6103 Project：Real or Not, NLP with Disaster Tweets

Weifei Wang

The George Washington University

Introduction

Twitter has played an important role in social media life. Nowadays, people like to post Tweets to share things and many people get information from Tweets, like news, weather. For example, some people may post a Tweet when they see disasters to alarm other people. However, not every Tweet is about real disasters, sometimes people use words like “ablaze” metaphorically to describe the beautiful view.

To help people avoid some unnecessary harm from this ‘disaster’ information, we found a dataset from Kaggle which is related to this topic. What’s more, we attempt to use this dataset to build some machine learning models that predict which Tweets are about real disasters and which one’s aren’t.

The data is from Kaggle, and here is the link:

https://www.kaggle.com/c/nlp-getting-started/overview

There are three csv file: train, test and sample submission. For the train set, there are 7613 rows by 5 columns. each column represents id - a unique identifier for each tweet, keyword- a particular keyword from the tweet (may be blank), location - the location the tweet was sent from (may be blank), text- the text of the tweet and target-this denotes whether a tweet is about a real disaster (1) or not (0). For the test set, there are 3264 rows by 4 columns, which contains id, keyword location and text, with no target.

Since the target are only 1 and 0, it’s a binary classification problem. Our goal is using the train set to train some suitable models and predicting the test texts by our models.

Experimental setup

1. Decision Tree

Decision tree is a prediction method, covering both classification and regression. It uses a tree-like model to make a prediction like its name. Each branch represents the result of the test, and each leaf node shows the class label. Therefore, a competed decision tree can explain how the prediction come.

The first step is to read data. The preprocessed data are read and by using the function of pandas. By checking the data, we notice that there is a column named ‘Unnamed: 0’, which is the index in the csv file, so we use ‘.drop()’ function to delete this column. After that, we defined the last column as ‘labels’ to show the result about whether the disaster is real or not and ‘1’ represents ‘real disaster’ while ‘0’ represents not. And the rest columns became the features. The second step is to import DecisionTressClassifier from sklearn.tree to model the data. After modeling, we got a result csv file. We submitted this result to Kaggle and got score of 0.69325(Fig.1).

Figure 1. Result before Modifying Parameters of Decision Tree.

The next step is to modify the parameters of decision tree to improve the result score. We made two lists of max depth and max leaf nodes to find the numbers with better accuracy by using function of KFold. Max depth is the maximum depth of the tree, which means that no matter how many features can be branched when the depth of the tree reaches max_depth, the decision tree will stop computing. Max leaf nodes represent the maximum number of leaf nodes. Unlimited when it is None. KFold is a cross validation method which can evaluate machine learning models and has a single parameter called k that can divide given data into k groups. As the figure2 shows, we chose max depth = 250 and max leaf nodes = 100 as the final parameters. After modified the parameters, we submitted the result to Kaggle again.

```
max depth=100, average accuracy=0.6950845396143638
max depth=150, average accuracy=0.6948873298861176
max depth=200, average accuracy=0.6951504151988102
max depth=250, average accuracy=0.696835613597779
max depth=300, average accuracy=0.6945150415198811
max depth=None, average accuracy=nan
max leaf nodes=100, average accuracy=0.7126872981922107
max leaf nodes=150, average accuracy=0.7108689681274909
max leaf nodes=200, average accuracy=0.6978214033958497
max leaf nodes=250, average accuracy=0.6914715636341463
max leaf nodes=300, average accuracy=0.6854736211496718
max leaf nodes=None, average accuracy=nan
```

Figure 2. Results of the Modified Parameters

2. Support Vector Machine

A support vector machine is a machine learning model which uses classification algorithms to solve the problems of two-group classification.

$$\max \frac{|w^T x + b|}{||w||}$$

Figure 3. The equation of SVM.

We import SVC to model the data and we got scores of 0.62167.

Figure 3. Result before Modifying Parameters of SVM.

And then we modified parameter of 'C' and 'gamma'.

```
best parameter for SVC {'C': 1.15, 'gamma': 'scale'}
best score for SVC 0.8024459815893488
```

Figure 4. Results of the Modified Parameters

Results

1. Decision Tree

1.1 Using all train data to train the model

By submitting the csv file to Kaggle, it returns the score and the screen shoot of it is shown as Figure 5:

Figure 5. Kaggle Accuracy Score of Decision Tree Model

The accuracy of the model is 0.74948, which is relatively low. Therefore, decision tree is not suitable for this data.

