Homework 01: Text Classification

CS 6501-005 Natural Language Processing

Updated on: September 13, 2019

1. Suppose you have a single feature x, with the following conditional distribution:

$$p(x \mid y) = \begin{cases} \alpha & X = 0, Y = 0 \\ 1 - \alpha & X = 1, Y = 0 \\ 1 - \beta & X = 0, Y = 1 \\ \beta & X = 1, Y = 1 \end{cases} \tag{1}$$

   Further suppose that the prior distribution is uniform, $P(Y = 0) = P(Y = 1) = 0.5$, and that both $\alpha > \frac{1}{2}$ and $\beta > \frac{1}{2}$. Given a Naive Bayes classifier with accurate parameters, what is the probability of making an error?

2. Suppose you have two labeled datasets D_1 and D_2, with the same feature set 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 logistic regression (LR) coefficients from training on dataset D_2,
   - Let $\boldsymbol{\theta}^{*}$ be the unregularized logistic regression (LR) coefficients from training on dataset $D_1 \cup D_2$.

   Under these conditions, prove that for any feature j,

$$\min(\theta_j^{(1)}, \theta_j^{(2)}) \leq \theta_j^{*} \leq \max(\theta_j^{(1)}, \theta_j^{(2)})$$

3. Let $\hat{\boldsymbol{\theta}}$ be the solution to an unregularized logistic regression problem, and let $\boldsymbol{\theta}^{*}$ be the solution to the same problem, with L_2 regularization. Prove that $\|\boldsymbol{\theta}^{*}\|_2^2 \leq \|\hat{\boldsymbol{\theta}}\|_2^2$.

4. Prove that F-measure is never greater than the arithmetic mean of precision and recall, $\frac{p+r}{2}$. Your solution should also show that F-measure is equal to $\frac{p+r}{2}$ if and only if $p = r$. [Hint: "if and only if" means that you need to prove the statement in both directions. In other words, your solution needs to show, with the definition of F-measure, both (1) $p = r \Rightarrow F = \frac{p+r}{2}$ and (2) $F = \frac{p+r}{2} \Rightarrow p = r$ hold.]

5. In this assignment, you will be asked to build a logistic regression classifier for sentiment classification with the following files

   - trn-reviews.txt: the Yelp reviews in the training set
   - trn-labels.txt: the corresponding labels of the Yelp reviews in the training set
   - dev-reviews.txt: the Yelp reviews in the development set
   - dev-labels.txt: the corresponding labels of the Yelp reviews in the development set

   The starting point of building a classifier is from the IPython notebook **demo.ipynb**. The first section of this notebook provides a simple code to load the training and development set. Your work starts from the *second* section.

- In the second section, you can implement the `CountVectorizer` function with different parameter settings, as shown in the two examples in this section.
- In the third section, try to pick different values of the parameters within function `LogisticRegression`

The task is to find the parameter setting used for both `CountVectorizer` and `LogisticRegression`, which can give the **best** accuracy on the development set. The baseline accuracy is 61.4% with uni-gram features and your results should be better than this number.

Your homework submission should include the IPython notebook with the name **[Your-ComputingID].ipynb**. Please keep the best parameter setting only in the notebook, so we can easily reproduce the results.