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?

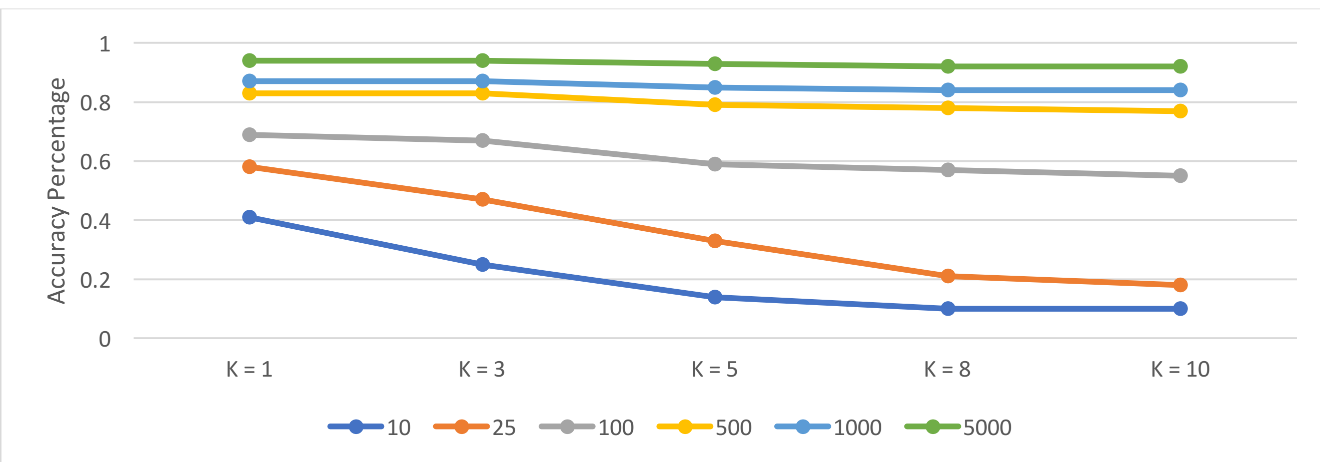
Answer:

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.

CSCI 5622: Machine Learning
Assignment Q3 Bias Variance Derivation
Laksh Advani

September 15, 2017

- Derive the bias-variance decomposition for k-NN regression in class. Specifically assuming the training set is fixed $S = (x_1, y_1), ..., (x_n, y_n)$, where the data are generated from the process $y = f(x) + \epsilon, E(\epsilon) = 0, Var(\epsilon) = \sigma^2$. k-NN regression algorithm predict the value for x_0 as $h_s = 1/k \sum_{l=1}^{k} y(l)$ where $x_(l)$ is the l-th nearest neighbor to x_0. $Err(x_0)$ is defined as $E((y_0 - h_s(x_0))^2)$.

  Prove that

  $Err(x_0) = (\sigma_\epsilon)^2 + [f(x_0) - 1/k \sum_{l=1}^{k} f(x_{(l)})]^2 + (\sigma_\epsilon)^2/k$

Answer: We have the error $Err(x_0)$ given as: $Err(x_0) = E((y - h(x_0))^2)$

$$Err(x_0) = E(f(x) + \epsilon - h(x_0))^2$$

$$Err(x_0) = f(x_0)^2 + \epsilon^2 + h(x_0)^2 + 2f(x_0).\epsilon - 2\epsilon.h(x_0) - 2f(x_0).h(x_0)$$

$$Err(x_0) = E(\epsilon^2 + 2\epsilon(f(x_0) - h(x_0))) + ((f(x_0) - h(x_0))^2)$$

$$Err(x_0) = E\epsilon^2 + 2E\epsilon * E(f(x_0) - h(x_0))) + E((f(x_0) - h(x_0))^2$$

The second term in the equation above will become '0' since $E\epsilon$ is zero.

$$2E\epsilon * E(f(x_0) - h(x_0))) = 0$$

$$Err(x_0) = (\sum \epsilon^2) + E(f(x_0) - h(x_0))^2$$

$$Err(x_0) = (\sigma_\epsilon^2) + (f(x_0) - E(h(x_0)))^2 + var(h(x_0))$$

Let us look at the $bias^2$, $((f(x_0) - Eh(x_0))^2)$
Here we replace $(h(x_0))$ with its value $1/k\sum_{l=1}^{k} y(l)$

We can move the terms around since the $bias^2$ is made up of squared terms.

$$bias^2 = ((f(x_0) - 1/k\sum_{l=1}^{k} y(l)))^2$$

Which is:

$$bias^2 = [(f(x_0) - 1/k\sum_{l=1}^{k} f(x_{(l)})]^2$$

Given that the set S is fixed, if we take the expectation of the bias over the test samples we will see that all the ϵ values disappear because the test sample has zero mean, so the value $E.1/k\sum_{l=1}^{k} y(l)))^2 = 1/k\sum_{l=1}^{k} f(x_{(l)}))^2$.

Now we will work on the variance section:

$$variance = var(h(x_0))$$

Here we will substitute the value of $h(x_0)$ with $1/k\sum_{l=1}^{k} y(l)$

$$variance = var(1/k\sum_{l=1}^{k} y(l))$$

$$variance = var(1/k\sum_{l=1}^{k} f(x_(l)) + \epsilon)$$

Here, we will use the following identity, $Var(aX) = a^2Var(X)$

$$variance = 1/(k^2)\sum_{l=1}^{k} var(f(x_l) + \epsilon)$$

$$variance = 1/(k^2)\sum_{l=1}^{k} var(f(x_l)) + var(\epsilon)$$

$$variance = 1/(k^2)\sum_{l=1}^{k} var(f(x_l)) + var(\epsilon)$$

Given that set S is fixed, the neighbors have no variance, which implies that:
$var(f(x_l)) = 0$

Hence, we get:

$$variance = 1/(k^2)\sum_{l=1}^{k} var(\epsilon)$$

$$variance = (1/(k^2))k(\sigma_\epsilon^2)$$

$$variance = (1/(k))(\sigma_\epsilon^2)$$

Now we have the following:

$variance = (1/(k))(\sigma_\epsilon^2)$

$bias^2 = [(f(x_0) - 1/k \sum_{l=1}^{k} f(x_{(l)})]^2$

Now when we substitute the values of $bias^2$ and variance we get our intended equation:

$$Err(x_0) = (\sigma_\epsilon)^2 + [f(x_0) - 1/k \sum_{l=1}^{k} f(x_{(l)})]^2 + (\sigma_\epsilon)^2/k$$