UCT MA Psychology Multivariate Statistics

Assignment 5

==========

1 RA Fisher devised the classification method known as 'linear discriminant analysis' (also known as Fisher discriminant analysis), and used a botanical data set recording measures of sepal and petal length and width in three species of iris. This dataset ('iris') is available in the base installation of R

a. Open the data set in R. Use summary statistics, and plotting methods to explore whether the iris species are likely to be easily discriminable on the basis of the variables in the dataset. (10)

b. Run a linear discriminant analysis, attempting to classify iris species on the basis of the four measured variables in the dataset (one package containing LDA functions is MASS). (5)

c. What is your success rate? Use the 'confusionMatrix' function in caret to compute success rate. (5)

d. Use the plotting methods built into LDA in MASS to explore the adequacy of the classification. What do they tell you? (5)

e. Note that LDA has a form of cross validation built into it as an option. Use it, and compare the classification rate success (NOTE: LDA is somewhat inconsistent in how it works, you will need to experiment with the predict method, the class object (element of the object created by LDA) etc. in order to accomplish this). (5)

f. Attempt b and c for the function qda, which implements quadratic discriminant analysis. (10)

2 You will find an Excel file on VULA called titanic3.xls These data record various attributes of passengers on the ill-fated Titanic voyage in 1912. A description of the data set is given overleaf.

a. Your task is to build a predictive model of survival status of passengers. You are free to consider any of the variables that appear in the data file. Split the data into training and test sets first, and then explore the training data to identify possible variables and their interactions for building a model. Keep the model relatively simple e.g. limit yourself to four predictors and some combinaton of their interaction(s). (25)

b. Test your model on the test set by computing and comparing MSE train and MSE test. (5)

c. Re-run the analysis above (i.e. split the dataset into training and test sets, use the model you developed in a) and b) to compute MSE of the model as applied to the test set only), 1000 times (inside a loop). Gather the accuracy rates and plot them as a histogram. What do you conclude? (10)

d. Conduct LOOCV using the cv.glm function (in package boot). Note the cost function that you need to use. How does the result compare to what you found in b)? (10)

e. Conduct 2 fold, 5 fold, and 10 fold cross validation using the cv.glm function. Set the random seed to 199 so that results are replicable. How do the results compare to what you found in c), b), and a)? (10)

DESCRIPTIVE ABSTRACT: The titanic3 data frame describes the survival status of individual passengers on the Titanic. The titanic3 data frame does not contain information for the crew, but it does contain actual and estimated ages for almost 80% of the passengers.

SOURCES: Hind, Philip. "Encyclopedia Titanica." Online. Internet. n.p. 02 Aug 1999. Avaliable http://atschool.eduweb.co.uk/phind

VARIABLE DESCRIPTIONS:

pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
survival	Survival (0 = No; 1 = Yes)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat
body	Body Identification Number
home.dest	Home/Destination

SPECIAL NOTES:
Pclass is a proxy for socio-economic status (SES)
1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)
If the Age is Estimated, it is in the form xx.5

Fare is in Pre-1970 British Pounds (£)
Conversion Factors: 1£ = 12s = 240d and 1s = 20d

With respect to the family relation variables (i.e. sibsp and parch) some relations were ignored. The following are the definitions used for sibsp and parch.

Sibling: Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic
Spouse: Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiancées Ignored)
Parent: Mother or Father of Passenger Aboard Titanic
Child: Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

Other family relatives excluded from this study include cousins, nephews/nieces, aunts/uncles, and in-laws. Some children travelled only with a nanny, therefore parch=0 for them. As well, some travelled with very close friends or neighbors in a village, however, the definitions do not support such relations.