CSCI 59000: Machine Learning

Assignment 3

Assigned: March 31, 2017

Due: April 13, 2017

Execute the following preprocessing steps with the competition training data set.

1. Convert each text fragment to a bag of words representation.
2. For each of the 10K features compute the mean and standard deviation. Then, from each feature, subtract the mean and divide by the standard deviation to obtain a standard set of features all with unit variance and zero mean.
3. Split your training data into 10 folds and make sure that text fragments from the same book are not split between multiple folds.
4. Download the liblinear package (http://www.csie.ntu.edu.tw/~cjlin/liblinear/) and install it on your machine. Liblinear has stand-alone executables but you can also use the Matlab interface provided in the package.

   You are going to use the following options.

   -s type : set type of solver (default 1)

   for multi-class classification

   0 -- L2-regularized logistic regression (primal)

   1 -- L2-regularized L2-loss support vector classification (dual)

   2 -- L2-regularized L2-loss support vector classification (primal)

   3 -- L2-regularized L1-loss support vector classification (dual)

   5 -- L1-regularized L2-loss support vector classification

   6 -- L1-regularized logistic regression

   7 -- L2-regularized logistic regression (dual)

See the next page.

Part 1. Using s=5 and only the most frequent 500 words fill out the following table. F1 score indicates the 10-fold mean F1 score.

C	F1 score	Training time	# of support vectors	# of nonzero coefficients in w	Margin (2/sqrt(w'*w))
0.01					
.1					
1					
10					

Independently investigate the correlation (if any) between the following pairs of variables.

a. F1 score vs. # of support vectors
b. F1 score vs. Margin
c. F1 score vs. # of nonzero coefficients in w
d. F1 score vs. C
e. C vs. training time
f. C vs. # of support vectors
g. C vs. # of nonzero coefficients in w
h. C vs. Margin

Part 2. Repeat the experiment in Part 1 using the most frequent 5000 words.

Part 3. Repeat the experiment in Part 1 this time using s=1.

Part 4. Repeat the experiment in Part 2 this time using s=1.

Discussion: Which configuration (which s, what c value, and what number of most frequent words) do you believe would perform the best in the competition test data and why?