UNIVERSITÉ BISHOP'S UNIVERSITY

Summer Project

Madjid Allili ∼ J125 ∼ ext. 2740 ∼ mallili@ubishops.ca

Project (Statistical Learning using R)

1. The data set `lowbwt.txt` contains information for a sample of 100 low birth weight infants. The variables are

   `sbp` : maternal systolic blood pressure

   `sex` : gender of the baby

   `toxemia` : toxemia during pregnancy (yes or no)

   `germ.hem` : germinal matrix hemorrhage (yes or no)

   `gest.age` : gestational age in weeks

   `apgar5` : five-minute `APGAR` score

   (a) Using germinal matrix hemorrhage as the response, fit a logistic regression model where the predictor variable x_1 is the 5-minute `APGAR` score. Write the equation and interpret β_1, the estimated coefficient of `Apgar` score.

   (b) What is the estimate and 95% confidence interval for the slope (coefficient for apgar5) in the odds ratio scale? Interpret the estimate (what does the odds ratio mean?).

   (c) At the 0.05 level of significance, test the null hypothesis: $H_0 : \beta_1 = 0$ where β_1 is the coefficient for `apgar5`.

   (d) If a new infant from the population has an `APGAR` score of 3 what is the predicted probability that this child will experience a brain hemorrhage? What is the probability if the child's score is 7?

2. In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the `Auto data` set.

   (a) Create a binary variable, `mpg01`, that contains a 1 if `mpg` contains a value above its median, and a 0 if `mpg` contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the `data.frame()` function to create a single data set containing both `mpg01` and the other Auto variables.

(b) Explore the data graphically in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting `mpg01`? `Scatterplots` and `boxplots` may be useful tools to answer this question. Describe your findings.

(c) Split the data into a training set and a test set.

(d) Perform `LDA` on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

(f) Perform logistic regression on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

(g) Perform `KNN` on the training data, with several values of K, in order to predict `mpg01`. Use only the variables that seemed most associated with `mpg01` in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?