Assignment 1

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

Winter 2022

Please consult the course website for current information on the due date, the late policy, and any data you need to download for this assignment. This assignment is designed to advance your understanding of text classification, feature design and selction, the evaluation of classifiers (e.g., the F_1 score), and the mathematics of some important classification models. ★ **problems are for CSE 517 students only.** Other problems should be completed by everyone.

Data: The data you need for this assignment is available at `https://nasmith.github.io/NLP-winter22/assets/data/A1.tgz`.

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 Text Classification – Eisenstein 4.6 (p. 89)

After you run `tar -xzf A1.tgz`, in the directory `review_polarity`, you will find a dataset of positively and negatively classified reviews that was used by Pang and Lee [2], a seminal paper about sentiment classification. Consult the readme file for more information. Hold out a randomly selected 400 reviews as a test set.

Sentiment lexicon-based classifier. Create a classifier using a sentiment lexicon. A lexicon from Hu and Liu [1] is provided in the directory `opinion_lexicon_English`, but you are welcome to find and use (with attribution, of course) another. 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. In your write-up, describe the features you have chose and explain the reasoning behind your choice.

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.

Breaking good. For each of the following, write a review document that you believe would be considered as *positive* by human English speakers, and:

- your lexicon classifier predicts it as *positive*, whereas your logistic regression classifier predicts it as *negative*.

- your lexicon classifier predicts it as *negative*, whereas your logistic regression classifier predicts it as *positive*.
- both of your classifiers predict it as *negative*.

For each of the above scenarios, briefly discuss why your classifier(s) would make incorrect predictions for the document you created.

Statistical significance (extra credit). Determine whether the differences in accuracy and F_1 score are statistically significant at $\alpha = 0.05$, 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 fully grasp the nuts and bolts of these important ideas. If you aren't sure about whether it's okay to import a particular library, please **ask** on the discussion board!

2 ★ Regularization – 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 XOR – Eisenstein 3.4 (p. 65)

Design a feedforward network to compute this function, which is closely related to XOR:

$$f(x_1, x_2) = \begin{cases} -1 & \text{if} \quad x_1 = 1 \wedge x_2 = 1 \\ 1 & \text{if} \quad x_1 = 1 \wedge x_2 = 0 \\ 1 & \text{if} \quad x_1 = 0 \wedge x_2 = 1 \\ -1 & \text{if} \quad x_1 = 0 \wedge x_2 = 0 \end{cases}$$

Your network should have a single output node that uses the "sign" activation function,

$$\text{sign}(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ -1 & \text{if} \quad x \leq 0 \end{cases}$$

Use a single hidden layer, with ReLU activation functions. Describe all weights and offsets.

4 Extra Credit: Initialization at Zero – Eisenstein 3.5 (p. 65)

Consider the same network as in problem 3 (with ReLU activations for the hidden layer), with an arbitrary differentiable loss function $\ell(y^{(i)}, \tilde{y})$, where $\tilde{y}$ is the activation of the output node. Suppose all weights and offsets are initialized to zero. Show that gradient descent will not learn the desired function from this initialization.

References

[1] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proc. of KDD*, 2004.

[2] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. of ACL*, 2004.