SathyaNarayanan Srinivasan

Sxs142031@utdallas.edu

Abstract

Implementing various classification techniques on PIMA dataset

CLAssification techniques

Machine Learning Assignment - 3

Table of Contents

[Using R 2](#_Toc432782414)

[Requirements 2](#_Toc432782415)

[Prerequisites 2](#_Toc432782416)

[Packages Used 2](#_Toc432782417)

[Compiling and Running Instructions 3](#_Toc432782418)

[Output 4](#_Toc432782419)

[Part 1. Exploratory Data Analysis 5](#_Toc432782420)

[A. Skewness and Kurtosis Tests 5](#_Toc432782421)

[B. Shapiro-Wilk Tests 6](#_Toc432782422)

[C. The Lilliefors (Kolmogorov-Smirnov) test 7](#_Toc432782423)

[D. Anderson – Darling Test 7](#_Toc432782424)

[Observation 7](#_Toc432782425)

[Part 2. Naïve Bayesian Classifier 8](#_Toc432782426)

[Part 3. SVM 9](#_Toc432782427)

[Part 4. KNN 9](#_Toc432782428)

[Result 9](#_Toc432782429)

# Using R

# Requirements

* **Operating System** – Windows 7 or Higher
* **Product Version :** [**R-3.2.2**](https://cran.r-project.org/bin/windows/base/R-3.2.2-win.exe)
* **PDF Processor** – Adobe PDF Reader

# Prerequisites

* Please ensure that R ver. 3.2.2 is installed properly and working in the machine.
* Open Command Prompt in ELEVATED PRIVILEGES mode.
* Please ensure that rscript is accessible in the command prompt.i.e. When rscript is typed in the command prompt the system recognizes that it is either an internal or an external command. If not, please follow the steps here : [Debug Link](http://stackoverflow.com/questions/17339438/r-script-from-command-line)
* The program runs in RStudio also.

# Packages Used

* caTools
* caret
* partykit
* klaR
* e1071
* stargazer
* ISLR
* DMwR
* Moments
* nortest
* fitdistrplus
* gplots
* gridExtra
* grid
* gtable

# Compiling and Running Instructions

* In a windows command prompt :
  + Navigate to the R folder
  + Type rscript <full path to the file Assignment3.R> with the .R extension followed by - Each of them is the absolute path to the file on the PC.
  + This assignment has four folders : ExploratoryDataAnalysis , kNN, NaiveBayesian and SVM
  + Navigate to each folder and apply the steps given above to run the project

For Eg : D:\DriveFiles\@UTD\Repositories\Assignment 3\ExploratoryDataAnalysis> rscript Expl\_Data.R

* + Press Enter
  + Wait for the packages to be installed – it will take a while depending on many factors
  + The output is ordered according the question of assignment 3.
* In RStudio:
  + Navigate to the solution folder
  + This assignment has four folders : ExploratoryDataAnalysis , kNN, NaiveBayesian and SVM
  + Navigate to each folder and open the .R file
  + Set the working directory to the folder opened
  + Source the R code and Run
  + Wait for the packages to be installed – it will take a while depending on many factors
  + The output is ordered according the question of assignment 3.

# Output

The output contains two files :

1. **Measurements.pdf** – Used to record the various tables and values generated by different functions
2. **Plots.pdf** – Used to plot the various graphs associated with the project.

# Part 1. Exploratory Data Analysis

1. Create following plots : Histogram and BarPlot

* Please refer to the **Plots.pdf** file created in the working directory. It has three types of graphs – Histogram, Q-Q Norm graph and a Barplot.
* The Q-Q norm plot is done to find out how far each variable is from a Normal Distribution. According to the property of the Q-Q Norm graph – more linear values of a Q-Q Norm graph, the distribution will be tending more towards a normal distribution.

1. Write a short note on the distribution of the variables. Are they normally distributed?

I have performed the following tests and recorded the results in separate pages in the **Measurements.pdf** file.

