MA598 Homework #2 Classification – Due 10/23 midnight

SVM

In particular, we will use Sklearn.svm.svc to study handwritten digits from the processed US Postal Service Zip Code data set. Download the data (extracted features of intensity and symmetry) for training and testing:

http://www.amlbook.com/data/zip/features.train

http://www.amlbook.com/data/zip/features.test

(the format of each row is: digit intensity symmetry). We will train a one-versus-one (one digit is class +1 and another digit is class -1) classifier for the digits '1' (+1) and '5' (-1).

Definition:

- Ein returns the in training sample error of the current svm model. It is the fraction of in training sample points which got misclassified.
- Eout returns the testing sample error of the current svm model. It is the fraction of testing sample points which got misclassified.
- Ecv returns the leave one out cross validation in training sample error of the current svm model.
- accuracy over the testing set = 1 - Eout

Practical remarks:

(i) For the purpose of this homework, do not scale the data when you use libsvm or other packages, otherwise you may inadvertently change the (effective) kernel and get different results.

(ii) In 10-fold cross validation, if the data size is not a multiple of 10, the sizes of the10 subsets may be off by 1 data point.

(iii) Some packages have software parameters whose values affect the outcome. ML practitioners have to deal with this kind of added uncertainty.

a. Consider the linear kernel K(x_n, x_m) = x_n^T x_m. Train and test using all of the points, writing the output to an output file hw2.txt. In addition to using all of the training examples, try subsets of the training data and print out accuracy over the testing set (1 - Eout (over all test examples), and the number of support vectors. Try with the first {50, 100, 200, 800} points with the linear kernel. The output of these experiments should be written in Markdown cells, as opposed to

output to a file. Only the first part (with all points) should be written to the file hw2.txt.

b. Consider the polynomial kernel K(x_n, x_m) = (1 + x_n^T x_m)^Q, where Q is the degree of the polynomial. Comparing Q = 2 with Q = 5, which of the following statements is correct?

i. When C = 0.0001, Ein is higher at Q = 5.
ii. When C = 0.001, the number of support vectors is lower at Q = 5.
iii. When C = 0.01, Ein is higher at Q = 5.
iv. When C = 1, Eout is lower at Q = 5.
v. None of the above

c. Consider the 1 versus 5 classifier with Q = 2 and C $\in$ {0.001, 0.01, 0.1, 1}.

Which of the following statements is correct? Going up or down means strictly so.

[a] The number of support vectors goes down when C goes up.

[b] The number of support vectors goes up when C goes up.

[c] Eout goes down when C goes up.

[d] Maximum C achieves the lowest Ein .

[e] None of the above

• Cross Validation

Read this: http://scikit-learn.org/stable/modules/cross_validation.html

In the next two problems, we will experiment with 10-fold cross validation for the polynomial kernel. Because Ecv is a random variable that depends on the random partition of the data, we will try100 runs with different partitions and base our answer on how many runs lead to a particular choice.

d. Consider the 1 versus 5 classifier with Q = 2. We use Ecv to select C $\in$ {0.0001, 0.001, 0.01, 0.1, 1}. If there is a tie in Ecv, select the smaller C. Within the 100 random runs, which of the following statements is correct?

[a] C = 0.0001 is selected most often.

[b] C = 0.001 is selected most often.

[c] C = 0.01 is selected most often.

[d] C = 0.1 is selected most often.

[e] C = 1 is selected most often.

e. Again, consider the 1 versus 5 classifier with Q = 2. For the winning selection in the previous problem, the average value of Ecv over the 100 runs is closest to

(a) 0.001
(b) 0.003
(c) 0.005
(d) 0.007
(e) 0.009

f. Consider the radial basis function (RBF) kernel K(x_n, x_m) = e^(-||xn - xm||^2) in the SVC approach. Which value of C $\in$ {0.01, 1, 100, 10^4, 10^6} results in the lowest Ein? The lowest Eout?

Put all Python code and results into one PDF file

Put your name and Purdue student ID in the first page of the PDF file

Submit it to blackboard by 10/23 midnight.