Assignment 1

# Unsupervised Learning

Shubham Londhe

10th October, 2021

# Introduction

Machine learning definition given by tom Mitchell is “If we want the Experience E on task T for performance measure p, If it is performance on task T ,for performance P so it increase with the experience E”

## Question 1

Explain the Importance of Machine Learning Algorithms. In detail explain the importance of Weka Tool. Any of the ML algorithm implementations with the Tool. List Pros and Cons of the tool.

## Answer 1

Machine learning (ML) is the area of Artificial Intelligence under computational science that focuses on analyzing and interpreting patterns and structures in data to enable learning, reasoning, and decision making outside of human interaction.

In simple words, machine learning allows the user to feed a computer algorithm an immense amount of data and have the computer analyze and make data-driven recommendations and decisions based on only the input data. If any corrections are identified, the algorithm can incorporate that information to improve its future decision making.

WEKA is a powerful tool used for developing Machine Learning models that provides implementation of several most widely used ML algorithms.

Before these algorithms are applied to our dataset, it also allows you to preprocess the data. The types of algorithms that are supported are classified under Classify, Cluster, Associate, and Select attributes.

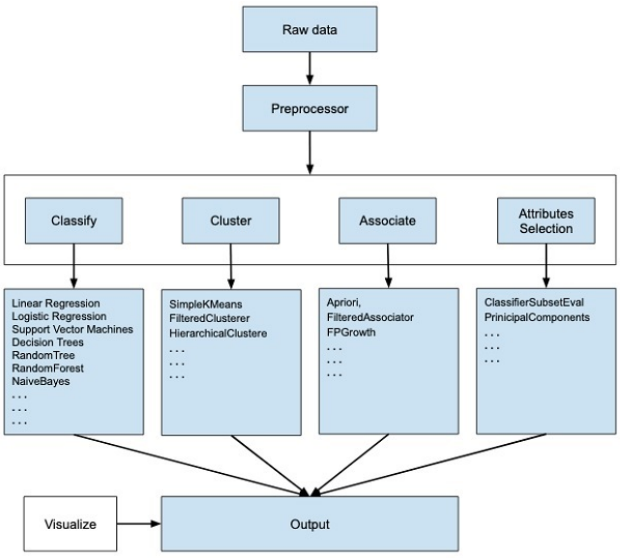
The result at various stages of processing can be visualized with a beautiful and powerful visual representation. This makes it easier for a Data Scientist to quickly apply the various machine learning techniques on his dataset, compare the results and create the best model for the final use.

## Pros of WEKA:

- The weka tool is an efficient tool with all the minimum amount of functionalities for implementing a Machine Learning model.
- The weka tool provides a great source of learning the algorithms and implementation of Machine Learning without learning Code.
- The weka tool interface is easy to use and performs the analysis like logistic regression, decision trees,etc .
- The most important benefit of the weka tool is the easy to analyse clusters and classify data giving us a better Understanding of the uses of various data sets.

## Cons of WEKA:

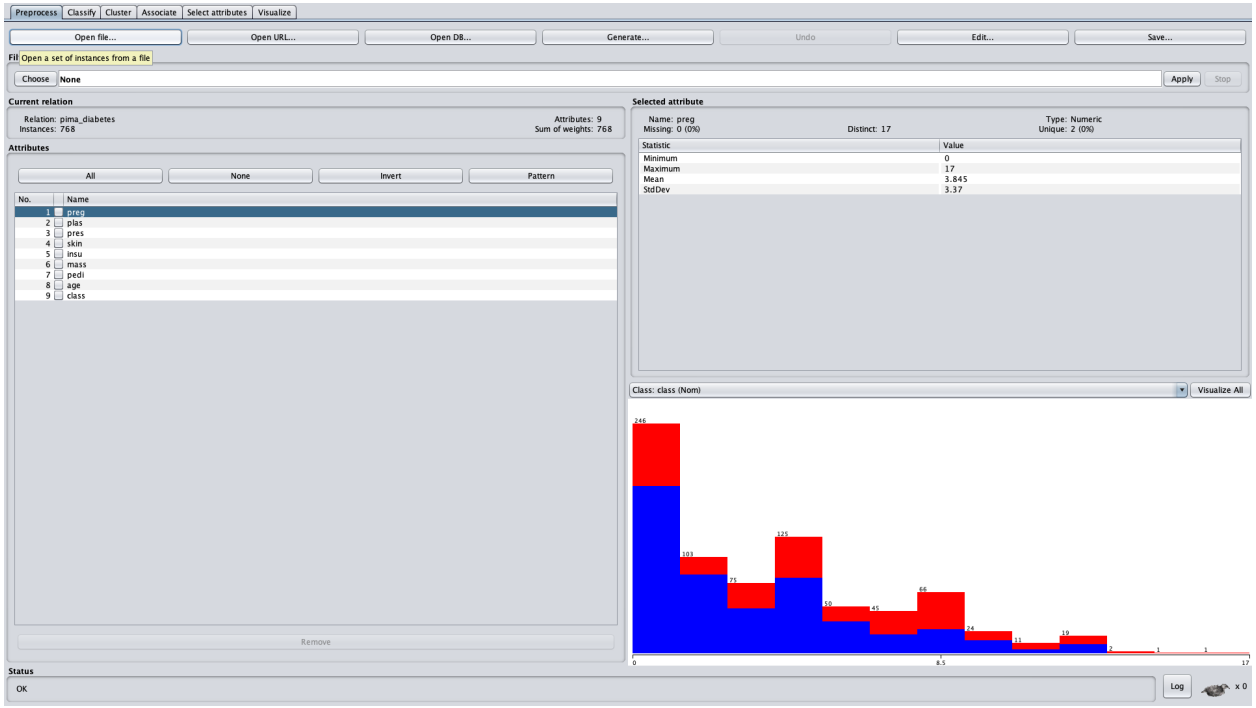
- The WEKA tool provides a limited amount of analysis features.
- Integrating WEKA with Python or R is a challenging task.
- WEKA can only handle small datasets, I tried giving a large dataset of images and the tool couldn't process them efficiently.
- Being not so popular in recent times, there is a limited amount of documentation and community support available on the internet on WEKA.



