ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Τομέας Επικοινωνιών, Ηλεκτρονικής & Συστημάτων Πληροφορικής
Εργαστήριο Διαχείρισης και Βέλτιστου Σχεδιασμού Δικτύων - NETMODE
Ηρώων Πολυτεχνείου 9, Ζωγράφου, 157 80 Αθήνα, Τηλ: 210-772.2503, Fax: 210-772.1452
e-mail: maglaris@netmode.ntua.gr, URL: http://www.netmode.ntua.gr

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ
(ΔΠΜΣ Επιστήμη Δεδομένων & Μηχανική Μάθηση)

Ο Ταξινομητής Naive Bayes – Δέντρα Αποφάσεων (Εξετάσεις 2020)

α) Ποια είναι η παραδοχή που κάνει ο αλγόριθμος Naive Bayes Classifier και ποια πλεονεκτήματα προσφέρει;

β) Βασισμένοι στον παρακάτω πίνακα, να εκπαιδεύσετε έναν Naive Bayes Classifier που θα ταξινομεί αν ένας ταξιδιώτης που καταφθάνει στο αεροδρόμιο της Αθήνας είναι θετικός στον ιό Covid-19 ή όχι. Το μέγεθος του training set ήταν 1000 ταξιδιώτες, από τους οποίους οι 600 ήταν γυναίκες και οι 400 ήταν άνδρες. Τα features είναι (i) το φύλο του ταξιδιώτη (Άνδρας, Γυναίκα), (ii) αν η θερμοκρασία του σώματός του είναι υψηλή ή όχι (Υψηλή, Χαμηλή) και (iii) το αν έρχεται από αεροδρόμιο της ελληνικής επικράτειας ή όχι (Ελλάδα, Εξωτερικό). Δίνεται ένας ταξιδιώτης για τον οποίο ισχύει: είναι Γυναίκα, έχει χαμηλή θερμοκρασία και προέρχεται από την Ελλάδα. Ποια είναι η εκτίμηση του μοντέλου που εκπαιδεύσατε για αυτόν τον ταξιδιώτη; (Για την εκπαίδευση του αλγορίθμου να υπολογίσετε μόνο τις πιθανότητες που είναι απαραίτητες για την ταξινόμηση του ταξιδιώτη)

Φύλο	**Θερμοκρασία**	**Προέλευση**	**Θετικοί στον Ιό**	**Αρνητικοί στον Ιό**
Γυναίκα	Υψηλή	Ελλάδα	40	100
Γυναίκα	Υψηλή	Εξωτερικό	170	50
Γυναίκα	Χαμηλή	Ελλάδα	10	150
Γυναίκα	Χαμηλή	Εξωτερικό	20	60
Άνδρας	Υψηλή	Ελλάδα	20	80
Άνδρας	Υψηλή	Εξωτερικό	100	20
Άνδρας	Χαμηλή	Ελλάδα	10	110
Άνδρας	Χαμηλή	Εξωτερικό	10	50

γ) Εξετάζετε την κατασκευή ενός Decision Tree για την επίλυση της παραπάνω ταξινόμησης. Να εξηγήσετε ποιο feature (Φύλο, Θερμοκρασία, Προέλευση) θα επιλέγατε στη ρίζα του Decision Tree. Να χρησιμοποιήσετε το Gini Index.

Λύση

α) Δείτε τις διαφάνειες του μαθήματος.

β) Για τις Prior πιθανότητες:

$$P(\text{ΘΕΤΙΚΟΣ/Η}) = \frac{40 + 170 + 10 + 20 + 20 + 100 + 10 + 10}{1000} = \frac{380}{1000} = 0.38$$

$$P(\text{ΑΡΝΗΤΙΚΟΣ/Η}) = \frac{100 + 50 + 150 + 60 + 80 + 20 + 110 + 50}{1000} = \frac{620}{1000} = 0.62$$

Για τις δεσμευμένες πιθανότητες:

$$P(\text{ΓΥΝΑΙΚΑ}|\text{ΘΕΤΙΚΟΣ/Η}) = \frac{40 + 170 + 10 + 20}{380} = \frac{240}{380} = 0.632$$

$$P(\text{ΓΥΝΑΙΚΑ}|\text{ΑΡΝΗΤΙΚΟΣ/Η}) = \frac{100 + 50 + 150 + 60}{620} = \frac{360}{620} = 0.5807$$

$$P(\text{ΧΑΜΗΛΗ}|\text{ΘΕΤΙΚΟΣ/Η}) = \frac{10 + 20 + 10 + 10}{380} = \frac{50}{380} = 0.1316$$

$$P(\text{ΧΑΜΗΛΗ}|\text{ΑΡΝΗΤΙΚΟΣ/Η}) = \frac{150 + 60 + 110 + 50}{620} = \frac{370}{620} = 0.5968$$

$$P(\text{ΕΛΛΑΔΑ}|\text{ΘΕΤΙΚΟΣ/Η}) = \frac{40 + 10 + 20 + 10}{380} = \frac{80}{380} = 0.2105$$

$$P(\text{ΕΛΛΑΔΑ}|\text{ΑΡΝΗΤΙΚΟΣ/Η}) = \frac{100 + 150 + 80 + 110}{620} = \frac{440}{620} = 0.7097$$

