CS59000: Machine Learning for Natural Language Processing

Homework 1

October 06, 2017

1 Task

Given the data consisting of reviews and corresponding sentiments (binary labels - good or bad), the objective is to build a classifier to automatically identify the sentiment of new reviews. Perceptron and a Multi-Layer Perceptron algorithms are implemented.

1.1 Features

Bag of words (BoW) with unigrams are used as features. The following procedure is followed for feature extraction:

1. Given the list of reviews in training, words that are observed and their frequencies are computed.
2. The top 100 words are ignored assuming that the articles, prepositions, etc bound to have higher frequencies. Next 2000 words are considered as feature words
3. Each review is represented in 2000 dimensional feature space. Each element indicating presence (1) or absence (0) of a feature word in the review.

1.2 MLP Configuration

1. λ (Regularization parameter) = 0.03
2. α (learning rate) = 0.01
3. Stopping criteria : max_iterations = 1000 ; min_thresh b/w weights = $2e-4$
4. No. of features = 2000
5. No. of hidden layers = 1 ; No. of nodes in the hidden layer = 10

1.3 Perceptron Configuration

1. α (learning rate) = 0.01
2. Stopping criteria : max_iterations = 1000 ; min_thresh b/w weights = $2e-4$
3. No. of features = 2000

1.4 Data Splits and Results

Training - 80% . Validation - 10% (For tuning hyper parameters such as regularization rate and learning rate.) . Testing - 10% (To verify the model on unseen data)

1. MLP: For 2000 features: F_1 score = 0.67
2. Perceptron: For 2000 features: F_1 score = 0.65

2 Comments

1. Training time and preprocessing time was drastically increased when attempted to use a bigram model. Hence I was restricted to unigram model.
2. It was noted that the, increase in the number of features did not lead to good results. The feature vectors are highly sparse, often containing only one or two ones meaning that only few words that are selected as feature words are present in the corresponding review.
3. When 2000, feature words are used, it was found that on an average only four elements that are one out of 2000 elements on an average.