# The implementation

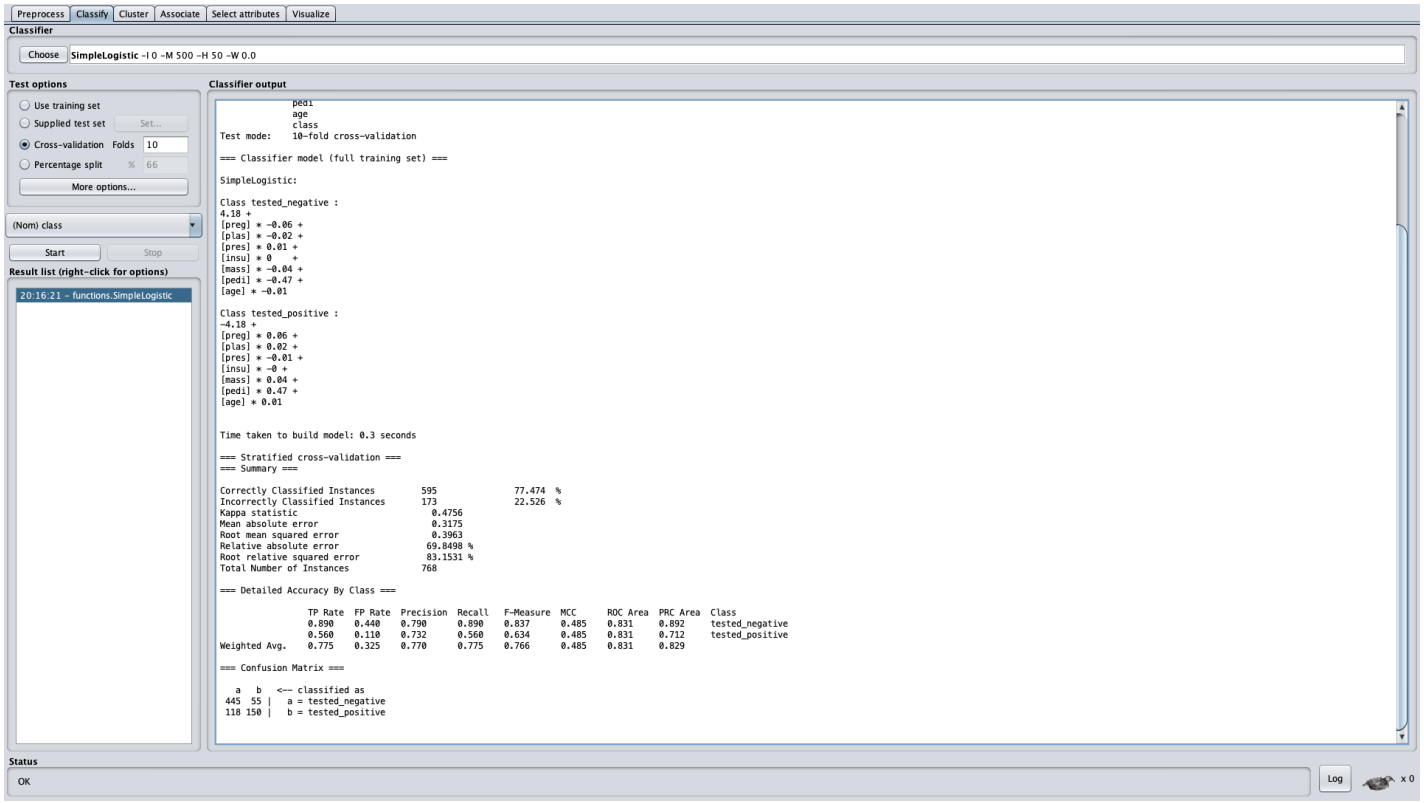
I applied Simple Logistic Regression on the dataset.