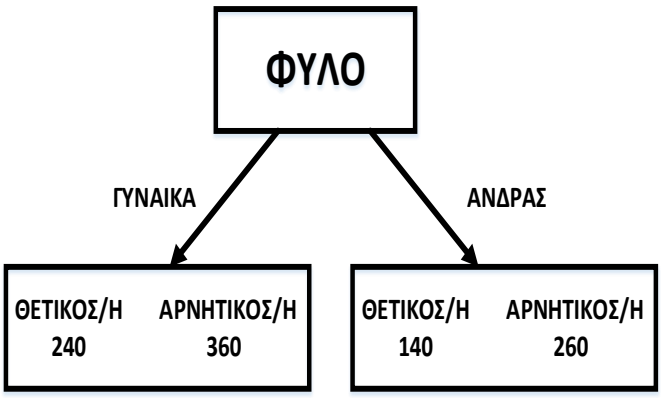
Για την ταξινόμηση του ταξιδιώτη που ζητείται από την άσκηση:

$$P(\text{ΘΕΤΙΚΟΣ/Η} \mid \text{ΓΥΝΑΙΚΑ}, \text{ΧΑΜΗΛΗ}, \text{ΕΛΛΑΔΑ}) = 0.38 \cdot 0.632 \cdot 0.1316 \cdot 0.2105 = 0.0067$$

$$P(\text{ΑΡΝΗΤΙΚΟΣ/Η} \mid \text{ΓΥΝΑΙΚΑ}, \text{ΧΑΜΗΛΗ}, \text{ΕΛΛΑΔΑ}) = 0.62 \cdot 0.5807 \cdot 0.5968 \cdot 0.7097 = 0.1525$$

Άρα η ταξιδιώτης είναι αρνητική.

γ) Για το feature "ΦΥΛΟ":

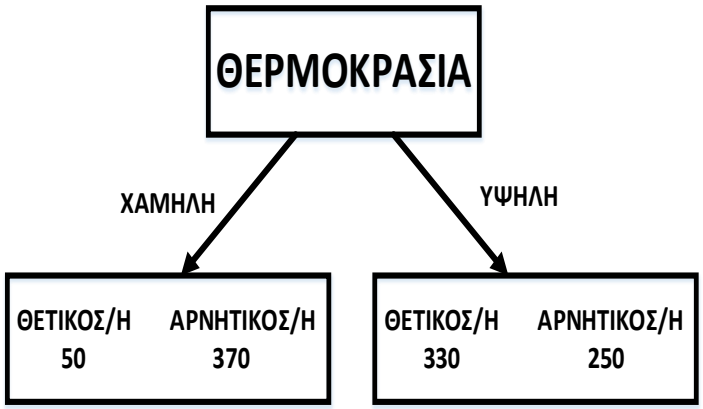


$$Gini_{left} = 1 - \left(\frac{240}{240+360}\right)^2 - \left(\frac{360}{240+360}\right)^2 = 0.48$$

$$Gini_{right} = 1 - \left(\frac{140}{140+260}\right)^2 - \left(\frac{260}{140+260}\right)^2 = 0.455$$

$$Gini_{weighted} = 0.48 \cdot \left(\frac{600}{1000}\right) + 0.455 \cdot \left(\frac{400}{1000}\right) = 0.47$$

Για το feature "ΘΕΡΜΟΚΡΑΣΙΑ":

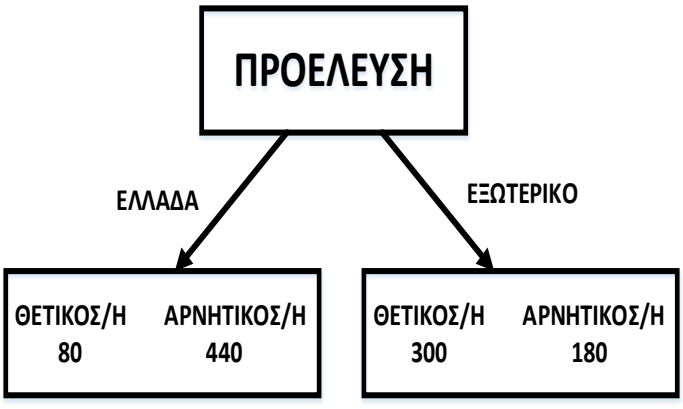


$$Gini_{left} = 1 - \left(\frac{50}{50+370}\right)^2 - \left(\frac{370}{50+370}\right)^2 = 0.20975$$

$$Gini_{right} = 1 - \left(\frac{330}{330+250}\right)^2 - \left(\frac{250}{330+250}\right)^2 = 0.49049$$

$$Gini_{weighted} = 0.20975 \cdot \left(\frac{420}{1000}\right) + 0.49049 \cdot \left(\frac{580}{1000}\right) = 0.37258$$

Για το feature "ΠΡΟΕΛΕΥΣΗ":



$$Gini_{left} = 1 - \left(\frac{80}{80+440}\right)^2 - \left(\frac{440}{80+440}\right)^2 = 0.26036$$

$$Gini_{right} = 1 - \left(\frac{300}{300+180}\right)^2 - \left(\frac{180}{300+180}\right)^2 = 0.46875$$

$$Gini_{weighted} = 0.26036 \cdot \left(\frac{520}{1000}\right) + 0.4688 \cdot \left(\frac{480}{1000}\right) = 0.36003$$

Επιλέγεται η "ΠΡΟΕΛΕΥΣΗ" γιατί έχει τη μικρότερη τιμή για το Gini Index.

Συντάχθηκε από τους υπεύθυνους εργαστηριακής υποστήριξης του μαθήματος Νίκο Κωστόπουλο και Δημήτρη Πανταζάτο, Υποψήφιους Διδάκτορες Ε.Μ.Π.