Cover page for answers.pdf
CSE353 Fall 2020 - Machine Learning - Homework 3

Your Name: Cynthia Cheung

Solar ID: 111736494

NetID email address: Cynthia.Cheung@stonybrook.edu

Names of people whom you discussed the homework with:

~~Peter Hwang~~ Peter Hwang
Aggelina Chatziagapi
Ajay Gopakrishna

Question 1

1.1 Show that the optimal Bayes risk for data point x is $r^*(x) = \min\{\eta(x), \alpha(1-\eta(x))\}$

* $\eta(x)$ is probability that x is positive

A: Case 1: $\eta(x) > 0.5$

Bayes classifier will predict x is positive

Error: The error is if x is actually negative

Probability of ~~att~~ x actually is negative: $1-\eta(x)$

Cost of error: $\alpha(1-\eta(x))$

Case 2: $\eta(x) \leq 0.5$

Bayes classifier will predict x is negative

Error: The error is if x is actually positive

Probability of x actually positive: $\eta(x)$

Cost of error: $1(\eta(x))$

When we combine case 1 and case 2. The optimal Bayes risk for data point x is ~~$r^*(x) = \min\{$~~ $r^*(x) = \min\{\eta(x), \alpha(1-\eta(x)$

1.2. Let $r(x)$ be the asymptotic risk of the 1-NN classifier for the data point x, express $r(x)$ in terms of α and $\eta(x)$

A: Let data point z be the nearest data point to x in training data.

$$r(x) = \eta(x)(1-\eta(z)) + (1-\eta(x))\eta(z)$$
$$= \eta(x)(1-\eta(z)) + \alpha(1-\eta(x))\eta(z)$$

→ where x is negative and z is positive

When n goes to infinity, z goes to x:

$$r(x) = \eta(x)(1-\eta(z)) + \alpha(1-\eta(x))\eta(z)$$
$$= \eta(x)(1-\eta(x)) + \alpha(1-\eta(x))\eta(x)$$
$$\boxed{r(x) = (1+\alpha)\eta(x)(1-\eta(x))}$$

1.3 Prove that $r(x) \leq (1+\alpha) r^*(x) (1 - r^*(x))$.

A: Let x be a data point: Let $r(x)$ be the asymptotic risk of the 1-NN classifier for data point x.

$$r(x) \leq (1+\alpha)\, \eta(x) (1 - \eta(x))$$

$$r(x) \leq (1+\alpha)\, r^*(x) (1 - r^*(x))$$

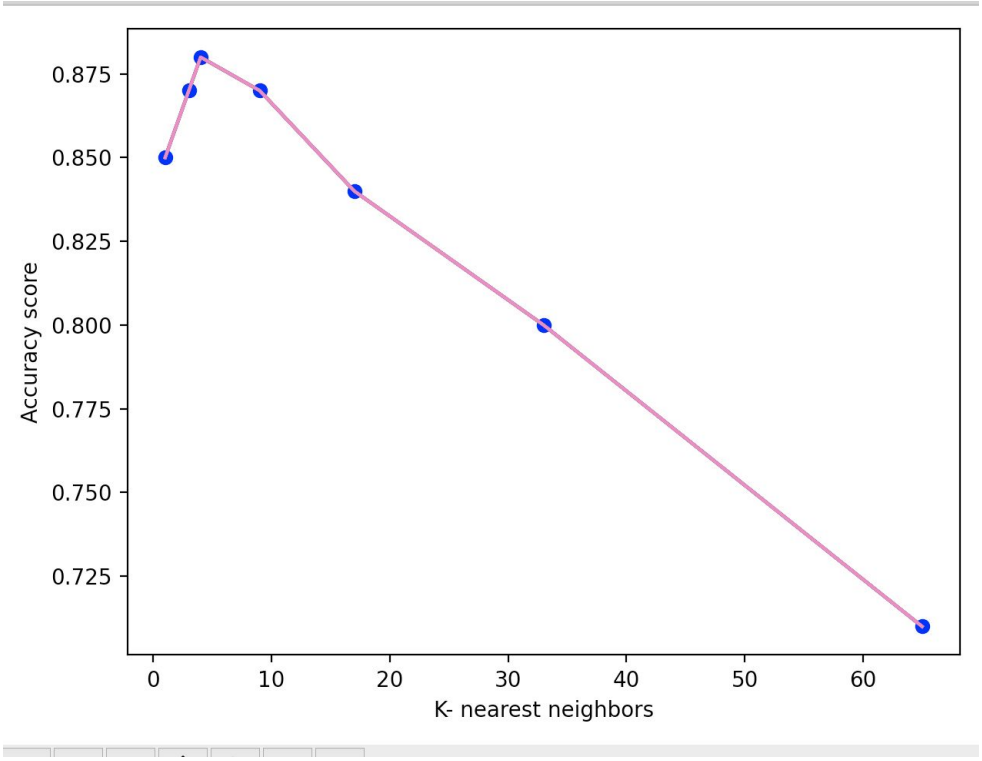
1.4 Let R be the asymptotic risk of the 1-NN classifier and R^* be Bayes risk. Prove that: $R \leq (1+\alpha) R^*(1-R^*)$

