CS 445/545

Machine Learning

Winter 2016

Addendum

- **Corrections:**
  - In the original homework writeup I said "Randomly shuffle test data." You don't need to do this. Just shuffle the training data.

  - In the original homework writeup, under Experiment 2, I said:
    For m = 1 to 57
    This should be:
    For m = 2 to 57

    That is, you need to have at least 2 features for the SVM to work.

- **Clarifications:** For Experiments 2 and 3 (Feature Selection), the original writeup said:

  - Train a linear SVM, SVM_m , on all the training data, only using these m features, and using C* from Experiment 1

  - Test SVM_m on the test set to obtain accuracy.

What this means is the following:

Once you have selected the m features you are going to use, create new training and test data files that contain only those m features.

For example, each row in my original training data file for SVM_light, called "spam.train" looks like this:

1 1:x_1 2:x_2 3:x_3 ... 57:x_{57}

where x_1 is the value of the first feature, x_2 is the value of the second feature, etc. (The 1 at the beginning denotes that this example is positive, i.e., spam.)

Now, suppose you have selected m = 3 features, and say the ones you selected are x_3 , x_{27} , and x_{41}. What you need to do is to create a new file, called something like "spam3.train", where each row contains only these three features:

1.0 1:x_3 2:x_{27} 3:x_{41}

Thus, this example is now represented by three features instead of 57. (SVM_light makes you put an index before each feature; thus the "1:", "2:", "3:". This will be different for other SVM packages.)

You would do the same for the test file for this value of m -- that is, get rid of all features on each

row except for the ones you have selected. The file would be called something like "spam3.test".

Then you would train an SVM on "spam3.train" and test it on "spam3.test", and record the accuracy.

You would need to do this for $m = 2, \ldots, 57$.

In class I mentioned something about setting features to zero, but that was incorrect.