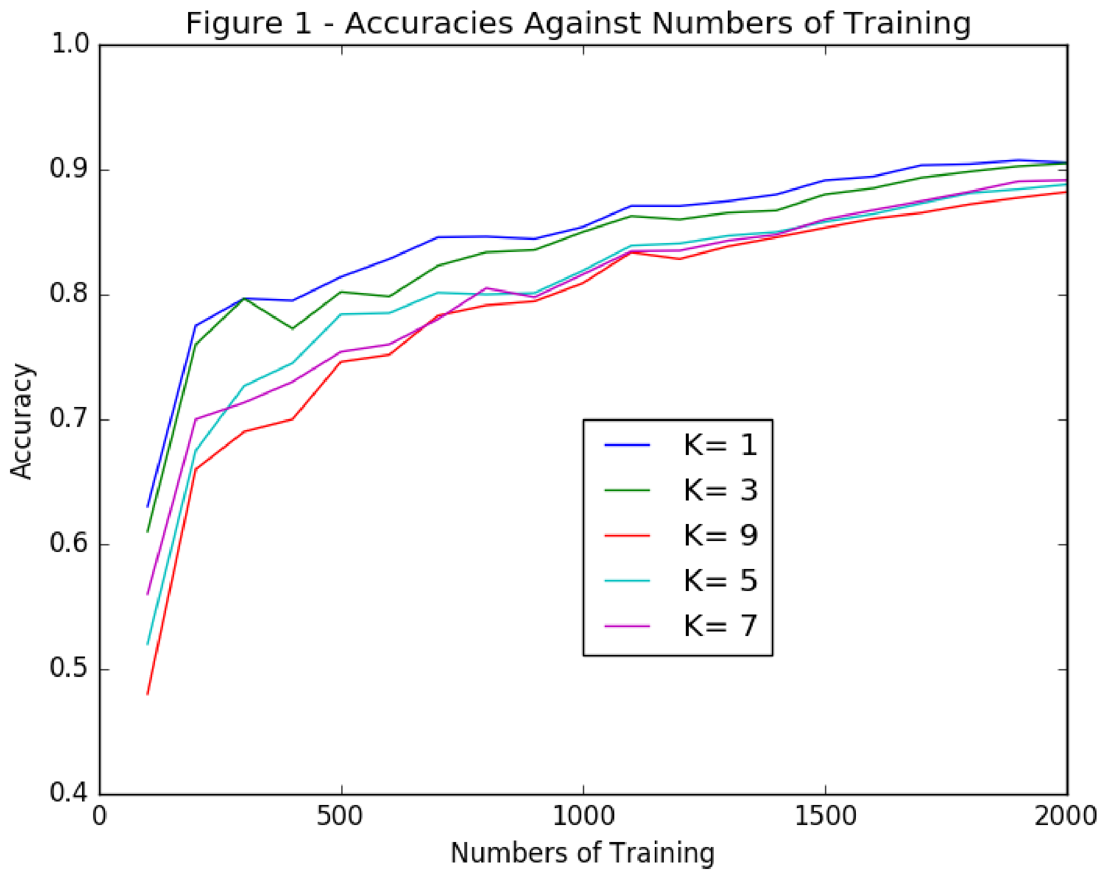
Amir Kashipazha
CSCI 5622 HW 1 Analysis

1.2 Analysis



This Figure shows accuracy of KNN against training instances.

1. What is the role of the number of training instances to accuracy (hint: try different "--limit" and plot accuracy vs. number of training instances)?
Accuracy on the KNN increases when the number of training instances increase. This can be seen in the above Figure. The accuracy range varies from 0.3 to 0.63 for lower instance of training. Then it will increase when the training number increase.

2. What numbers get confused with each other most easily?
According to confusion matrix, numbers 4 and 9 got confused the most with each other, for a total of19 times. Then 7 and 2 as well as 5 and 8 got confused 18 times.

3. What is the role of k to training accuracy?
The Figure above shows the role of k to training accuracy. As this Figure shows k of values 1 and 2 have a higher accuracy than larger values of k. The accuracy of higher k is lower because of averaging accuracies of more data points. Besides, the possibility of an incorrect classification is higher with larger k.