A: $R \leq (1+\alpha) E[r^*(x)] - (1+\alpha) E[r^*(x)]^2$

$(1+\alpha) E[r^*(x)] (1 - E[r^*(x)])$

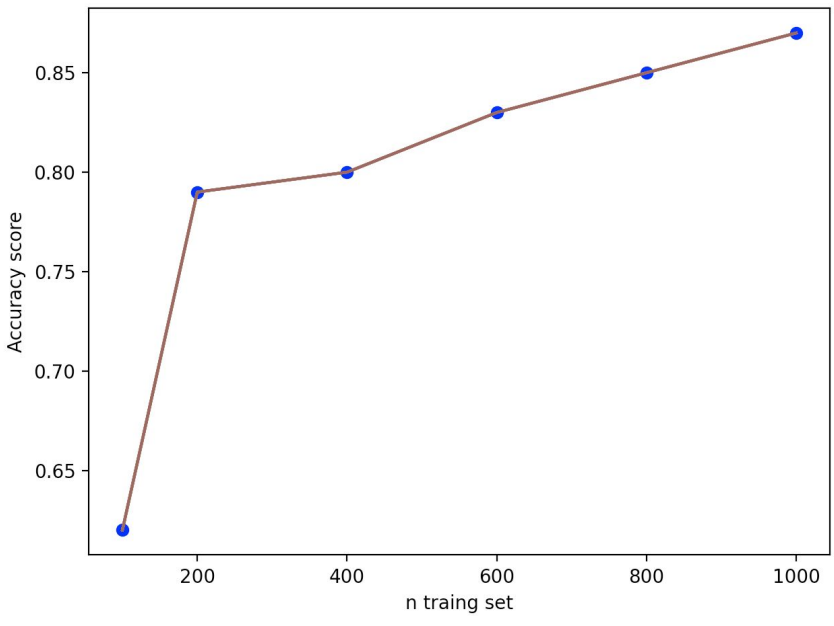
$\boxed{R \leq (1+\alpha) R^* (1-R^*)}$

2.2.1



K does affect the performance of the classifier. For different k values, there are different accuracy scores with k = 5 having the highest accuracy score. The accuracy score increases from k = 1 to k = 5, but the accuracy score decreases past the k = 5 value. This probably corresponds with the testing error increasing after k = 5.

2.2.2

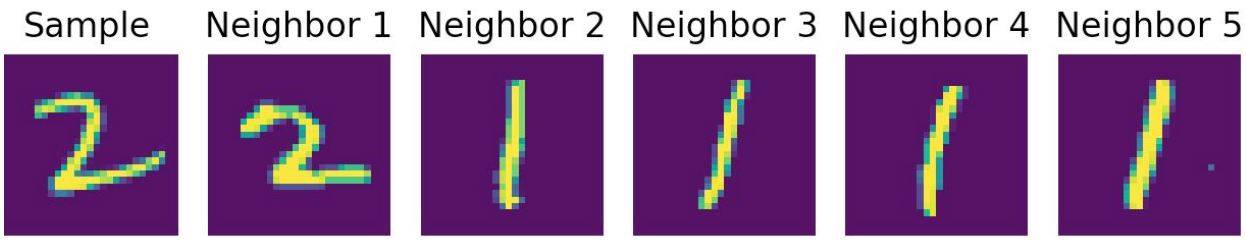


The number of training data does affect the performance of the classifier because as the number of training data increases, the accuracy performance increases as well.

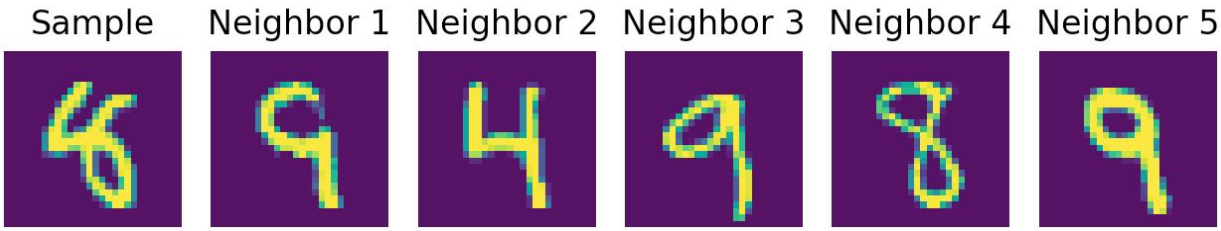
2.2.3
Accuracy of the test data set for k = 3 and using “Manhattan” distance: 0.83
Accuracy of the test data set for k = 3 and using “Euclidean” distance: 0.87

The result is better using the Euclidean distance due to the greater accuracy.

2.2.4
Test sample 1



Test sample 2



Test sample 3