BITS F464 – Machine Learning (2019-2020 Sem II)

Assignment 1 (Due Date – 21/04/2020)

General Instructions:

- Download the datasets here (use your BITS email to access it).
- Your report, code, and visualizations must be uploaded to CMS by any group member, and will be checked for plagiarism. The report should describe your design decisions, results, and any conclusions that can be drawn (e.g., why an algorithm performed good/bad).
- The algorithms must be written in C++/Java/Python. High level libraries (e.g., Scikit-learn, Gensim, etc.) may NOT be used. Data manipulation libraries (like NumPy and Pandas) are allowed (and recommended). For visualizations, any library can be used.
- Try using vector operations (like NumPy matrix multiplications) whenever possible to improve your implementation's scalability. **This is recommended but not mandatory**.
- Please email Vamsi Aribandi (f20160803@hyderabad.bits-pilani.ac.in) for any queries.

1) Fisher's Linear Discriminant Analysis:

- Use `a1_d1.csv` and `a1_d2.csv` as datasets to test your implementation on.
- Each row corresponds to a data point, with the last column being its class label and the rest of the columns containing its feature values.
- Once the points are transformed to one dimension, plot a normal distribution for each class and use the intersection point of both curves as the classification threshold. Report your implementation's accuracy and F-score.
- Once the points are transformed to one dimension, visualize them as points on a line and colour points of both classes separately.

2) Naïve Bayes:

- Use `a1_d3.txt` as the dataset to test your implementation on.
- Each row corresponds to a review followed by its sentiment (0 or 1).
- Stemming, n-gram usage, and/or lemmatization are NOT mandatory. However, it is recommended that you do some basic text preprocessing (like removing symbols). Report the steps you took and whether they improved your model's performance.
- Use 5-fold cross validation to evaluate your implementation and report your implementation's accuracy and F-score as the "mean ± standard deviation" across the 5 folds (e.g., if the test-folds' accuracies are 0.52, 0.52, 0.53, 0.54 and 0.54; report the accuracy as 0.53 ± 0.01). Shuffling the data points is NOT mandatory.
- Make sure that all parameters are learned using only the train set. The model's vocabulary should only be determined using the train set. Handling out-of-vocabulary words is NOT mandatory.

Suggested Reading:

- Bishop's Ch. 4.1.
- Stanford Lecture notes for Naïve Bayes.