Pattern Recognition and Machine Learning

Lab - 4 Assignment
Bayes Classification with Non-Linearity

Early Bird Submission Deadline: Tuesday Batch: 7 Feb, 11:59 PM
Thursday Batch: 9 Feb, 11:59 PM
Late Submission Deadline: Tuesday Batch: 8 Jan, 2023, 12:00 Midnight (20% penalty)
Thursday Batch:Feb 10, 2023, 23:59 (20% penalty)
Final deadline: Tuesday Batch: Feb 9, 2023, 23:59 (addl. 20% penalty, total penalty = 40%)
Thursday Batch: Feb 11, 2023, 23:59 (addl. 20% penalty, total penalty = 40%)

Guidelines for submission

1. Perform all tasks in a single colab file.
2. Create a report regarding the steps followed while performing the given tasks. The report should not include excessive unscaled preprocessing plots.
3. Try to modularize the code for readability wherever possible
4. Link for In-Lab Submission: Link
5. Submit the colab[.ipynb], python[.py] and report[.pdf] files here : Link
6. Plagiarism will not be tolerated

Guidelines for Report:

1. The report should be to the point. Justify the space you use!

2. Explanations for each task should be included in the report. You should know the 'why' behind whatever you do.

3. Do not paste code snippets in the report.

[In-Lab Submission : Question 2, first 2 subparts]

Question 01: [60 marks] : Gaussian Bayes Classifier

Dataset : Iris Dataset. Preprocess the dataset, and also perform exploratory analysis. Split the data into training and testing dataset in the ratio 70:30.

1. Implement a **Gaussian Bayes Classifier class** from scratch.(You are not allowed to use the inbuilt scikit function, you are only allowed to use numpy and pandas). The classifier class must have 3 variants defined using its constructor, for each of the cases given below. [10 marks]

Case $\Sigma_i = \sigma^2.I$ (I stands for the identity matrix)
Case $\Sigma_i = \Sigma$ (covariance of all classes are identical but arbitrary!)
Case Σi = actual covariance

2. The Gaussian Bayes Classifier class should also have the following function:
    a. **Train**: Takes x,y (training data) as input and trains the model.
    b. **Test**: Takes testing data, testing labels as input, and outputs the predictions for every instance in the testing data, and also the accuracy.
    c. **Predict**: Takes a single data point as input, and outputs the predicted class.
    d. **Plot decision boundary**: Takes input the training data points, and their labels, and plots the decision boundary of the model with the data points superimposed on it. (Consider only two features while plotting the decision boundary) [10 marks]
3. Train the Bayes model on the training dataset and plot the decision boundary for each case implemented in Q1. Comment on the decision boundaries obtained in all the 3 cases. Compare the three models and report how well they perform on the dataset. [15 marks]
4. Perform 5 fold cross validation on the training dataset and report the accuracies on each validation set as well as comment on the generalizability of each model. [10 marks]
5. Create a synthetic dataset which has 2 features, and data is generated from a circular distribution: $x^2 + y^2 = 25$. The data has 2 classes, points which have distance <=3 have class=1 and points having distance>3 and distance<= 5 have class=2. {Note that the classes are thus, not linearly separable)
   Train the implemented Gaussian Bayes Classifier(case 3) on such a synthetic dataset, and plot the decision boundary for the same. [15 marks]

<u>Question 02:</u> [40 Marks]: Mahalanobis Distance

Reference Links: [Mahalanobis Distance]()

Given the covariance matrix, $\Sigma = \begin{bmatrix} 3/2 & 1/2 \\ 1/2 & 3/2 \end{bmatrix}$, $\mu = [0\ ,\ 0]$, sample random points from the multivariate normal distribution and use it as a set of datapoints (say X).

1. Calculate the covariance matrix of the sample X , say Σ_s. Find the eigenvectors and eigenvalue of Σ_s and plot it superimposed on the datapoints X. [5 marks]
2. Perform the transformation $Y = \Sigma_s^{-1/2}X$ on the datapoints X. Calculate the covariance matrix of transformed datapoints Y , say Σ_Y. Comment on the obtained covariance matrix and infer what was the purpose of the transformation.[10 marks]
3. Uniformly sample 10 points on the curve $x^2 + y^2 = 25$. Let these set of points be called P. Plot points in P along with the datapoints in X. Make sure to give each point a different color [Hint :

You can use properties like hue to do this] for better visualization. Report the eucledian distance of each point from μ using barplot. [10 marks]

4. Perform the transformation $Q = \Sigma_s^{-1/2} P$ on the datapoints P. Calculate the eucledian distance of transformed datapoints Q from μ and report it using barplot . Plot points in Q along with datapoints in Y. Make sure that the color of point before and transformation doesn't change. Comment on the difference in eucledian distance before and after transformations of the points in P. [15 marks]