What is Data Science?

Notes by José A. Espiño P. [1]

Summer Semester 2023–2024

[1]The content in these notes is sourced from what was covered in the course the document is named after. I claim no autorship over any of the contents herein.

Contents

1 Defining Data Science

Data Science, in short, is the usage of massive amounts of data to extract patterns and outliers from it; these may confirm existing hypotheses or reveal completely new knowledge.
Data Science is a multidisciplinary field that combines statistics, mathematics, computer science, and domain knowledge. It is a field that is constantly evolving and is driven by the increasing amount of data and the increasing computational power that is available to process and analyze this data. Data Science is used in a wide range of applications, from business to healthcare to social sciences.
The cloud is a key enabler of data science. It allows data scientists to access large amounts of data and computational resources without having to invest in expensive hardware by using the computational resources of the cloud provider. This provides advantages such as scalability, instant access to resources, and cost savings. Some popular cloud providers are Amazon Web Services (AWS), Google Cloud Platform (GCP), IBM Cloud, and Microsoft Azure.

2 Data Science Topics and its Role in Business

Big data is a term that refers to the large volume of data that is generated by businesses and organisations. It is normally characterised by the five v's: volume, velocity, variety, veracity, and value. Data is being generated at an extremely fast and continuous rate, in very big quantities (especially through the use of different sources like sensors)— these are the velocity and volume. Data comes from a variety of sources, such as sensors and social media, and can be neatly formatted (*structured*) or not (*unstructured*)— this is the variety. The veracity refers to the quality of the data, and it is an extremely relevant concept nowadays, when so much information could be AI–generated and inaccurate. Finally, the value refers to the insights that can be extracted from the data—being able to turn data into actionable insights is the main goal of data science.

Hadoop is a collection of software utilities that allows for large datasets to be processed across a distributed computing environment. Instead of storing and processing the entire data in a local computer, