**Name:** Laksh Advani

**Course:** CSCI 5622 - Machine Learning, Fall 2017

**Date:** Friday, September 15, 2017

**Assignment:** 01 (Part 1)

**Title:** Analysis for Question 1

**Q1. What is the role of number of training instances to accuracy?**

From the plot above we observe that accuracy increases with an increase in the number of training instances. So, it’s easy to conclude the role of the training instances helps improve the accuracy of the model. We also observe as the training set increases in size, the number of unseen test values decrease and so do the outliers.

**Q2. What numbers get confused with each other most easily?**

Answer:

On analyzing the confusion matrix for a model of accuracy at 83% we can observe that the numbers **4 and 9 were most confused** with each other, followed by the pairs (5, 3) and (7, 9). 4 and 9 are appear more confused because they look quite similar when written by hand.

**Q3. What is the role of ‘k’ to training accuracy?**

From the ‘k’ vs accuracy plot above we can observe that when **‘k’ increases the accuracy drops**, so we can conclude that ‘k’ is inversely proportional to the accuracy of our model.

**Accuracy is well maintained (degrades marginally) when we have higher # of training points (say 5000).**

**Q4. In general, does a small value for k cause “overfitting” or “underfitting”?**

Answer:

Overfitting occurs when our model performs well on the training data but does not perform well on the evaluation data. This is because our model is memorizing the data and is unable to generalize for unseen examples.

On the other hand, underfitting occurs when our model performs poorly on the training data. This is because the model is unable to capture the relationship between the input examples and the target values.

With a small value of ‘k’ we have only a small number of data points to compare our target with, this acts as a model that has memorized the data. **So, a small value of ‘k’ causes overfitting**.

**Assignment:** 01 (Part 2)

**Title:** Analysis for Question 2

**Q1. What is the best k chosen from 5-fold cross validation with “—limit 500”?**

**Answer:** The best-chosen k with “—limit 500” is ‘3’ (k = 3).

**Q2. What is the best k chosen from -fold cross validation with “—limit 5000”?**

**Answer:** The best-chosen k with “—limit 5000” is ‘3’ (k = 3).

**Q3. Is the best k consistent with the best performance k in problem 1?**

**Answer:**

No, the best k for cross validation is not consistent with the best performance k in problem 1.