Predicting Rating Star Based on Sentiment Analysis of Yelp Review Text

Jiao Chen
Computer Science and Engineering
chenjiao@msu.edu

Shaohua Yang
Computer Science and Engineering
yangshao@msu.edu

Abstract

In our project, we apply multi-class SVM, Loggistic Regression and Neural network to classify the image samples that belong to 164 classes. We use machine learning tool Vowpal Wabbit (VW) to perform these three classification methods. VW is a online learning based tool, which provides reasonable classification accuracy and fast training and testing speed on large dataset. We tune the hyperparameters of these three classification models on part of our training dataset, and then train three classification models based on the whole training dataset. Overall, logistic regression achives an accuracy of 37.46% on our testing data with 250,000 samples, while the accuracy of SVM is only 30.45% on the same testing dataset. Neural network is much slower than SVM and logistic regression, but provides the highest accuracy (39.30%) and MAP on our testing dataset. For evaluation, results of these three models are assembled together with equal weight. We get a evaluation MAP score of 72.81%, which ranks first among 18 teams in this class.

1 Our Approach

1.1 Feature extraction

After we downloaded the Yelp reviews dataset, first we preprocess the dataset before extracting features. The original dataset is in json format, we use python *json* module to extract review texts and their corresponding rating stars. All the words in review text are converted to lower case. Stop words are elimitated from the reviews by *nltk.corpus.stopwords* and punctuations are also deleted by using regular expression. After preprocessing, we generate bag of words and bag of pharases feature vectors for each review text. The number of each unigram, bigram, trigram in a review text are counted and saved in the corresponding vectors.

1.1.1 bag of words

Bag of words (unigram) models are widely used in sentiment analysis of texts. Here we generate the unigram vocabulary from all the 1,600,000 review texts, resulting a dictionary with 414,197 unique words. However, most words in this dictionary are not related to sentiment and of low frequency. To improve the efficientcy of downstream analysis, we only keep the top 20,000 most frequenct words as candidate features. To make sure that sentiment related words will be in our features, we downloaded a sentiment vocabulary from previous work(**?**), which includes 4,783 negative words, 2,006 positive words. So the total length of our unigram feature vector will be 26,789.

1.1.2 bag of pharases

In bag of pharases model (n-gram), we generate bigram and trigram features. Here for the size of our review texts, trigram feature vectors have already been very sparse, so we did not generate n-grams with n¿3 features. For bigram features, the total number of unique bigrams from the rviews is 17,721,951. Noticing that a lot of bigrams occure only once in the reviews, we only keep bigrams occuring at least four times as our features. Thus, the bigram feature vector size is reduced to 2,858,126. For trigram features, the total number of unique tri-

grams is 68,126,384. Similar to bigram, we also only keep trigrams which occure at least four times as features. In this way, the trigram feature vector size is reduced to 2,145,801.

1.2 Classification models

Our goal is to predict the star rating (1, 2, 3, 4, 5) from the review text based on sentiment analysis. If we treat the star ratings as class labels, the problem can be treated as a mlti-class classification problem. We use Vowpal Wabbit (vw) (http://hunch.net/ vw/) machine learning tool to perform Logistic Regression, Support Vector Machine (SVM) and Neural Network classification methods on our datasets. VW is a fast online learning tool which was started at Yahoo! Research and continuing at Microsoft Research. Its default learning algorithm is a variant of online gradient descent.

In vw, for different classification models, the loss function can be written in the uniform way:

$$\sum_i (L(\mathbf{x}_i, y_i, \mathbf{w}) + \lambda_1 ||\mathbf{w}||_1 + 1/2 \cdot \lambda_2 ||\mathbf{w}||_2^2 \quad (1)$$

$\mathbf{x}$ is the feature vector and $\mathbf{w}$ is the parameter vector. λ_1 and λ_2 are the coefficients to specify the level of L1 and L2 regularization, respectively. Different classification models are identified by the loss function type. For example, the loss function for Logistic Regression is logistic loss: $ln(1 + exp(-y\mathbf{w}^T\mathbf{x}))$, while loss function for SVM is hinge loss: $max(0, 1 - y\mathbf{w}^T\mathbf{x})$. Another important parameter for vw is the learning rate λ. Let y_t be the ground truth class label for tth smaple, and $\hat{y_t}$ be the predicted class label. In online learning algorithm, the classification vector $\mathbf{w}$ is updated in the way: If $y_t \neq \hat{y_t}$, then

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \lambda y_t \mathbf{x}_t \quad (2)$$

Neural network is an interconnected group of nodes (neurons) which can compute values from inputs. A set of input neurons can be activated by the input data, the activations of these neurons are then passed to the other neurons. Repeat this process until finally an output neuron is activated. Here, we use neural network with 10 hidden layers for classifying the review samples.

1.3 Regression model

Considering that their is order between rating stars, that is 5>4>3>2>1. So rating 1 and 2 is more close than rating 1 and 5. We use linear regression to perform ordinal classification on our dataset. In general, for each sample feature vector $\mathbf{w}$, it is mapped to a real value $y(\mathbf{x}, \mathbf{w})$:

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\mathbf{x} \quad (3)$$

And $\mathbf{w}$ is learned by minimizing the loss function for linear regression:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2} \sum_{n=1}^{N} \left\{ y_n - \mathbf{w}^T\mathbf{x}_n \right\}^2, \quad (4)$$

The value $y(\mathbf{x}, \mathbf{w})$ will be converted to an integer between 1 to 5 to represent for the star ratings.

2 Results