Logistic Regression for Mobile Phone Price Interval Prediction

The Logistic Regression is a popular algorithm in classification problems. It is also a linear classifier that can output probabilities for two or more classes which the instance should be classify. From the course slides (Classification2\_handout, Page 102), in a two-class classification problem, a formula about the probability P of an instance X and a solution weight w is given:

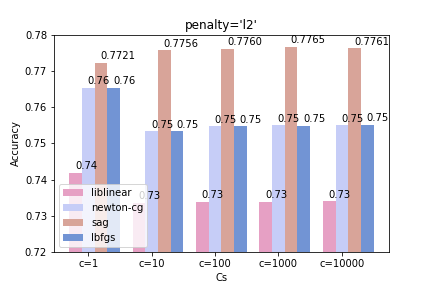
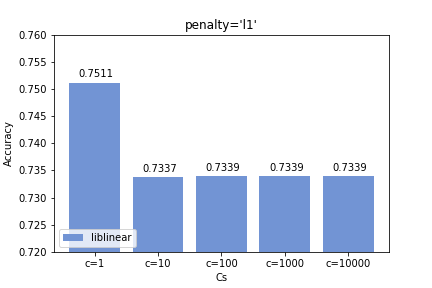
In the case of our model (Mobile Phone Price Interval and its price), the specifications of mobile phones are considered as features. We will use features of mobile phones to predict the price of them. According to the graph of sigmoid, we can get Logistic Regression can do well in classification (especially in two-class classification). Therefore, the continuous mobile phone prices are put into different intervals in our model. In this case, we have 3 classes for prices: [0, 300), [300, 700], and (700, +inf). Then we can use Logistic Regression Model to predict which the price interval should a mobile phone in. In the case of three-class classification problem, we can get the equation for probability P for price Y which is in class 1:

and the sum of P(Y=1|X), P(Y=2|X) and P(Y=3|X) is 1. This model would output the prediction class with a highest score(probability).

In this study, the above diagram shows the workflow and algorithm of our model. First, we do data preparing step. The program read a csv file and clean the data to drop rows with “null” value. Then we get 1035 complete rows with 16 different columns(features). By observing the head of features, the meaningless “key\_index” column is dropped from feature. Because the values of each features are sparse, we then use Standard Scaler to normalize these 15 features. We also do the price classify in this step which is necessary to build logistic regression model easily.

The second step is Running Logistic Regression model. We decide to try different parameters to get an optimal combination. So, we use Grid Search module to try different loss functions, normalization coefficients and solvers. In this step we can get a matrix of accuracy for each parameter combinations. Since “newton-cg”, ”sag” and” lbfgs” methods can only use loss function “l1”, there are totally 25 test accuracy in the matrix.

In the third step, we use bar graphs to compare the accuracy for different combinations. The test results are divided into two groups by different loss function. The left graph only contains the accuracies for “liblinear” method with different normalization coefficients C. The right graph compares all four methods for different C.



According to the results of grid search, we noticed that the “sag” method with penalty “l2” has high accuracy. The influences of C with value 10, 100, 1000 or 10000 is not significant. Therefore, we use “C=10000” to calculate the accuracy score, which is about 70.05%.

Conclusion:

Logistic Regression Model can deal classification better rather than value prediction. The class prediction accuracy has negative correlation with the number of class. Although we divided the price value into three classes, the accuracy score is only about 70%. More data pre-processing may improve the accuracy of Logistic Regression Model.

Future Works:

There are some features include continuous values, and it may increase the convergence of logistic regression. We could try to set intervals to divide these values and put them into different classes.