Linear Regression Report

20180396 오지오

. This is the linear regression equation, in which we approach to model a relationship between a dependent variable and independent variables. We tend to find the function that best explains the current datasets and try to predict the values, when new datasets come into the model. We want to find the most accurate model, thus we try to minimize the objective function (the squared sum of the difference of true value and predicted value), , and find the coefficient that minimizes the objective function. Thus, we are trying to solve . are both scalar values and one is a transpose of another, thus they have the same values. If we differentiate this w.r.t to and set it to zero, we can derive the unique solution, , thus the ordinary least square estimator is .

B. (Result3/result\_5.png)

Regarding all the average MSE() and the standard deviation, I derived from the model, I believe the model with the dimensionality of 3 is the best. In order to evaluate the model, I used the coefficient of determination, also known as R2, which as the formula of . This tell us how well the observed outcomes are expressed by the model, based on the proportion of total variance. As this coefficient value gets closer to one, we determine that the model well expresses the whole dataset. As we can see in the output text file, and the graph below, the R2 value was the closest to 1 (0.7907), for the model with dimension 3. Plus, both the average MSE and average standard deviation is the lowest for the model with dimension 3 and we say that the model with the lowest MSE and lowest variance is the best model.

**

(output.txt) (R2 value for each dimension)data\_compare.py

C.

As we can see in the above output.txt file, the best average MSE was 0.0947 and the deviation for it was 0.0396.

D. (Result9/result\_5.png)

It seems that the model is overly optimized for only the training dataset, causing overfitting. The model does reflect the tendency and movement of the training datasets, but doesn’t do so for newly arrived datasets (test data). We can see that the average MSE and standard deviation for this particular dimension are extremely large compared to the optimal model. (approximately 7 times bigger average and 22 times bigger deviation) Plus, the value of R2 is negative, which tells us that the model is even worse than a horizontal line (. Generally speaking, if the model has too much dimensions, or overfitted, then the model has a high degree of flexibility and captures every change (noise) in the training set (high variance), leading to express an incorrect result. On the other hand, if the model’s dimension is too low, or underfitted, it has low variance, but since it is too simplified and less flexible it cannot account for the data (high bias). The model should have some proper number of dimensions (in this case 3) in order to have low bias and low variance, thus express the model optimally.