PROJECT ΘΕΩΡΙΑ ΑΠΟΦΑΣΕΩΝ

Pattern Analysis

Παναγιώτης Τριανταφυλλόπουλος
1054367
Ανδρέας Κάλλιστρος
1054351

Ερώτημα 1ο

Αρχικά τα χαρακτηριστικά κάθε δείγματος είναι 10 και η τελευταία στήλη περιέχει την ετικέτα κατηγορίας κάθε δείγματος (1 για ασθενές, 2 για υγιές άτομο). Τα δείγματα εκπαίδευσης είναι 583 και τα υγιή άτομα είναι 167 και τα ασθενή 416. Για να εισάγουμε το αρχείο στο MATLAB χρησιμοποιούμε το import tool του MATLAB και μέσω αυτού αντικαθιστούμε τις κενές τιμές με NaN. Αφού αλλάξαμε τις τιμές Male, Female σε 0, 1 αντίστοιχα παρατηρήσαμε μερικές κενές τιμές στο $10^{ο}$ χαρακτηριστικό κάποιων δειγμάτων, τις οποίες αντικαταστήσαμε με την μέση τιμή των υπόλοιπων τιμών αυτής της στήλης. Στην συνέχεια κανονικοποιήσαμε τις 10 πρώτες στήλες με την συνάρτηση normalize() της MATLAB η οποία χρησιμοποιεί τον παρακάτω τύπο:

$$z = \frac{(x - \mu)}{\sigma}.$$

όπου μ η μέση τιμή και σ η τυπική απόκλιση κάθε στήλης

Ερώτημα 2ο

Όταν αναφερόμαστε στον Αφελή Bayes (Naïve Bayes) ταξινομητή, εννοούμε ένα σύνολο αλγορίθμων που όλοι βασίζονται στην υπόθεση ότι κάθε χαρακτηριστικό συνεισφέρει, ανεξάρτητα από πιθανές συσχετίσεις με τα υπόλοιπα χαρακτηριστικά. Δηλαδή θεωρούμε ότι όλα τα χαρακτηριστικά:

- Είναι **ανεξάρτητα** μεταξύ τους.
- Έχουν την **ίδια βαρύτητα** στην τελική απόφαση.

Θα εξετάσουμε πρώτα την περίπτωση που τα δεδομένα μας θα είναι διακριτά. Τότε υπολογίζουμε την εκ των υστέρων(a posteriori) πιθανότητα για κάθε πιθανή έξοδο-κλάση η οποία είναι ίση με το γινόμενο της εκ των προτέρων(a priori) πιθανότητας με το γινόμενο της πιθανοφάνειας (likelihood) κάθε χαρακτηριστικού, δεδομένης της αντίστοιχης εξόδου. Στην συνέχεια κανονικοποιούμε, διαιρώντας το παραπάνω γινόμενο με το γινόμενο των πιθανοτήτων όλων των χαρακτηριστικών.

Εκ των υστέρων: $P(\omega_j|x)$

Εκ των προτέρων: $P(\omega_j)$

Πιθανοφάνεια: $P(x_i|\omega_j)$

Όπου ω_j είναι η έξοδος-κλάση j και x_i είναι το χαρακτηριστικό i. Αφού υπολογίσουμε τα παραπάνω, επιλέγουμε την κλάση με την μεγαλύτερη εκ των υστέρων πιθανότητα (για τα δεδομένα χαρακτηριστικά).

Στην περίπτωση που τα δεδομένα μας είναι συνεχή τότε θεωρούμε ότι αυτά ακολουθούν κανονική κατανομή (Για αυτό ο αλγόριθμος είναι γνωστός και ως Gaussian Naïve Bayes). Τότε ακολουθούμε την ίδια διαδικασία μόνο που σε αυτή την περίπτωση η πιθανοφάνεια κάθε συνεχούς χαρακτηριστικού υπολογίζεται από τον εξής τύπο:

$$P(x_i|\omega_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \cdot e^{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}}$$

Όπου x_i είναι το χαρακτηριστικό i , ω_j η κλάση j και μ_j,σ_j είναι η μέση τιμή και η διασπορά του χαρακτηριστικού x_i για την κλάση ω_j.

