Homework 4 (10 pts)

Topic: SVM classification and regression

Note: you need to study/understand/apply SVM classification and SVM regression software

Problem 1:

(a) Compare the prediction performance of k-nearest neighbors and SVM methods for classification, using a 20 dimensional data set, specified as follows:

- 20 input variables are uniformly distributed in [0, 1] hypercube;
- the class label is y = *Sign*(x_1+ x_2+ x_3+…+x_{10}-5)

Note: the output depends only on the first 10 inputs variables, and the other 10 inputs contribute to noise.

Use 50 training samples, 50 validation samples (for model selection), and 1,000 test samples. For comparisons, use linear SVM and k-nearest neighbors classifiers. Both methods have a single tuning parameter estimated using a validation data set.

Make sure you repeat the experiments using 3-5 different realizations of training + validation data, and record (training, validation and test) errors of your optimal SVM models.

(b) Recall two analytical bounds for SVM shown as (9.9) and (9.10) in the textbook.

$$E_n[\text{error\_rate}] \le \frac{E_n[\text{number of support vectors}]}{n}. \qquad (9.9)$$

$$h \le \min\left(\frac{r^2}{\Delta^2}, d\right) + 1, \qquad (9.10)$$

Which one of these bounds is more suitable (works better) for this data set? To answer this question, you may find useful to show/analyze estimated SVM models using the 'histogram-of-projections' technique.

Problem 2

(a) Apply SVM regression software (using RBF kernel) to estimate regression model, according to the following experimental procedure.

Six-dimensional regression data is generated according to:

$$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 0x_4 + 5x_5 + 0x_6$$

where **x** is uniform in [-1.5,1.5].

Generate three data sets: training (100 samples), validation (100 samples) and test (800 samples).

The following procedure must be used for tuning SVM regression complexity parameters:

- Set the value of C using the following analytic prescription for all experiments.

$$\max(|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|)$$

- Select optimal values of (epsilon, gamma) that minimize MSE validation error (on independent validation set). You can present model selection results in a format similar to the following table.

	$\varepsilon = 0$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 6$	$\varepsilon = 8$
$\gamma = 2^{-5}$	xxxx	xxxx	xxxx	xxxx	xxxx
$\gamma = 2^{-4}$	xxxx	xxxx	xxxx	xxxx	xxxx
$\gamma = 2^{-3}$	xxxx	xxxx	xxxx	xxxx	xxxx
$\gamma = 2^{-2}$	xxxx	xxxx	xxxx	xxxx	xxxx
$\gamma = 2^{-1}$	xxxx	xxxx	xxxx	xxxx	xxxx
$\gamma = 2^{0}$	xxxx	xxxx	xxxx	xxxx	xxxx
$\gamma = 2^{1}$	xxxx	xxxx	xxxx	xxxx	xxxx
$\gamma = 2^{2}$	xxxx	xxxx	xxxx	xxxx	xxxx
$\gamma = 2^{3}$	xxxx	xxxx	xxxx	xxxx	xxxx
$\gamma = 2^{4}$	xxxx	xxxx	xxxx	xxxx	xxxx
$\gamma = 2^{5}$	xxxx	xxxx	xxxx	xxxx	xxxx

Report the NRMS and MSE test error of your regression model.

(b) Compare your SVM results (test error) in part (a) to the test error obtained using Projection Pursuit Regression (PPR) where an optimal number of terms (complexity parameter) is estimated using the same validation data as in part (a). Discuss your comparisons and comment SVM performance (considering that PPR method is most appropriate for this data set).