BST 764: Applied Statistical Modeling
Breheny

Assignment 2: Discriminant analysis
Due: Tuesday, October 4

Mathematical concepts and derivations

1. Prove the lemma stated on slide 4 of the 9-15 notes: that for any symmetric matrix $\mathbf{A}$,

$$\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{y}^T\mathbf{A}\mathbf{y} = (\mathbf{x}+\mathbf{y})^T\mathbf{A}(\mathbf{x}-\mathbf{y})$$

2. (a) Show that $\frac{1}{n-1}\sum_i(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ is an unbiased estimator of the variance of a vector $\mathbf{x}$. Recall that the multivariate definition of variance is

$$\mathrm{Var}(\mathbf{x}) = \mathrm{E}\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\},$$

   where $\boldsymbol{\mu}$ is the expected value of the random variable $\mathbf{x}$.

   (b) Using the result of part (a), show that if each class has the same variance $\boldsymbol{\Sigma}$, then the pooled estimate $\hat{\boldsymbol{\Sigma}}$ (as defined on slide 6 of the 9-15 notes) is an unbiased estimate of $\boldsymbol{\Sigma}$.

3. Suppose that $\pi_k = \pi_l$ and that f_k and f_l are both multivariate normal with equal variance. Show that if $\mathbf{x}$ lies halfway between μ_k and μ_l, then $\mathbf{x}$ is on the decision boundary[1] between classes k and l.

Simulation

4. Consider a classification setting with $K = 5$ classes, all with equal marginal probabilities. For each observation, in addition to its class, we also observe $\mathbf{x}$, a 15-dimensional vector of potential explanatory variables. Consider two settings:

   - Setting 1: $x_{ij}|G = k \overset{\text{iid}}{\sim} Z + \mu_{jk}$
   - Setting 2: $x_{ij}|G = k \overset{\text{iid}}{\sim} C + \mu_{jk}$

   where Z is a standard normal random variable and C is a standard Cauchy random variable. Note that μ_{jk} is the mean, median, and mode of $x_{ij}|G = k$ in setting 1, and the mode and median of $x_{ij}|G = k$ in setting 2, but that the mean of the Cauchy distribution is undefined.

   Consider the quantity

$$D = \sqrt{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)},$$

[1] The true decision boundary, not an estimate of the decision boundary arising from some procedure.

called the *Mahalanobis distance*. Set up $\{\boldsymbol{\mu}_k\}$ so that, in Setting 1, the Mahalanobis distance between any two classes is equal to 4. Use the same $\{\boldsymbol{\mu}_k\}$ for Setting 2, even though its Mahalanobis distance is undefined.

Conduct a simulation comparing the error rates of linear discriminant analysis and multinomial logistic regression in these two settings. To evaluate the error rate, use a sample size of 100 to estimate each model's parameters and a sample size of 10,000 to evaluate the predictive accuracy of each model.

Application

5. The course website contains data from an orthopedic research study of two spinal conditions: *spinal disc herniation* and *spondylolisthesis*. Both conditions involve the displacement of vertebra from their normal positions, and it can be difficult to distinguish the two from each other, and from normal variation in the spine.

   The data set `vertebra.txt` contains measurements of five different angles pertaining to the orientation of the spine and pelvis, taken on a sample of 100 normal patients (`NO`), 60 patients with disc herniation (`DH`), and 150 patients with spondylolisthesis (`SL`).

   (a) Randomly split the data set into a training set and a testing set, both with an equal number of observations. Fit an LDA model to the training set, then see how well it predicts the outcomes in the testing set. Report the misclassification error.

   (b) Repeat part (a) for QDA.

   (c) Consider changing the angle of lordosis, while all the other angles remain at their sample averages. Make a plot of how the probabilities of the three classes change as the angle of lordosis changes from 10 to 100.

   (d) Obviously, this study intentionally oversampled patients with spinal disorders. In most external settings, the marginal probabilities of disc herniation and spondylolisthesis are going to be much lower than their proportions in this sample. Suppose the marginal probability of having a normal spine was 80%, and 10% for each of the two disorders. Repeat part (c) in this setting.