Homework Assignment 1- CMPS242: Machine Learning

Huimin Yan, Pegah Ojaghi
hyan90@ucsc.edu, pojaghi@ucsc.edu

April 10, 2018

1 Cross validation

Cross-validation is a method to assess predictive performance of the models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. The model is used to judge how they perform outside the sample to a new data set. As we know, training data is expensive and generating more of it for testing is not easy. It is a good way in such cases to use a part of the available data for training and a different part for testing the model. The part of the data used for testing is also called a validation dataset.

In this homework, we repeat the sampling of a validation set multiple times and use different samples each time for the validation set to evaluate the performance. We create 10 different validation sets and 10 different models on the remaining training datasets. we have 10 folds or experiments and we chose 10-Fold validation, then this would break up the training dataset into 10 folds. In the end we get just average those 10 different test errors and will end up with a better estimate which is less dependent on the actual sample of the test set.

2 10-fold cross validation with bias term

The goal of this section is to use 10-fold cross validation to find the best choice of λ and report the loss on the test set. The motivation to use cross validation method for us is that when we fit a model, we are fitting it to a only training dataset. Without cross validation we only have information on how does our model perform over our sample data. Ideally we are trying to see how does the model perform when we have a new data in terms of accuracy of its predictions. In this homework, the performance is judged by its errors.

2.1 basic algorithm

The algorithm 1 indicates how we implemented the 10-fold cross validation with bias term. In cross-validation, we run our modeling process on different subsets of the data to get multiple measures of model quality. We divide the data into 10 pieces, each being 10 percentage of the full dataset. There two types of cross validation we can perform: leave one out and k fold. In this section, we used k fold cross validation with bias term. For finding the best choice of λ, we choose the λ candidates range from logspace(-2, 2, 50), which can make the λ numbers spaced evenly on a log scale. For solving the problem that we don't know how to measure that the validation set was not particularly easy for the model. We use random sample to solve this problem.

2.2 result of 10-fold cross validation with bias term

Using a validation set from our training data in order to calculate the test data is an excellent way to get a much more reliable estimation on the future accuracy of a model. We run this experiment in the way that we use the first fold as a validation set, and everything else as training data. Therefore, we run a second experiment, where we hold out data from the second fold. We repeat this process, using every fold once as the validation set. Putting this together, 100 percentage of the data is used as a holdout at some point. We implement this process for every

Data: train.txt, test.txt
Result: use 10-fold cross validation on train data, and get best predict of test data

```
/* read data from file */
read train dataset from train.txt;
read test dataset from test.txt;
/* use cross validation to find best λ */
for λ ← logspace(-2 to 2) do
    repeat
        shuffle train dataset;
        split train dataset into 10 fold;
        for fold ← 1 to 10 step 1 do
            get actual train data matrix X from other 9 fold data;
            get valid data matrix valid_X from this fold data;
            get ω by solving function ω* = (XX^T + λI)^{-1} Xt;
            get root mean square error using ω to valid data matrix;
        end
    until 1000 times;
end
best λ = the λ corresponding to the minimum average root mean square error;
/* calculate best ω using best λ */
get train data matrix X from original train dataset;
get best ω by solving function ω* = (XX^T + λI)^{-1} Xt, where λ = bestλ;
/* calculate test error */
get test data matrix test_X from original test dataset;
get test root mean square error using best ω to test_X;
```

