**MACHINE LEARNING FOR DATA SCIENCE**

**Q1. List at least two differences (or pros and cons) between discriminative and generative models.**

**Solution:**

**Discriminative Models:**

1. A discriminative model learns the **conditional** probability distribution p(y|x). It takes the data as given and models only the conditional probability of the class.
2. Determining the difference in the models without learning the models and then classifying the class types.
3. They give high accuracy performance. The overall gist is that discriminative models generally outperform generative models in classification tasks with more amount of training data.
4. Examples – Logistic regression, SVMs

**Generative Models:**

1. A generative model learns the joint probability distribution p(x,y) and tries to maximize this joint likelihood using the Bayes theorem.
2. Generate two models and looks for a boundary in space where one model becomes more plausible/likely than another one. Learning each model and then classifying it using the knowledge you just gained.
3. Can perform efficiently with less amount of data.
4. Examples – Naïve Bayes, Gaussian Discriminative Analysis.

**Q2. In 2-4 sentences, describe how GDA and logistic regression are related and different.**

**Solution:**

The outcome of the Gaussian Discriminative Analysis and logistic regression is same, but the way we go about it, is different. GDA is more restrictive as it assumes the gaussian distribution and can perform efficiently with the less amount of data. Logistic regression is more liberal and more robust and can work with any types of data distribution but it requires lots of data to perform well. If we are performing GDA on certain data, then we can also do logistic regression on the same data, whereas the vice versa is not true.

**Q3. Use any 4 of the 5 given datasets to build a spam filter and use that filter to check the accuracy on the remaining dataset.**

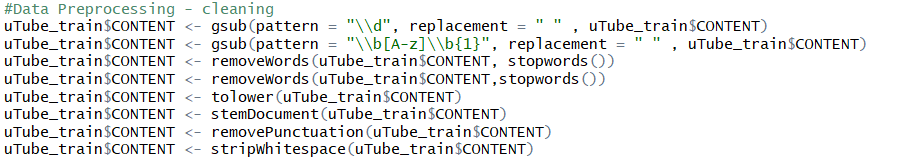
**Solution:**

**Steps followed to build the spam filter with the given datasets**

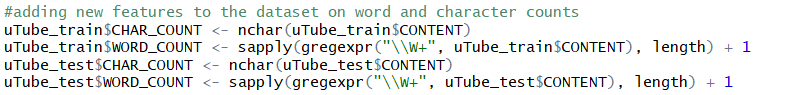
1. Load the 4 datasets into the R studio and merge as training dataset



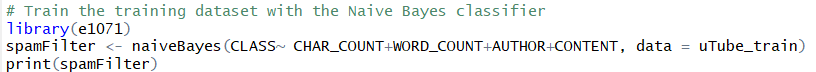
1. Load the remaining one dataset and keep it as a testing dataset
2. In both the datasets, convert the output variable(CLASS) as factor variables
3. In-order to build the spam filter, we should do the text mining with the given datasets
4. Let us first do the data preprocessing by cleaning the data
5. Cleaning the data includes removing numbers, punctuations, stop words, etc., from the CONTENT variable



1. As the features available in the given dataset is not providing enough information, we can create our own features by calculating the character counts and word counts for both training and testing datasets



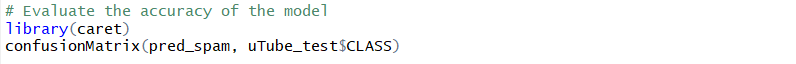
1. Now with new features included in the dataset, let us build the model with the Naïve Bayes classifier



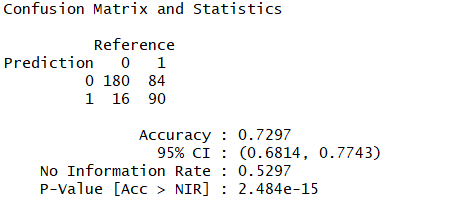
1. Next, we can make the predictions to our test dataset with the above trained spam filter Naïve Bayes model.



1. Based on the prediction, the accuracy of the model can be calculated with the confusion matrix as given below:



1. The accuracy of the model is turned out to be (0.7297) 73%, which can be considered as a good model. Also, the confusion matrix gives the details about how many times the predictions were made correct and incorrect. Example with result, Non-spam (CLASS 0) is predicted correctly as 0 for 180 times.



\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*