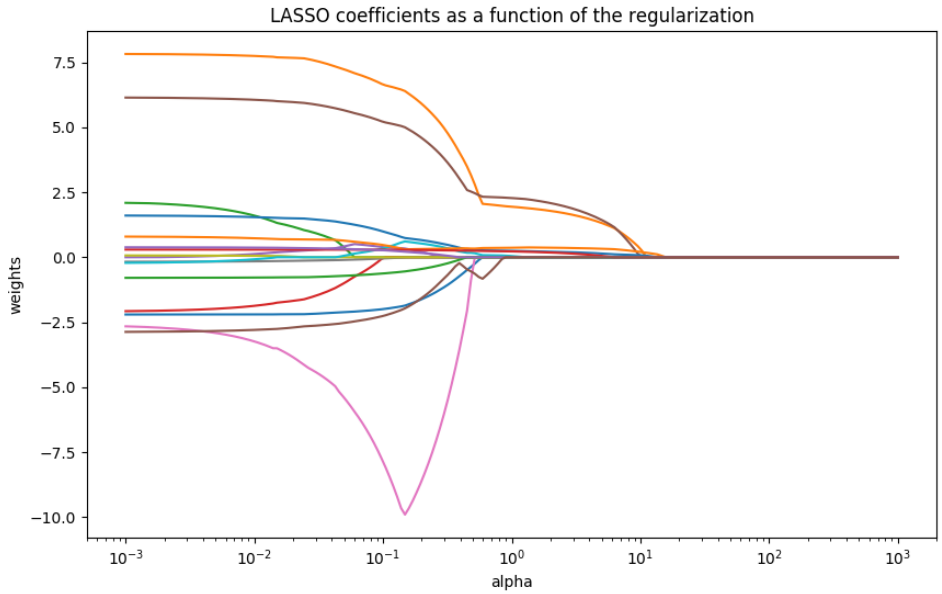
Kornraphop Kawintiranon (Ken)
kk1155@georgetown.edu

Homework 0

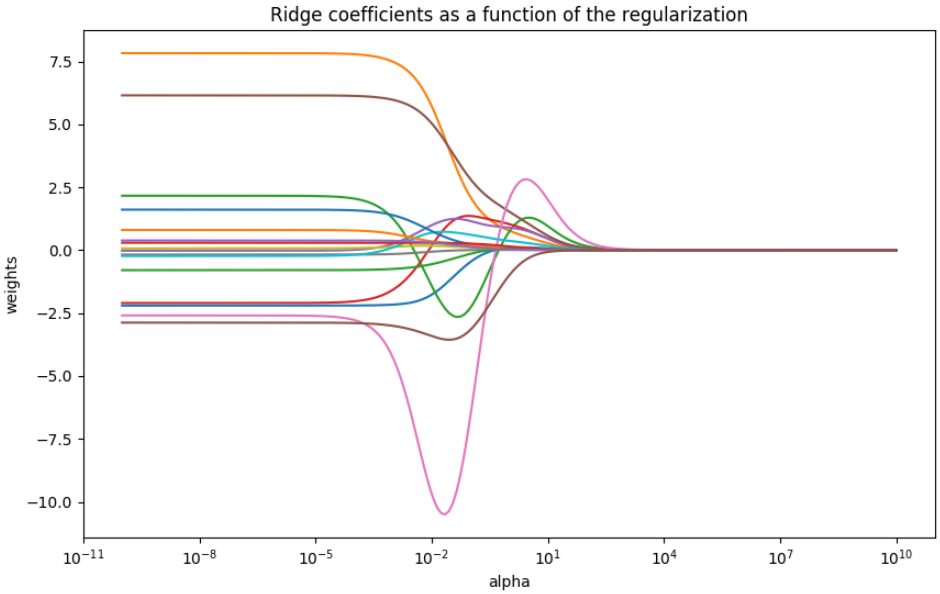
1. Regularization

1.1 LASSO Regression



- Final three predictors are Hits, CRuns, CRBI
- After using cross-validation, 14 predictors are left in the model

1.2 Ridge Regression



- There are no last three features, all features are still in use since their coefficients are non-zero
- After using cross-validation, all 16 predictors are left in the model