4. In general, does a small value for k cause "overfitting" or "underfitting"?
I believe a small value of k causes overfitting. We can smooth the prediction by using large k as long as it is not too large to prevent noise in the model. But small k may introduce noise to the model.

2.2 Analysis
1. What is the best k chosen from 5-fold cross validation with "--limit 500"?
The best K for 5-folds cross validation was 3 and its accuracy was 0.87. The accuracy was 0.83 for k=3 when it was run with knn.py.

2. What is the best k chosen from 5-fold cross validation "--limit 5000"?
The best K for 5-folds cross validation was 1 and its accuracy was 0.946. The accuracy was 0.938 for k=4 when it was run with knn.py.
The second highest accuracy was 0.944 which for k=3. This is only 0.002 smaller than k=1.

3. Is the best k consistent with the best performance k in problem 1?
Figure 1 shows the k=1 has the highest accuracy. Therefore the result of cross validation is consistent with the best performance of the problem. Accuracy difference between k values of 1 and 3 of cross validation is 0.0002. I believe these two k values could be considered as the best k. Figure 1 also shows that 1 and 3 have the highest accuracy.

3. Bias-variance tradeoff

We have:

(1) $E((y_0 - h_0(x_0))^2)$
(2) y = f(x) + ε
(3) E(ε) = 0
(4) $h_s(x_0) = \frac{1}{k}\sum_{l=1}^{k} y_{(l)}$

y of the equation (1) is replaced by equation (2):

$E((\varepsilon + \mathrm{f}(x_0) - h_0(x_0))^2) = E((\varepsilon + (\mathrm{f}(x_0) - h_0(x_0)))^2) =$
$E((\varepsilon^2 + 2\varepsilon(\mathrm{f}(x_0) - h_0(x_0)) + (\mathrm{f}(x_0) - h_0(x_0))^2) =$
$E(\varepsilon^2) + 2\mathrm{E}\varepsilon * \mathrm{E}\big(\mathrm{f}(x_0) - h_0(x_0)\big) + E(\mathrm{f}(x_0) - h_0(x_0))^2)$
$E(\varepsilon^2) = \sigma_\varepsilon^2$
Because E(ε) = 0 so $2\mathrm{E}\varepsilon(\big(\mathrm{f}(x_0) - h_0(x_0)\big))$ = 0
So will have
(5) $\sigma_\varepsilon^2 + 0 + E(\mathrm{f}(x_0) - h_0(x_0))^2$

We will continue with $E(\mathrm{f(x)} - h_0(x_0))^2$
$E(\mathrm{f}(x_0) - h_0(x_0))^2 = E(\big(\mathrm{f}(x_0) - E\big(h_0(x_0)\big)\big) + \big(E\big(h_0(x_0)\big) - h_0(x_0)\big))^2 =$

(6) E $(f(x_0) - E(h_0(x_0)))^2$ +

(7) $2E\left(\left(\mathrm{f}(x_0) - E\left(h_0(x_0)\right)\right) * \left(E\left(h_0(x_0)\right) - h_0(x_0)\right)\right)$ +

(8) $E\left(E\left(h_0(x_0)\right) - h_0(x_0)\right)^2$

Solving (6)

$E(\mathrm{f}(x_0) - E(h_0(x_0)))^2$ =

$E[f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f\left(x_{(l)}\right) + \varepsilon]^2$ =

$$[f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f\left(x_{(l)}\right)]^2$$

Solving (7)

$(7) = 0$ because E($\left(E\left(h_0(x_0)\right) - h_0(x_0)\right)$) = 0

Solving (8)

$E\left(E\left(h_0(x_0)\right) - h_0(x_0)\right)^2$=

E$[\frac{1}{k}\sum_{l=1}^{k} f\left(x_{(l)}\right) - \frac{1}{k}\sum_{l=1}^{k} f\left(x_{(l)}\right) + \varepsilon]^2$ =

$\frac{6_{\varepsilon}^{2}}{k}$

Finally from (5) and solutions of (6), (7) and (8) we can prove that:

$Err(x_0)$ = 6_{ε}^{2} + $[f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f\left(x_{(l)}\right)]^2$ + $\frac{6_{\varepsilon}^{2}}{k}$