**HW01 REPORT**

When handling data with continuous distribution, Naive Bayes classifier considers that the big data is generated through a Gaussian process with normal distribution.

The aim of the Naive Bayes classifier to compute the following probability, where y is the classes, x is the data points and c is the number of classes :

P (y=c | x)

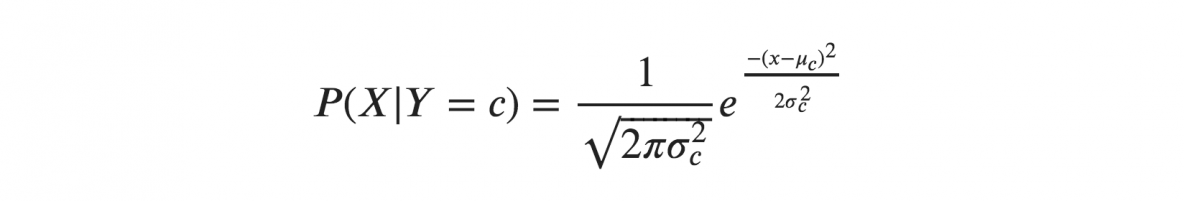
Applying Naive Bayes rule, it can be represented as follows from Bayes’ Theorem:

P (y=c | x) = P (x | y=c) . P (y=c)

I split data to train\_data and test\_data. Each data set is a 200x4096 matrix. I divide train\_data to female and male data set according to their class labels. Using sapply() function, I calculated means matrix which is 4096x2. Columns represent classes and each row represents mean for each pixel according to their class labels. Again, using sapply() function, I calculated prior probabilities, which are scalars, for each class. I calculated standard deviations for each class separately, then using cbind() function, I put both vectors in a matrix (4096x2), since I could not find the exact standard deviations same with the given values using sd() function.

Since I found means and standard deviations according to train\_data set, to find score function for each class, I applied below function to find P (x | y=c) using sapply() function, and made row sum to get 200x2 matrix. To make the process faster, I converted all the vectors to a matrix using matrix() function. I calculated two score functions both for train\_data set and test\_data set. Each score function is a 200x2 matrix that consists of two vectors holding score values for each image in train\_data set according to class labels.

log[P (y=c | x)] = sum[log[P (x | y=c)] + log[P (y=c)] where log[P (y=c | x)] is a score function.

The reason that taking log is to avoid underflow.

After calculating score functions for both train\_data set and test\_data set, using which.max inside apply() function, I took the highest score for each image. Then, put that estimations to a confusion matrix to make comparisons for both train\_data set and test\_data set.