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`.

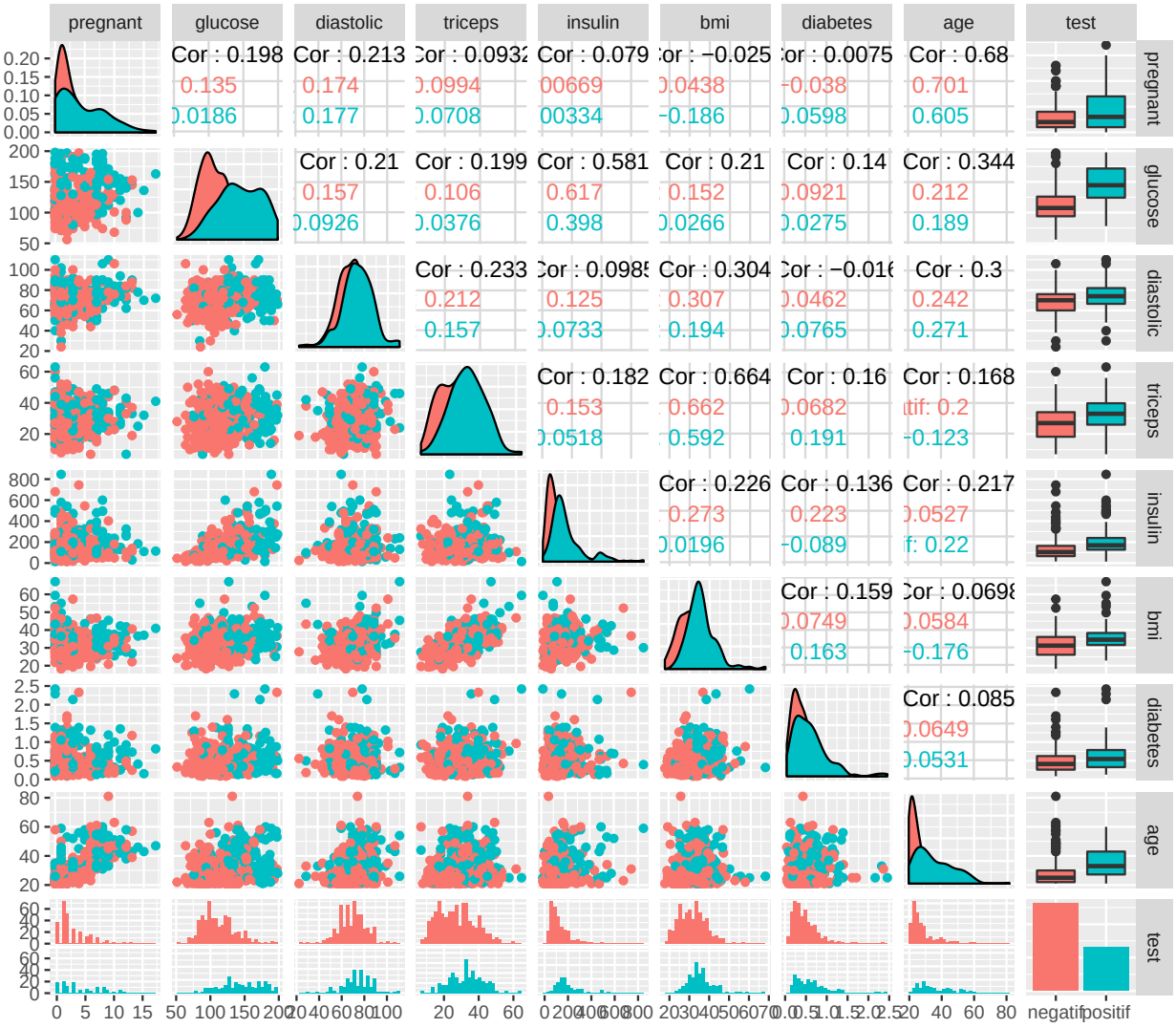
- Choose a subset of the variables that seem most associated with `test`. This subset may change depending on the type of model. (Hint: Which type of model(s) can handle categorical predictors?)

