CS 445/545

Machine Learning

Winter 2016

Homework 3: SVMs and Feature Selection

Due Tuesday, February 16, 2016, 2pm

In this homework you will do experiments with linear SVMs and feature selection.

Please follow the steps below. Read carefully!

- **Data processing:**
  - Download data from https://archive.ics.uci.edu/ml/datasets/Spambase
  - Create a subset of the data that has equal numbers of positive and negative examples.
  - Put data into format needed for the SVM package you're using
  - Split data into ~ ½ training, ½ test (each should have equal numbers of positive and negative examples)
  - Scale training data using standardization (as in HW 2)
  - Scale test data using standardization parameters from training data
  - Randomly shuffle training data

- **Experiment 1:** Cross-validation using linear SVM to find best "C" parameter value
  - Use SVM$^{\text{light}}$ (http://svmlight.joachims.org/) or any SVM package
  - Use linear kernel (default in SVM$^{\text{light}}$)
  - Use 10-fold cross-validation to test values of C parameter in {0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1}
    - Split training set into 10 approx equal-sized disjoint sets S_i
    - For each value of the C parameter j :
      - For i = 1 to 10

        Select S_i to be the "validation" set

        Train linear SVM using C=j and all training data except S_i.

        Test learned model on S_i to get accuracy $A_{j,i}$
      - Compute average accuracy for C=j: $A_j = \sum_{i=1}^{10} A_{j,i}$
    - Choose value C= C* that results in highest A_j
  - Train new linear SVM using all training data with C=C*

- Test learned SVM model on test data. Report accuracy, precision, and recall (using threshold 0 to determine positive and negative classifications)

- Use results on test data to create an ROC curve for this SVM, using about 200 evenly spaced thresholds.

- **Experiment 2:** Feature selection with linear SVM

  - Using final SVM model from Experiment 1:
    - Obtain weight vector **w**. (For SVM^{light}, see https://www.cs.cornell.edu/people/tj/svm_light/svm_light_faq.html)

  Select features:
  - For m = 2 to 57
    - Select the set of m features that have highest $|w_m|$

    - 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.

  - Plot accuracy vs. m

- **Experiment 3:** Random feature selection

  Same as Experiment 2, but for each m, select m features at random from the complete set. This is to see if using SVM weights for feature selection has any advantage over random.

What to include in your report:

- **Experiment 1:**
  - Which SVM package you used
  - The C* (best value of C) you obtained
  - Accuracy, Precision, and Recall on the test data, using final learned model
  - ROC Curve

- **Experiment 2:**
  - Plot of accuracy (on test data) vs. m (number of features)
  - Discussion of what the top 5 features were (see https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.names for details of features)
  - Discussion of the effects of feature selection (about 1 paragraph).

- **Experiment 3:**
  - Plot of accuracy (on test data) vs. m (number of features)
  - Discussion of results of random feature selection vs. SVM-weighted feature selection (about 1 paragraph).

What code to turn in:

- Your code / script for performing cross-validation in Experiment 1
- Your code for creating an ROC curve in Experiment 1
- Your code / script for performing SVM-weighted feature selection in Experiment 2
- Your code / script for performing randomfeature selection in Experiment 3

How to turn it in (read carefully!):

- Send these items in electronic format to mm@pdx.edu by 2pm on the due date. No hard copy please!
- The report should be in pdf format and the code should be in plain-text format.
- Put "MACHINE LEARNING HW 3" in the subject line.

If there are any questions, don't hesitate to ask me or e-mail the class mailing list.

Policy on late homework: If you are having trouble completing the assignment on time for any reason, please see me before the due date to find out if you can get an extension. Any homework turned in late without an extension from me will have 5% of the grade subtracted for each day the assignment is late.