CPH 636 — Spring 2009 — Dr. Charnigo

Written Assignment 4

Written Assignment 4 is due on Monday 30 March at the end of lecture. You are encouraged to work in groups of two or three, though you may work individually if you prefer.

The first exercise is a theoretical problem whose solution is anticipated to be rewarding to a graduate student in statistics.

[10] 1. Derive formula (2) in Lecture 7. Then substantiate my claim in Lecture 7 that a simple comparison of Z scores is warranted only if $p_1 = \sigma_1/(\sigma_1 + \sigma_2)$.

Hints: Begin by producing an expression for

$$P\left(X \text{ originated from } N(\mu_1, \sigma_1^2) \mid X \in [x - \delta, x + \delta]\right),$$

where x is an arbitrary but fixed real number and δ is a small positive number. Then evaluate the limit of your expression as $\delta \to 0$. This will yield formula (2). Finally, prove that

$$\{\ |x - \mu_1|/\sigma_1 = |x - \mu_2|/\sigma_2 \Rightarrow \text{formula (2)} \ = 0.5\ \} \Rightarrow p_1 = \sigma_1/(\sigma_1 + \sigma_2).$$

To understand why this substantiates my claim in Lecture 7, note what the part in {brackets} says: an observation whose absolute Z score with respect to $N(\mu_1, \sigma_1^2)$ equals its absolute Z score with respect to $N(\mu_2, \sigma_2^2)$ has the same probability of having originated from $N(\mu_1, \sigma_1^2)$ as from $N(\mu_2, \sigma_2^2)$.

Remarks: Formula (2) is used in the famous *expectation maximization algorithm* of computational statistics to estimate parameters when data arise from a probability density function of the form

$$\frac{p_1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] + \frac{p_2}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right],$$

where $0 \leq p_1 = 1 - p_2 \leq 1$. Can you convince yourself that the above probability density function is precisely that which we have when $100p_1$% of individuals belong to a subpopulation described by a $N(\mu_1, \sigma_1^2)$ distribution and $100p_2$% of individuals belong to a subpopulation described by a $N(\mu_2, \sigma_2^2)$ distribution?

In the next three exercises, you will apply *discriminant analysis* to the classification problem in which ArrhythmiaAny (or some misspelling thereof) is the response variable and candidate explanatory variables are Age, Female, Height, Weight, QRSDur, PRInt, QTInt, TInt, PInt, and HeartRate. Of particular interest will be whether better predictions can be made from discriminant analysis than from logistic regression.

[10] 2. Which candidate explanatory variables are suggested for discriminant analysis by backward elimination based on the training subset? Refer to this collection of variables as Group I. Which are suggested by forward selection and stepwise selection based on the training subset? Refer to these collections of variables as Group II and Group III, respectively. Groups I, II, and III may or may not be distinct. Finally, let Group IV contain all of the candidate explanatory variables.

[20] 3. For each of the four Groups identified in exercise 2, apply discriminant analysis to the training subset. In each instance, please respond to the following items.

[05] a. Use diagnostic plots to assess whether the multivariate normality assumption is tenable. Explain why, for Group IV, you *know* that the candidate explanatory variables do not arise from a multivariate normal distribution.

Remarks: From a pragmatic data mining perspective, failure of the normality assumption is only a serious concern insofar as our predictions suffer because of it. Can you reconcile this perpsective with your own experiences as a student in (and a teaching assistant for) introductory methods courses, in which a normality assumption is emphasized as a requirement for a T test or an F test?

[05] b. Identify a linear combination of standardized variables that is optimal for separating those with arrhythmia from those without arrhythmia.

[05] c. What is the prior-weighted error rate on the training subset using cross-validation?

[05] d. What is the prior-weighted error rate on the validation subset?

[10] 4. Based on the results of exercise 3, decide which of the four Groups identified in exercise 2 is preferred for making predictions. For this Group, what is the prior-weighted error rate on the test subset? How does this compare to the misclassification rate on the test subset for your preferred logistic regression model in Written Assignment 3? For a fair comparison between discriminant analysis and logistic regression, please use a threshold of 0.50 in determining the misclassification rate on the test subset for your preferred logistic regression model.

In the last four exercises, you will apply a *classification tree* to the classification problem in which ArrhythmiaAny (or some misspelling thereof) is the response variable and candidate explanatory variables are Age, Female, Height, Weight, QRSDur, PRInt, QTInt, TInt, PInt, and HeartRate. Of particular interest will be whether better predictions can be made from a classification tree than from discriminant analysis or logistic regression.

[10] 5. Apply a classification tree to the training subset, using misclassification rate on the validation subset to determine the number of leaves in the tree. Furnish a diagram like that in {SAEx.mdi}.

[10] 6. What is your tree's misclassification rate on the test subset? How does it compare to the misclassification rate from your logistic regression model and the prior-weighted error rate from your discriminant analysis in exercise 4? Among the three supervised learning techniques presented in Lectures 6 through 8, which makes the best predictions on the test subset?

[20] 7. Please respond to the following items.

[05] a. According to your tree, what is the estimated probability of arrhythmia for someone whose profile on the candidate explanatory variables matches that of the person with ID = 5?

[05] b. What about for someone whose profile matches that of the person with ID = 317?

Remarks: Presumably you removed this person from the data set in Written Assignment 2, but you can still answer the question.

[05] c. If your preferred logistic regression model from Written Assignment 3 included Height as an explanatory variable, what does your preferred logistic regression model supply as the fitted probability of arrhythmia for someone whose profile matches that of the person with ID = 317? Does this answer *seem* reasonable?

If your preferred logistic regression model from Written Assignment 3 did *not* include Height as an explanatory variable, then suppose for the sake of discussion that Height has been added to that model.

What do you *think* that the fitted probability of arrhythmia will turn out to be for someone whose profile matches that of the person with ID = 317? Does such an answer *seem* reasonable?

[05] d. Neglecting the possibility of missing values necessitating invocation of surrogate rules, which candidate explanatory variables are *never* consulted by your classification tree?

[10] 8. Until now, your classification tree has used a threshold of 0.50 for the estimated probability of arrhythmia to determine for whom arrhythmia should be predicted. To the nearest 0.05, what threshold should be used if you want to maximize specificity on the test subset while requiring at least 90% sensitivity on the test subset?