**Project Report**

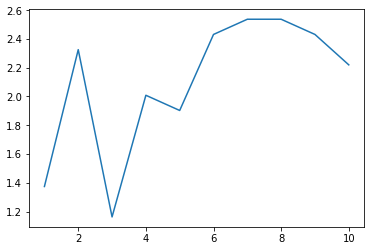
**Q3:** **K Nearest Neighbours:** The testing error rates of k-nearest-neighbors using the data in the testing folder with varying

k from 1 to 10 are as follows:

1.374207188, 2.325581395, 1.162790698, 2.00845666, 1.902748414, 2.431289641, 2.536997886, 2.536997886, 2.431289641 and 2.21987315

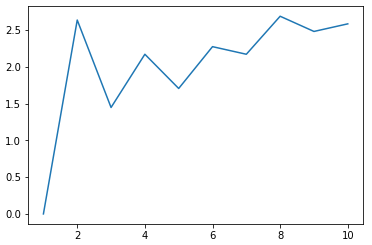
*Plotting the values in K Nearest Neighbours*

The test error rates w.r.t to k



For this dataset, which k = 3 works the best

The training error w.r.t to k



Both trends observe the error rate increase for even no. of nearest neighbours, this is intuitive since it accounts to the condition of splitting equally to two different classes and picking one of them results in a error. If we use training error to select the k value, we will choose the least k value, i.e. the model with highest model complexity. The training error gradually increases with k value, where as testing error first increases and then decreases.

The testing error rate for model averaging would be 1.7970

* Compare them with the corresponding error rates from NBC.

The error rates of the K NN are comparatively low in this case to the error rates of Naive Bayes. This is mostly because of the higher dimensional space.

* Briefly discuss pros and cons of KNN and NBC.

It is computationally and memory wise expensive to find the expensive to find the nearest neighbors, for large data set, it requires a lot of memory. KNN is conceptually simple and it decision boundaries are very flexible and susceptible to outliers and wrongly labelled training sets. The algorithms also doesn’t give a probabilistic output. The output of the classifier depends a lot on distance function.

Naive Bayes models can be used to tackle large scale classification problems for which the full training set might not fit in memory. Naive Bayes is relatively simple and computationally fast. Its easily trained, even with a small dataset and its fast and needs much less memory. Although, it assumes every feature is independent which isn’t the case always. It gives us a point estimate which doesn’t account for the uncertainty in the outcome.

* Discuss possible ways to improve the performance of NBC.

NBC can be improved by using Gaussian Distribution for modelling the independent features (Bayesian Naive Bayes). One more method of improving Naive Bayes is introducing Smoothing for the features. Also, there are advanced techniques for improving Naive Bayes for including dependent features.

**Q4:** For this specific application I don’t see any improvement for PCA in Naive Bayes and KNN.