Machine learning algorithm –Part1:

1. Use spark.mllib package to build different machine learning models including logistic regression, decision trees, random forests, naive Bayes to predict tweets sentiments
2. Preprocess the class label of original data set, which is 3 class labels including positive, neutral and negative and compare the accuracy of different models.
3. Again preprocess the class label of original data set, which is 12 class labels this time including positive, neutral and 10 different kinds of negative reasons. Compare the accuracy of different models and the accuracy of previous part.

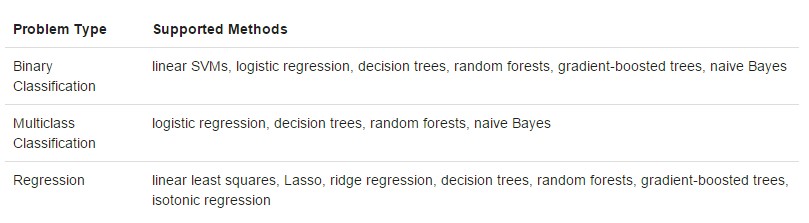
Data preprocess:

logistic regression, decision trees, random forests this three use the libsvmfile to read data, so we use a python file to preprocess the input matrix and get the corresponding files to build model.

Technique:

In the previous part, we have preprocess the data set and build a matrix of word bag for each tweets with class label, so in the Machine learning stage, we could directly use the matrix(in csv file) to train the models

The spark.mllib package supports various methods for binary classification, multiclass classification, and regression analysis.



For our tweets sentiments classification problem, which is a multiclass classification problem. So we use the 4 supported methods mentioned above to build our model. And we will explain it in detail in the next part.

Model:

Logistic regression:

Logistic regression using a logistic function to determine the relationship between the variables by estimating probabilities. And for Our particular case, which is the multiclass classification. So we need to generalize the binary logistic regression into multinomial logistic regression. For example, for N possible classes, one of the classes can be chosen first and the other N−1 classes can be separately regressed against the first class.

Decision trees:

Decision tree is a greedy algorithm which runs a binary partitioning among features recursively. Each partition greedily select among all the possible splits, by this way we could maximize the information gain so that at every tree node we could get the best split to construct the tree. And the recursive tree algorithm will stop at a node when it meets specific conditions.

Random forests:

Random forests are ensembles of decision trees algorithm. They combine many decision trees to form a forest in order to gain accuracy while reduce the risk of overfitting.

Naïve Bayes:

Naïve Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

Accuracy:

|  |  |  |
| --- | --- | --- |
|  | 3 class | 12 class |
| Decision trees | 0.6207 | 0.4143 |
| Naïve Bayes | 0.6550 | 0.3785 |
| Random forests | 0.6332 | 0.2869 |
| Logistic regression | 0.5252 | 0.4993 |

Conclusion:

As can be seen in the accuracy, our model has a significant better result in 3 class labels than that of 12 class labels. Which is also a common situation, since the probability of random guess for 3 class label data is 0.34, while the probability of random guess for 12 class label data is 0.125. It’s hard to get good accuracy for 12 class label data. Because, for the 3 class label situation, it generally has a huge different between tweets which have different sentiments, and it is much easier for both human and computer to label them properly. But for 12 class label situation, even human will have a hard time to label different tweets, especially to classify the reason of negative sentiments. Since a negative Tweets may have complained 2 reasons at the same time. So in the original data set this is a feature named airline sentiment confidence and negative reason confidence

, people who created the files cannot perfectly label them. So it is extremely easy for a machine learning algorithm to mismatch negative reasons for an ambiguity tweet. That’s why we got a low accuracy towards 12 class label situation.