Student Number:231033

1. Introduction

Data are from photos which present whether or not they were taken in the sun. Training data have predict labels, 1 for sunny, 0 for not sunny and confidence label for each sample. They have done their feature extraction work where the task is to predict the sunny or not. There are three steps taken for classification process, in initial step, the data-processing is used on training data. Then in second step, the classifier was trained on training data and predicts the test data at end.

In this lab, I choose the logistic regression classification model to finish this binary classification task.

2. Methods

2.1. Logistic Regression Classification Model

Logistic regression (LR) is a commonly used probabilistic statistical classification model and often used in binary classification cause it is simple. In order to map predicted values to probabilities, The LR model uses the Sigmoid function[1] to squeeze the output of a linear equation between 0 and 1.

$$y = \frac{1}{1 + e^{-(w^T x + b)}} \tag{1}$$

Assuming a set of labeled samples(x1,y1),...,(xn,yn) obeys Bernoulli distribution, then use maximum likelihood estimation to estimate the parameters and use gradient descent to solve the parameters to achieve the purpose of binary classification. Therefore, the idea of logistic regression is to first fit the decision boundary (not limited to linear, but also polynomial), and then establish the relationship between the boundary and the probability of classification, thereby obtaining the probability in the case of two classifications. In order to map this to a discrete class (true/false), we select a threshold value or tipping point above which we will classify values into class 1 and below which we classify values into class 2.

The loss function is called the cross entropy[2] loss function. Entropy represents the uncertainty of an event, and cross entropy represents the difference between prediction and actual probability distributions. The accuracy of the logistic regression classifier can be maximized by minimizing the cross-entropy loss function.

$$Loss = -\frac{1}{N}\sum[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] \tag{2}$$

Key parameter: penalty is regularization, it is important in logistic regression modeling, Without regularization, logistic regression would keep driving loss towards 0 in high dimensions and suffer immensely from the curse of dimensionality. It generally adopts L1 L2 regularization.

To minimize our cost, LR use Gradient Descent[3] which use the loss function Loss to find the first-order partial derivative of the parameter w to determine the direction, and determine the step size to update w, Stop until abstract of Loss(wk+1)-Loss(wk) is less than a certain threshold or reaches the maximum number of iterations.

$$w_i^{k+1} = w_i^k - \alpha \frac{\partial Loss}{\partial w_i} \tag{3}$$

Input the dataset, find an optimal parameter vector to minimize Loss function and put it into function[4], according to a threshold, to get the classification of the sample. If model is working, cost decrease after every iteration.

$$f(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{4}$$

2.2. Pre-processing data

According to Guyon and Elisseeff (2013) , feature selection is the process of selecting a subset of relevant features for use in model construction , thereby improving the prediction performance of the predictor. In this task, Data have confidence labels, in order to get a more accurate prediction, the low confidence labelled data were deleted. Then I calculate the mean of features importance with data from CNN and GIST, scores can provide insight into the dataset. The result is importance of CNN is 0.00024 and GIST score is 0.00196 which shows GIST features is a bit important than CNN features.

Rescale is an important step in pre-processing, many machine learning algorithms perform better when digital input variables are scaled to the standard range. Witten(2015) shown one way which is dividing all values by the maximum value encountered or by subtracting the minimum value, then dividing by the range between the maximum and minimum values called Min Max Scaler.

Dataset could have missing data and Witten(2015) described the possible reason such as faulty measuring instruments or changes in experimental design during data collection. These values may be represented in a variety of ways, including an empty string, the explicit string NULL or unknown or N/A or NaN, and the number 0, among others. (McCallum,2012) Therefore, data imputation approach entails using approximate the value of a column from the current values and then replacing all missing values in the column with the estimated statistics. It computes statistics quickly and is normally very accurate. The scikit-learn machine learning library provides the Simple Imputer method with the mean strategy was used in lab.

Last step in processing is dimensionality reduction. Guyon (2013) emphasizes it can produce a more interpretable representation of the target concept and focusing on the most relevant variables. The most popular technique for dimensionality reduction in machine learning is Principal Component Analysis(PCA). It can be thought of as a projection method. Through evaluating the score with different Number of components to keep in PCA, the best result can be determined.

2.3. Select, train and test model

Take the dataset and split it into a training dataset and a test dataset. I used Repeated Stratified K-Fold method which Stratified K-Fold is repeated n times with separate randomization of each iteration.

