Final Project

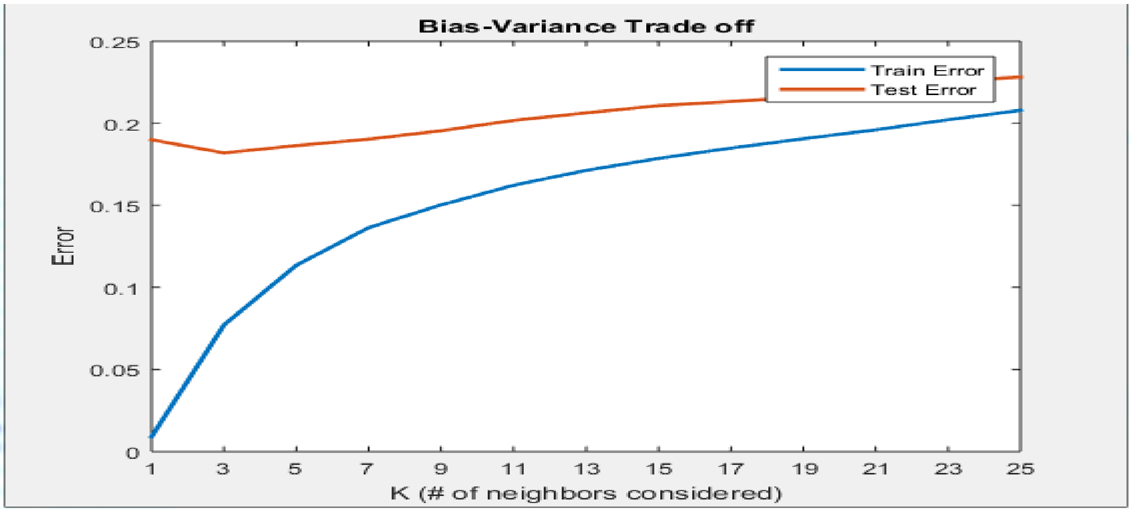
Niloufar Hosseini Pour

861242801

12/08/2016

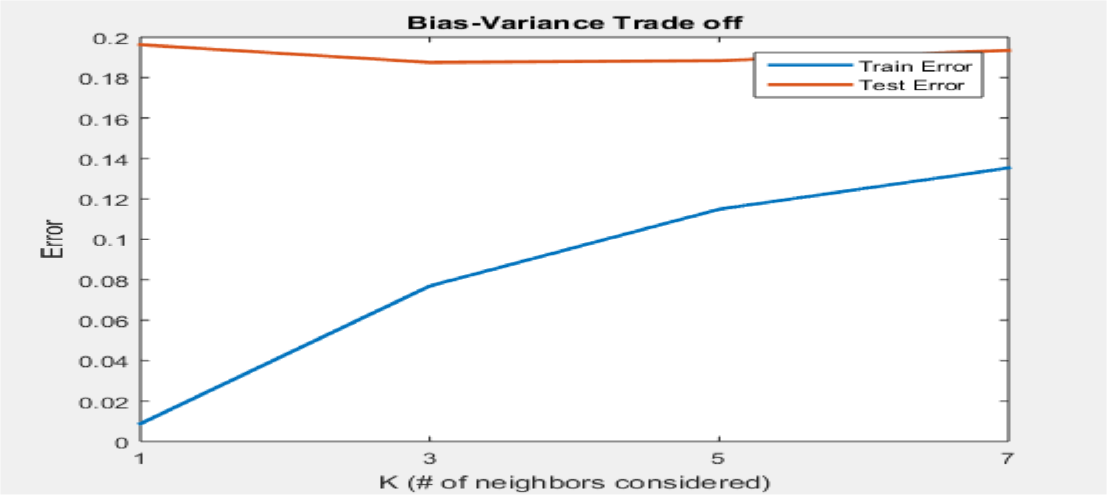
CS229

I am using nearest neighbor classifier for this character handwriting recognition problem. In order to pick the optimum K in K-Nearest Neighbor I have calculated both training error and test error for different number of K's. This process is written in "Initial_Approach.m" file. I chose the first 30000 records as training set (about 70% of data) and the rest for test set. The result is shown in the following plot:

