Intelligent Data Analysis

Problem Set 3 **Due 4/7/2016**

1. Naïve Bayesian model assumes that all inputs are *independent* for each class. Using the estimates, the discriminant function becomes

$$g_i(\mathbf{x}) = -\frac{1}{2}\sum_{j=1}^{d}\left(\frac{x_j - m_{ij}}{s_j}\right)^2 + \log \hat{P}(C_i), \quad i = 1, 2, \ldots, K.$$

Simplify the above discriminant function and show that the decision boundary between any two classes C_i and C_j is a hyperplane. In the case of d=2, a hyperplane is a line on the plane. In the case of d=3, a hyperplane is a plane.

2. Prove the identity $\frac{\partial(\mathbf{x}^T\mathbf{A}\mathbf{x})}{\partial\mathbf{x}} = (\mathbf{A}^T + \mathbf{A})\mathbf{x}$.

3. Let $\mathbf{A}$ be an m by n real matrix. Show that $N(\mathbf{A}) = N\left(\mathbf{A}^T\mathbf{A}\right)$, where $N(\mathbf{A})$ denotes the nullspace (also called kernel) of $\mathbf{A}$.

4. This problem is to prove that R^2 of a multiple regressor is the square of the correlation coefficient between $\mathbf{r}$ and $\mathbf{y}$, where $\mathbf{y} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{r}$. A simple way to prove this statement is by geometric interpretation.
   (a) First prove that the correlation coefficient between $\mathbf{r}$ and $\mathbf{y}$ is equivalent to $\cos\theta$, where θ is the angle between vectors $\mathbf{r} - \bar{r}\mathbf{1}$ and $\mathbf{y} - \bar{r}\mathbf{1}$, where $\mathbf{1}$ denotes the vector with all entries 1. The definition of sample correlation coefficient is given in Chapter 5.
   (b) Use the identity $\sum_{t=1}^{N}\left(r^t - \bar{r}\right)^2 = \sum_{t=1}^{N}\left(y^t - \bar{r}\right)^2 + \sum_{t=1}^{N}\left(r^t - y^t\right)^2$ to prove that

$$\cos^2\theta = \frac{\sum_{t=1}^{N}\left(y^t - \bar{r}\right)^2}{\sum_{t=1}^{N}\left(r^t - \bar{r}\right)^2} = R^2.$$

5. This problem is to prove

$$\frac{\mathbf{e}^T\mathbf{e}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{t=1}^{N}\left(e^t\right)^2 \sim \chi^2\left(N - d - 1\right),$$

where σ^2 is the variance of random noise.
   (a) A matrix is called *idempotent* if $\mathbf{M}^2$=$\mathbf{M}$. What are the eigenvalues of $\mathbf{M}$? Is

M diagonalizable? Show that the rank of an idempotent matrix **M** is equal to its trace, i.e., rank(**M**)=tr(**M**).

(b) Let **M** be an N by N symmetric, idempotent matrix of rank(**M**)= r . Consider the quadratic form $\mathbf{z}^T\mathbf{M}\mathbf{z}$, where $\mathbf{z}\sim N(\mathbf{0},\mathbf{I})$. Let **Q** be the real orthogonal matrix ($\mathbf{Q}^T = \mathbf{Q}^{-1}$) formed by the eigenvectors of **M** and define $\mathbf{y}=\mathbf{Q}^T\mathbf{z}$. Show that E[**y**]=**0** and Var(**y**)=**I**. Then, prove that

$$\mathbf{z}^T\mathbf{M}\mathbf{z} \sim \chi^2(r).$$

(c) If **X** is an N by (d+1) matrix with independent columns, prove that rank(**M**)=N-d-1, where $\mathbf{M}=\mathbf{I}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

(d) Now you can prove the statement on the top of this problem by using the relationship

$$\mathbf{e}^T\mathbf{e} = \boldsymbol{\varepsilon}^T\mathbf{M}\boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}\sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

6. Program Assignment

The goal of this assignment is to study multivariate parametric classification using Fisher's Iris data. With four input variables, namely, sepal length (x_1), sepal width (x_2), petal length (x_3), petal width (x_4), use the first 150 data points to build classifiers.

(a) Use x_1 and x_2 to build classifiers. Use 5-fold cross-validation (CV) to select models. That is, partition the data set into 5 equal-sized subsets. Use 4 subsets to estimate the unknown parameters of the four classifiers: QDA, LDA, Naïve Bayesian, and NMC, as described in Chapter 5. Apply the remaining subset to each classifier, report the classification matrix and compute the accuracy. Plot the average CV-accuracy verse model complexity (the four models). See Chapter 4, pp 23.

(b) Use all the inputs x_1, x_2, x_3, x_4 to build classifiers. Do the same thing as (a).