Path to data science *

Manuel Gijón Agudo

*Collection of blog's articles

Índice

1. Introduction: Welcome to the blog and motivations

It was a year ago when I had heard for the first time the term "Machine learning". I was talking with two friends in the bar of my faculty. They were talking about their new project for the next Hackaton (UPCHack 2018 in spring), to be true I didn't understand so much about the conversation. Then I asked and my friend explained me for the first time what is machine learning and why it's so useful today.

I did not have and idea at that moment of what it's the real importance of this technology, but now I know it's everywhere. He explain to me (in the best way he could, he did his best and know that I know how difficult it is I really appreciate the effort he made) what is a neural network by explaining a example of its use.

He says that you can train a neural network to recognize birds in a photograph. At that time i hadn't think never about the complexity of the problem. That conversation opened my eyes to a new world of amazing technology and applications. Since then I've wanted to know more. Unfortunately at that time I was finishing my bachelor's degree at mathematics and I haven't had so much time.

Now, while I'm studying my Master degree in mathematics, I have time to learn, to study and to practice. I'm willing to study every day and to write down my notes and other interesting stuff in this blog at least two times per week. I'm convince that this is going to help me to understand better the materia. Trying to explain something is very useful because you put in a place where you have to figure out what is not complete clear. As Albert Einstein said (well, he really didn't say that) "You do not really understand something unless you can explain it to your grandmother".

I pretend to share part of the knowledge I'm getting doing my final master thesis in words embeddings and some of the discoveries I'm doing too. This world is exciting and I pretendt to share with you all how is being my path to become and expert in data science.

Apart from this, you'll find in my blog information and little tutorials about statistics, some related mathematical topics and data display.

2. Machine Learning generals

2.1. Introduction to ML

BREVE RESEÑA HISTÓRICA DEL TEMA

There are a lot of ways to explain what machine learning is, but the intuitive idea is to have a program to do some task that is able to improve by itself. A more formal definition was given by Tom Mitchell: 'À computer program is said to learn from experience E respect to some class of task T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E".

A simple example: linear regression model. In simple linear regression model we have two parameters to improve. We'll go deeper in this in next articles, by know the only thing we need two know is that θ_0 and θ_1 are the parameters to optimize and we can define and error using them. To be true, the parameter to improve is the error, our goal is to reduce it. In this case E, the experience, would be the act of use the algorithm to reduce the error once and again. The task T is to reduce the error and the measure P is the value of the error.

In general, we can divede the machine learning problems into two main categories:

- Supervised learning: we have labeled output samples. It generates a inferred function from the training examples that is used to predict new outputs.
- Unsupervided learning: we infer a function to describe structure from unlabeled data that is not obvious.

Mention to the REINCOSA LEARNING Y DECIR QUE TAMBIÉN LO ESTURIARÉ EN EL FUTURO.

All allong different post, I'll present my study process of this methods, theorical work, implementations and a lot of examples.

2.2. Notation and Model Representation

We'll use $x^{(i)}$ to denote the 'input' variables and $y^{(i)}$ for the 'output' variable, the variable to predict. The set of all pairs $(x^{(i)}, y^{(i)})$ for all $i = 1, ..., m$ is called a **training set**, where each one of the pairs is called a **training example**. Note that the superscript (i) is simply and index for a certain training example into the training set. We also use X to denote the space of input variables and Y to denote the space of the output variables.

When we have more than one features (or input variables) we use subscripts to refer each one of these $(x_1, x_2, ...)$.

In the context of the supervised learning we'll call **hypothesis function** (by historical reasons) $h(x)$ to the function that we obtain by applying a learning algorithm to the training set. It's a function from that leads a $x \in X$ from the input space to a element of the output space $y \in Y$.

When we have a input space Y continuos we call the supervision problem a **regression problem** and when it's discrete we call it a **classification problem**.

3. Clasification - Supervised learning

4. Clustering - Unsupervised learning

5. Neuronal networks

6. Miscelanius

Referencias