FINAL PROJECT

Logistic Regression & Clustering

Charles Aevedo Díaz
Ingeniería de Sistemas
Universidad Tecnológica de Bolívar
Cartagena, Colombia
chad9591@gmail.com

I. ABSTRACT

This particular paper will be about implementing logistic regression for sorting out some data based in samples with known outputs which is known as supervised learning; but also it has a program that classify data with only known features but unknown outputs, it is known as unsupervised learning, and it will group the data into clusters where all the similar data will be.

keywords: Logistic regression, python, machine learning, program, artificial intelligence, scripts, clustering, k-mean, elbow method

II. INTRODUCTION

This project is intended to implement the methods of Artificial intelligence we have learned in this course, therefore, we have two different datasets available that were supplied by the professor with educational purposes.

The first dataset contains samples with two features and their respective outputs. We are going to use the logistic regression to generate a model that will be capable of classifying new data incomes that may appear, then we are going to measure the resulting model. Also, we are going to modify the model by implementing *poly transformation*.

Besides, there is the clustering program that will sort out the data into clusters by using the k-means method, which will group them into the number of clusters that were given. For choosing the right amount of clusters, we are going to use the elbow method.

III. THEORY

Logistic Regression:

This is a method that helps finding discrete results, or binary outputs, which means that is intended to be used in circumstances where the results can be classified into a certain amount of groups like "Yes/No", "To buy, To sell", etc. The output here is a probability that takes a particular value based on combination of values taken by the predictor.

Its name is based on the function used at the core of the method, *the logistic function,* also called *sigmoid function* was developed by statisticians in order to describe properties of population growth in ecology. This is an S-shaped curve that take any real value number and gives as result a value between 0 and 1, but never exactly at those limits.

Clustering:

This is a Bottom-up or Hierarchical method, that has n clusters, and the entities are grouped into their own clusters based in the distance between entity and the centroid of clusters. In this case, the distance metric is the Euclidean distance.

The entity will be *send* to the closest cluster, and when we finish with all clusters every entity will have a cluster, and each cluster will have their entities. For initializing the clusters it will be random.

Cost Fuction

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y_i(\log h(x_i)) + (1-y_i)log(1-h(x_i)) + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

IV. PROCEDURE

Logistic Regression:

- The dataset needs to be splitted in two groups, for training purpose; in our case we have decided to split it with 70/30 percent relationship.
- We test the model by predicting results with it.
- We verify if the predictions are good or not. We do this by applying some metrics to the results like: accuracy, recall, precision and f1 score.

Clustering:

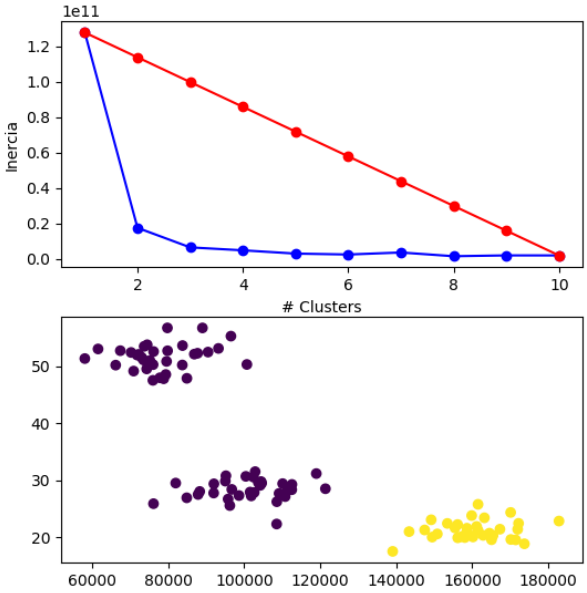
- We initialize the centroids
- We calculate the distance between entities and clusters
- We group the entities into clusters

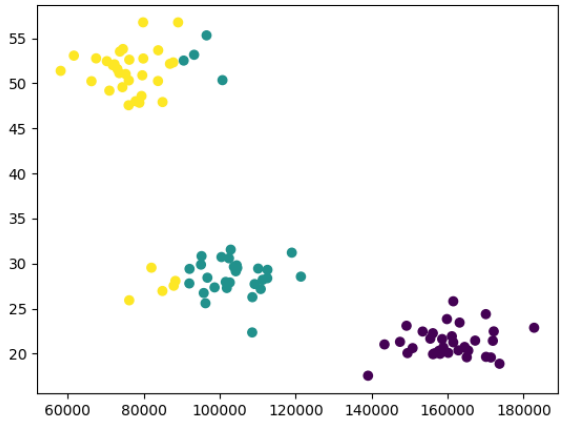
- Re-calculate the new clusters centroids.
- Repeat until clusters has no changes.
    - This method will be used 10 times with different amount of clusters, to perform the elbow method and find the optimal amount.

V. RESULT

```
##############################################
#########     Regresion Logistica     ########
##############################################


       Lineal Model

[0.03410168665528504, 0.24908892771562155, -0.26678985926750376]
Error:  0.1432323563651696
##########################################
##########################################
       CONFUSION MATRIX
         Negative     Positive
Negative   20           0
Positive   3            22
##########################################
##########################################
Precision:  1.0
Recall:  0.88
f_score:  0.9361702127659575



       Poly Model

[0.07396019884359308, 0.18288328531969725, 0.11083300253253117, 0.2920539906061656, -0.24932074082030295]
Error:  0.05858682521044482
##########################################
##########################################
       CONFUSION MATRIX
         Negative     Positive
Negative   19           1
Positive   0            25
##########################################
##########################################
Precision:  0.9615384615384616
Recall:  1.0
f_score:  0.9803921568627451
```

For testing, the program has two outputs, the first one is the lineal model and the second one is a poly model, the grade is asked during execution but in this case is 2. Each model has their evaluation metrics.





The clustering program uses the elbow method to choose the amount of clusters that suit better for the dataset, but it also ask if the method worked well, and ask for a posible number of clusters that will suit in the best way.

VI. CONCLUSSION

The method implementations work pretty good, the metric shows good scores and the confusion matrix shows that lineal model suits better for predicting actual positives but in a general way the poly model (second grade) has a better performance with an F1-Score of 0.98 over 0.93 that resulted in lineal model.
Also, the error is higher in lineal model (0.14% over 0.058%)

In clustering, the elbow method works fine, but it could be improved by setting a custom amount of clusters after seeing the dispersion of the data using the graphic provided by the program.

VII. REFERENCES

[1] Medium.com/@aprendizaje.maq/regresion-lineal-con-gradiente-decendiente-c3b5ca97e27c

[2] Jason Brownlee, “Logistic Regression for Machine Learning” Phil. machinelearningmastery.com/logistic-regression-for-machine-learning/ April 2016.

[3] Data Science: Theories, models, algorithms, and analytics; By Sanjiv Ranjan Das; 2013 (Chapter 17.2: Clustering using k-means).