BITS F464 - Machine Learning

Assignment-1A: Fischer's Linear Discriminant

Deadline: ~~TBD~~ 2359Hrs 15th March 2021

1 General Instructions

1. This assignment is a coding project and is expected to done in groups. Each group can contain at most **three** members. Make sure that all members in the group are registered to this course and please try to maintain the same group for all the assignments.
2. This assignment is expected to be done in Python using standard libraries like NumPy, Matplotlib and Pandas. You can use Jupyter Notebook and any other python in-built data structure or library. No other ML libraries like scikit/sklearn, TensorFlow, Torch etc. should be used.
3. Refrain from directly copying codes/snippets from other groups or the internet as all codes will be put through a plagiarism check.
4. All deliverable items (`.py` files, `.ipynb` files, report, images) should be put together in a single `.zip` file. Rename this file as `FD_`⟨`id-of-first-member`⟩ (preferably ID number of the student submitting) before submission.
5. Submit the zip file on CMS/GForms on or before the aforementioned deadline. Please note that this is a hard deadline and no extensions or exemptions will be given. The demos for this assignment will be held on a later date which shall be conveyed to you by the IC. All group members are expected to be present during the demo.
6. Dataset for this assignment can be found here.

2 Problem Statement

1. In this problem, you will implement Fischer's Linear Discriminant from scratch as learnt in class, i.e. given the higher dimensional data reduce the data to one dimension while maximizing difference of means and minimizing sum of variances of the clusters. Finally, calculate the intersection point of both the normal distributions corresponding to the collapsed clusters and find the discriminant vector in 1-D and 3-D.
2. Note that, for this question, you need not make any train-test split and you can use the entire data for the procedure.
3. Try to vectorize your code as much as possible to make your computations faster and efficient. Do not hard code any parts of the implementation unless it is absolutely necessary.

3 What needs to be documented

1. A very brief description of your model and its implementation.
2. Plot of the higher dimensional data. (you can use Matplotlib's 3D plotting feature for this)
3. Plots of the reduced clusters and their corresponding normal distribution in two separate plots. It is recommended that you use two different colors *(say red and blue)* to represent the two classes. Also, do visualize the discriminant line in your plots.
4. The intersection point of both the normal distributions and unit vector along the discriminant line in 1-D and 3-D.

BITS F464 - Machine Learning

Assignment-1B: Naive Bayes Classifier

Deadline: ~~TBD~~ 2359 Hrs 15th March 2021

1 General Instructions

1. This assignment is a coding project and is expected to done in groups. Each group can contain at most **three** members. Make sure that all members in the group are registered to this course and please try to maintain the same group for all the assignments.

2. This assignment is expected to be done in Python using standard libraries like NumPy, Matplotlib and Pandas. You can use Jupyter Notebook and any other python in-built data structure or library. No other ML libraries like scikit/sklearn, TensorFlow, Torch etc. should be used.

3. Refrain from directly copying codes/snippets from other groups or the internet as all codes will be put through a plagiarism check.

4. All deliverable items (`.py` files, `.ipynb` files, report, images) should be put together in a single `.zip` file. Rename this file as `NB_`⟨`id-of-first-member`⟩ (preferably ID number of the student submitting) before submission.

5. Submit the zip file on CMS/GForms on or before the aforementioned deadline. Please note that this is a hard deadline and no extensions or exemptions will be given. The demos for this assignment will be held on a later date which shall be conveyed to you by the IC. All group members are expected to be present during the demo.

6. Dataset for this assignment can be found here.

2 Problem Statement

1. In this problem, you will implement a simple Naive Bayes classifier to classifiy mails as spam or not. You will need to create a 7-fold cross validation to train and test your model. You may choose to discard various stop words, commas, fullstops, numbers, hyphens, brackets, exclamation marks and any other single/double letter words (such as a, an, the, be etc) which do not contribute to the sentiment of the text.

2. Use laplace smoothening to avoid the problem of divison by zero.

3. Try to vectorize your code as much as possible to make your computations faster and efficient. Do not hard code any parts of the implementation unless it is absolutely necessary.

3 What needs to be documented

1. A very brief description of your model and its implementation.
2. Accuracy of your model over each fold and the overall average accuracy.
3. Major limitations of the Naive Bayes classifier.
