

Ασκήσεις μελέτης της 9^{ης} διάλεξης

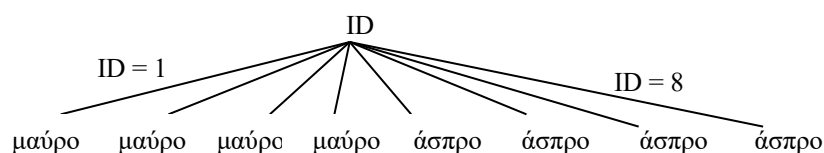
9.1. (α) Βάσει των παραδειγμάτων εκπαίδευσης του πίνακα, πόση είναι η εντροπία $H(C)$ της κατηγορίας C και γιατί;

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

Απάντηση: $P(C=\text{μαύρο}) = P(C=\text{άσπρο}) = 1/2$. Έχουμε δύο ισοπίθανα ενδεχόμενα $C = \text{μαύρο}$ και $C = \text{άσπρο}$ 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 (κάντε τους).

(β) Σχεδιάστε το δέντρο απόφασης που θα κατασκευάσει ο ID3, 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.

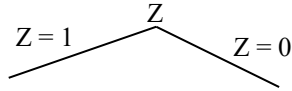
Απάντηση: Αν ξέρουμε την τιμή της ιδιότητας ID, τότε ξέρουμε με βεβαιότητα την τιμή της κατηγορίας C όλων των παραδειγμάτων. Βάσει, δηλαδή, των παραδειγμάτων του πίνακα, $H(C | ID) = 0$ και επομένως $IG(ID, C) = H(C) - H(C | ID) = 1$. (Το ίδιο συμπέρασμα προκύπτει από τον ορισμό του IG, με αριθμητικούς υπολογισμούς. Κάντε τους.) 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 κάθε υποδέντρου (μόνο ένα παράδειγμα ανά υποδέντρο) 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.



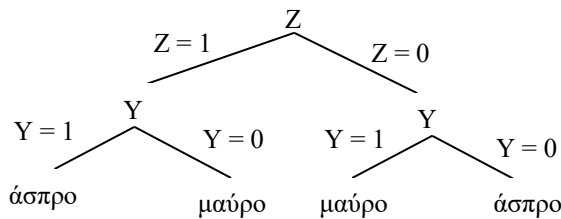
(γ) Σχεδιάστε τώρα το δέντρο απόφασης που θα κατασκευάσει ο ID3, αν του επιτρέψουμε να χρησιμοποιήσει ως ιδιότητες μόνο τις X, Y, Z (το δέντρο προβλέπει πάλι την κατηγορία C). Δεν χρειάζονται αριθμητικοί υπολογισμοί, μόνο προσεκτική παρατήρηση των δεδομένων. Εξηγήστε γιατί θα προκύψει το δέντρο που σχεδιάσατε.

Απάντηση: Τα δεδομένα εκπαίδευσης δείχνουν ότι αν πληροφορηθούμε πως $Z = 1$, είναι πολύ πιθανό (πιθανότητα $3/4$) ότι $C = \text{μαύρο}$ και αν πληροφορηθούμε πως $Z = 0$, είναι πολύ πιθανό (πιθανότητα $3/4$) ότι $C = \text{άσπρο}$. 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 $IG(C, Z)$ (υπολογίστε το). Αντίθετα, τα δεδομένα εκπαίδευσης δείχνουν ότι αν πληροφορηθούμε πως $Y = 1$, η πιθανότητα να έχουμε $C = \text{μαύρο}$ παραμένει $1/2$ και ίση με την πιθανότητα να έχουμε $C = \text{άσπρο}$: ομοίως, αν πληροφορηθούμε ότι $X = 0$, η πιθανότητα να έχουμε $C = \text{μαύρο}$ παραμένει $1/2$ και ίση με την πιθανότητα να έχουμε $C = \text{άσπρο}$. Επομένως, όποια κι αν είναι η τιμή της Y, εξακολουθούμε να

έχουμε δύο ισοπίθανα ενδεχόμενα $C = \text{μαύρο}$ και $C = \text{άσπρο}$ 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 .



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 . 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 .



9.2. 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:

- σίγουρα 100%
- X σίγουρα μικρότερο του 100%
- τίποτα από τα παραπάνω (το αποτέλεσμα εξαρτάται από τις τυχαιότητες των δεδομένων).

Απάντηση: 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%.