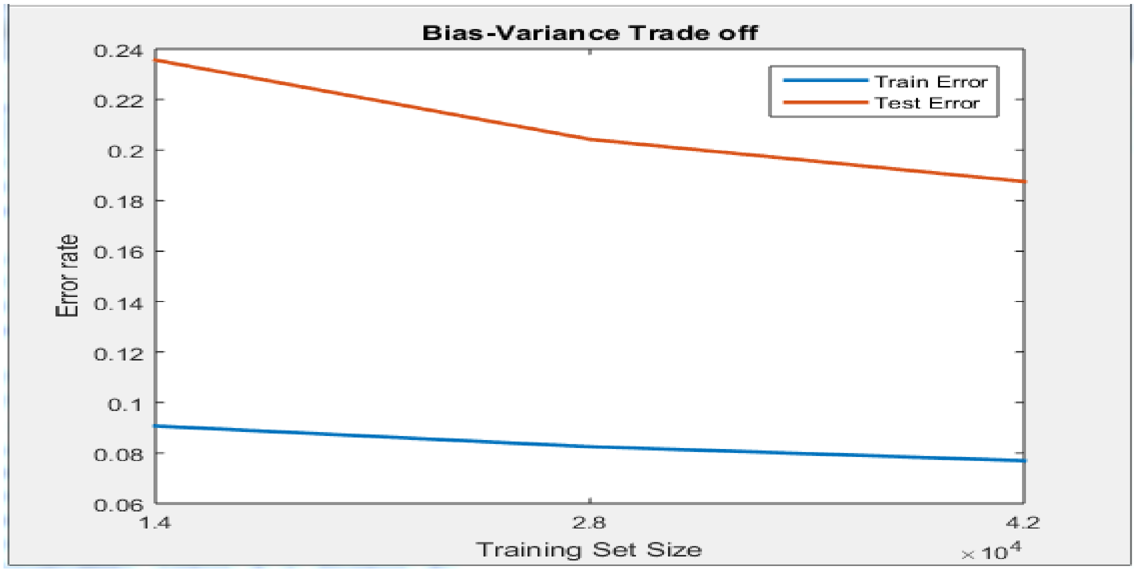


As it is shown in the above diagram, we see that we have smallest test error when k is 3 so best K is k=3. As number of k increases we have more bias. And when K is small we have high variance.

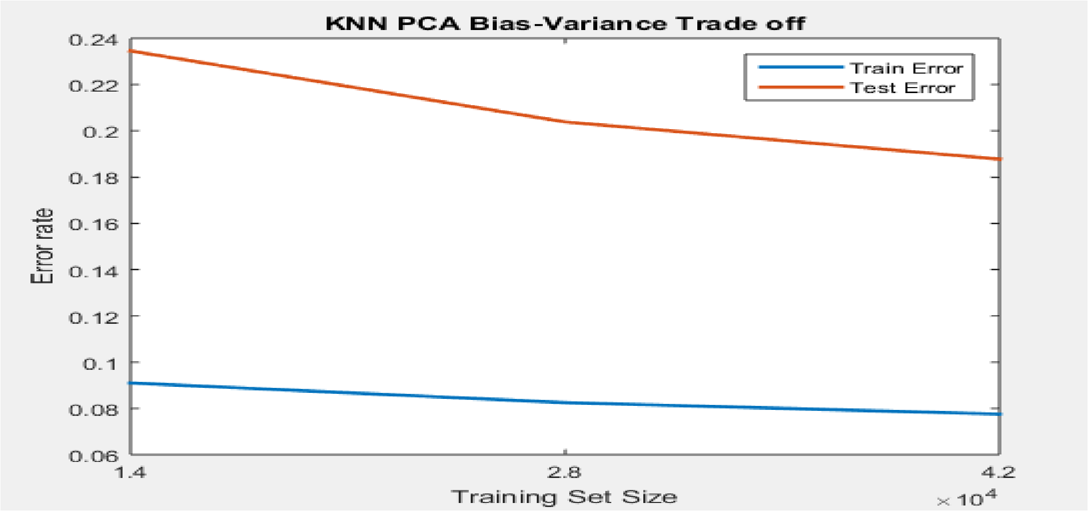
In order to have a more accuarate error rate, I use K fold cross validation method. In "Initial_Approach_CVE.m" file I used 3 fold cross validation. So each time training on 2/3 of data and testing on 1/3. And I got the same results. It is depicted in the following picture.



So I chose K=3. Now I want to analyze my initial approach to see if it is a good classifier for my problem or not. In "Initial_Approach_trainSize.m" file I have computed training error and testing error while training set size increases. This time my K is fixed and it is K=3. The result is as depicted in the following diagram:



As it is shown in the above picture, by increasing training set size, training error is decreasing and there is a large gap between test and training error which shows high variance. So this classifier has the problem of overfitting. In order to fix the high variance problem, I will reduce the set of features. I will apply a dimensionality reduction technique called principal component analysis (PCA). I got the idea of using PCA in conjunction with KNN algorithm by reading **"**Feature extraction for object recognition using PCA-KNN with application to medical image analysis" [1] paper. Again I applied 3 nearest neighbor classifier but this time on the reduced size feature set and calculated training and test error while training set size increases. The result is depicted in the following picture:



As it is shown in the picture, even after applying dimensionality reduction technique this classifier still has high variance. The accuracy of this classifier is about 82 percent on test set.

In order to increase accuracy we can take advantage of multi class SVM or Neural Network classification techniques.

In summary, as my initial approach I chose K_nearest neighbor technique. I had an experiment to choose optimal K, then I analyzed my initial approach by calculating training and test error (3-fold cross validation)and I identified the problem of overfitting. To fix high variance problem, I applied PCA, a dimensionality reduction technique. But even after that I still have high variance problem and also my accuracy rate is low. So we have to use another classifier to improve accuracy. I tried Matlab pattern recognition and classification in nnstart toolbox to apply neaural network method. It is written in ‘nn.m’ file. Initially I set 70 percent for train, 15 percent validation and 15 percent test and 1 hidden layer with 10 hidden units. It didn’t perform well. So we can increase number of hidden layers.

1- P. Kamencay, R. Hudec, M. Benco and M. Zachariasova, "Feature extraction for object recognition using PCA-KNN with application to medical image analysis," *2013 36th International Conference on Telecommunications and Signal Processing (TSP)*, Rome, 2013, pp. 830-834.