Εργαστηριακή Άσκηση

για το μάθημα Θεωρία Αποφάσεων

2022-2023

Ονοματεπώνυμο: Πύργας Αθανάσιος **ΑΜ: 1041866 (παλαιός: 236189)**

Ερώτημα 1. Προεπεξεργασία δεδομένων

a) Να αναφέρετε, πόσα είναι τα χαρακτηριστικά κάθε δείγματος και πόσα δείγματα εκπαίδευσης περιέχει το αρχείο.

Απάντηση: Το αρχείο περιέχει συνολικά **583 δείγματα εκπαιδευσης**, και κάθε δείγμα εκπαίδευσης περιέχει **10 χαρακτηριστικά**. Επίσης κάθε δείγμα περιέχει και την πληροφορία για το αν ο ασθενής πάσχει ή όχι από ασθένεια στο συκώτι.

b) Η δεύτερη στήλη περιέχει το φύλο του ανθρώπου που συμμετείχε στο δείγμα. Στο αρχείο όμως είναι σημειωμένη με Male για αρσενικό και Female για θηλυκό. Προκειμένου να την χρησιμοποιήσουμε σαν είσοδο θα πρέπει να αντιστοιχίσετε το Male με την τιμή 0 και το Female με την τιμή 1.

Απάντηση: Η διαδικασία εκτελείται και περιγράφεται στο αρχείο **e1b.py** . Το παραπάνω πρόγραμμα κάνει και κάποιες ακόμη μικρές αλλαγές στα δεδομένα. Συγκεκριμένα αλλάζει τον τρόπο απεικόνισης της ασθένειας **απο 1 και 2 σε 0 και 1** . Δεν υπάρχει κάποιος ίδιαίτερος λόγος για την αλλαγή αυτή, ωστόσω, σε προβλήματα classification, ταιριάζει καλύτερα το labeling να γίνεται με τις τιμές 0 και 1. Επίσης διορθώνει τυχόν ελλείψεις τιμών (missing values), είτε με την αντικατάσταση τους από τον αριθμητικό μέσο της εκάστοτε στήλης (αν πρόκειται για αριθμητικά δεδομένα) είτε με τυχαία ανάθεση μίας τιμής (αν πρόκειται για δεδομένα κατηγοριοποίησης)

c) Το εύρος τιμών των δεδομένων που σας έχουν δοθεί διαφέρει σημαντικά ανά χαρακτηριστικό. Για αυτό τον λόγο, για να μην υπερεκτιμηθεί η συνεισφορά κάποιου χαρακτηριστικού έναντι άλλων, θα πρέπει πριν την επεξεργασία των χαρακτηριστικών εισόδου να κανονικοποιηθούν στο εύρος [-1,1]. Χρησιμοποιήστε το matlab (ή όποια άλλη εφαρμογή θέλετε) τόσο για το διάβασμα του αρχείου που σας δίνεται όσο και για την κανονικοποίηση των δεδομένων εισόδου στο εύρος τιμών [-1,1].

Απάντηση: Η διαδικασία εκτελείται και περιγράφεται στο αρχείο **e1c.py** .

Ερώτημα 2. Στο μάθημα συζητήθηκε εκτεταμένα ο ταξινομητής Bayes. Στη βιβλιογραφία, υπάρχει μια παραλλαγή του που λέγεται Αφελής Ταξινομητής Bayes (Naïve Bayes), με την υπόθεση ότι τα χαρακτηριστικά είναι στατιστικά ανεξάρτητα. Αναζητήστε τη σχετική βιβλιογραφία στο Internet, και να κάνετε μια σύντομη παρουσίαση του αλγορίθμου. Στη συνέχεια να κάνετε μια σύγκριση με τον Ταξινομητή Bayes.

Απάντηση: Ο ταξινομητής Bayes, είναι ένας ταξινομητής που βασίζεται στο θεώρημα Bayes το οποίο περιγράφεται από την παρακάτω σχέση:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Για πρακτικούς λόγους, χρησιμοποιούμε μία παραλλαγή του παραπάνω ταξινομητή η οποία ονομάζεται Naive Bayes Ταξινομητής και κάνει την εξής "αφελή" παραδοχή:

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y)$$

Με απλά λόγια, κάνει την παραδοχή ότι κάθε ένα χαρακτηριστικό είναι ανεξάρτητο από όλα τα υπόλοιπα, και μετατρέπει το θεώρημα του Bayes στην παρακάτω σχέση:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Επίσης, εφόσον η πιθανότητα $P(x_1,\ldots,x_n)$ είναι σταθερή δεδομένης της εισόδου, μπορούμε να χρησιμοποιήσουμε τον παρακάτω κανόνα ταξινόμησης:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

Ερώτημα 3. Με χρήση της μεθόδου 5-fold cross validation, να εκπαιδεύσετε τον Naïve Bayes ταξινομητή, να παρουσιάστε και να σχολιάσετε την απόδοσή του. Μπορείτε να χρησιμοποιήσετε κατάλληλες συναρτήσεις του matlab ή οποιαδήποτε εφαρμογή επιθυμείτε (ή να υλοποιήσετε δικό σας κώδικα). Για την αξιολόγηση της απόδοσης του ταξινομητή να χρησιμοποιήσετε τις μετρικές του ερωτήματος 4, παρακάτω.

Απάντηση: Η διαδικασία εκτελείται και περιγράφεται στο αρχείο **e3.py** . Μιας και το σετ δεδομένων μας περιέχει συνεχή δεδομένα (βιοϊατρικές μετρήσεις), θα χρησιμοποιήσουμε την Gaussian έκδοση του Naive Bayes που μας προσφέρει η βιβλιοθήκη Scikit Learn της Python. Παρουσιάζονται παρακάτω τα αποτελέσματα του αλγορίθμου σύμφωνα με τις μετρικές Sensitivity, Specificity και Geometric Mean για κάθε Fold αλλα και η μέση τιμή αυτών.

Fold 1:
- Sensitivity: 0.5244
- Specificity: 0.7714
- Geometric Mean: 0.6360

Fold 2:
- Sensitivity: 0.4130
- Specificity: 1
- Geometric Mean: 0.6427

Fold 3:

Sensitivity: 0.2125
Specificity: 0.9460
Geometric Mean: 0.4483

Fold 4:

Sensitivity: 0.3157
Specificity: 1
Geometric Mean: 0.5620

Fold 5:

Sensitivity: 0.5581
Specificity: 1
Geometric Mean: 0.7471

Mean Values:

Sensitivity: 0.4048
Specificity: 0.9435
Geometric Mean: 0.6072

```
Sensitivity per fold
[0.52439024 0.41304348 0.2125     0.31578947 0.55813953]

Mean sensitivity
0.404772546146248

Specificity per fold
[0.77142857 1.         0.94594595 1.         1.        ]

Mean Specificity
0.9434749034749036

Geometric Mean per fold
[0.63602643 0.64268459 0.44834531 0.56195149 0.74708737]

Mean Geometric Mean
0.6072190350620777
```

Παρατηρούμε ότι ο αλγόριθμος για το συγκεκριμένο σύνολο δεδομένων, ενώ έχει πολύ υψηλό Specificity, έχει και πολύ χαμηλό Sensitivity. Δηλαδή αυτό σημαίνει ότι εντοπίζει αρκετά καλά τους ανθρώπους που δεν αντιμετωπίζουν πρόβλημα αλλά δεν κάνει καλή δουλειά στον εντοπισμο των ανθρώπων που είναι πραγματικά ασθενείς.