Downloading the dataset was done from here and then imported in the weka tool.




The dataset was easy to understand and I decided to use Simple Logistic Regression over it.

There are some estimated unique values from a set of independent variables. So It helps Us to predict the Probability of a task by fitting data to a logic function. So This process is called logistic regression.




The accuracy of the model and confusion matrix.

The WEKA tool gives all the details in a fast and efficient way making it a bit easier for the beginners to understand the concepts of logistic regression.

```
Time taken to build model: 0.3 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         595               77.474  %
Incorrectly Classified Instances       173               22.526  %
Kappa statistic                          0.4756
Mean absolute error                      0.3175
Root mean squared error                  0.3963
Relative absolute error                 69.8498 %
Root relative squared error             83.1531 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.890    0.440    0.790      0.890    0.837      0.485    0.831     0.892     tested_negative
                 0.560    0.110    0.732      0.560    0.634      0.485    0.831     0.712     tested_positive
Weighted Avg.    0.775    0.325    0.770      0.775    0.766      0.485    0.831     0.829

=== Confusion Matrix ===

   a   b   <-- classified as
 445  55 |   a = tested_negative
 118 150 |   b = tested_positive
```

## Question 2

List the importance of Dimensionality Reduction Techniques. Apply PCA on the following data set. State the Limitations of PCA Suppose two columns X and Y be the 2 feature

## Answer 2

Dimensionality reduction helps us with problems like when we are trying to store the most of the important information in the data so it is needed to learn The accurate and The predictive models.

There is too much information about the data to the prediction of the good model so this type of The factors are called features.

When There is a huge amount of data we have to work on and it's hard to get the visualization and the training of the data.

When the number of variables is in a very huge amount relative to the number of observations in your dataset, this type of the algorithm has difficulty training effective models So This is called the “Curse of Dimensionality.”

- Dimensionality Reduction reduces the time and storage space required.
- Dimensionality Reduction helps us with the Removal of the multicollinearity which helps us in the interpretation of the given model.
- Dimensionality Reduction technique Removes the unwanted data because it reduces the accuracy of the data and based on that makes our model to train on the unwanted data.

# Principal component Analysis

| X | Y |
|---|---|
| 1 | 4 |
| 2 | 3 |
| 3 | 4 |
| 4 | 6 |
| 5 | 8 |

**Step: 1**

Find the mean of X

$$= \frac{1+2+3+4+5}{5} = \frac{15}{5}$$

Find the mean of Y $= 3$

$$= \frac{4+3+4+6+8}{5} = \frac{25}{5}$$

$= 5$

**Step: 2**

Construct the covariance matrix

Formula,

$$C = \begin{bmatrix} cov(x,y), & cov(x,x) \\ cov(y,x), & cov(y,y) \end{bmatrix}$$

| X | Y | x-mean | $x^2$-mean | $\Sigma x^2$ mean |
|---|---|---|---|---|
| | | | $n-1$ | |
| 1 | 4 | $1-3=(-2)$ | 4 | |
| 2 | 3 | $2-3=(-1)$ | 1 | |
| 3 | 4 | $3-3=0$ | 0 | $\frac{4+1+0+1+4}{5-1}$ |
| 4 | 6 | $4-3=1$ | 1 | |
| 5 | 8 | $5-3=2$ | 4 | $=\frac{10}{4}$ |

$\therefore Cov(x,x) = 2.5$

| X | Y | x-mean | y-mean | $\Sigma \frac{\bar{x}-\bar{y}}{n-1}$ |
|---|---|---|---|---|
| 1 | 4 | $1-3=-2$ | $4-5=-1$ | |
| 2 | 3 | $2-3=-1$ | $3-5=-2$ | $-2\times-1=2$ |
| 3 | 4 | $3-3=0$ | $4-5=-1$ | $(-1)\times(-2)=2$ |
| 4 | 6 | $4-3=1$ | $6-5=1$ | $0\times(-1)=0$ |
| 5 | 8 | $5-3=2$ | $8-5=3$ | $1\times1=1$ |
| | | | | $3\times2=6$ |

