4033/5033 Assignment: k-Nearest Neighbor

Sasank Sribhashyam

In this assignment, we will implement a compressed variant of the kNN classifier, which only stores a randomly picked $\tau \in [0, 1]$ fraction of training instances and use them to make classification. We will evaluate the performance this CkNN classifier on the Diabetes data set. Split the data set into a training set S and a testing set T. (As an example, $\tau = 0.1$ means we only store and use 10% of S to make classification.)

Task 1. Implement the CkNN classifier from scratch.

Task 2. Set $\tau = 1$. Evaluate the classifier on testing set and report testing error versus k in Figure **??**. Pick 10 values of k yourself but the first one must be $k = 1$.

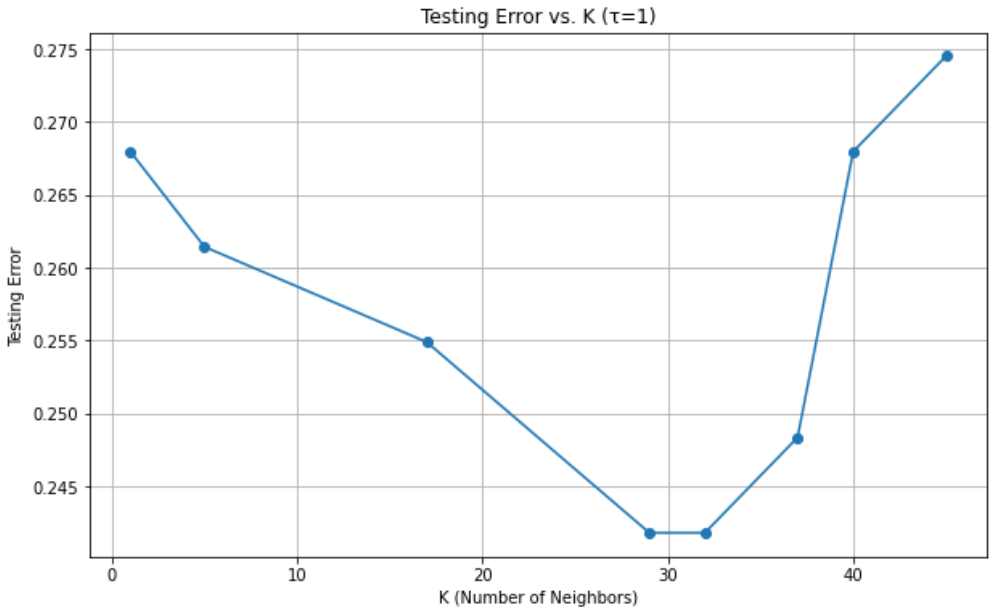

Fig. 1. A Sample Graph

Task 3. Pick three values of k by yourself. For each k, evaluate the classifier on testing set and report testing error versus τ in Figure **??**. Pick 10 values of τ yourself but the last one must be $\tau = 1$. Note: Figure **??** should contain three curves, each for one value of k. You should add a legend in the figure to clarify the k value for each curve.

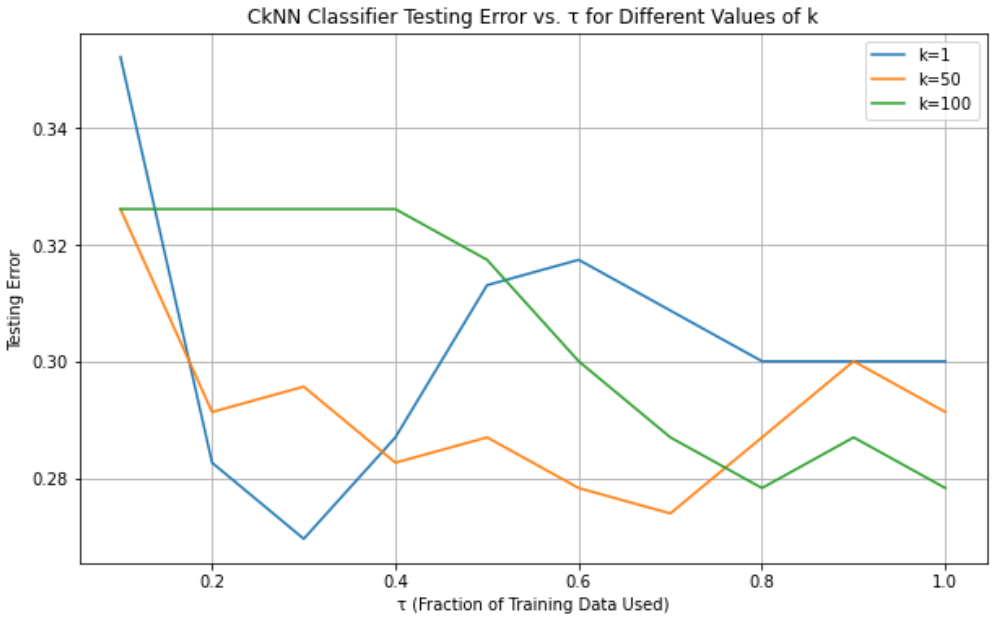

Fig. 2. A Sample Graph