Then evaluate the top 7 algorithms with Spot-checking algorithms for training dataset. Choose best four models and do algorithm tuning for them to improve the performance of model. It is best to use grid search and pipeline cause this method is simple to apply, does not rely on randomness, and covers the entire space. however, it employs a large number of points.(Kochnderfer and Wheeler, 2019)

After searched the hyper-parameters of classifiers, use the best classifier to predict the test dataset and output the prediction results.

3. Results

Based on preliminary investigation results [Figure 1], I selected Naive Bayes Classifier, Logistic Regression Classifier and Multi-layer Perceptron Classifier to update their hyper-parameters.

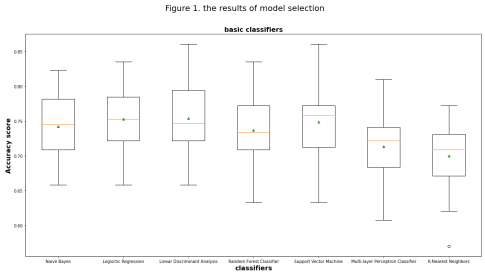

Figure 1. results of model

In second exploration, I focused on LR. The reason for it is I found LR performance well on binary classification problems and the result confirm my assume[Figure 2].Besides that I also test how different hyper-parameters affect performance of the classifier.[Figure3]

The accuracy of final LR model is 0.751163, but I wonder to know if there has better performance on other model such as AdaBoost classifiers with Ensemble Algorithms. However compare to their learning curves, I found AdaBoost overfits obviously.[Figure 4, Figure 5]

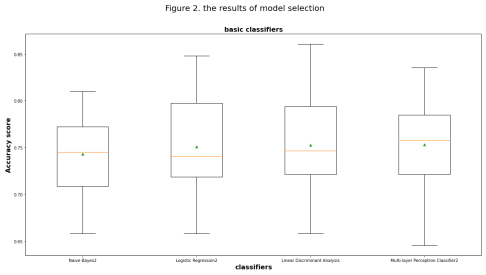

Figure 2. results of model selection

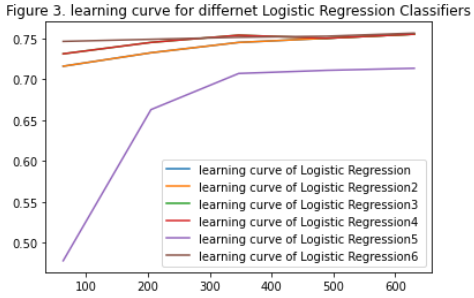

Figure 3. learning curve for different Logistic Regression model

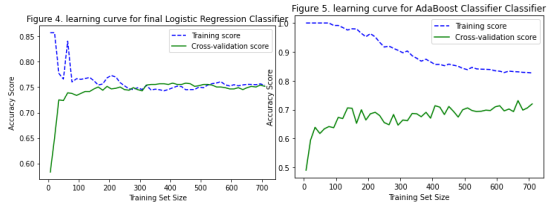

Figure 4. learning curve for Logistic Regression model and Adaboost model

At last, I conducted a more in-depth comparison of various training sets. Accuracy score of test data without low confidence is 0.751163 and score of test data include low confidence is 0.750741. It seems the impact of whether to consider the training label confidence is minimal. But I suppose this might have an effect on the predictions.

4. Discuss

There might be ways of getting better performance. Get more higher-quality data is a good idea. Also, there should have a better method to deal with the confidence label instead of delete the low confidence labeled data directly.

In this task, I have learnt how to pre-processing data and decide the best suitable model for data with grid search and pipeline method.

References

[1] Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), pp.1157-1182.

[2] Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2005. Practical machine learning tools and techniques. Morgan Kaufmann, p.578.

[3] McCallum, Q.E., 2012. Bad data handbook: cleaning up the data so you can get back to work. " O'Reilly Media, Inc.".

[4] Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y. and Zhou, Z.H., 2008. Top 10 algorithms in data mining. Knowledge and information systems, 14(1), pp.1-37.

[5] Kochnderfer, M. and Wheeler, T., 2019. Algorithms for optimization. Cambridge, Massachusetts: The MIT Press. 2019. Algorithms for optimization. Cambridge, Massachusetts: The MIT Press.