Homework I

Statistics 640 / 444 Fall 2014

Assigned: September 23

Due: October 9

Stat 640:

1. *Computations for Fisher's Discriminant Analysis.*
   Assume we observe data, $\mathbf{X}$, with n observations belonging to K classes and p features. Assume that the columns of $\mathbf{X}$ have previously been centered. Let $\boldsymbol{\mu}_k$ be the group mean or centroid for class k.
   (a) Derive an expression for the between-class covariance, $\boldsymbol{\Sigma}_B$, that is a function of $\boldsymbol{\mu}_k$, $k = 1, \dots K$.
   (b) Let $\mathbf{Y} \in \{0,1\}^{n \times K}$ be the indicator matrix of class membership. Derive an expression for $\boldsymbol{\Sigma}_B$ that is a function of only $\mathbf{X}$ and $\mathbf{Y}$.
   (c) Show that the covariance of $\mathbf{X}$, $\boldsymbol{\Sigma}_T$, can be decomposed into $\boldsymbol{\Sigma}_T = \boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_W$ where $\boldsymbol{\Sigma}_W$ is the within-class covariance.
   (d) Let $\{\boldsymbol{\alpha}_1, \dots \boldsymbol{\alpha}_{K-1}\}$ be the set of Fisher's discriminant vectors. Show that $\{\boldsymbol{\alpha}_1, \dots \boldsymbol{\alpha}_{K-1}\} = span\{\boldsymbol{\Sigma}_W^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j) : \ k \neq j \in 1, \dots K\}$.

2. *Discriminant Analysis*: Prove the equivalence of Linear Discriminant Analysis (Bayes classifier for the multivariate normal model), Fisher's Discriminant Analysis, and Optimal Scoring when the number of training samples in each class are equal.
   Hint: Write the gradient equations for all methods in terms of $\boldsymbol{\Sigma}_W$ and $\boldsymbol{\Sigma}_B$ and show that Fisher's discriminant vectors, the optimal scoring coefficients, and the normal model discriminant functions are proportional to each other.

3. *SVMs and the hinge loss*: Show that the following two optimization problems are equivalent:

$$\begin{aligned} &\underset{\beta_0,\beta}{\text{minimize}} \ \ \frac{1}{2}||\beta||_2^2 + \gamma \sum_{i=1}^{n} \xi_i \\ &\text{subject to} \ \ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \ \& \ \xi_i \geq 0 \ \ \forall \ i = 1 \dots n. \end{aligned} \tag{1}$$

$$\underset{\beta_0,\beta}{\text{minimize}} \ \ \sum_{i=1}^{n} \left(1 - y_i(x_i^T\beta + \beta_0)\right)_+ + \frac{\lambda}{2}||\beta||_2^2. \tag{2}$$

4. *Kernel Methods*: Suppose you wanted to perform kernel logistic regression. That is, you want to fit the model, $logit(Y) = f(X)$ for some infinite dimensional $f(X) = \sum_{j=1}^{\infty} c_j \phi_j(X)$.
   - Derive an expression for the kernel logistic log-likelihood using the kernel trick.
   - Compute the gradient and Hessian for this model. Is there a way to simplify computations?

   Note: 640 students have the option of completing only 3 of the above proofs (you choose!) and completing one of the applied data analyses below.

Stat 444:

1. Digits data: 3 and 8.
   For this problem, use the training and test sets of the zip code data available from the ESL textbook website. Limit consideration to classifying between the digits "3" and "8".

   (a) Compare and contrast the following classifiers in terms of test misclassification error:

      i. Naive Bayes Classifier.
      ii. KNN Classifier.
      iii. LDA.
      iv. Quadratic Discriminant Analysis.
      v. Logistic regression.
      vi. Regularized logistic regression. Which type of regularization did you choose? Why?
      vii. Linear SVMs.
      viii. Kernel SVMs - polynomial and radial kernels.

      If there are tuning parameters associated with the model, you can use built in default methods (i.e. cross-validation) to select these.

   (b) Reflection. Interpret the results. Which method performed the best? Why?

2. Digits data.
   For this problem, use the training and test sets of the zip code data available from the ESL textbook website. Use all the digits for this problem. Compare the following two methods for multi-class SVMs:

   (a) SVMs: One vs. All.

   (b) SVMs: One vs. One.

   Reflection. Which method performs better in terms of test error? Why? Show the confusion matrix for multi-class misclassification. Which classes are most often misclassified by the two methods? Why? Interpret the results.

3. 14-cancer microarray data.
   For this problem, use the training and test sets of the 14-cancer microarray data available from the ESL textbook website.

   (a) Compare the training and test error rates over the range of regularization parameters for the following methods:

      - Multinomial ridge regression.
      - Multinomial elastic net.
      - Multinomial lasso.

   (b) Visualize the regularization paths for these methods.

   (c) How do the misclassification rates compare to those of the following methods?

      - Naive Bayes Classifier.
      - KNN.
      - Linear SVMs.

   (d) Reflection. Interpret the results. Which method has the lowest test error rates? Why? Interpret the regularization paths. Which classifier would you recommend for this data if all you desired was the best error rates? Which would you recommend if you also wanted to interpret the results of the classifier? That is, assume you want to know which genes or sets of genes best separate the cancer types.