**Estructura de datos y algoritmos I**

**Entrega final**

**Daniel Andrés Hernández Oyola**

**Jose Miguel Blanco Vélez**

**Dónovan Castrillón Franco**

**What is an estimation?**

When we talk about the concept “Estimate” we mean to reference about something that is going to happen or happened in the past, and even if this sounds like a very mathematic-statistic concept it is something that it is already in our everyday life, for example when we say how much time are we going to spend doing something or when we approximate the time that it’s going to take us to get to an specific place, and it is not only to estimate the time, but also distances and results of any kind.

Even thought it is a concept that we can apply in a lot of fields, we have to difference when are we referring to an estimation of want and infer.

Why do we have to difference this? It’s because when we talk about want, are things that have nothing to do in the statistic and has no way to be “calculated” with any precision, but, when we talk about infer, we are talking about a process that has an apparent result coming from parameters previously analyzed.

**How do we do an estimation?**

First of all, we need to know that inside of the statistical estimation there are 2 kinds of estimation: the point estimation and the interval estimation.

To difference them, let’s see the upcoming example.

Imagine that you want to estimate the approximate heigh that the people have in a specific place, for this you could say a number, for example the average heigh of the people is 1,70m in this case we are talking about a point estimation because we are only pointing one data, but, in the other side, we could say that the average height of the people in this place is 1,65m and 1,75m and here we would be referring as an interval estimation, because, as the name says we are giving an interval in where the people’s height could be located or not.

**Estimation in data science.**

When we are going to work with an estimation, probabilities, and in general of statistic in data science there must be 5 basic activities.

1. **Analysis design: This activity involves the planning of the details to obtain the data that we need and the generation of the hypothesis to be evaluated.**
2. **Data exploration: in this activity we are dedicating to play with the data, we describe it, we summary it, make graphs to see it from different angles. This exploration helps us to make sure that the data that we got has no parts missing and the design was correct.**
3. **Model assembly: in this activity we try to make a model that explains the behavior of our data and make it able to make predictions of himself. The idea is that the model is able to describe the fundamental properties of our data.**
4. **Make estimates: Here, we are going to try to make estimations based on the model that we assembled previously. We are also going to estimate the error size that our model could have in its predictions.**
5. **hypothesis contrast: this activity is the one that is going to produce the final choice of if the predictions of the model were right and help us conclude if the data that we have confirm or deny the hypothesis that we generate in the activity**

* There also exists mathematic concepts that we have to consider depending on the necessities of our project.
* Arithmetic average: The arithmetic average is the value obtained from adding all the data and divide it by the result between the total elements.
* Variance: The variance is the arithmetic average of the square of the deviations respect the average of a statistic distribution.
* Typical deviation: The typical deviation is the square root of the variance.
* Mode: The mode is the value that has the most absolute frequency.
* Median: The median is the value that fits the central space of all the data when they are ordered by lower to higher.

And these are only of the some of the many statistic concept that exist.

Each one of these has its way to be calculated, however, to facilitate the things to the developers in the programming language Python, there are some libraries that make us able to work with probabilities and statistics, moreover that they are free access.

* numpy: The popular mathematic packet of Python, it is used that much that some people consider it integral part of the language. It gives us some statistic functions that we can apply easily on numpy’s arrays.
* scipy.stats: This submodule of the scientific paquet scipy is the perfect complement for numpy, the statistical functions that we don’t find in one of them, we will find them in the other one.
* statsmodels: This library gives us a lot of tools to explore data, estimate statistics models, make statistic tests and a much more.
* matplotlib: It is the most popular library in Python for visualization and graphs.

It is going to make us able to make the graphs of the various distributions of the data.

* [seaborn](https://stanford.edu/~mwaskom/software/seaborn/): This is the ideal complement of matplotlib for making statistical graphs.
* pandas: This is the most popular library for analysis of financial data.

It has some useful functions to make statistical description of some of our data and this helps us at the time of working with time series.

* pyMC: pyMC is a python’s module that implements statistical Bayesians models, including the chain of Markov Monte Carlo (MCMC). pyMC offers functionalities to make the Bayesians analysis the simplest possible.

**.33333What is the artificial intelligence?**

The artificial intelligence (AI) summarized it he combination of algorithms planned with the purpose of making machines that represents the same capabilities than the human being.

The AI it’s present in the facial detection of the phones, in the virtual voice assistants like Siri from Apple, Alexa from Amazon, or Cortana from Microsoft and it is integrated in our devices through bots (summarized from robots) or applications for mobile phones.

**What is machine learning?**

The machine learning is one the many disciplines that are found inside of the area of the informatic sciences, This one is focused mainly on the development of the artificial intelligent, and that works to create new systems that could learn by themselves.

This is a technology that allows us to automatize tasks that are mainly a series of operations, this, with the purpose of reduce the necessity of humans. This supposes a big advantage when it is about control a big amount of information for it to being realized in a way more effective way.

Inside of the machine learning, exists some different kinds of algorithms, and these are classified in mainly 2 types: supervised learning and unsupervised learning.

**Supervised learning:** When we refer to this type of learning is when the operator teaches the machine with examples, what does this mean?, the expected entrys and outputs are given to the machine and the machine has to look for a way to get to the output from the entry.

