CEEL 82B, Data Science, 2022

Lab 3: Data Classification

**Objective:**

(i) Apply Naive bayes classifier for given data see *Naive\_bayes\_handout.zip* file in edmodo.

(ii) Apply and compare Linear Discriminating Analysis with Naive bayes classifiers w.r.t. error, specificity and sensitivity

Load the construct an python code for:

1. Convert the textual meta-data into a suitable (e.g. corpus) object.
2. Triage some of the irrelevant punctuation and other symbols in the corpus document,change all text to lower case, etc.
3. Tokenize the job descriptions into words. Examine the distributions of two important features
4. Classify attributes in two categories.
5. Graphically visualize the difference between low and high value of targeted feature graph.
6. Transform the features into categorical data
7. Ignore those low frequency words and report the sparsity of your categorical data matrix with or without delete those low frequency words. Note that the sparsity of a matrix is the fraction:

Sparsity(A) = number of zero-valued elements / total number of matrix elements (m×n).

1. Apply the Naive Bayes classifier to original matrix and lower dimension matrix, what do you observe?
2. Apply and compare LDA and Naive Bayes classifiers with respect to the error, specificity and sensitivity.

Refer: https://www.analyticsvidhya.com/blog/2019/08/11-  
important-model-evaluation-error-metrics/

**Conclusion:** (Write in own words)

**Note:** Complete your write-up with conclusion and upload your outputs on your github account