CS 6140 Machine Learning: Homework 1

Jayantha Nanduri

September 2023

Question 1

- (a) The sample size n is extremely large, and the number of predictors p is small - **Flexible statistical learning method** would be better as it would have less bias when compared to the non-flexible model and with a large amount of data, the risk of overfitting is reduced, which is a common concern with flexible models.
- (b) The number of predictors p is extremely large, and the number of observations n is small - With a small number of data points, the risk of overfitting increases for a flexible model(high variance). **Non-flexible model** should be picked based on bias-variance tradeoff.
- (c) The relationship between the predictors and response is highly non-linear - A non-flexible model can never correctly estimate non-linear data, implying that this model has a high bias. **Flexible model** should be picked based on bias-variance tradeoff.

Question 2

- (a) - (d) present in Jupyter Notebook
- (e) Increases K makes a KNN mode less flexible - Since KNN is a type of instance-based or lazy learning algorithm that classifies data points based on the majority class among their K nearest neighbors, when K is small, the model considers only a few nearest neighbors, which can result in more complex and jagged decision boundaries. But as K gets larger, the model takes into account a greater number of neighbors, leading to a more generalized and smoother decision boundary.
- (f) Bias: As K increases, bias generally increases - This happens because when K is large, the model considers a larger number of neighbors, resulting in a more generalized prediction. However, it might not capture local patterns in the data, leading to a higher bias.

Variance: As K increases, variance generally decreases - With a larger K, the model relies on more neighbors to make predictions, which results in a smoother and more stable prediction. This reduces the variance because the predictions become less sensitive to individual data points.

Training MSE: Training MSE tends to increase with larger values of K - When K is large, the model is more likely to underfit the training data. It becomes overly generalized, causing it to make predictions that are farther from the actual training data points, resulting in higher training MSE.

Test MSE: Test MSE initially decreases and then increases with K, following a U-shaped curve - Initially, as K increases, the model captures more global patterns in the data, reducing test MSE. However, when K becomes too large, the model loses the ability to capture important local patterns, leading to an increase in test MSE due to underfitting.

Irreducible Error: Irreducible error remains constant, regardless of the value of K - Irreducible error represents the inherent noise or randomness in the data that cannot be reduced by the model. Changing K does not affect this source of error.

Question 3

- (a) - (h) will be in jupyter notebook
- (i) plots in notebook. As the degree of the polynomial model increases, there is a steady decline in squared bias and mse, indicating that the model has become more flexible. However as the model becomes more flexible, it is more sensitive to changes in data points and this is shown in the variance plot. There is a line parallel to x-axis in the graph which is the irreducible error. Irreducible error remains constant, regardless of the value of degree.

Question 4

- (a) in jupyter notebook
- (b) From the plot we can infer that the generated data is non-linear.
- (c) Results shown in Jupyter Notebook
- (d) Results shown in Jupyter Notebook. I didn't get the same results for (c) and (d) because I used a different seed in (c) and (d).

- (e) Polynomial with degree 5 has the smallest LOOCV error. Yes, since the model used to generate the synthetic data is a polynomial model of degree 5.

- (f) const (Intercept):

  Coefficient Estimate: -0.0053 Statistical Significance: The coefficient for the intercept is statistically significant with a p-value of 0.034. This indicates that the intercept is not likely to be zero, meaning there is a statistically significant constant term in the model.

  x1: Coefficient Estimate: 0.1436 Statistical Significance: The coefficient for x1 is statistically significant with a low p-value of 0.005. This suggests that changes in x1 have a significant effect on the dependent variable, and the estimated coefficient is unlikely to be zero by chance.

  x2: Coefficient Estimate: -0.8814 Statistical Significance: The coefficient for x2 is statistically significant with a p-value of 0.006. This indicates that x2 has a significant negative impact on the dependent variable, and this relationship is unlikely to be due to random chance.

  x3: Coefficient Estimate: 3.1448 Statistical Significance: The coefficient for x3 is highly statistically significant with a p-value of 0.000. This suggests a strong positive relationship between x3 and the dependent variable, and the estimated coefficient is very unlikely to be zero by chance.

  x4: Coefficient Estimate: -4.2798 Statistical Significance: The coefficient for x4 is highly statistically significant with a p-value of 0.000. This indicates a strong negative relationship between x4 and the dependent variable, and the estimated coefficient is very unlikely to be zero randomly.

  x5: Coefficient Estimate: 1.8827 Statistical Significance: The coefficient for x5 is highly statistically significant with a p-value of 0.000. This implies a strong positive association between x5 and the dependent variable, and the estimated coefficient is very unlikely to be zero by chance.

  In summary, all the coefficients in the model have estimated values that are different from zero. The statistical significance of the coefficients is assessed using p-values, and all of them have p-values well below the common significance threshold of 0.05, indicating their statistical significance.