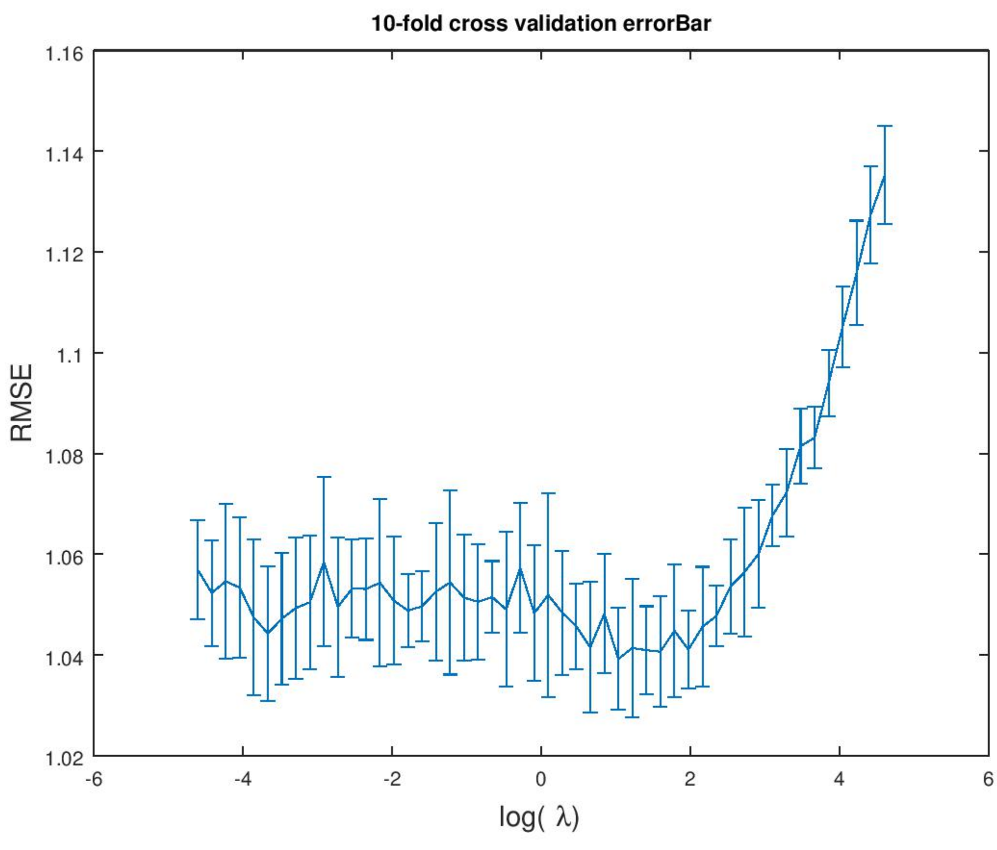
Algorithm 1: Proposed algorithm for 10-fold cross validation with bias term

λ in the range of logspace(-2, 2, 50). The best λ that is chosen is: $\lambda = 2.8118$
The best weight ω that is obtained for the best $\lambda = 2.8118$ given by:

$$\omega = \begin{bmatrix} 1.1768e{+}00 \\ -3.0218e{-}01 \\ -1.1149e{-}01 \\ 2.4050e{-}01 \\ -9.5471e{-}03 \\ -3.5887e{-}02 \\ 1.2731e{-}03 \\ 1.8577e{-}03 \\ -3.5859e{-}05 \\ -3.1714e{-}05 \end{bmatrix}$$

For evaluation we use root mean square error as a common metric used to evaluate the difference between estimated and observed values. The root mean square error of test set is: 1.0901
The following Fig 2.2 is shown the ln(λ) vs RMSE errorbar plot.

Figure 1: The mean square error with RMSE errorbar plot for 10-fold cross validation with bias term



As it is shown in Fig 2.2, when ln(λ) belongs to the range of -4 to 2, the error is around 1, which is pretty well. When ln(λ) is bigger than 2, error will increase very rapidly. However, the variance of RMSE is very small. We reduce variance without increasing bias by repeating cross-validation with the same K folds but different random folds and then averaging the results.

