STAT 315 Assignment 4

May 31, 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)

   We need to find where $f_1(x) = f_2(x)$. This is given as $x = 1.5936$ in the question. For $x > 1.5936$ we have $e^{-x} > 0 = f_2(x)$ so we should classify to class 1 if $x > 1.5936$. Similarly, classify to class 2 if $x < 1.5936$.

   (b) Calculate the Bayes rate. (2)

   The Bayes rate is the proportion of errors made by the optimal classifier. One kind of error is to misclassify an observation into class 2 when it's actually in class 1. The probability of this happening is

$$\int_0^{1.5936} e^{-x} dx = -e^{-x}|_0^{1.5936} = -e^{-1.5936} - (-e^{-0}) = 1 - e^{-1.5936} = 0.797.$$

   The probability of the other type of error is

$$\int_{1.5936}^{2} (1 - 0.5x) dx = x - x^2/4|_{1.5936}^{2} = (2 - 2^2/4) - (1.5936 - (1.5936)^2/4) = 0.0413.$$

   Half of the population consists of each class, so the overall rate of misclassifications is

$$0.5(0.797) + 0.5(0.413) = 0.419.$$

   (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)

   0.445 for Harvard and 0.44 for logistic. They are bigger than the Bayes rate (as they should be!)

(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)

Answers will vary. I choose 41,42,43,... 50 as the seeds and get error rates of:

Harvard	Logistic	difference
0.46	0.46	0
0.48	0.485	-0.05
0.42	0.43	-0.01
0.43	0.435	-0.05
0.47	0.46	0.01
0.47	0.47	0
0.55	0.55	0
0.47	0.46	0.01
0.44	0.44	0
0.45	0.45	0

The p–value for a t–test that the mean is different from 0 looks to be 0.23, which is huge. So there is no evidence of a difference.

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)

$0.230, 0.245, 0.205$

(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)

In the *test data* 0.1121 of healthy are misclassified as diabetic. But 0.3853 of diabetics are misclassified as healthy. So more diabetics are misclassified.

(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)

$0.230, 0.265, 0.2324$.

(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)

I recommend LDA. The LDA model performs better on unseen data, both under cross-validation and on an unseen test set. The QDA model is more complicated, but it is fitting the "noise" in the data too closely, whereas the LDA model is capturing the "signal".