EEE3142 Introduction to Machine Learning

Final Project

Notice

- Carefully read the instructions in the project announcement in Cyber Campus
- Skeleton codes will be provided
- Complete the skeleton codes by following the instructions given by TODOs
- Answer to the questions (b) at the end of the codes
- Your answers could be written in Korean if you want
- Submit your python files to Cyber Campus
- Deadline : 11:59pm on Dec 23, 2022 (Team setup due is 8:59am on Dec 9)
- Late policy: No late submission

All problems use the dataset (`result.csv`) used in Assignment 6. Part of the codes can be reused in this project.

Question 1.

(a) [20 pts] Write codes to reconstruct the projected data (done by PCA) in the original data space. Then compare the root-mean squared error (RMSE) of the reconstructed data when you use the first two principal components (PCs) and the last two PCs. The RMSE is computed for each feature. The result will look like where [...] will be filled with actual values:

```
*** Reconstruction error (RMSE) when using the first two PCs ***
assignment_1     ...
assignment_2     ...
assignment_3     ...
assignment_4     ...
assignment_5     ...
midterm          ...
final            ...
bonus            ...
dtype:  float64

*** Reconstruction error (RMSE) when using the last two PCs ***
assignment_1     ...
assignment_2     ...
assignment_3     ...
assignment_4     ...
assignment_5     ...
midterm          ...
final            ...
bonus            ...
dtype:  float64
```

(b) [10 pts] Discuss why you have the result in (a)

Question 2.

(a) [20 pts] Given the PCA result in Assignment 6, write codes to find the parameters of the Gaussian Mixture Model for the projected data onto the 2D subspace spanned by the first two principal components. Assume we have $K = 3$ components in the GMM. Then plot the three distributions using their means and covariance error ellipses (using the drawing part given in the skeleton code). Your result will look like Fig.1.

(b) [10 pts] In (a), you could compute the means and covariances of the three components as well as the mixture weights. What are the actual values of the parameters? Please give them for all $K = 3$ components. Note that the values could be different slightly in each execution as we use numerical methods.

Question 3.

(a) [30 pts] Write codes to train an SVM model to classify the projected data from Assignment 6 into three classes. In the original data, the labels of the classes are A, B, and C. In the skeleton code, they are transformed to 2, 1, and 0 to have numerical labels. Once you train the model, test it with the test set. Print the accuracy score and the ground truth and predicted values of the labels.

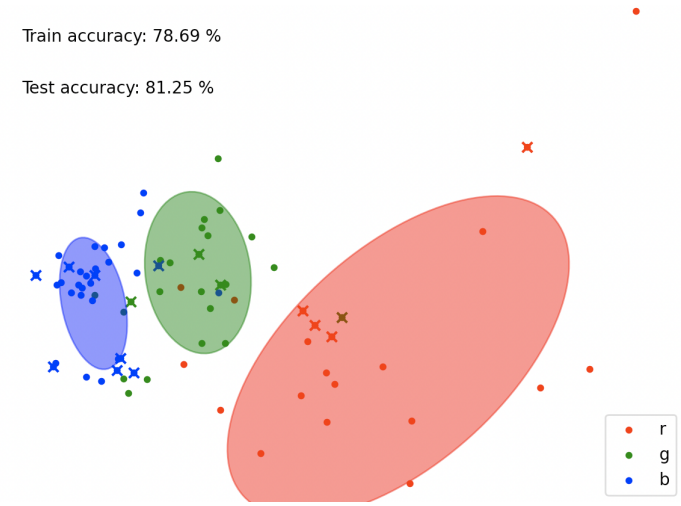

Figure 1: Question 2

(The accuracy score computed in the codes shows the percentage of the true positive and true negative to all data points.) The result will look like:

```
Accuracy:  81.25 %
Test:        [2, 2, 1, 0, 2, 0, 1, 2, 2, 2, 1, 0, 2, 2, 0, 1]
Prediction:  [2, 2, 1, 1, 2, 0, 2, 2, 2, 2, 1, 1, 2, 2, 0, 1]
```

(b) [10 pts] What kernel did you use in (a)? Why do you choose that kernel?