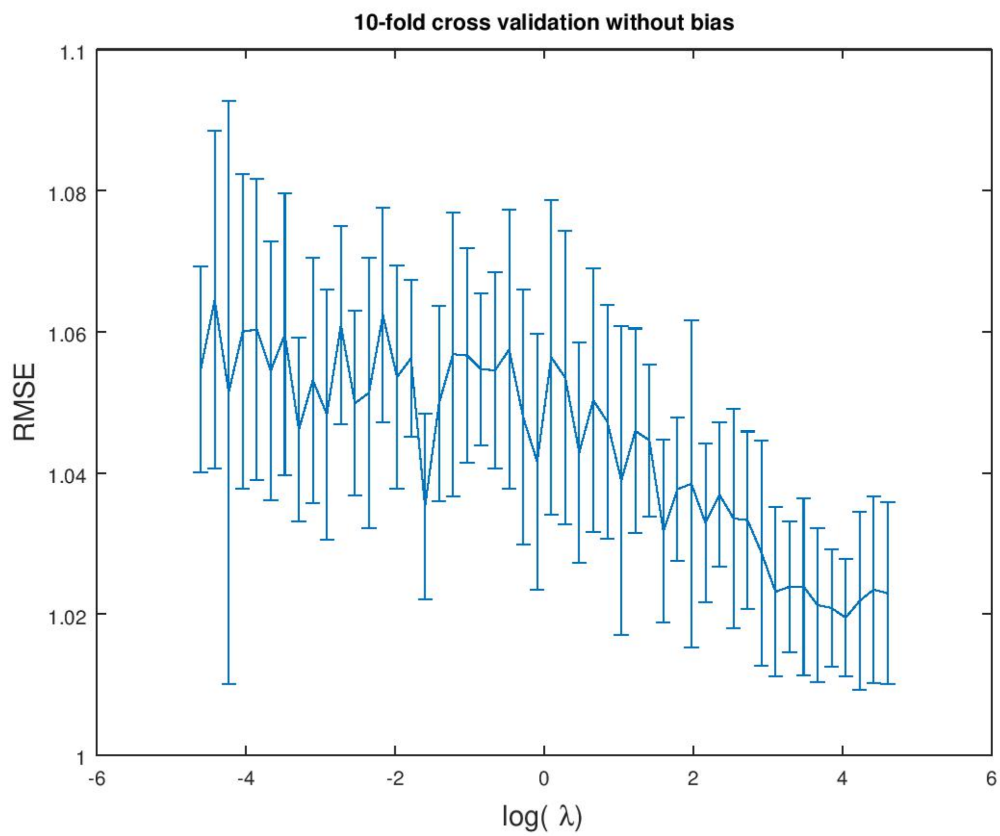
3 10-fold cross validation without bias term

In this section, we implemented the 10-fold cross validation without a bias correction for the error rate in 10-fold cross-validation. Repeating the algorithm 1 without the bise term $biasterms = 0$.
The best λ that is achived is: $\lambda = 56.899$
The best weight ω that is obtained for the best λ given by

$$\omega = \begin{bmatrix} 1.2368e{+}00 \\ -2.6890e{-}02 \\ -8.5475e{-}02 \\ 1.1356e{-}01 \\ -1.6702e{-}02 \\ -1.9041e{-}02 \\ 1.7368e{-}03 \\ 1.0086e{-}03 \\ -4.4700e{-}05 \\ -1.7279e{-}05 \end{bmatrix}$$

For evaluation we use root mean square error as a common metric used to evaluate the difference between estimated and observed values as previous. Minimizing the prediction error against reference samples in the form of root mean square error (RMSE) is commonly used as an objective for model selection. The root mean square error of test set is: 1.0858
The following Fig 3 is shown the $\ln(\lambda)$ vs RMSE errorbar plot.

Figure 2: The mean square error with RMSE errorbar plot for 10-fold cross validation without bias term



In comparison to the previous cross validation the RMSE is a little smaller than using bias term, but the variance bar is bigger than cross validation with bias.

