Statistical Foundations of Business Analytics
46-883

Carnegie Mellon University
Tepper School of Business

Homework 4

The data in `diabetes.csv` - also hosted at `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database` - contains information about female patients of Pima Indian heritage who are at least 21 years old. The data contains the following variables:

- `Pregnancies`: number of pregnancies experienced by the patient
- `Glucose`: the plasma glucose concentration measured from an oral glucose tolerance test (in mg/dL)
- `BloodPressure`: the patient's diastolic blood pressure (in mmHg)
- `SkinThickness`: the skin fold thickness of the patient's triceps (in mm)
- `Insulin`: the patient's serum insulin level (in μU/ml)
- `BMI`: the patient's Body Mass Index (in kg/m^2)
- `DiabetesPedigreeFunction`: a measure of the likelihood that the patient will develop diabetes based on family history
- `Age`: the patient's age (in completed years)
- `Outcome`: whether or not the patient was diagnosed with diabetes (1: diagnosed with diabetes, 0: not diagnosed with diabetes).

Question 1 - 5 points

Load the data contained in the `diabetes.csv` file in R.

Question 2 - 10 points

Replicate the logic used in the `class8.r` file to divide the data in a train, validation and test set. Use a 40% - 30% - 30% split.

Question 3 - 5 points

Using all available predictors, fit to the training set:

- a classifier based on logistic regression
- an LDA classifier
- a QDA classifier
- a Naive Bayes classifier.

Question 4 - 15 points

A group of physician asks you to produce a classifier that achieves 85% Sensitivity when used to test new Pima Indian female patients for diabetes. Using the validation set

- plot the ROC curves for the models you built
- use the `roc` function of the `pROC` library to find - for each of the models you built - the largest threshold t that makes your model achieves at least 90% Sensitivity (just in case, we build some extra margin here to stay a little conservative and make it more likely that we can hit the target Sensitivity goal)
- which model performs best (i.e., achieves the largest Specificity) under these conditions?

For the second and third bullet point, you can use this kind of logic:

```
# set target sensitivity
target_sensitivity <- 0.90

# calculate the ROC curve
logistic_diabetes_roc <- roc(
    diabetes_validation$Outcome,
    predict(logistic_diabetes, diabetes_validation, type = "response")
)
# find the largest threshold t that achieves the target sensitivity
logistic_diabetes_roc_index <- (
    which.max(logistic_diabetes_roc$sensitivities < target_sensitivity) - 1
)
logistic_diabetes_t <- logistic_diabetes_roc$threshold[
    logistic_diabetes_roc_index
]
# find the specificity of the model at this threshold
logistic_diabetes_roc$specificities[logistic_diabetes_roc_index]
```

Question 5 - 15 points

How different are the ROC curves of the classifier obtained by means of logistic regression and of the LDA classifier? Are you surprised by this result? Explain.

Question 6 - 10 points

Evaluate the winner model of Question 4 on the test set using the `confusionMatrix` function. You will need to use the threshold that you computed for this model in Question 4. Does this model seem to satisfy the Sensitivity requirement that the physicians shared with you?

Question 7 - 5 points

What is your best estimate about the Specificity that your model will achieve on future patients?

Question 8 - 20 points

Fit a knn classifier to the training data and tune the parameter k of your knn classifier using the validation set in such a way that k maximizes the Sensitivity of the classifier on the validation set. You can look back at the `class8.r` file that we discussed in class and adapt the code from there.

Question 9 - 5 points

What is the best value of k on these data based on your tuning?

Question 10 - 10 points

Evaluate the knn model on the test set using the `confusionMatrix` function. Does this knn model perform better or worse than the winner model of Question 4? Which model will you share with the physician to help them diagnose diabetes on future female Pima Indian patients?