Ανάπτυξη Naïve Bayes Classifier (NBC)

Ο NBC που θα αναπτύξετε θα πρέπει να ικανοποιεί τις ακόλουθες απαιτήσεις:

1. Τα πρότυπα εκπαίδευσης (training patterns) θα δίνονται με τη μορφή αρχείου csv χωρισμένα σε γραμμές, με μία γραμμή ανά πρότυπο. Η πρώτη γραμμή θα περιγράφει τα χαρακτηριστικά (attributes) των προτύπων. Το τελευταίο χαρακτηριστικό που θα δίνεται θα αφορά την κλάση στην οποία ανήκει το πρότυπο. Η περιγραφή των χαρακτηριστικών (attributes) μπορεί επίσης να δίνεται στην αρχή σε μία χωριστή γραμμή για κάθε χαρακτηριστικό. Τα χαρακτηριστικά θα δίνονται με τη σειρά που δίνονται και οι τιμές τους στα πρότυπα.

2. Κάθε χαρακτηριστικό θα ακολουθείται από ένα διακριτικό **D** ή **C** ανάλογα με το αν πρόκειται για χαρακτηριστικό με διακριτές ή με συνεχείς τιμές. Θα ήταν δυνατό μετά από κάθε χαρακτηριστικό να δίνεται και η λίστα των διαφορετικών διακριτών τιμών ή το διάστημα από τη μικρότερη έως τη μεγαλύτερη πραγματική τιμή. Οι τιμές των συνεχών μεταβλητών που θα δίνονται με τα πρότυπα θεωρούμε ότι ακολουθούν την Gaussian κατανομή οι παράμετροι της οποίας είναι άγνωστες. Το τελευταίο χαρακτηριστικό που θα δίνεται θα αφορά τις κλάσεις και θα αποτελεί μια μεταβλητή με διακριτές τιμές.

Παράδειγμα το Iris classification dataset

```
"sepal.length",C,"sepal.width",C,"petal.length",C,"petal.width",C,"variety"
5.1,3.5,1.4,.2,"Setosa"
4.9,3,1.4,.2,"Setosa"
4.7,3.2,1.3,.2,"Setosa"
…
7,3.2,4.7,1.4,"Versicolor"
6.4,3.2,4.5,1.5,"Versicolor"
6.9,3.1,4.9,1.5,"Versicolor"
…
6.3,3.3,6,2.5,"Virginica"
5.8,2.7,5.1,1.9,"Virginica"
7.1,3,5.9,2.1,"Virginica"
```

3. Ο NBC θα φορτώνει τα δεδομένα και θα υπολογίζει τις a priori πιθανότητες για τις κλάσεις καθώς και τις πιθανοφάνειες των κλάσεων για τις διαφορετικές τιμές των attributes που δίνονται με τα πρότυπα. Ο NBC θα επιλύει το zero frequency problem όπου χρειάζεται και θα υπολογίζει τις τιμές των παραμέτρων για όλα τα attributes με συνεχείς τιμές που θεωρούνται ότι ακολουθούν τη Gaussian κατανομή. Οι παράμετροι κάθε Gaussian κατανομής θα τυπώνονται χωριστά για κάθε χαρακτηριστικό και για κάθε κλάση.

4. Ο NBC θα δίνει τη δυνατότητα στο χρήστη να διατυπώνει ένα ερώτημα το οποίο θα πρέπει να επιλυθεί και θα απαντά δίνοντας τις a posteriori πιθανότητες για τις κλάσεις σχετικά με το ερώτημα που έθεσε ο χρήστης.

5. Η ανάπτυξη θα πρέπει να γίνει σε περιβάλλον MATLAB/Octave, python, R ή κάποιο άλλο περιβάλλον και γλώσσα προγραμματισμού για το οποίο να αιτιολογηθεί επαρκώς ότι χρησιμοποιούνται για αλγορίθμους Μηχανικής Μάθησης.

6. Για την εργασία θα πρέπει να ετοιμαστεί πλήρες κείμενο (έκθεση) που θα περιγράφει το πρόβλημα, όλες τις επιλογές και τον κώδικα του προγράμματος. Η γραπτή περιγραφή και ο κώδικας θα πρέπει να παραδοθεί το αργότερο έως 16 Μαΐου 2025 στην Ασύγχρονη Τηλεκπαίδευση. Η παρουσίαση και εξέταση της εργασίας θα καθοριστούν σε μεταγενέστερη ημερομηνία.