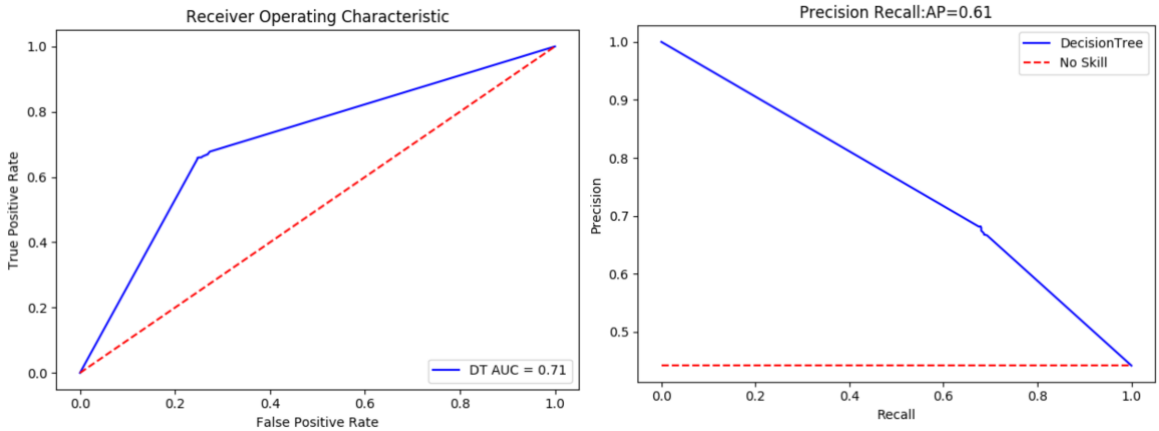

Figure 6. ROC Curve and P-R Curve of Decision Tree Model

ROC Curve (the left figure in Figure 6) and P-R Curve (the right figure in Figure 6) are both diagnostic tools that help in the interpretation of probabilistic forecast for binary classification predictive modeling problems.

ROC Curves summarize the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds, while Precision-Recall curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds.

AUC is the area under the curve. Here, AUC of the ROC curve is 0.71 And values on the y-axis of the plot is large, indicating a high true positives and low false negatives. And values on the x-axis of the plot is small, thus it has a low false positives and high true negatives. This is same as confusion matrix showing.

2 Support Vector Machine

2.1 Using all train data to train the model

By submitting the SVM test csv file to Kaggle, it returns the score and the screen shoot of it is shown as Figure 7:

Figure 7. Kaggle Accuracy Score of Support Vector Machine Model

The accuracy of SVM model is 0.81595 which means SVM performs well.

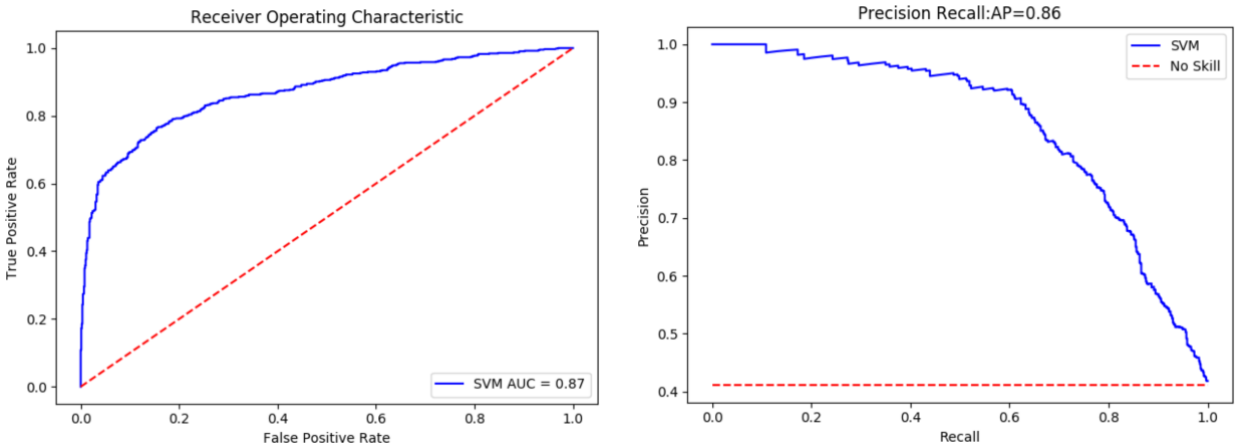

Figure 8. ROC Curve and P-R Curve of SVM Model

Here AUC of ROC curve (the left image of Figure 8) is 0.87. The larger the AUC value, the more likely the classification algorithm is to rank positive samples in front of negative samples, which means a better classification. Therefore, we can conclude that support vector machine is a better classification compared with decision tree in this case.

Summary and conclusion

After modified parameters, the accuracy of decision tree improved from 0.69325 to 0.74948 while support vector machine increased from 0.62167 to 0.81595. In conclusion, support vector machine performs better than decision tree in this case.

7. Reference

1. Brownlee, J. (2016, November 9). How To Implement The Decision Tree Algorithm From Scratch In Python. Retrieved from https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/
2. Brownlee, J. (2018, May 23). A Gentle Introduction to k-fold Cross-Validation. Retrieved from https://machinelearningmastery.com/k-fold-cross-validation/
3. Brownlee, J. (2020, January 14). ROC Curves and Precision-Recall Curves for Imbalanced Classification. Retrieved from https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/
4. Brownlee, J. (2016, April 20). Support Vector Machines for Machine Learning. Retrieved from https://machinelearningmastery.com/support-vector-machines-for-machine-learning/
5. Fawcett, Tom. “An Introduction to ROC Analysis.” *Pattern Recognition Letters*, vol. 27, no. 8, 2006, pp. 861–874., doi:10.1016/j.patrec.2005.10.010.
6. Gandhi, Rohith. “Naive Bayes Classifier.” *Medium*, Towards Data Science, 17 May 2018, towardsdatascience.com/naive-bayes-classifier-81d512f50a7c.
7. Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *Plos One*, *10*(3). doi: 10.1371/journal.pone.0118432
8. Yiu, Tony. “Understanding Random Forest.” *Medium*, Towards Data Science, 14 Aug. 2019, towardsdatascience.com/understanding-random-forest-58381e0602d2.
9. ZHOU, Z. H. I.-H. U. A. (2020). *Machine Learning*. S.l.: SPRINGER VERLAG, SINGAPOR.