COMP 441 — Large Scale Machine Learning. April 13th, 2016

Assignment 4

Due By: 1st May 2016 *Due By: 1st May 2016*

This assignment is worth total 25 points. Your assignment is due by 12:00pm, 1st May, 2016 either by email or in class.

1 Active Learning (25 Point)

(Use your favorite SVM code, suggested: liblinear https://www.csie.ntu.edu.tw/ cjlin/liblinear/).

We will compare active learning with random sampling.

1. Download the 20news-bydate-matlab.tgz data file from http://qwone.com/ jason/20Newsgroups/
2. Use their test partition for comparing the accuracy (TEST).
3. Randomly sample 20 data vectors from the train set (TRAIN). This will form our initial set S for training with 20 points. We will slowly increase it.
4. Repeat the following: (Rounds)
   - Use criteria C (described below) to remove 5 points from the training set (TRAIN) and add them to S.
   - Train SVM on the training set S and test it on the fixed full test set (TEST), use default SVM parameter c.
   - Plot the progress in the accuracy as we proceed, i.e. accuracy on TEST set with size of S increasing (20, 25, 30, ...).

Compare the following 3 criterions for C.

1. Randomly select 5 points from TRAIN. (RAND)
2. Select 5 points x from TRAIN with smallest value of $|w^t x|$. Here || is the absolute value and w is the weight vector (model) learned from the previous round. (SMALL DIST)
3. Select 5 points with maximum value of $|w^t x|$. (LARGE DIST)

Outputs: Plot the accuracy (on the full TEST set) with increase in S for all the three criterions. Write a small report (less than a page) summarizing your observations and conclusions.