CS 532 Project Update1

Topic: Machine learning approaches to predict forest fires

Name: Sicheng Fu

Email: sfu42@wisc.edu

For this project update, we mainly focus on the Data pro-processing and Linear regression approach to prediction.

Data Pro-processing

Our forest fire dataset combine 12 features (X location and Y location, month and day, 4 FWI components and 4 weather variables) and 1 labels (burned area) for 517 samples. Firstly, we need transform the nominal variables (month and day) to numeric variables first. Using each corresponding numbers to represent month and day.

And then assign the label is required. As observed from burned area, majority of fire burned area presenting a small size, which is less than 1 hectare. Based on forest fire level, related department usually classify 4 cases for fire damage: normal (burned area is smaller than 1 hectare), high (burned area is larger than 1 hectare but smaller than 100 hectare), very high (burned area is larger than 100 hectare but smaller than 1000 hectare) and extreme (burned area is larger than 1000 hectare). Consider our data is not involved many cases for extreme condition, so we just assign binary labels for our data, burned area <1 h as normal fire damaged level {y = -1} (it involved 243 samples) and burned area > 1 h as high fire damaged level {y = 1} (it involved 274 samples).

Next step is to split the testing data and training data, we assign 70% of data to training data and remaining 30% data for testing data, and save them in csv document.

Linear regression approach

We use four ways to do linear regression prediction (least square method, ridge regression, LASSO regression and Truncated SVD method). But based on the forest fire data and feature information, it can be expected that these feature will not strictly follow a linear distribution, and the linear model may not lead a good result. At first, we need normalize the data, all attributes should normalized to a zero mean and one standard deviation (or use min-max normalization).

We split the training data into 5 subsets and complete a 5-fold cross validation for each methods, find the average error rate and mean square error and best weighted vector, and utilize the best classifier to test.

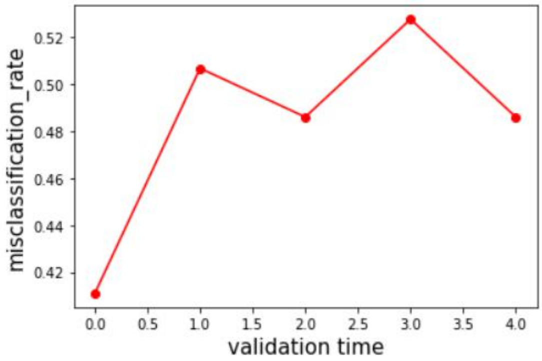
1) Least square method

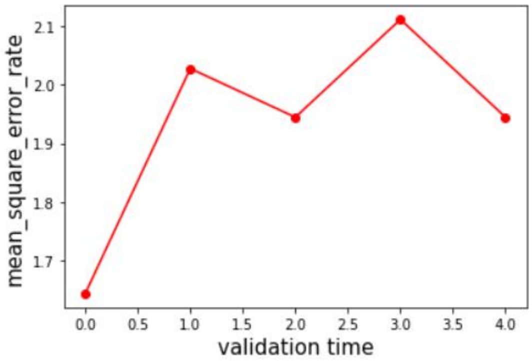
$$\arg\min_w = \| Xw - y \|_2^2$$

$$w = (X^T X)^{-1} X^T y$$

This least square solution is used to find the optimal weighted vector.

The plot of error rates and mean square error for each validations





Validation average error rate: 0.48
Validation Mean square error: 1.93
Testing error rate: 0.54
Testing mean square error: 2.14

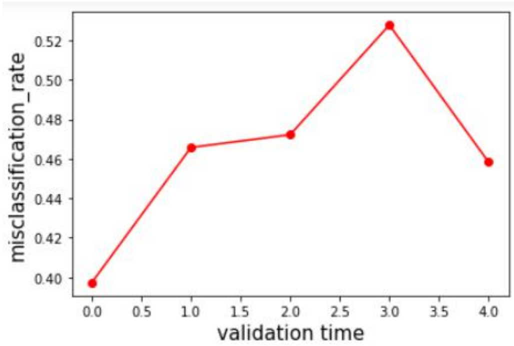
2) Ridge regression

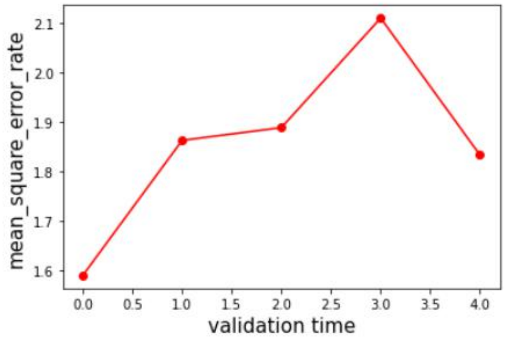
$$\arg\min_w =\| Xw - y \|_2^2 + \lambda \| w \|_2$$

$$w = (X^T X + \lambda I)^{-1} X^T y$$

This ridge regression solution is used to find the optimal weighted vector. In order to find the best λ value, we assign the values of λ range from 10^{-10} to 10, spaced logarithmically. Use the solution that has smallest error rate to find best λ， for each validation it will produce different best λ.

