονται αριθμητικοί υπολογισμοί, μόνο προσεκτική παρατήρηση των δεδομένων. Εξηγήστε γιατί θα προκύψει το δέντρο που σχεδιάσατε.

Αν ξέρουμε την τιμή της ιδιότητας ID, τότε ξέρουμε με βεβαιότητα την τιμή της κατηγορίας C όλων των παραδειγμάτων. Βάσει, δηλαδή, των παραδειγμάτων του πίνακα, $H(C \mid ID) = 0$ και επομένως $IG(ID, C) = H(C) - H(C \mid ID) = 1$. Αντίθετα, καμία από τις άλλες ιδιότητες (X, Y, Z) δεν προβλέπει με απόλυτη βεβαιότητα την κατηγορία όλων των παραδειγμάτων, επομένως το κέρδος πληροφορίας που παρέχουν είναι μικρότερο από 1. Επομένως ο ID3 θα προτιμήσει να τοποθετήσει στη ρίζα του δέντρου απόφασης την ερώτηση για την ιδιότητα ID. Θα υπάρχουν 8 κλαδιά κάτω από τη ρίζα, ένα για κάθε δυνατή τιμή της ID που εμφανίζεται στα παραδείγματα εκπαίδευσης. Στο υποδέντρο κάτω από κάθε κλαδί θα καταλήξει ακριβώς ένα από τα παραδείγματα, εκείνο με την αντίστοιχη τιμή ID. Επομένως τα παραδείγματα (είναι μόνο ένα) κάθε υποδέντρου θα ανήκουν σε μία μόνο (ανά υποδέντρο) κατηγορία. Άρα ο ID3 θα σταματήσει και θα επιστρέψει το παρακάτω δέντρο, που δεν χρησιμοποιεί τις άλλες ιδιότητες.

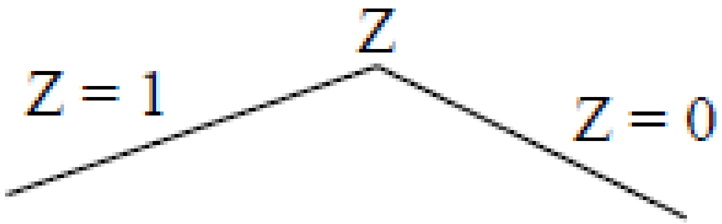
ID	X	Y	Z	C
1	0	0	1	Μαύρο
2	1	0	1	Μαύρο
3	0	0	1	Μαύρο
4	1	1	0	Μαύρο
5	0	1	1	Άσπρο
6	1	0	0	Άσπρο
7	0	0	0	Άσπρο
8	1	0	0	Άσπρο



γ) Σχεδιάστε τώρα το δέντρο απόφασης που θα κατασκευάσει ο ID3, αν του επιτρέψουμε να χρησιμοποιήσει ως ιδιότητες μόνο τις X, Y, Z (το δέντρο προβλέπει πάλι την κατηγορία C). Δεν χρειάζονται αριθμητικοί υπολογισμοί, μόνο προσεκτική παρατήρηση των δεδομένων. Εξηγήστε γιατί θα προκύψει το δέντρο που σχεδιάσατε.

