STAT 315 Assignment 4

May 25, 2014

Lab: 26 May. Due Date: 30 May, 4pm. Submit via Learn. Maximum possible marks: 12.
Save the files `Pima.txt` and `Pima_test.txt` and the SAS file `a4code.sas` in the folder `P:\stat315`. Note: please attempt the questions before the lab, as otherwise you won't have much time. Questions 1 (a) and (b) do not require SAS.

1. In a two-class classification problem, the prior probabilities for class membership are $\pi_1 = \pi_2 = 0.5$. The density of class 1 is

   $$f_1(x) = e^{-x} \qquad x \geq 0$$

   and the density of class 2 is

   $$f_2(x) = \begin{cases} 1 - \frac{1}{2}x, & 0 \leq x \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

   The two solutions to the equation $f_1(x) = f_2(x)$ are $x = 0$ and $x = 1.5936$.

   (a) Find the Bayes optimal classifier. Your answer should have the form "Classify to class 1 if (condition) else classify to class 2." Sketching $f_1(x)$ and $f_2(x)$ will help. (2)

   (b) Calculate the Bayes rate. (2)

   (c) The SAS code simulates a data set of 50 observations from both classes and an independent test set of 200 observations. It fits the "Harvard model" (linear probability model) and a logistic regression and calculates the error rate for each model. What are the error rates? How do they compare with the Bayes rate? (1)

   (d) Calculate the error rates of the two models for 10 different values of the random seed by changing the number in `call streaminit`. Is there evidence of a difference in the predictive performance of the two models? (2)

2. The script `Pima.sas` loads in training and test datasets relating to diabetes in Pima women. The variables are: number of pregnancies `npreg`, glucose `glu`, blood pressure `bp`, and others. The category is `type`=1 if diabetes is present and 0 otherwise.

   (a) The script fits an LDA model and calculates the training error (called resubstitution error in the output), cross-validation error, and error on a test set. Write down the three errors. (1)

   (b) Which kind of misclassification is more common in the test data: patients with diabetes misclassified as healthy, or healthy patients misclassified as having diabetes? (1)

   (c) Fit a QDA model by changing the line `pool=yes` in `proc discrim` to `pool=no`. Write down the same three errors as in part (a) for the QDA model. (1)

   (d) A health organisation wants you to recommend one of these models for diagnosing diabetes. What would you tell them? Explain your decision in a way that a non-statistician could understand. (2)