CSE 517A Milestone4 Readme

Yuxiang Wang
Ashwin Kumar

`https://github.com/tiger12055/cse517a_example_application_project`

May 4, 2018

1 Summary on Milestone4

We compare the different methods by performing 10 fold cross validation on the predictions and performing summary statistics and then using a statistical test such as the t-test. Linear Regression (A) vs Gaussian Process (B). The box captures the middle 50% of the data, outliers are shown as '+' and the red line shows the median.

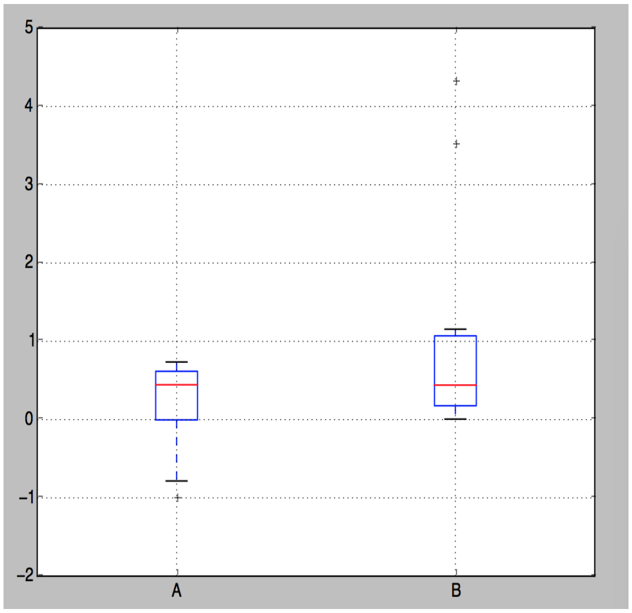

Figure 1: Comparison between Linear Regression and Gaussian Process

We can see the data indeed has a similar spread from both distributions and is not symmetric about the median. We see that A (Linear regression) performs better in handling outliers.

We cannot use the Student t-test or the Welchs t-test if our data is not Gaussian.An alternative statistical significance test we can use for non-Gaussian data is called the Kolmogorov-Smirnov test.This test can be used on Gaussian data, but will have less statistical power and may require large samples.Using this test we find The p-value is very small, suggesting a near certainty that the difference between the two populations is significant.And also see that the Gaussian Process is better statistically than Linear Regression.

Next, we compared the performance of Semi-Supervised Learning using Linear Regression and Gaussian Process. By observing the following image

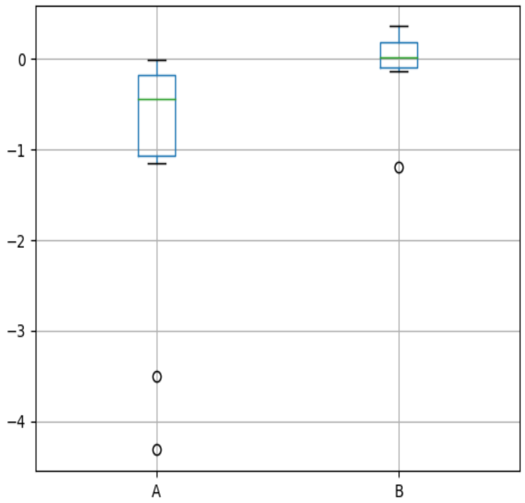

Figure 2: Comparison between Semi-Supervised Learning using Linear Regression and Gaussian Process

We can find B (Semi-Supervised Learning) has a better spread compared to A (Gaussian Process). Also, we can see B performs better in handling outliers.

After applying Normality Test on both data-set, we found both data-set are not form a normal distribution. Thus, we cannot use the t-test. We applied Kolmogorov-Smirnoy test on them and we found the p-value is very small. As a result, we have a near certainty that the difference between the means is statistically significant. Based on the image above, we can conclude Semi-Supervised Learning will be a better method comparing to the Gaussian Process even through the Mean Square Error we got in GP (MSE = 1.12) is better than Semi-Supervised Learning (MSE = 11.01.

2 Option: Semi-Supervised Learning

We used first 100 Boston housing data and its target as our labeled data. Then we used sklearn library function LabelSpreading to learn the rest of unlabeling data through labeled data. Finally, apply linearly regression to new dataset and we got 11.01 as mean square error compared to 989.48 if we only used the first 100 labeled to train our model then predict the rest of data. Moreover, if we compare this result to the milestone1 which we got 25.74 as mean square error, the semi-suprevised still preforms better. Thus, by adding cheap and abundant unlabeled data, we are able to build a better model than using supervised learning alone.