Statistical Learning-Classification
STAT 441 / 841, CM 762

Assignment 3
Department of Statistics and Actuarial Science
University of Waterloo

Due: Thursda November 9, at 8:30 am

Policy on Lateness: Slightly late assignments (up to 24 hs after due date) are accepted with 10% penalty. No assignment are accepted after 24 hs after the due date.

1. Download the Ionosphere dataset from the course webpage :

   a) Write a program to fit a single hidden layer neural network via back-propagation and weight decay.

   b) Apply your program in part *a*) to the data . Chose Ion.test as the test set, and Ion.trin as the training set. Plot the training and test error curves as a function of the number of epochs for four different values of the weight decay parameter. Discuss the overfitting behavior in each case.

   c) Set the value of weight decay equal to zero, then vary the number of hidden units in the network (starting from 1 unit, and determine the minimum number needed to perform well for this task. Plot the training and test error curves as a function of the number of hidden units.

   d) Select the best model (the optimum number of hidden nodes or the best value for weight decay) and classify the test data using the network and report the observed misclassification error rate. Construct a 2 by 2 table of the form

	$\hat{h}(x) = 0$	$\hat{h}(x) = 1$
$y = 0$	?	?
$y = 1$	?	?

   Note 1: Attach your code to your assignment as an appendix and submit the code to the assignment drop box in D2L as well.

2. *a*) Write a program to fit an RBF network. In implementing RBF, you need to cluster the data and find the center and spread of each cluster. You don't need to implement a clustering algorithm yourself.

[1].

b) Use the Ionosphere dataset (Ion.mat) . Use the standard cross validation, standard Leave one out cross validation, and Leave one out cross validation as expressed in (1) (The method explained in Question 6 shows how LOO can be performed without iteration.) and find the optimum number of basis function for each model. Compute the test error in each case and complete the following table.

In this table
CV is standard cross validation
LOO is standard leave one out cross validation
CLOO is Leave one out cross validation as expressed in (1).

	Target function	
Method	$TrainingError$	$TestError$
CV		
LOO		
CLOO		

Note: Attach your code to your assignment as an appendix and submit the code to the assignment drop box in Learn as well.

3. Support Vector Machine

a) Write a function $[b, b_0] = HardMarg(X, y)$ which takes a $d \times n$ matrix X and $n \times 1$ vector of target labels y and returns: a $d \times 1$ vector of weights b and a scalar offset b_0, corresponding to the maximum margin linear discriminant classifier.

b) Write a function $[b, b_0] = SoftMarg(X, y, \gamma)$ which takes an additional scalar argument γ and returns b and b_0 corresponding to the maximum soft margin linear discriminant classifier.

c) Write a function $[yhat] = classify(Xtest, b, b_0)$ which takes a $d \times m$ matrix $Xtest$, a $d \times 1$ vector of weights b, and a scalar b_0, and returns a $m \times 1$ vector of classifications $yhat$ on the test patterns.

d) For each of the datasets linear, noisylinear, and quadratic on Piazza solve for each kind of discriminant function: $[bh, b_0h] = HardMarg(X, y)$, $[bs, b_0s] = SoftMarg(X, y, 0.5)$, produce a 2D plot of the training data and the two hypotheses corresponding to bh, b_0h and bs, b_0s and report the mean misclassification error (i.e., the sum of misclassification errors divided by the number of data points) that each of the two hypotheses obtained on the training data and on the test data.
Hand in a plot and two tables for each dataset.

[1] Alternately you can use any clustering algorithm and any clustering routine in any langauge based on your preference

Note 1: Submit your code to Learn drop box.
Note 2: Your function must be able to handle arbitrary d, n, γ, and m.

4. Let $\hat{f}$ be an estimator of the quantity f, show that its mean-squared error can be decomposed as follows:

$$\begin{aligned} E(\hat{f}-f)^2 &= E[\hat{f}-E(\hat{f})]^2 + [E(\hat{f})-f]^2 \\ &= Var(\hat{f}) + Bias^2(\hat{f}) \end{aligned}$$

Only for Grad Students

5. Given a set of data points $\{\mathbf{x}_i\}$, we can define the *convex hull* to be the set of all points $\mathbf{x}$ given by

$$\mathbf{x} = \sum_i \alpha_i \mathbf{x}_i$$

where $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$. Consider a second set of points $\{\mathbf{y}_i\}$ together with their corresponding convex hull. By definition, the two sets of points will be linearly separable if there exist a vector $\hat{\mathbf{w}}$ and a scaler w_0 such that $\hat{\mathbf{w}}^T\mathbf{x}_i + w_0 > 0$ for all $\mathbf{x}_i$, and $\hat{\mathbf{w}}^T\mathbf{x}_i + w_0 < 0$ for all $\mathbf{y}_i$.

Show that if their convex hulls intersect, the two sets of points cannot be linearly separable.

6. **Leave-one-out cross validation.** Consider the model $y_i = f(\mathbf{x}_i) + \epsilon_i$. When $f(\mathbf{x}_i) = \beta_0 + \beta^T\mathbf{x}_i$, this is known as an ordinary least square (OLS). Let H be the hat matrix associated with OLS. Show that

$$y_i - \hat{f}^{(-i)}(\mathbf{x}_i) = \frac{y_i - \hat{f}(\mathbf{x}_i)}{1 - H_{ii}} \tag{1}$$

where H_{ii} denote the i-th diagonal element of H; and $\hat{f}^{(-i)}(\mathbf{x}_i)$ denotes estimating $\hat{f}(\mathbf{x}_i)$ using an $\hat{f}$ that is obtained without using the i-th observation. Thus show that the leave-one-out cross validation can be computed without iteration.