$\therefore Cov(x,y), Cov(y,x) = 2.75$ $\quad \frac{2+2+0+1+6}{4} = 2.75$

| X | Y | $Y-\bar{Y}$ | $\bar{Y}2$ | $\Sigma \frac{y^2}{n-1}$ |
|---|---|---|---|---|
| 1 | 4 | $4-5=1$ | 1 | |
| 2 | 3 | $3-5=1$ | 4 | |
| 3 | 4 | $4-5=-1$ | 1 | $\frac{1+4+1+1+9}{4}$ |
| 4 | 6 | $6-5=1$ | 1 | |
| 5 | 8 | $8-9=3$ | 9 | $=4$ |

$\therefore Cov(Y,Y) = 4$

$$C = \begin{array}{c} \\ x \\ y \end{array} \begin{array}{c} \begin{array}{cc} x & y \end{array} \\ \begin{bmatrix} 2.5 & 2.75 \\ 2.75 & 4 \end{bmatrix} \end{array}$$

Step 3:

$$C - \lambda I = 0$$

$$\begin{bmatrix} 2.5 & 2.75 \\ 2.75 & 4 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

$$\begin{bmatrix} 2.5 & 2.75 \\ 2.75 & 4 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = 0$$

$$\begin{bmatrix} 2.5 - \lambda & 2.75 \\ 2.75 & 4 - \lambda \end{bmatrix} = 0$$

$$(2.5 - \lambda)(4 - \lambda) - (2.75)(2.75) = 0$$

$$10 - 2.5\lambda - 4\lambda - \lambda^2 - 7.5625 = 0$$

$$\lambda^2 - 6.5\lambda + 2.4375 = 0$$

$$a^2 + bx + c$$

$$a = 1$$

$$b = 6.5$$

$$c = 2.4375$$

$$\frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

$$= \frac{6.5 + \sqrt{(6.5)^2 - 4(1)(0.4375)}}{2(1)}$$

$$= \frac{6.5 + \sqrt{32.5}}{2(1)}$$

$$= \frac{6.5 + 5.7009}{2}$$

$$\lambda 1 = 6.100$$

---

$$\frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

$$= \frac{6.5 - \sqrt{32.5}}{2(1)}$$

$$= \frac{6.5 - 5.7009}{2}$$

$$\lambda 2 = 0.399$$

Step: 4

$$\begin{bmatrix} 2.5 & 2.75 \\ 2.75 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0.399 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$$2.5x_1 + 2.75y_1 = 0.399x_1 \quad \text{—①}$$

$$2.75x_1 + 4y_1 = 0.399y_1 \quad \text{—②}$$

①

$$2.5x_1 + 2.75y_1 = 0.399x_1$$

$$2.5x_1 - 0.399x_1 = -2.75y_1$$

$$y_1 = 1$$

$$2.101x_1 = -2.75 \quad (1)$$

$$x_1 = -1.31$$

★ Eigen vector for $0.4 = \lambda \begin{bmatrix} -1.31 \\ 1 \end{bmatrix}$

for

$$\lambda = 6.100$$

$$\begin{bmatrix} 2.5 & 2.75 \\ 2.75 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 6.100 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$$2.5x + 2.75y = 6.10x \quad -1$$

$$2.75x + 4y = 6.10y \quad -2$$

(1)

$$2.5 + 2.75 = 6.10x$$

$$2.75 = 6.10x - 2.5x$$

$$2.75 = 3.6x$$

$$x = \frac{2.75}{3.6} = 0.76$$

Eigen Vector $Y = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0.76 \\ 1 \end{bmatrix}$

for $\lambda = 6.10$

so Eigen vector are

$$\begin{bmatrix} -1.31 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.76 \\ 1 \end{bmatrix}$$

Step 6

for $\begin{bmatrix} -1.31 \\ 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$

$$\sqrt{(-1.31)^2+(1)^2} = \sqrt{2.7161} = 1.6481$$

New eigen values,

$$\frac{-1.31}{1.6481} = 0.6067$$

for $\begin{bmatrix} 0.76 \\ 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$

$$\sqrt{(0.76)^2+(1)^2} = \sqrt{1.5776} = 1.2560$$

$$\frac{0.76}{1.2560} = 0.6050 = \begin{bmatrix} 0.6050 \\ 0.7961 \end{bmatrix}$$

$$\frac{1}{1.2560} = 0.7961$$

## Limitations of PCA

To See in the Principal Component Analysis and produce meaningful output, we must first check whether some assumptions here hold limitation of PCA That the person running the analysis.

In that situation the PCA must be unique otherwise It The lack meaning.

Another limitation of the Principal Component Analysis Is the assumption of orthogonality.

If there is a large variance dataset in the structure, sometimes the structure is found to be in places with low variance.

Principal Component Analysis has the hard with working with not given data and outliers

## Conclusion

This assignment helped me to understand concepts like PCA and Dimensionality Reduction. The exposure of Weka tool is also an addition to my AI/ML Journey and it seems like a great tool to begin the data science journey.