Assignment 1

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

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

Submit: You will submit your writeup (a pdf) and your code (do not include data) via Gradescope. Instructions can be found here. Note that you will make two submissions: one for the pdf, one for the code.

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 and report 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 report its accuracy and F_1 score (as above) on the test set. Do not use pretrained word vectors or any features implemented or constructed by anyone else. 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 bootstrap for the difference in macro F_1 score. Report the results.

Important note: You should implement all parts of this problem from scratch (you may use `numpy`). Do not use existing implementations for text tokenization, feature construction, logistic regression, stochastic gradient descent, automatic differentiation, or statistical significance testing. In general, it's a good idea to use existing, trusted implementations, but in this assignment we want you to experience attempting them on your own, even if your implementation is not the best in the world, so that you will learn more. If you aren't sure about whether it's okay to import a particular library, please **ask** on the discussion board.

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 $\|\boldsymbol{\theta}^*\|_2^2 \leq \|\hat{\boldsymbol{\theta}}\|_2^2$.