Some supervised learning algorithms:

* Linear regression.
* Polynomial regression.
* Decision trees.
* Random forest.

**Unsupervised learning**: In this case the algorithm of automatically learning has to study the data to identify any type of pattern or patterns. Any type of hint or order by the operator is nonexistent. The machine has to determine the relationships that exist in the available data.

Some unsupervised learning algorithms:

* K-Means Clustering
* Hierarchical Clustering

In this project we are going to deep in the supervised learning, specifically the decision trees.

**Decision trees:**

A decision tree is a map of the possible results of some series of choices related to each other. It makes an individual or an organization compare possible actions between them, according to costs, probabilities and benefits. They can be able to be used to lead a trade of informal ideas, or propose and algorithm that anticipates mathematically the best option.

The decision trees generally start with a root node and start branching our in its possible results that generate extra nodes that generates more possibilities making therefore a figure like a tree.

Inside of the nodes exist 3 different types: Probability, decision, and terminal nodes.

The decision trees are very popular because we find different advantages in them.

* Easy to understand.
* They can be useful with, or without reliable data, and any data requires a minimum preparation.
* The value at the time to select the best of numerous options.
* New options can be added to existent trees.
* It can be mixed easily with other decision-making tools.

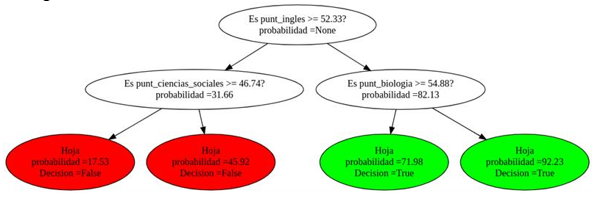
However, in some cases this can become very complex, making its understanding complex. When this happens, new types of diagrams should be considered for the problem solving

**Project:**

In the project, the information gain was chosen instead of Gini for practicality reasons, this answers the doubt of: if using Gini was necessary, the answer, obviously is no. There are different ways to get something similar and with a similar precision.

As demonstration of the project’s result we are going to be using the dataset with the name “TRAIN(4)”. The Dept that we chose was 3 but is good to mention that you can use whatever.

**Generated tree:**

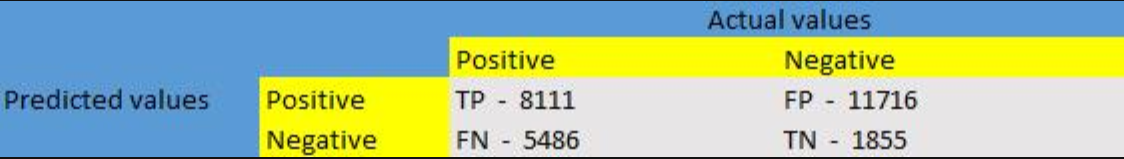
****

The red color means the mode is not true which means that anything that comes into this, came because it lost, so, if it goes to the green one it means it won.

In each existent node we put the probability that a person has to win in that point.

The accuracy of this generated tree is 72,98%

**Confusion matrix:**

****

**Used auxiliar methods:**

* partition
* targetcalculate
* entropía
* mejorDiv
* seleccionarDatos
* predicciones
* calcularMejorDiv
* crearMatriz
* correr

Bibliography:

* Manzano Arrondo V. (2012-2014). Estimación estadística. Recovered from: [EstimacionEstadistica.pdf (us.es)](http://asignatura.us.es/dadpsico/apuntes/EstimacionEstadistica.pdf)
* San Nicolás, N.L. (2009). Hacer estimaciones estadísticas. Recovered from:

[Microsoft Word - 10 Hacer estimaciones (uanl.mx)](http://eprints.uanl.mx/12506/1/A10.pdf)

* (2015). Probabilidad y estadística con Python. Recovered from:

[Probabilidad y Estadística con Python (relopezbriega.github.io)](https://relopezbriega.github.io/blog/2015/06/27/probabilidad-y-estadistica-con-python/)

* [Redacción APD](https://www.apd.es/author/redaccion/) (2019). ¿Qué es machine learning y cómo funciona? Recovered from: [Qué es Machine Learning, cómo funciona y a qué se aplica | APD](https://www.apd.es/que-es-machine-learning/)
* [Redacción APD](https://www.apd.es/author/redaccion/) (2019). ¿Cuáles son los tipos de algoritmos de machine learning? Recovered from: [¿Cuáles son los tipos de algoritmos del machine learning? | APD](https://www.apd.es/algoritmos-del-machine-learning/)
* Que es un diagrama de árbol de decisión. Recovered from: [¿Qué es un diagrama de árbol de decisión? | Lucidchart](https://www.lucidchart.com/pages/es/que-es-un-diagrama-de-arbol-de-decision)
* Ligdi Gonzales (2018). Clasificación de machine learning | #4 Curso de introducción al machine learning. Recovered from: [(3) CLASIFICACIÓN DE MACHINE LEARNING | #4 Curso de Introducción a Machine Learning - YouTube](https://www.youtube.com/watch?v=R4SZGkLIp88&list=PLJjOveEiVE4Cbbx1dVjydfmPPpjl0pg86&index=4&ab_channel=AprendeIAconLigdiGonzalez)