CSCE 550 Project1

Due: September 26, 2019

1- Objective

In this assignment, we build a classifier using Naïve Bayesian technique that we investigated in the class.

2- Dependencies

Python > 2.7
Pandas toolkit (If it is needed)
Numpy toolkit (If it is needed)

Note: Using Python's Scikit-learn tool **is not allowed** for this assignment

3- Dataset

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Note: You do not need to download the data. Two subsets of data for training and test is created and posted for downloading. The original data set also is included for your further testing and experimenting.

The list of the fields in order of columns in the data file is:
PregnanciesNumber of times pregnant
GlucosePlasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressureDiastolic blood pressure (mm Hg)
SkinThicknessTriceps skin fold thickness (mm)
Insulin2-Hour serum insulin (mu U/ml)
BMIBody mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunctionDiabetes pedigree function
Age (years)
OutcomeClass variable (0 No Diabetes or 1 Diabetes) 268 of 768 are 1, the others are 0.

For more information see
https://www.kaggle.com/uciml/pima-indians-diabetes-database

Publication related to this data
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/pdf/procascamc00018-0276.pdf

It will be beneficial if you review this publication.

4- Classifier program guideline

The guideline to develop your classifier is as follows. You may use your own way to break the problem in any format/function you like.

a. A function that uses Pandas API to load files (training and test) as a Panda's data frame
b. A function that calculates μ and σ for each column
c. A function for calculating normal distribution likelihood
d. A classifier function
e. An accuracy function to count/estimate accuracy (and hence error)

5- What we need to generate?

Sets of training and test data are posted online and you should use the given data for your training and test. Your classifier needs to generate the following results:

a- An evaluation of accuracy by counting classified and misclassified points
b- An evaluation of accuracy by using confusion matrix as follows:

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. An example of a confusion matrix is shown below.

n=192	**Predicted: 0**	**Predicted: 1**	
Actual: 0	TN = 118	FP = 12	130
Actual: 1	FN = 47	TP = 15	62
	165	27	

After you construct the confusion matrix, calculate the values below:

Classification Error: Overall, how often is the classifier correct?

Accuracy = (TP + TN) / (TP + FP + TN +FN)

Classifier Error: Overall, how often is the classifier correct?

Error = (FP + FN) / (TP + FP + TN +FN)

Classifier Sensitivity: When the actual value is positive, how often is the prediction correct?

Sensitivity = TP / (FN + TP)

Classifier Specificity: When actual value is negative, how often is the prediction correct?

Specificity = TN / (TN/FP)

What is your interpretation of these values?