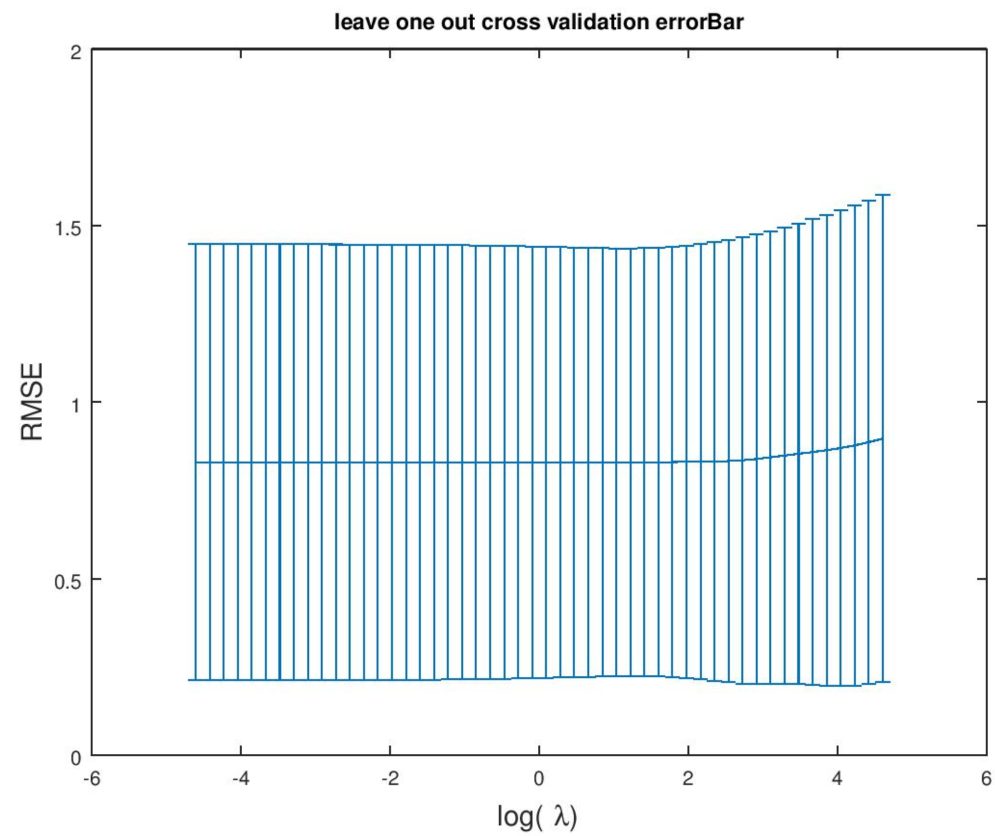
4 leave one out cross validation with bias term

The model may be more computationally demanding. It could be that the random sample we selected is not so random after all, especially if you we have small training datasets available like our example. The same logic can be applied using leave one out cross validation. The best λ that is achieved is: $\lambda = 1.9307$
The best weight ω that is obtained for the best λ given by

$$\omega = \begin{bmatrix} 1.2331e{+}00 \\ -3.2638e{-}01 \\ -1.4366e{-}01 \\ 2.5041e{-}01 \\ -4.7243e{-}03 \\ -3.7114e{-}02 \\ 1.0103e{-}03 \\ 1.9166e{-}03 \\ -3.1144e{-}05 \\ -3.2675e{-}05 \end{bmatrix}$$

The root mean square error of test set is: 1.0830
The following Fig 4 is shown the $\ln(\lambda)$ vs RMSE errorbar plot.

Figure 3: The mean square error with RMSE errorbar plot for leave one out cross validation with bias term



We implemented leave one out cross validation and showed that in RMSE sense this is better than 10-fold cross validation. The RMSE is a little smaller than 10-fold cross validations, the reason is that leave one out use more training data than 10-fold. As we know leave-one-out cross-validation is a particular case of leave-k-out cross-validation with k = 1. This methods is widely used when the available data are very rare, but in our case because our data is not very rare the RMSE difference is very small.

