# CS532 Project Update1

Topic: Asteroids Classification

## 1. Current Progress

*Data preprocessing*

The raw data has 4687 rows and 40 columns, which includes 39 features and 1 label. First, the columns with non-numeric values as well as the columns with a single common value are dropped. Then, a correlation matrix of the features is created. Most of the highly correlated features are dropped to reduce the dimension of the features. The number of the features is reduced to 21. In addition, the features are normalized, and the Boolean label is convert to the format of (+1)/(-1) as the indication of the sign. Last but not the least, 70% of the data is divided into the training set and the rest 30% of the data is divided into the testing set.

*Algorithms used and observations*

a. Linear regression (completed)

The linear regression algorithm is almost completed. The models are trained using the simple least-squares methods and the ridge regression. For the two methods, both the regular LS formula and the SVD formula are used. For the ridge regression methods, a k-fold cross-validation is used to find the optimal $\lambda$ by making a proper tradeoff between the bias and the variance. The code was posted on GitHub repository.

The following table shows the results of the least-squares methods. Overall, the method using the regular formula has a better performance.

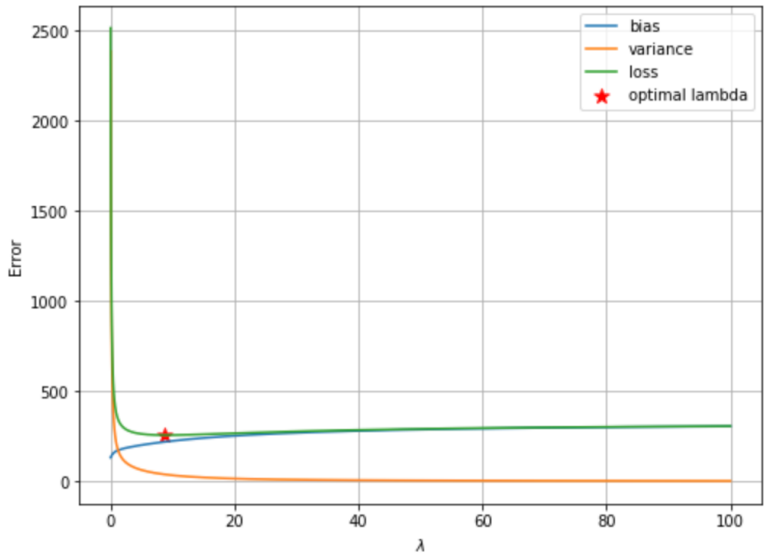
| Accuracy (%) | Least-squares | |
|---|---|---|
| | Regular formula | SVD |
| Training accuracy | 86.46 | 48.87 |
| Testing accuracy | 90.33 | 48.19 |

The following table shows the results of the ridge regression. The two methods have similar performance.

| | Ridge regression | |
|---|---|---|
| | Regular formula | SVD |
| Optimal loss | 255.4 | 287.96 |
| Optimal testing accuracy (%) | 82.53 | 82.50 |
| Optimal lambda | 8.6851 | 6.8665 |

To find the optimal value of $\lambda$, the k-fold cross-validation is used. The series of $\lambda$ is from 0.01 to 100 spaced uniformly on a log scale. For each value of $\lambda$, the training set is divided into k groups, where one of the k groups is the validation set and the rest groups are the training set to calculate the weight matrix $w$. A new model is trained with every different set of training set. The trade-off between bias and variance is applied. The model with the least validation loss is the best model. The value of $\lambda$ is the optimal value.

The following figure shows an example of the result of the ridge regression. The optimal value of $\lambda$ is picked based on the trade-off.




In addition, the LASSO regression is also used. The k-fold cross-validation is also used. The average testing accuracy is 90.68%, which represents a decent performance among all the linear classifiers trained.

b. SVM classifier (ongoing)
Since the course introducing the gradient descent for SVM was just finished, the SVM algorithm is still in development and will be completed in this week. The ideal method is the gradient descent.

c. Neural networks (ongoing)
The neural networks will be developed from next week and will be completed before the second update.

## 2. Next step

The sequence of the application of SVM and neural networks is switched since the courses introducing the SVM has just finished. So, the original plan is changed, and the planned progress is a little delayed. I plan to catch up the schedule in this week. The application of SVM models should be completed in this week. Starting from the next week, the neural networks should be designed and optimized. In addition, the completed linear regression models can be fine-tuned.

| Week | Tasks |
|---|---|
| 10/19 – 10/25 | ~~Data preprocessing; feature selection.~~ |
| 10/26 – 11/01 | ~~Build, train and test linear classifier.~~ |
| 11/02 – 11/08 | ~~Optimize linear classifier~~; ~~Build neural networks. (Changes to SVM)~~ |
| 11/09 – 11/15 | Train and test neural networks. (Changes to SVM) First update 11/17. |
| 11/16 – 11/22 | Optimize neural networks. Build SVM. (Changes to neural networks) |
| 11/23 – 11/29 | Train, test and optimize SVM. (Changes to neural networks) Second update 12/1. |
| 11/30 – 12/06 | Validation and evaluation. Visualization and documentation. |
| 12/07 – 12/13 | Write project report 12/12. |

**GitHub Link:** https://github.com/keshuw726/CS532-Course-Project-Asteroids-Classification