Predicting the Yelp Star Rating Based on Text Analysis of User Reviews

Junyi Wang
Department of Statistics, Stanford University

Task Definition

We apply machine learning algorithms to predict the Yelp star rating based on the user review.
Data Source: Yelp Dataset Challenge
Positive feedback: a star rating of 4 or 5
Negative feedback: a star rating of 1, 2 or 3
Input: some text of user review
Output: +1 indicating a positive feedback and -1 a negative feedback

Example:
Input: {"What a cool bar restaurant.. I will no doubt be visiting again. The service and prices were great.. and the restrooms were clean. I had the buffalo chicken sandwich and it was delicious. The menu consists of typical bar food, however; there's a few different items on there which stand out on the menu. A cool bar off the beaten path that is a worth a trip. Cheers ! !"}
Output: {"sentiment": +1}

Training set size: 600
Test set size: 400

Error Metric

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

TP: Number of true positives
FP: Number of false positives
FN: Number of false negatives

Model

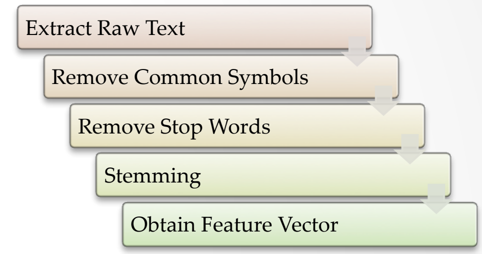
1. Perceptron Learning Algorithm

Run stochastic gradient descent on weight vector until convergence

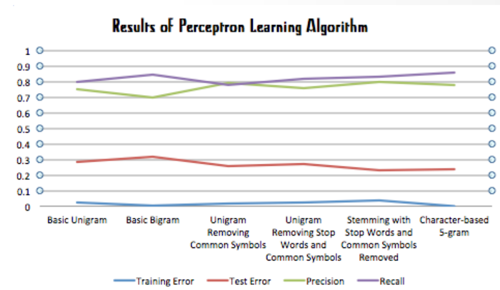
$$g(\theta^T x) = \begin{cases} 1 & \text{if } \theta^T x \geq 0 \\ -1 & \text{if } \theta^T x < 0 \end{cases}$$

$$\theta \leftarrow \theta - \eta(\nabla_\theta \max\{0, 1 - \theta \cdot \phi(x) y\})$$

x: a vector of features
θ: corresponding weight vector.



Language Model	Training Error	Test Error	Precision	Recall
Basic Unigram	0.023	0.282	0.749	0.799
Basic Bigram	0.005	0.318	0.697	0.845
Unigram Removing Common Symbols	0.018	0.260	0.788	0.778
Unigram Removing Stop Words and Common Symbols	0.022	0.268	0.759	0.816
Stemming with Stop Words and Common Symbols Removed	0.040	0.228	0.799	0.833
Character-based 5-gram	0.000	0.237	0.774	0.858



2. Naïve Bayes

$$p(y=1|x) = \frac{(\prod_{i=1}^n p(x_i|y=1))p(y=1)}{(\prod_{i=1}^n p(x_i|y=1))p(y=1) + (\prod_{i=1}^n p(x_i|y=-1))p(y=-1)}$$

Language Model	Training Error	Test Error	Precision	Recall
Basic Unigram	0.092	0.460	0.908	0.262
Basic Bigram	0.002	0.408	0.882	0.372
Unigram Removing Common Symbols	0.132	0.454	0.920	0.269
Unigram Removing Stop Words and Common Symbols	0.072	0.404	0.896	0.372
Stemming with Stop Words and Common Symbols Removed	0.116	0.424	0.868	0.349
Character-based 5-gram	0.006	0.464	0.926	0.249