The dataset under analysis contains various medical measurements from 392. The objective is predict if someone shows signs of diabetes (`test` variable) via the other predictor variables: Pregnancy, glucose, diastolic blood pressure, triceps measure, insulin levels, BMI, family diabtese history, and age. Scatterplot matrices of the bivariate relationships between all variables is presented below.

The distributions of all the predictor variables has some difference when splitting the distributions based off diabetes status. Densities of the distributions when split by diabetes status are along the main diagonal. These differences are relatively weakest for diastolic blood pressure and triceps. Overall, those with diabetes show significant right skew across all predictor variables.

Overall, glucose levels (unsurprisingly), number of pregnancies, insulin levels (again, unsurprisingly), and age seem to have .

The absolute worst variable which shows no perceivable difference in distribution between the positive and negative diabetes group is the diastolic blood pressure.



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)

For prediction of whether an individual shows signs of diabetes or not, four different model types were tried: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression, and K-Nearest Neighbor (KNN).

To this end, the data were split into two equal sized groups, a training and a testing set. Model selection will be done via testing error.

Experimentation was done for each type of model to determine which variables produce the best testing error. Your choice a variables and why they were used may vary.

For the LDA model, we assume that the predictors follow a multivariate normal distributions in the two diabetes diagnosis groups; additionally, the covariance matrices between the two groups are assumed equivalent.

For this reason, the number of pregnancies was excluded from modeling since it had a very right skewed distribution, was discrete, and was bounded below by 0. The best LDA model found was predicting the diabetes diagnosis using glucose and log of age. A log transformation was used because the distribution was right skewed and the transformation made the age distribution much more normal.

The QDA model is similar to LDA but we drop the assumption of equal covariance matrices. The only resulting difference in chosen predictor variables is that in addition to the glucose levels and log of age, the tricep measurement was also used.

For the logistic regression model, the model chosen used glucose levels, insulin levels, BMI, and age. We are able to use more variables in the logistic regression model because no strong assumptions are made about the distribution of the predictors.

For the KNN model, the model chosen used the number of pregnancies, glucose levels, BMI, and diabetes family history.

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?

Error.Type	LDA	QDA	Logistic	KNN
Training	0.250	0.250	0.245	0.250
Testing	0.179	0.179	0.204	0.209

The above table summarizes the training and testing errors for each model type. The best training error of 0.245 was produced by the KNN algorithm, but the best testing error of 0.179 (our model selection criterion) was produced by the LDA and QDA models. With regards to the principle of parsimony, the selected final model will be the LDA model which makes use of only two predictors.

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.

Below is a summary of testing error and accuracy of the LDA model. The confusion matrix shows that out of the 58 individuals that were truly showing signs of diabetes, only 38 of them were correctly identified by the model. The sensitivity of the model is then 0.655 which is a bit better than predicting if someone has signs of diabetes by chance, but not that much better. The model performs much better at predicting if someone does not show signs of diabetes. Of the 138 individuals that did not show signs, 123 were correctly categorized which gives us a specificity of 0.891. The overall accuracy was 0.821

If we were to predict that no one showed signs of diabetes, the model accuracy would still be 0.701, which is not all that much lower than the accuracy of the LDA model at 0.821. A 95% confidence interval for the LDA model accuracy is 0.761 to 0.872, and for the null model is 0.635 to 0.767. There is a slight overlap which is concerning. Overall, our model is not very good classifying individuals with diabetes or not.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction negatif positif
##    negatif     123      20
##    positif      15      38
##
```

```
##                Accuracy : 0.821
##                  95% CI : (0.761, 0.872)
##     No Information Rate : 0.704
##     P-Value [Acc > NIR] : 0.000117
##
##                   Kappa : 0.56
##
##  Mcnemar's Test P-Value : 0.498962
##
##             Sensitivity : 0.655
##             Specificity : 0.891
##          Pos Pred Value : 0.717
##          Neg Pred Value : 0.860
##              Prevalence : 0.296
##          Detection Rate : 0.194
##    Detection Prevalence : 0.270
##       Balanced Accuracy : 0.773
##
##        'Positive' Class : positif
##
```

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)