STAT-S 432: Homework 5

Due April 12, 2019 by 11:59 PM

Instructions: You must submit this homework by pushing a file named "hw5.Rmd" file to your team's repo. Note that that is the **only** file you will be allowed to push. Commit early and often.

Data is given by the file PimaIndians.csv

You will be investigating data pertaining to the Pima Indians. The Pima are a group of native Americans living in what is now central and southern Arizona. The Pima Indians of Arizona have the highest rate of obesity and diabetes ever recorded, and since they have the willingness to help the research process, the National Institute of Diabetes and Digestive and Kidney have been able to collect the data about the Pima's group (only women are included in this study).

Variable list and descriptions:

- `pregnant` It represents the number of times the woman got pregnant during her life.
- `glucose` It represents the plasma glucose concentration at 2 hours in an oral glucose tolerance test.
- `diastolic` The diastolic which is in the fact the pressure in (mm/Hg) when the heart relaxed after the contraction.
- `triceps` It is a value used to estimate body fat (mm) which is measured on the right arm halfway between the olecranon process of the elbow and the acromial process of the scapula.
- `insulin` It represents the rate of insulin 2 hours serum insulin (mlIU/ml).
- `bmi` It represents the Body Mass Index (weight in kg / (height in meters squared), and is an indicator of the health of a person.
- `diabetes` It is an indicator of history of diabetes in the family.
- `age` It represents the age in years of the Pima's woman.
- `test` It can take only 2 values ('negatif' or 'positif') and represents if the patient shows signs of diabetes.

1. As usual, perform exploratory data anaylsis. Explore the data graphically in order to investigate the association between `test` and the other variables. Which of the other variables seem most likey to be useful in predicting `test`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings. It is best to be thorough. Make sure to explain your findings. Only include graphs and explanations related to variables that seem best for `test`.

- Choose a subset of the variables that seem most associated with `test`. This subset may change depending on the type of model. (Hint: Which type of model(s) can handle categorical predictors?)

2. Split the data into a training and test set. Use a 50/50 (random!) split. Using the selected subset(s) of variables on the training data, use the following methods for predicting whether a Pima woman shows signs of diabetes or not:

- LDA
- QDA
- Logistic Regression
- KNN (use CV to choose best value for k)

3. For each method:

- Get the misclassification rates for the test and training data.
- Which model gives the best training error?
- Which model gives the best testing error?

4. For the best model:

- Report the confusion matrix.
- Explain the sensitivy and specificity of the model.
- Compare the model to the model that simply predicts that no one shows signs of diabetes.

- Describe how well or poorly the model predicts signs of diabetes.

Rubric

Scoring will be done according to the following criteria.

Each question or bullet point has been addressed. Questions are answered accurately. Any plot or table produced is explained and relationships are described accurately.

If document does not knit (assuming required packages are installed) -5

Problem 1 (7 points)

Problem 2 (7 points)

Problem 3 (4 points)

Problem 4 (7 points)