The plot of error rates and mean square error for each validations



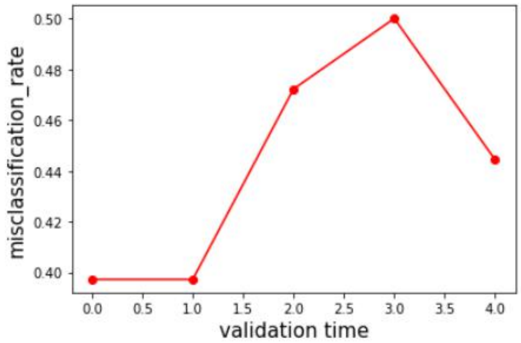


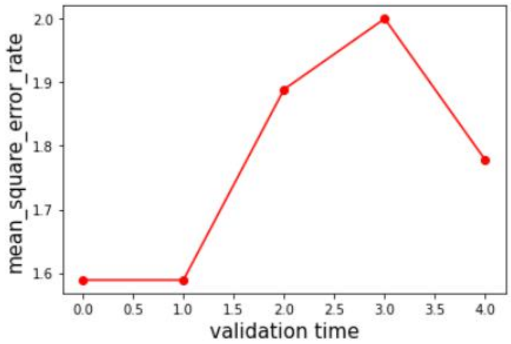
Validation average error rate: 0.46
Validation Mean square error: 1.86
Testing error rate: 0.52
Testing mean square error: 2.06

3) LASSO regression

$$\arg\min_w = \| Xw - y \|_2^2 + \lambda \| w \|_1$$

We implement iterative soft thresholding via proximal gradient descent to solve the LASSO problem. A hot start procedure is used to find solutions with different λ. λ also ranges from 10^{-10} to 10, spaced logarithmically.



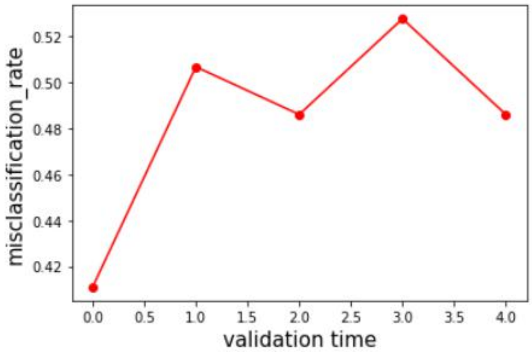


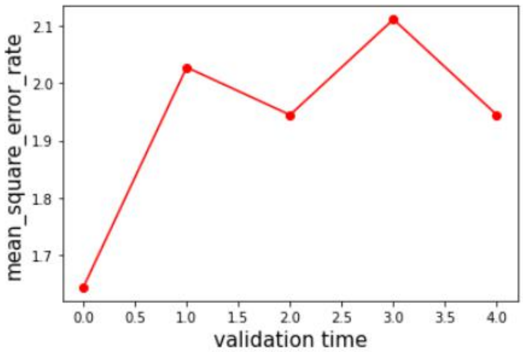
Validation average error rate: 0.44
Validation Mean square error: 1.77
Testing error rate: 0.53
Testing mean square error: 2.17

4) Truncated SVD

Singular vector decomposition: $X = U\sum V^T$,where $\sum$ is a diagonal matrix with singular values.

Truncated SVD, use the pseudo-inverse $w = V\sum^{-1} U^T y$ to estimate w for each choice of the regularization parameter r , $r = 1,2,...,12$, which means find the best rank-approximation.



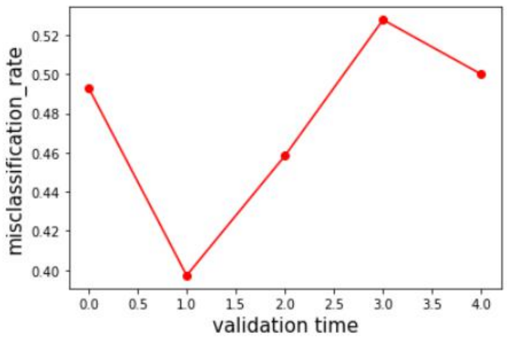


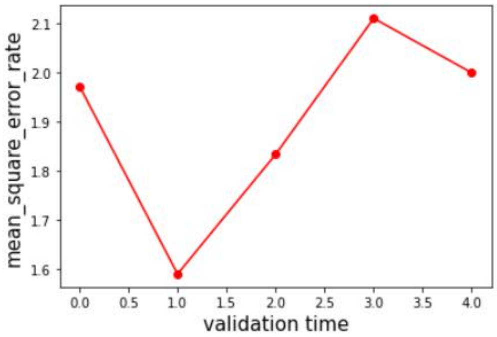
Validation average error rate: 0.48
Validation Mean square error: 1.93
Testing error rate: 0.52
Testing mean square error: 2.09

5) Feature selection

As we can see, these linear regression approach obviously did not present a good result. The fire burned area may do not have a simple linear relation with weather condition and FWI components. I select several more weighted features which may influence burned area to analyze again. Selected features are [“month”,”FFMC”,”DC”,”temp”,”RH”,”rain”].

The ridge regression results are shown below:





Validation average error rate: 0.47
Validation Mean square error: 1.90
Testing error rate: 0.46
Testing mean square error: 1.85

The result is better than before, but it is still not enough to satisfy a successful prediction.

Next step:
Linear regression is a classical model has been widely used, but it can only learn linear mappings. To eliminate this drawback, more nonlinear functions like NN and SVM should be considered. So next step, I will implement another two nonlinear method to do prediction. And feature selection analyse is required, since the FWI components and weather condition may not follow a strictly rule with the burned area.