Orçun Özdemir

54020

21.12.2018

COMP341 Report of Assignment 5

Answer of Question 1

Without normalization data can be too complex to process, in addition in order to interpret histogram faster and more optimized, normalization is crucial. The purpose of normal distribution is to handle the data segments as probabilities. Consequently, normalization of data is needed due to the transform the data into normal distribution.

Answer of Question 2

The overfitting problem is occurring when k values are in between 0 and 5. Thus, as we learned in the course, overfitting can be solved by increasing k value and subsequently, to observe the underfitting training and test errors should be both high, however there is not proper and signifying point to clearly let me see this situation. When we think about ideal k value, 17 can be that value since, training and test errors are low.

Answer of Question 3

According to my observations, overfitting occurs at c = 10 and continues until the end of the histogram. The best alternative of c value to 10 is 1, because both error rates are lower than the rest.

Answer of Question 4

I did not observe any fundamental differences between the error rate of KNN and error rate of Log-Reg, therefore I would choose best k for KNN and best c for regression. On the other hand, if the rest of the hyper parameters are added to the algorithm, I would choose KNN, since even in smaller k's, the result is pretty decent.

Answer of Question 5

I believe that beginning of graph can be seen as overfitting situation, however after the lambda goes through 0.1, test and train error rates decreases and get stabilized. Specifically, selected parameters are not the same for ridge and linear regression. In addition, ridge regression is more beneficial than the linear one, because the average rate of error in linear regression doubles the average rate of error in ridge regression and standard deviation is smaller in ridge regression as well.

Answer of Question 6

When the parameter is getting increased, both error rates increases as well. Therefore, if the parameter keeps increasing, underfitting occurs. I could not observe any clear and specific point for overfitting situation and the parameter selection is different as well. On the other hand, their standard deviation and average error rate are almost the same. Thus, comparison of these two cases might not be the ideal and optimal decision.

Answer of Question 7

One thing for certain is that this case is an example of overfitting, since all values in horizontal axis are the representation of overfitting. The testing error rate is quite high on the other hand training error rate is low. Even though the different data selection occurs, their standard deviation and their average error rate is enormously look alike. Therefore, comparison of these two cases might not be the ideal and optimal decision.