## Skewness and Kurtosis Tests

Skewness: indicator used in distribution analysis as a sign of asymmetry and deviation from a normal distribution.   
  
Interpretation: 

* Skewness > 0 - Right skewed distribution - most values are concentrated on left of the mean, with extreme values to the right.
* Skewness < 0 - Left skewed distribution - most values are concentrated on the right of the mean, with extreme values to the left.
* Skewness = 0 - mean = median, the distribution is symmetrical around the mean.

Kurtosis - indicator used in distribution analysis as a sign of flattening or "peaked nature" of a distribution.   
  
Interpretation: 

* Kurtosis > 3 - Leptokurtic distribution, sharper than a normal distribution, with values concentrated around the mean and thicker tails. This means high probability for extreme values.
* Kurtosis < 3 - Platykurtic distribution, flatter than a normal distribution with a wider peak. The probability for extreme values is less than for a normal distribution, and the values are wider spread around the mean.
* Kurtosis = 3 - Mesokurtic distribution - normal distribution for example.

## Shapiro-Wilk Tests

* These test the NULL hypothesis that the samples came from a Normal distribution. This means that if your *p-value <= 0.05*, then you would *reject* the NULL hypothesis that the samples came from a Normal distribution. The shapiro-wilk test tests **against** theassumption of Normality.
* In case of the null hypothesis of shapiro-wilk, a p-value <= 0.05 would reject the null hypothesis that the samples come from normal distribution. To put it loosely, there is *a rare chance* that the samples came from a normal distribution. The side-effect of this hypothesis testing is that this rare chance happens very rarely.

## The Lilliefors (Kolmogorov-Smirnov) test

* If p-value is below < .05 the data does NOT come from a normally distributed population.  The null hypothesis is rejected and that the data is normally distributed.

## Anderson – Darling Test

* The Anderson-Darling test is used to test if a sample of data came from a population with a specific distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution.
* P-value < 0.05 = not normal. normal = P-value >= 0.05

## Observation

**Normal Distribution - Identification Per Attribute**

According to the tests performed above and comparing recorded values against the threshold values from various tests – I concluded that these variables are close to a normal distribution: **[V2 (or) Glucose] and [V3 (or) Blood] are very close to a normal distribution, whereas the [V6 (or) BMI] is closer to a normal distribution but exhibits positive skewness.**

**Type of Distribution**

According to the Cullen – Frey Graph – between square of skewness and kurtosis values – each attribute has been identified to belong to a distribution as denoted in the **Plots.pdf**

**However, I also understand that it is not ACCURATE as the graph gives only AN APPROXIMATION to the nearest possible distribution.**

**Correlation between all the attributes and the class variable**

The data has missing values. I replaced them with NA and while computing correlation – the results had NA. Hence, I took the dataset and omitting missing attributes I computed the correlation. Hence, I record two results.

I have performed the correlation and recorded the results separately in the **Measurements.pdf** file.

Highest Correlation considering missing attributes: **[V8 (or) AGE]**

Highest Correlation NOT considering missing attributes: **[V2 (or) GLUCOSE]**

**Correlation between pairs of attributes using for loop**

I have performed the correlation and recorded the results separately in the Measurements.pdf file.

Maximum Mutual - Correlation occurred between **[V1 (or) PREGNANT] and [V8 (or) AGE]**

# Part 2. Naïve Bayesian Classifier

All the steps were performed according to the assignment and the results are recorded in a separate file called **NBAccuracy.text**

# Part 3. SVM

All the steps were performed according to the assignment and the results are recorded in a separate file called **SVMAccuracy.text**

# Part 4. KNN

All the steps were performed according to the assignment and the results are recorded in a separate file called **KNNAccuracy.text**

# Result

I prefer the **SVM with Radial Kernel** – classification technique as it has an overall higher accuracy than the others.