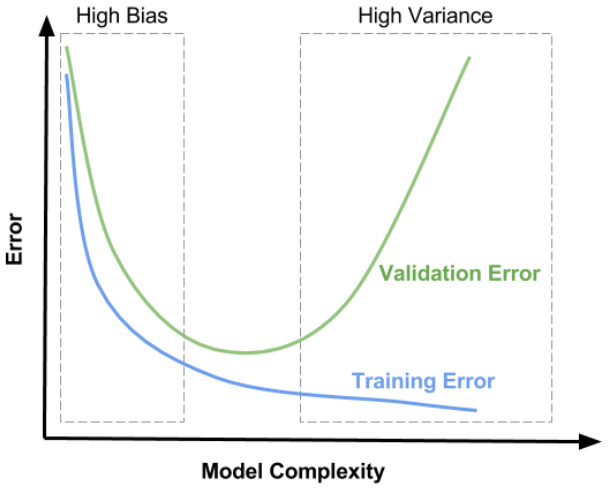
Kornraphop Kawintiranon (Ken)
kk1155@georgetown.edu

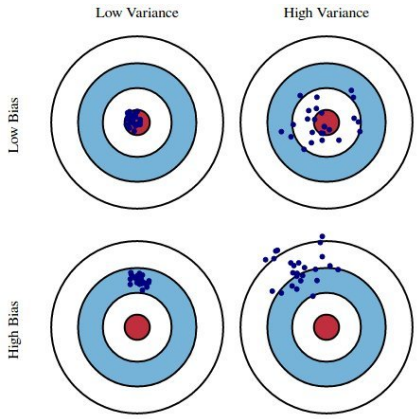
2. Bias-Variance Tradeoff

2.1 Explain in your own words the bias-variance tradeoff

Bias is something like how well a model can predict given the training set. High bias means too simple model, it does not work well even with training set. Low bias, conversely, means the model is complex enough, it works well in training set.

Variance means how well a model can predict unseen observations (test set or validation set) given the training set. In the other words, variance means how large different error a model can perform given train and test set. This statement is quite confusing but it is shown clearly in the figure on the right. Hight variance means the different error between validation error and training error is high. This simply means the model is so complex that it starts memorize the training data and works well only on it, does not work well on test set or in general.





Ideally, if we can build a model that is good in both (low) bias and (low) variance, it should be the best but actually it's almost impossible. That's why it is called tradeoff. Realistically, when we build a very complex model, it gives a very good performance based on training set, but it is usually not good when validate the model with test set, called *overfitting* (low-bias but high variance as shown in the previous figure). However, if a model is not complex enough or too simple, its performance is generally in both training and test set. This indicates high bias and high variance. To understand all cases including low-hight for both bias and variance, the figure below does really well in visualization.

In conclusion, Low Bias/Low Variance means the model is well tuned and work well in general. Low Bias/High Variance represents *overfitting*, it works well only in training set. High Bias/Low Variance represents *underfitting. The* model does poorly on any given dataset. Finally, High Bias/High Variance is like the first case, it's almost impossible because in Machine learning, the model learns something to predict unseen observations in the future but in this case the model just randomly gives prediction. No strategy, no algorithm, no knowledge in the model. In general, machine learning model will gives us high bias/low variance or low bias/ hight variance. That's why this is called tradeoff.

Kornraphop Kawintiranon (Ken)
kk1155@georgetown.edu

2.1 What role does regularization play in this tradeoff? Make reference to your findings in number (1) to describe models of high/low bias and variance.

The regularization gives coefficient scores of features which can make better performance in building a model. In the other words, it tells us how important each feature is. Once we get the coefficients scores, we can decide which features should be considered and which features are noise and should be removed from our model creation. By this, we can build a better model since noise features are removed. For example, in the experiment from Q1, the model built using Lasso regression performs better than Ridge regression in term of Mean Square Error (MSE=96243:113323).

I have compute we can compare this as the figure we mentioned before in Q2.1. When models for every alphas and compute average MSE for both train and test set. The plot is shown below. This explicitly demonstrates that the figure shown in Q2.1 reflects the real world situation, more complex model (low alpha) performs better (low error). However, once the model is too complex, it still gives a better result for training set but it goes worst when validate with test set. On the right side of the figure below or simple model means *High Bias* since both train/test error are close to each other and still high, called *Underfitting*. On the left side of the figure or more complex model, shows *High Variance* because train error still goes down but test error starts rising up. This indicates that the model starts to be *Overfitting*. Both two figures of Lasso and Ridge illustrate this in the same way.