PA-4 Report

Colab Link

Problem 1

Task 2:

LDA projection vector was plotted using the GetLDAProjectionVector() function written in task-1. Along with it all the data points are shown, data points related to class 0 are shown in blue colour while class 1 in red.

Task 3:

Dataset was split into train and test dataset using 70-30 split. Then first KNN (K=1) is used to fit on the train dataset and accuracy on original data is calculated using test dataset.

After this LDA projection vector is computed using the train dataset, then both train and test dataset were projected using this projection vector and again KNN (K=1) was used to calculate the accuracy on projected data.

Accuracy on original data : 89.0%

Accuracy on projected data : 88.5%

Problem 2

Task 0:

Iris dataset was loaded from Github and split using test size as 2 and train size as 12. Samples with index 9 and 11 are in the test dataset.

Task 1:

Prior probabilities for (Play = yes) and (Play = no) are calculated by dividing the no of samples with (Play = yes) by total size of the dataset and similarly for (Play = no).

Prior probability for (Play = yes) : 0.58

Prior probability for (Play = no) : 0.42

Task 2:

Likelihood probabilities were calculated for each combination of feature-value and were stored in the dictionary. The values can be seen in the attached image.

```
P(Outlook = Rainy | Play = yes) = 0.29
P(Outlook = Rainy | Play = no) = 0.60
P(Outlook = Overcast | Play = yes) = 0.43
P(Outlook = Overcast | Play = no) = 0.00
P(Outlook = Sunny | Play = yes) = 0.29
P(Outlook = Sunny | Play = no) = 0.40
P(Temp = Hot | Play = yes) = 0.29
P(Temp = Hot | Play = no) = 0.40
P(Temp = Cool | Play = yes) = 0.43
P(Temp = Cool | Play = no) = 0.20
P(Temp = Mild | Play = yes) = 0.29
P(Temp = Mild | Play = no) = 0.40
P(Humidity = High | Play = yes) = 0.29
P(Humidity = High | Play = no) = 0.80
P(Humidity = Normal | Play = yes) = 0.71
P(Humidity = Normal | Play = no) = 0.20
P(Windy = f | Play = yes) = 0.71
P(Windy = f | Play = no) = 0.40
P(Windy = t | Play = yes) = 0.29
P(Windy = t | Play = no) = 0.60
```

Task 3:

Posterior probabilities for each sample were first initialised to prior probabilities and then were multiplied with the likelihood probability for each combination of feature-value to calculate the final value of posterior probability for both (Play = yes) and (Play = no).

Sample 1: Posterior Probability for (Play=yes): 0.0243, Posterior Probability for (Play=no): 0.0053

Sample 2: Posterior Probability for (Play=yes): 0.0058, Posterior Probability for (Play=no): 0.0000

Task 4:

Outcome both the samples were predicted using the posterior probabilities that if posterior probability for (Play = yes) is greater than posterior probability for (Play = no) than the outcome is Yes otherwise No.

Predicted outcome for sample 1: Yes

Predicted outcome for sample 2: Yes

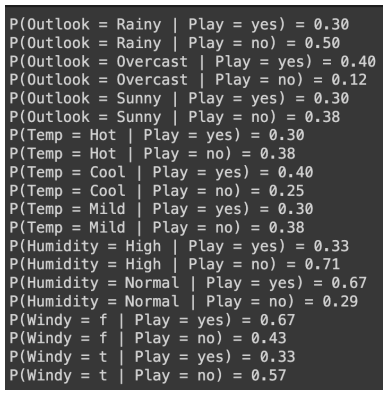
Accuracy : 100%

Task 5:

In the case where particular feature-value combination is not present in dataset, likelihood probability for that combination will be 0. For example, likelihood probability for Outlook = Overcast and Play = no is 0.

Now while calculating posterior probability this combination will never be considered and to avoid this we introduce the concept of Laplace smoothing

in which we increase the increase the numerator with alpha and denominator with alpha * no of unique features. Here alpha is the smoothing parameter and increasing its value leads the likelihood probability towards 0.5, usually value of alpha is set as 1.



We again calculate the likelihood probabilities using Laplace smoothing (alpha=1) and the results can be seen in the attached image. It can be observed that likelihood probability for Outlook = Overcast and Play=no is not 0 anymore.

Corresponding posterior probabilities for both the samples are:

Sample 1: Posterior Probability for (Play=yes): 0.0233, Posterior Probability for (Play=no): 0.0072

Sample 2: Posterior Probability for (Play=yes): 0.0078, Posterior Probability for (Play=no): 0.0080

Predicted outcome:

Sample 1 : Yes

Sample 2 : No

Accuracy : 50%

Note that the accuracy here has decreased after the introduction of Laplace smoothing due to the very small size of dataset but in general with the large datasets this concept can be very helpful as it gives a small amount of probability to every possibility.