Assignment 1

CSE 447 and 517: Natural Language Processing – University of Washington

Winter 2021 – Due: January 13, 2021, 11:59 pm

1 CSE 447 and CSE 517 Students: Based on Eisenstein 4.6 (p. 89)

Download the Pang and Lee movie review data, currently available from `http://www.cs.cornell.edu/people/pabo/movie-review-data/`. Hold out a randomly selected 400 reviews as a test set.

Sentiment lexicon-based classifier. Download a sentiment lexicon, such as the one currently available from Bing Liu at `https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon`. Tokenize the data, and classify each document as positive if and only if it has more positive sentiment words than negative sentiment words. Compute the accuracy and F_1 score (on detecting positive reviews) on the test set, using this lexicon-based classifier.

Logistic regression classifier. Train a (binary) logistic regression classifier on your training set using features of your own choosing, and compute its accuracy and F_1 score (as above) on the test set. Do not use an existing implementation of logistic regression, stochastic gradient descent, or automatic differentiation.

Statistical significance (extra credit). Determine whether the differences in accuracy and F_1 score are statistically significant, using two-tailed hypothesis tests: binomial for the difference in accuracy and bootstarp for the difference in macro F_1 score.

2 CSE 517 Students: Eisenstein 2.5 (p. 44)

Suppose you are given two labeled datasets D_1 and D_2, with the same features and labels.

- Let $\boldsymbol{\theta}^{(1)}$ be the unregularized logistic regression (LR) coefficients from training on dataset D_1.
- Let $\boldsymbol{\theta}^{(2)}$ be the unregularized LR coefficients (same model) from training on dataset D_2.
- Let $\boldsymbol{\theta}^{*}$ be the unregularized LR coefficients from training on the combined dataset $D_1 \cup D_2$.

Under these conditions, prove that for any feature j,

$$\theta_j^* \geq \min\left(\theta_j^{(1)}, \theta_j^{(2)}\right)$$
$$\theta_j^* \leq \max\left(\theta_j^{(1)}, \theta_j^{(2)}\right).$$

3 CSE 517 Students: Eisenstein 2.6 (p. 44)

Let $\hat{\boldsymbol{\theta}}$ be the solution to an unregularized LR problem, and let $\boldsymbol{\theta}^*$ be the solution to the same problem, with ℓ_2 regularization. Prove that $\|\boldsymbol{\theta}^*\|_2^2 \leq \|\hat{\boldsymbol{\theta}}\|_2^2$.