Τα δεδομένα εκπαίδευσης δείχνουν ότι αν μάθουμε πως Z = 1,
είναι πολύ πιθανό (πιθανότητα ¾) ότι C = μαύρο· και αν μάθουμε πως Z = 0, είναι πολύ πιθανό (πιθανότητα ¾) ότι C = άσπρο. Επομένως, η γνώση της τιμής της ιδιότητας Z μειώνει την εντροπία (αβεβαιότητα για την τιμή) της C, εντροπία που αρχικά ήταν μέγιστη, δηλαδή IG(C, Z) > 0. Το ίδιο συμπέρασμα προκύπτει υπολογίζοντας αναλυτικά το IG(C, Z).
Αντίθετα, τα δεδομένα εκπαίδευσης δείχνουν ότι αν μάθουμε πως Y = 1,
η πιθανότητα να έχουμε C = μαύρο παραμένει ½ και ίση με την πιθανότητα να έχουμε C = άσπρο· ομοίως, αν μάθουμε ότι X = 0, η πιθανότητα να έχουμε C = μαύρο παραμένει ½ και ίση με την πιθανότητα να έχουμε C = άσπρο. Επομένως, όποια κι αν είναι η τιμή της Y, εξακολουθούμε να έχουμε δύο ισοπίθανα ενδεχόμενα C = μαύρο και C = άσπρο και, επομένως, μέγιστη εντροπία (αβεβαιότητα) H(C) = 1. Άρα η γνώση της τιμής της Y δεν μειώνει καθόλου την εντροπία (αβεβαιότητα για την τιμή της) C, δηλαδή IG(C, Y) = 0. Το ίδιο συμπέρασμα προκύπτει υπολογίζοντας αναλυτικά το IG(C, Y). Ομοίως IG(C, X) = 0. Επομένως ο ID3 θα επιλέξει να τοποθετήσει στην κορυφή του δέντρο απόφασης την ερώτηση για την ιδιότητα Z.

ID	X	Y	Z	C
1	0	0	1	Μαύρο
2	1	0	1	Μαύρο
3	0	0	1	Μαύρο
4	1	1	0	Μαύρο
5	0	1	1	Άσπρο
6	1	0	0	Άσπρο
7	0	0	0	Άσπρο
8	1	0	0	Άσπρο

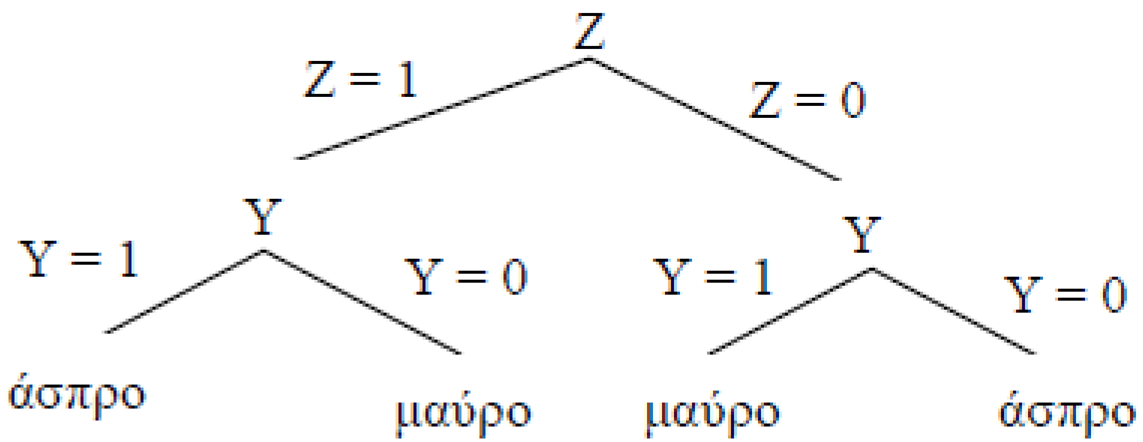


Στο υποδέντρο για Z = 1, θα καταλήξουν τα παραδείγματα με ID = 1, 2, 3, 5, ενώ στο υποδέντρο για Z = 0 τα παραδείγματα με *ID = 4, 6, 7, 8.*
Και στα δύο υποδέντρα, αν μάθουμε κατόπιν την τιμή της ιδιότητας Y, πετυχαίνουμε πλήρη διαχωρισμό (πρόβλεψη) των κατηγοριών, ενώ αντίθετα δεν συμβαίνει το ίδιο αν μάθουμε την τιμή της ιδιότητας X.
Επομένως και στα δύο υποδέντρα η ιδιότητα Y παρέχει μεγαλύτερο κέρδος πληροφορίας απ΄ ό,τι η X.

