Matric number: s1632798

# Task 2

Task 2.2

The elapsed time for my_bnb_classify () is around 2.628459 seconds.

The requested information for every k from Kb:

| Number of test samples | Number of wrongly classified test samples | Accuracy |
|---|---|---|
| 7800 | 2881 | 0.6306 |

Task 2.3

This task is about naive Bayes classification with multivariate Bernoulli distributions. We are working only with binary vectors. Similar to the previous task we have training data matrix Xtrn, labels to the training data Ctrn and test data matrix Xtst. This time we also have a threshold. This is a scalar value for binarasation. We form different vectors in the Xtrn using different thresholds, because every value which is smaller than the given threshold is 0 and if it is greater or equal to the threshold then it is 1. The bnb_classification algorithm is straightforward. We convert the matrices into binary matrices using the threshold. We create a new matrix which contains the number of vectors for each class. Then using a for loop going through each of the 26 classes we create a matrix with the vectors for each class and estimate the likelihoods into a probability matrix. It is good idea to use the log function when working with big data, because the probabilities can become really small and cause numerical underflow. But using log we have a slight problem with the zero probability, because log (0) = -Inf. So we replace the 0 probability with a really small number close to 0 – for example 1.0E-10. That way we no longer have a problem with infinity. The next step is to go through each test vector and compute the likelihood for class Ck where k is between 1 and 26. We sort the likelihood matrix and take the first column which is the classification using the algorithm. When we change the threshold we can see that the accuracy is also chaining. We can see that the accuracy is like a parabola. Investigating the threshold, we can see that when we increase the threshold the accuracy increases and peak of the parabola is when the threshold equals 115 (the accuracy is 0.6429). After that number the accuracy slowly declines until it gets almost 0 with threshold >= 256, because the Xtrn and Xtst will become two zero matrices, because the values of the elements in the matrices are between 0 and 255. The peak is 115, because that number best distinguishes the values and forms the best delimitation.