Lab - 4

PRML
AY 2020-21 Trimester - III

March 30, 2021

Boosting and Bayes Classification

Dataset:- Consider the dataset containing 50 samples of 3 different species of iris (150 samples total). The dataset contains 5 columns. Sepal length, sepal width, petal length, petal width are to be taken as input features and the last column “Species” contains 3 classes “Iris Setosa, Versicolor, or Virginica” which is to be used as the target. All the input features are in cm.

Perform the following tasks for this dataset:-

Question-1 (Boosting): (Total 20 Marks)

1. Preprocessing the data. (5 Marks)
    a. Plot the distribution of the target variable.
    b. Visualize the distribution of data for every feature.
2. Perform boosting-based classification using Decision Tree as the base classifier. (5 Marks)
3. Perform cross validation over the data and calculate accuracy for a weak learner. (5 Marks)
4. Build the AdaBoost model using the weak learner by increasing the number of trees from 1 to 5 with a step of 1. Compute the model performance. (5 Marks)

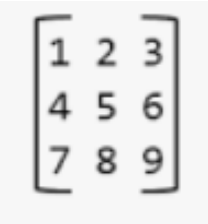
Question-2 (Bayes classification) : (Total 20 Marks)

1. Estimate the accuracy of Naive Bayes algorithm using 5-fold cross validation on the data set. Plot the ROC AUC curve for different values of parameters. (5 Marks)

2. Use linear discriminant function to calculate the accuracy on the classification task with 80% training and 20% testing data. (5 Marks)

3. Calculate the Bayes risk

    Given: $\lambda =$



(5 Marks)

Question 3: Visualisation in Bayesian Decision Theory

DATASET 1:

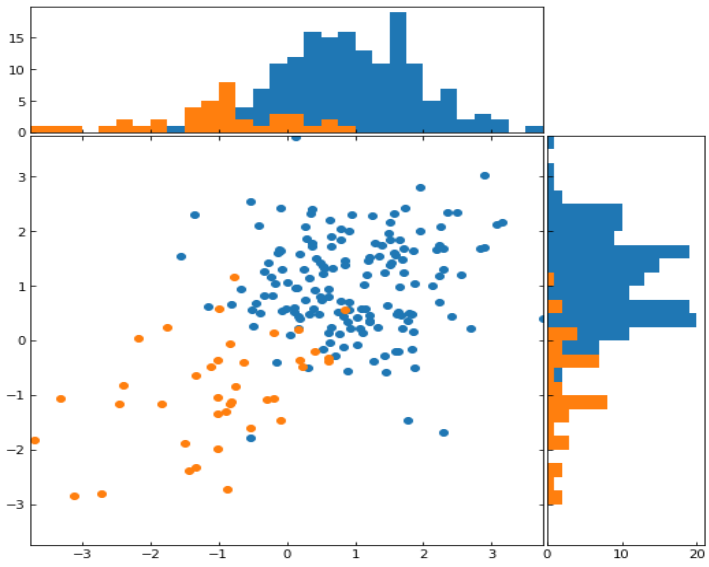
Consider the height of the car and its cost is given. If the cost of a car > 550 then the label is 1, otherwise 0.

a. Create the labels from the given data.
b. Plot the distribution of samples using histogram.
c. Determine the prior probability for both the classes.
d. Determine the likelihood / class conditional probabilities for the classes. (Hint : Discretize the car heights into bins, you can use normalized histograms)
e. Plot the count of each unique element for each class. (Please mention in the report why this plot is different from the distribution)
f. Calculate the P(C1|x) and P(C2|x) i.e posterior probabilities and plot them in a single graph. [5 marks]

DATASET 2:

Now for the second dataset there are two files c1 and c2 . c1 and c2 contain two features each for class 1 and 2 respectively. Read the dataset and repeat all the above steps for Dataset 2.

Note : Plot the data distribution and the histogram of feature 1 and feature 2 in the X axis and Y axis respectively. The distribution of feature 1 will be along the top of X axis and feature 2 along the right of Y axis. An example is shown below. [5 marks]



Real Life Dataset:

Now it's time to visualise a real life dataset. Take any one feature from the above IRIS dataset and take the class labels. In this dataset there are three class labels. Extend all the visualisation mentioned previously for this dataset. (5 Marks)

Note: Only the calculations for the posterior probabilities will be graded (except those mentioned already). All the remaining ones will be shown in the lab session.

Here is the Colab notebook attached for your reference.

Notebook used for visualisation demo : Link

Instructions:-

Please Submit the necessary code(s) (Notebook) and a PDF explaining and analyzing the steps in both the questions along with necessary plots/figures.