Υπάρχουν επίσης οι εξής παραλλαγές:

- Multinomial Naïve Bayes (για δεδομένα που αντιπροσωπεύουν συχνότητες γεγονότων μιας πολυωνυμικής κατανομής)
- Bernoulli Naïve Bayes (για Boolean μεταβλητές)

Παρά τις απλουστευμένες υποθέσεις του, ο Αφελής ταξινομητής Bayes (τελείως ανεξάρτητα δεδομένα) έχει αποδειχθεί καλύτερος από τον απλό Bayes σε πολλά προβλήματα (όπως για παράδειγμα σε ταξινομήσεις με πολλές κλάσεις, για ιατρικά δεδομένα, και σε προβλέψεις σε πραγματικό χρόνο κ.α.). Επιπλέον απαιτεί μικρό αριθμό δειγμάτων εκπαίδευσης για την εκπαίδευσή του και είναι

αρκετά γρήγορος (σε σύγκριση με τον Bayes αλλά και άλλους περίπλοκους ταξινομητές). Ωστόσο στην πραγματικότητα η υπόθεση περί ανεξαρτησίας των δεδομένων είναι πολύ δύσκολο να ισχύει, καθώς τα δεδομένα σχεδόν πάντοτε σχετίζονται μεταξύ τους. Από την άλλη ο Bayes είναι πολύ πιο ακριβής και ελαχιστοποιεί το ρίσκο της απόφασης, υστερώντας όμως στον χρόνο ταξινόμησης.

Ερώτημα 3ο

Αφού εκπαιδεύσαμε τον ταξινομητή με χρήση των συναρτήσεων της MATLAB (fitcnb(), crossval() για Naïve Bayes και 5-fold cross validation αντίστοιχα) υπολογίσαμε τον πίνακα σύγχυσης (confusion matrix). Υπολογίσαμε τα specificity και sensitivity από το confusion matrix:

$$SPEC = \frac{TN}{TN + FP} \qquad SENS = \frac{TP}{TP + FN}$$

Confusion matrix	Predicted ill	Predicted healthy
ILL	TP	FN
HEALTHY	FP	TN

sensitivity≈0,39 specificity≈0,95

Παρατηρήσαμε ότι το specificity είναι πολύ κοντά στη μέγιστη τιμή του (1) που σημαίνει ότι ο ταξινομητής ταξινόμησε με μεγάλο ποσοστό επιτυχίας τους ασθενείς που δεν ήταν άρρωστοι (λίγοι υγιείς που ταξινομήθηκαν ως ασθενείς).

Όσον αφορά το sensitivity παρατηρούμε ότι η τιμή του είναι αρκετά χαμηλή που σημαίνει ότι ο ταξινομητής απέτυχε να ταξινομήσει πολλούς ασθενείς που πράγματι ήταν άρρωστοι (πολλοί άρρωστοι ταξινομήθηκαν ως υγιείς).

Ο γεωμετρικός μέσος που υπολογίζεται από τον παρακάτω τύπο:

Geometric Mean = sqrt (Sensitivity * Specificity) είναι περίπου ίσος με 0,62. Δεδομένου ότι ο γεωμετρικός μέσος έχει μέγιστη τιμή το 1, ο ταξινομητής μας δεν είναι και πολύ καλός. Πιο συγκεκριμένα επειδή τα δεδομένα μας είναι ιατρικά, είναι προτιμότερο να έχουμε σχετικά υψηλό sensitivity και χαμηλό specificity, ώστε να μην υπάρχουν πολλοί ασθενείς που θα διαγνωστούν ως υγιείς, ενώ δεν μας επηρεάζει τόσο πολύ να έχουμε υγιή άτομα που διαγνώστηκαν λάθος. Δηλαδή το κόστος της λάθος ταξινόμησης ασθενούς είναι σημαντικά μεγαλύτερο από το κόστος της λάθος ταξινόμησης υγιούς ατόμου.