Social Media Analytics

Final Project Report

1. Introduction

Nowadays, we can use machine learning to analyze the social media in order to find some useful information, the airline companies can use this method to find out what they should do to improve the customers' satisfaction and why the customers choose them.

In final project, we need use machine learning to classify the Tweets, find the noncompliant Tweets and check the precision of our model.

2. Model building and model selection

Before building the model, first, I clean the data and split it into training data (80%) and validation data (20%). I use '1' to represent the complaint Tweets and use '0' to represent the noncompliant Tweets, this is really helpful for my following steps. Then, using training data to build the model and validation data to check the accuracy and select the model.

After data cleaning and data splitting, I start to build the model. I tried six different models and choose the one that has the highest accuracy and without severe over-fitting problem. These are my models: SVM (linear kernel), SVM (polynomial kernel), Naïve Bayesion, Maximum Entropy, Random Forest and KNN.

1). SVM (linear kernel)

SVM is a powerful model to deal with the classification, since the output of this mode l is numeric numbers, I have the following rules, if the prediction number is larger tha n 0.1, it represents complaint Tweets, otherwise, it represents noncompliant Tweets. T he accuracy of this model is 0.726.

2). SVM (polynomial kernel)

Then I change the SVM kernel to 'polynomial' and validate the accuracy, the accurac y is dropped down to 0.706

3). Naïve Bayesion

This method is used widely in social media analytics, so I select it as one of my model s. Since this model can only predict factor variables, I need change ‘y’ into factor vari able first. The accuracy of this model is 0.671, which is worse than the SVM models.

4)Maximum Entropy

The principle of Maximum Entropy is to select the statistical properties of random var iables that best fit the objective situation. Because we have hundreds of features, this model may select the most useful feature and better predict our data. The accuracy of t his model is 0.697, which is better than the Naïve Bayesion.

5). Random Forest

Random Forest is a useful way to predict some digital value, the accuracy of this mod el is 0.692

6) KNN

KNN is the simplest way to do classification in machine learning, it’s classified by me asuring the distance between different observations, I set the K to 7 and distance to 1, the accuracy of this model is really high, which reaches 0.859, even though the accura cy is really high, I did not choose this model at last, because I found this model have s evere over-fitting problem, which cannot predict the test data correctly. Furthermore, t his model is not stable, if I change my training dataset, the accuracy of this model will change dramatically.

3. Word frequency

After building these six models, I choose SVM (linear kernel) as my final model, I us e it to predict my test data and I get 688 “noncompliant” Tweets dataset, in order to m ake my prediction more precisive, I use the word frequency to delete some Tweets fro m my dataset. I compared the complaint Tweet word frequency to the noncompliant T weet word frequency, “fuck|shit|worst|bad|hate|awful|wtf|worse|disappoint|frustrat|rud e|never|poor|sad|disgust|bother|hell|suck|stuck|unacceptable|upset|ruin|shame|unprofes sional|complaint|nothappy|kill|waste|screw” are usually appeared in the complaint Tw eets and do not appear in the noncompliant Tweets, so I delete them from my 688 “no ncompliant” Tweets dataset and I finally get 303 observations.

4. Calculate the accuracy of test data

After getting my 303 observations “noncompliant” dataset, I need evaluate whether th e classification is correct manually. In my data, 209 of the observations are exactly no ncompliant Tweets and 94 of them are complaint Tweets, the precision is 0.689769.