Statistical Learning-Classification
STAT 841 / 441, CM 763

Assignment 2
Department of Statistics and Actuarial Science
University of Waterloo

Policy on Lateness: No assignment are accepted after the due date.

Note: Matlab is not mandatory. You can use any programming language.

1. **Face detection**:
   Download the faces.mat from the course webpage.
   The file *faces.mat* is composed of *train_faces, train_nonfaces, test_faces*, and *test_nonfaces*.
   Make a training and a test set as follows:

   ```
   training_data=[train_faces' train_nonfaces'];
   % (This will be a 361 by 4858 matrix.)

   test_data=[test_faces' test_nonfaces'];
   % (This will be a 361 by 944 matrix.)
   ```

   - Write a program to fit a logistic regression model to the training data. Report the first 5 components of the optimum value of the logistic parameter β, as well as the training error and the test error.

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

3. In a maximum likelihood problem, we can define an error function by taking the negative logarithm of the likelihood. Show that the error function for the logistic regression model is a convex function of β, and hence show that it has a unique minimum value.

4. Let $u_1, u_2, \dots, u_p$ be the first p eigenvectors with largest eigenvalues of $S = \frac{1}{n}XX^T$ i.e., the principal basis vectors.

   We will construct a method for finding the u_j sequentially. As we showed in class, u_1 is the first principal eigenvector of S, and satisfies $Su_1 = \lambda u_1$. Now define $\tilde{x}_i$ as the orthogonal projection of x_i onto the space orthogonal to u_1:

   $\tilde{x}_i = P_{\perp u_1} x_i = (I - u_1 u_1^T) x_i$

   Define $\tilde{X} = [\tilde{x_1} \dots \tilde{x_n}]$ as the *deflated matrix* of rank $d - 1$, which is obtained by removing from the d dimensional data the component that lies in the direction of the first principal direction: $\tilde{X} = (I - u_1 u_1^T)^T X = (I - u_1 u_1^T) X$

   a) Using the facts that $XX^T u_1 = n\lambda_1 u_1$ (and hence $u_1^T XX^T = n\lambda_1 u_1^T$) and$u_1^T u_1 = 1$, show that the covariance of the deflated matrix is given by

   $$\tilde{S} = \frac{1}{n}\tilde{X}\tilde{X}^T = \frac{1}{n}XX^T - \lambda u_1 u_1^T$$

   b) Let v be the principal eigenvector of $\tilde{S}$. Show why $v = u_2$. (You may assume v is unit norm.)

Only for Grad Students

5. Consider a multiclass logistic regression model (multilogit model) applied to d-dimensional data with K classes. Let β be the $(d+1)(K-1)$-vector consisting of all the coefficients. Define a suitably enlarged version of the input vector x to accommodate this vectorized coefficient vector. Derive the Newton-Raphson algorithm for maximizing the log-likelihood, and describe how you would implement this algorithm.

6. Consider a classification model for two classes with prior class probabilities $\pi_k, k = 1, 2$. Suppose that the class-conditional densities are given by Gaussian distributions with a shared covariance matrix. Suppose we are given a training data set $\{(x_i, y_i)\}$ where $i = 1 \dots n$, and $y \in \{0, 1\}$ are class labels. Assume that the data points are drawn independently from this model.

   a) Compute the maximum-likelihood estimation for the prior probabilities.

   b) Compute the maximum-likelihood estimation for the mean of the Gaussian distribution for each class.

   c) Compute the maximum-likelihood estimation for the shared covariance matrix.