# CPH 636 — Spring 2009 — Dr. Charnigo

## Written Assignment 4 Solutions

*Note*: Your answers to exercises 2 through 8 may differ from mine, depending on which observations you removed. I removed the observations with ID = 317, ID = 142, ID = 61, ID = 127, ID = 214, ID = 321, and ID = 394.

1. By Bayes' Theorem, we have

$$P\left(X \text{ originated from } N(\mu_1, \sigma_1^2) \mid X \in [x-\delta, x+\delta]\right)$$
$$= \frac{p_1\left\{\Phi\left[\frac{x+\delta-\mu_1}{\sigma_1}\right] - \Phi\left[\frac{x-\delta-\mu_1}{\sigma_1}\right]\right\}}{p_1\left\{\Phi\left[\frac{x+\delta-\mu_1}{\sigma_1}\right] - \Phi\left[\frac{x-\delta-\mu_1}{\sigma_1}\right]\right\} + p_2\left\{\Phi\left[\frac{x+\delta-\mu_2}{\sigma_2}\right] - \Phi\left[\frac{x-\delta-\mu_2}{\sigma_2}\right]\right\}},$$

where $\Phi[\cdot]$ is the standard normal cumulative distribution function. By Taylor's Theorem,

$$\Phi\left[\frac{x+\delta-\mu_1}{\sigma_1}\right] - \Phi\left[\frac{x-\delta-\mu_1}{\sigma_1}\right] = 2\delta\ \phi\left[\frac{x-\mu_1}{\sigma_1}\right]/\sigma_1 + O\left(\delta^2\right)$$

and

$$\Phi\left[\frac{x+\delta-\mu_2}{\sigma_2}\right] - \Phi\left[\frac{x-\delta-\mu_2}{\sigma_2}\right] = 2\delta\ \phi\left[\frac{x-\mu_2}{\sigma_2}\right]/\sigma_2 + O\left(\delta^2\right),$$

where $\phi[\cdot]$ is the standard normal probability density function. Combining the above results and letting $\delta \to 0$ yields

$$\text{limit}_{\delta\to 0}\ P\left(X \text{ originated from } N(\mu_1, \sigma_1^2) \mid X \in [x-\delta, x+\delta]\right)$$
$$= \frac{p_1\ \phi\left[\frac{x-\mu_1}{\sigma_1}\right]/\sigma_1}{p_1\ \phi\left[\frac{x-\mu_1}{\sigma_1}\right]/\sigma_1 + p_2\ \phi\left[\frac{x-\mu_2}{\sigma_2}\right]/\sigma_2},$$

which, upon cancellation of common $\sqrt{2\pi}$ factors, is formula (2) in Lecture 7.

If $|x-\mu_1|/\sigma_1 = |x-\mu_2|/\sigma_2$, then formula (2) simplifes to

$$\frac{p_1/\sigma_1}{p_1/\sigma_1 + p_2/\sigma_2}.$$

This equals 0.5 if and only if $p_1 = \sigma_1/(\sigma_1+\sigma_2)$.

2. Groups I, II, and III consist of Female, Weight, QRSDur, TInt, and PInt.

3a. Groups I, II, and III: The multivariate normality assumption is not tenable because both quantile-quantile plots (one for ArrhythmiaAny = 0 and one for ArrhythmiaAny = 1) exhibit marked departures from straight-line patterns.

Group IV: The multivariate normality assumption is not tenable because both quantile-quantile plots (one for ArrhythmiaAny = 0 and one for ArrhythmiaAny = 1) exhibit marked departures from straight-line patterns. (These plots look worse than the ones for Groups I, II, and III.) In fact, the multivariate normality assumption cannot possibly be true because Female is dichotomous and hence not normally distributed within either stratum of ArrhythmiaAny.

3b. Groups I, II, and III: Let $S_X$ denote the standardized version of $X$. The first canonical variable is

$$0.4198 S_{Female} + 0.3238 S_{Weight} - 0.5897 S_{QRSDur} - 0.4033 S_{TInt} + 0.3870 S_{PInt}.$$

Group IV: The first canonical variable is

$$-0.0818 S_{Age} + 0.4024 S_{Female} - 0.1280 S_{Height} + 0.3886 S_{Weight} - 0.5749 S_{QRSDur}$$

$$+0.1207 S_{PRInt} - 0.0687 S_{QTInt} - 0.3947 S_{TInt} + 0.3428 S_{PInt} - 0.2698 S_{HeartRate}.$$

3c. Groups I, II, and III: The prior-weighted error rate on the training subset using cross validation is 0.3394.

Group IV: The prior-weighted error rate on the training subset using cross validation is 0.3136.

3d. Groups I, II, and III: The prior-weighted error rate on the validation subset is 0.3507.

Group IV: The prior-weighted error rate on the validation subset is 0.3025.

4. Since the prior-weighted error rate on the validation subset is smaller for Group IV than for Groups I, II, and III, we will use Group IV to make predictions. The prior-weighted error rate on the test subset is 0.3786, which is not as good as the misclassification rate of 0.3393 obtained from my preferred logistic regression model in Written Assignment 3.

5. [Furnish your diagram.] My classification tree had six terminal nodes and consulted the following explanatory variables: QRSDur, HeartRate, Age, TInt.

6. My classification tree had a misclassification rate of 0.2679 on the test subset, which was somewhat less than the misclassification rate from my logistic regression model and much less than the prior-weighted error rate from my discriminant analysis. Among the three supervised learning techniques presented in Lectures 6 through 8, the classification tree makes the best predictions on the test subset.

7a. The estimated probability of arrhythmia for someone whose profile on the explanatory variables matches that of the person with ID = 5 is 0.18447. This had to be determined using a surrogate split since the person with ID = 5 had a missing value for HeartRate.

7b. The estimated probability of arrhythmia for someone whose profile on the explanatory variables matches that of the person with ID = 317 is 0.88889.

7c. My preferred logistic regression model from Written Assignment 3 did not include Height as an explanatory variable. If Height had been included, then the fitted probability of arrhythmia would have been very close to 1 or 0 for someone whose profile matched that of the person with ID = 317, depending on whether the estimated partial slope coefficient for Height had been positive or negative.

This does not seem reasonable, as we should feel somewhat uncertain rather than almost completely certain about a prediction for someone who had an obviously incorrect value on one of the explanatory variables.

7d. Neglecting the possibility of missing values necessitating invocation of surrogate rules, the following explanatory variables are never consulted: Female, Height, Weight, PRInt, QTInt, PInt.

8. We can achieve a sensitivity greater than 90% (in fact, equal to 100%) by choosing a threshold less than or equal to 0.15, but this yields a specificity of 0%. Raising the threshold to 0.20 yields a sensitivity less than 90%.

Why does this occur? Sensitivity and specificity are not smooth functions of the threshold for a classification tree because there are large jumps in sensitivity and specificity at the fitted probabilities appearing inside the terminal nodes (18%, 69%, 80%, 83%, and 89% for my classification tree).