ID	X	Y	Z	C
1	0	0	1	Μαύρο
2	1	0	1	Μαύρο
3	0	0	1	Μαύρο
5	0	1	1	Άσπρο

ID	X	Y	Z	C
4	1	1	0	Μαύρο
6	1	0	0	Άσπρο
7	0	0	0	Άσπρο
8	1	0	0	Άσπρο

Άρα ο ID3 θα προτιμήσει να προσθέσει και στα δύο υποδέντρα της ερωτήσεις για την ιδιότητα Y. Σε κάθε κλαδί κάτω από τις δύο ερωτήσεις Y, καταλήγουν παραδείγματα μίας μόνο κατηγορίας. Επομένως ο ID3 θα σταματήσει και θα επιστρέψει το παρακάτω δέντρο, που δεν χρησιμοποιεί την ιδιότητα X.



ID	X	Y	Z	C
1	0	0	1	Μαύρο
2	1	0	1	Μαύρο
3	0	0	1	Μαύρο
5	0	1	1	Άσπρο

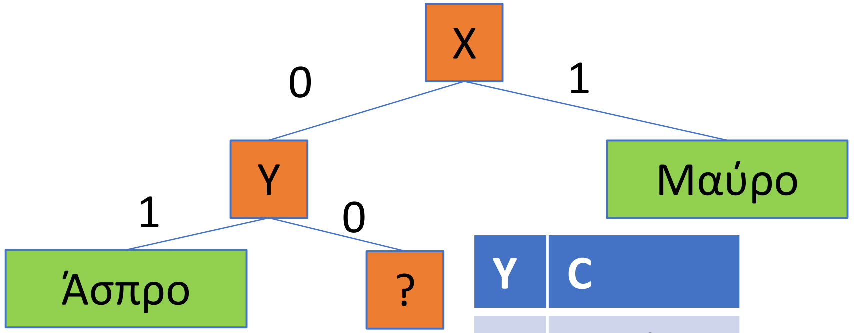
ID	X	Y	Z	C
4	1	1	0	Μαύρο
6	1	0	0	Άσπρο
7	0	0	0	Άσπρο
8	1	0	0	Άσπρο

Άσκηση 16.5

Αν αξιολογήσουμε έναν ταξινομητή ID3 (χωρίς πριόνισμα) στο ίδιο σύνολο διανυσμάτων στο οποίο τον εκπαιδεύσαμε, αλλά στα διανύσματα εκπαίδευσης περιλαμβάνονται και ασυνεπή παραδείγματα, το ποσοστό ορθότητας του ταξινομητή θα βρεθεί να είναι:

Τα ασυνεπή διανύσματα εκπαίδευσης δεν είναι δυνατόν να διαχωριστούν,
αφού έχουν τις ίδιες τιμές σε όλες τις ιδιότητες, οπότε καταλήγουν στα ίδια φύλλα και (αφού ανήκουν σε διαφορετικές κατηγορίες) δεν συμφωνούν οι κατηγορίες όλων τους με τις κατηγορίες των φύλλων στα οποία κατέληξαν. Αφού αξιολογούμε χρησιμοποιώντας τα ίδια διανύσματα που χρησιμοποιήθηκαν κατά την εκπαίδευση, κάθε διάνυσμα αξιολόγησης καταλήγει στο ίδιο φύλλο όπου κατέληξε και κατά την εκπαίδευση και κάποια από τα ασυνεπή διανύσματα αξιολόγησης (και εκπαίδευσης) καταλήγουν πάλι σε φύλλα των οποίων οι κατηγορίες είναι διαφορετικές από εκείνες των ασυνεπών διανυσμάτων. Επομένως, θα υπάρχουν σίγουρα λάθη κατάταξης και άρα το ποσοστό ορθότητας θα είναι σίγουρα μικρότερο του 100%.

Y	C
0	Μαύρο
0	Άσπρο
1	Άσπρο



Y	C
0	Μαύρο
0	Άσπρο

X	Y	C
0	0	Μαύρο
1	0	Μαύρο
0	0	Άσπρο
0	1	Άσπρο