Pattern Recognition and Machine Learning

(Winter 2022)

Assignment 5

Submission Deadline : 27 Feb 2022, 23:59

Guidelines for Submission:

- Perform all tasks in a single collab file.
- Create a report regarding the steps followed while performing the given tasks.
- The report should not include excessive unscaled preprocessing plots.
- Try to modularize the code for readability wherever possible
- Submit the collab file [.ipynb] and report [.pdf] (separately) on the classroom assignment
- Submit the .py file on the floated form for the laboratory (it will be floated on 27th, so please submit at the time of final submission only)
- Plagiarism will not be tolerated

Guidelines for the Report :

- The report should be to the point and explanations for each task you performed should be included in the report.
- Report should include any sort of visualization along with results.
- DO NOT copy paste code snippets in the report.

Q1. Quadratic Discriminant Analysis (QDA) is a Baysian generative model that comes up with a non-linear (quadratic, specifically) decision boundary with the assumption that each class follows a normal distribution and has its own covariance matrix (contrary to Linear Discriminant Analysis, which assumes each class having same covariance matrix). ***[45]***
Dataset : iris dataset

A. Preprocess the data and split in training and test set in 70:30 ratio and make sure all classes are present in test data in approximately the same number. Visualize the training data with scatterplot, using all possible combinations of two attributes. ***[3]***
B. Choose any three pairs of features (which you think will give good results), Train QDA models for each pair, for the classification task. You can use the QuadraticDiscriminantAnalysis function from sklearn. ***[10]***
C. Report the mean and covariance of the distributions found from each QDA model. ***[2]***
D. Plot the decision boundary given by the QDA models on top of the corresponding scatterplot visualization of the data. ***[10]***

E. Predict the test data and report error rate for each case. Which pair of features do you think gives the best result and why? ***[5]***
F. Take the pair of features that has given the best result and train LDA model with same training data. You can use the LinearDiscriminantAnalysis function from sklearn. ***[5]***
G. Plot the decision boundary given by the LDA model on top of the scatterplot visualization of the data. ***[2]***
H. Report the error rate on the test data for LDA model. Which one between LDA and QDA has performed better do you think? Justify your answer. ***[3]***
I. Visualize the gaussian distributions obtained from QDA and LDA. (you can draw ellipse as shown in demo or you can use any other function to visualize.) ***[5]***

Q2.. Use the iris dataset as given in the above question and assume that the data follows the gaussian distribution. Do the preprocessing of the data and consider only petal length and petal width features for the below questions. ***[30]***

a) From the data, find out the sample mean and sample covariance matrix of each class and visualize the gaussian distribution using the obtained sample mean and sample covariance matrix. ***[5]***
b) Write a function compute_likelihood to compute the likelihood of data given the parameters mean and covariance matrix assuming gaussian distribution. ***[10]***
c) Write a function to perform maximum likelihood estimation over the training dataset to determine mean and covariance and classify it using Bayes classifier. Report the parameters obtained for each class. ***[10]***
d) Visualize the gaussian distribution from mean and covariance of both the above parts (a and c) in a single plot. ***[3]***
e) Predict the test data and compare the performance obtained in question1(with QDA).***[2]***

Q3. Consider the following dataset that contains: ***[25]***
(https://drive.google.com/drive/folders/13-1uWVfoDHoxioDcKQw9ISYliKhxHzeq?usp=sharing)
You will need to implement the Naive Bayes algorithm to classify a news corpus into 20 different categories.

- train.data - Which contains bag-of-words data for each training document. Each row of the file represents the number of occurrences of a particular term in some document. The format of each row is (docId, termId, Count).
- train.label - That contains a label for each document in the training data.
- test.data - Contains bag-of-words data for each testing document. The format of this file is the same as that of the train.data file.
- test.label - Contains a label for each document in the testing data.

Perform following operations:

1. Compute the likelihood of the training data-set(without using any standard library).

[5]

2. Smoothing is a technique that helps tackle the problem of zero probability in the Naïve Bayes algorithm. Apply Laplace smoothing to solve the zero observations problem on a training data-set from scratch (without using any standard library). ***[10]***

3. naiveBayesClassify(trainData, trainLabels, testData) - Classifies the data using the Naive Bayes algorithm. ***[10]***