An Introduction to Data Science

**Milad Goudarzi . Arash Mehrzadi**

**Introduction**

Data science is an extensive title on its own. There are, however, some basic terms which are frequently being used among data scientists, regardless of their field of focus. Hence, having a good insight into them could lay the foundations of an ample understanding of more advanced concepts. This article tries to elaborate on some basic concepts, which you might hear oftentimes while working in this field.

**Why there is lots of statistics present in Data Science?**

Before moving on with the definitions, let’s get familiar with a pre-requisite, without which you won’t be able to understand more complex concepts of Data Science in the future. Merely one episode of a beginner-friendly tutorial is enough to perceive the significance of statistics in data science. Thus, it is natural to face

a lot of statistics formulas in the very beginning. Explaining each and every of them needs its own tutorial, but let us make it a bit clearer on why having this particular knowledge is a must for a data scientist.

Consider the census process, in which an individual shows up in front of your house and asks for details like your name, number of your family members, your age, etc. It is obvious that some people may not be available at that particular time. What does the census agent do? He/she comes back another time. But what if he/she could not reach out to you anyway? There are, therefore, almost always some missing information. This missing information could completely be dropped out of the census, which ends in lower accuracy, or could be replaced by the mean or mode of other records, which brings about more accuracy than the previous method. This occasion happens many times in a Data Science project and the aforementioned technique is a method of handling missing values, which you can see is pure statistics.

**Different Types of Datasets:**

One of the most important parts of Data Science are Datasets. In fact, we collect a set of data in various standard structures, namely records, graphs and networks, ordered, spatial, image and multimedia, to be able to use them to analyze the world around us.

These different dataset types are elucidated below:

**Record:**

It can be said that, this type is one of the most well-known types of datasets. Just take a look at the tables in the image above [img.1], most likely you are familiar with this format of tables. All that exist in this type are a number of rows and columns that each column is a feature and each row is an entity that contains one or more attributes.

**Graph and Network:**

To explain this type, we consider an online store that offers a variety of food products and we want to implement a system that categorizes our customers and recommends products which the customer is more likely to buy, according to their interests and needs.

The solution is to design a dataset containing rows and columns with the same number of goods in warehouse. This dataset contains cells that determine the relationship of each product to other products, and the value of each cell indicates the strength of relationship. So, higher value indicates a stronger connection between goods and we can use this value to recognize what products to recommend the customer, for which the chance of being purchased is higher (the value of each cell is determined by customers purchase information).

This is an example of a dataset in which data are related to each other and can be represented as a graph or data network.

**Ordered:**

Basically, ordered dataset records have a special sequence, for example consider a video data that is created by placing multiple sequences or consider a temporal data in which the data are recorded in a specific time series (weather information etc.).

In addition, we have other formats of data like sequential data and genetic sequence data that all of them are considered as ordered datasets.

**Spatial, image and multimedia:**

There is another format of data such as maps data, image data and video data, every one of which is one of the most challenging data in data science. Fortunately, in recent years, powerful tools have been developed to use this type of data.

**Good data metrics:**

After getting acquainted with types of datasets, now it’s time to know the problems of working with these datasets. The datasets may have problems such as outliers, noises, missing values, inconsistencies, numerosity and etc. Fortunately, there are solutions to all these problems and we can use them to adjust the data according to the requirements of the project. Resolving any of these problems can bring us closer to a good dataset. On the other hand, it can also have a negative impact on our project. Therefore, the correct use of each of aforementioned items is crucial and can be effective on the outcome of the project.

**Data Cleaning:**

You cannot come up with a high-accuracy result with dirty data, meaning that the data could be incorrect, incomplete, irrelevant, duplicated or improperly formatted. Data, before doing any further processing, needs to be cleaned. For instance, when the census agent asks for someone’s age, one person could answer directly with the number and another person could answer with the date of birth. The agent should decide which format to register, owing to the fact that registering records with both formats could lead to more calculations and lower performance. In addition, if someone was not available at the moment the agent arrives, there would be some **missing data**. This missing data should be handled in some manner, like replacing them with the average of other records. Because waiving them, results in lower accuracy.

These were some examples of what we exactly need in Data Science: “a clean data”. Dirty data could be due to faulty instrument, human or computer error, transmission error, etc.

**Data integration:**

Data integration is nothing but bringing data from different sources in one place. Imagine a company, consisting of different departments. This company has to have every department data to find out how each department affects the other ones and how the whole company could grow.

**Data Reduction:**

If you have ever worked with a huge database, you perfectly know that even running a simple query takes longer than usual. Now what happens if you are planning to analyze these data? This is where the Data Reduction techniques come to play. We reduce the data not only to make the process faster, but also to increase the accuracy, prevent overfitting, etc. Data reduction techniques help you condense the data in terms of size, but still keep the critical information. In some techniques we tend to drop out some unnecessary data. For instance, have you ever cropped a photo? Why did you do that? Because you didn’t need those areas of the photo. Huge datasets undergo exactly the same process in Data Science. In addition, in some other techniques, which are more advanced, we don’t drop out any data, but we reduce the number of values of a continuous attribute by converting them into a finite set of intervals and assigning each interval a specific data value.

**Data Transformation:**

The process of transforming data to a format which is easier to mine, is called Data Transformation. As a computer end-user, for instance, you have probably converted a jpg file format into a png, or a word document to a pdf file based on your needs. When you are doing so, you are transforming data. Data transformation can be simple or complex based on the requirements of the project you are working on. In addition, different data analytics tools, data volumes, data complexity, etc. do exist which make it a bit trickier for a data scientist to decide between many available tools and methods. Speech-to-text transformation, character encoding and CSV to XML data transformation are a few good examples of Data Transformation applications.

**Classification:**

Let’s explain it this way. Try to guess what is it. It is an animal, has four legs and a tail, is bigger than human, human rides it and it runs fast. What is it? Yes... It is a horse. But it doesn’t matter if you guessed it right or wrong. This is the concept behind classification. When you train your model with some classification algorithm, it tries to classify the training data based on their features, which are available in each and every dataset type described before. Afterwards, when it is trying to predict the result, looks at the input data and the classification it has made from the training data and checks the features you are giving as input: it is an animal, has four legs, is bigger than human, human can ride it, it runs fast. Finally, after processing these features the model shows us “horse” as the result of the prediction. This was a high-level definition of classification and a basic concept of how it works. Classification is all about making predictions based on past data, where we have known what the correct prediction was. Now these are some real-world and of course well-known examples of classification:

* For each order Amazon receives, Amazon would like to predict: is this order fraudulent?
* Doctors would like to know: does this patient have cancer?
* Politicians would like to predict: are you going to vote for them?

**Regression:**

In regression, as opposed to classification, we do not classify our training set (part of the whole dataset which is used to train our model). Instead we try to find a function to predict an outcome variable Y, based on the value of some input variables X1, X2, …. You will completely get a good perception after reading the two following examples.

Let’s say we are trying to make a model to predict horse breeds, based on some input features such as the horse height, color, mass, lifespan, etc. Therefore, we try to find a function which can predict the correct breed based on the input features (height, color, etc.). Here, the outcome variable Y is the horse breed.

As another example, consider having some features related to houses, namely their size, their location, their facilities, etc. and we know the price for each house. Now we want to predict house prices in the future regarding this information. For such problem we could use regression techniques. According to the mentioned definition, we are trying to find a function to calculate (predict) the price of each house based on some inputs as size, location, facilities, etc.

**Conclusion:**

Understanding basic concepts of Data Science deeply is vital for every data scientist. In this article, we tried to explain some of the most basic and important terms of this field. We will try to go over deeper concepts in future articles.