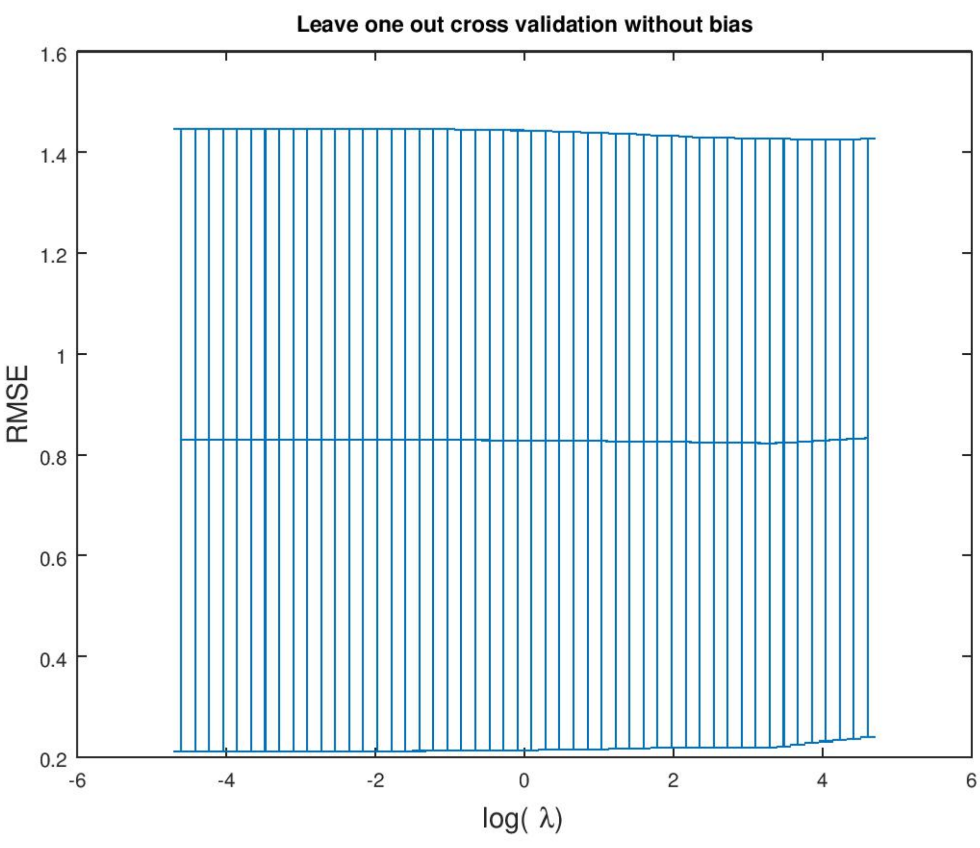
5 leave one out cross validation without bias term

In this section, we implemented the leave one out cross validation without bias term. The best λ that is achieved is: $\lambda = 26.827$
The best weight ω that is obtained for the best λ given by

$$\omega = \begin{bmatrix} 1.2797\text{e+}00 \\ \text{-}7.6108\text{e-}02 \\ \text{-}1.2808\text{e-}01 \\ 1.4077\text{e-}01 \\ \text{-}9.3155\text{e-}03 \\ \text{-}2.2900\text{e-}02 \\ 1.3094\text{e-}03 \\ 1.2094\text{e-}03 \\ \text{-}3.6815\text{e-}05 \\ \text{-}2.0749\text{e-}05 \end{bmatrix}$$

The root mean square error of test set is: 1.0897
The following Fig 5 is shown the $\ln(\lambda)$ vs RMSE errorbar plot.

Figure 4: The mean square error with RMSE errorbar plot for leave one out cross validation without bias term



As the Fig shown the variance is the same on all the λ. The root mean square error of test set is bigger than the leave one out cross validation with bias term.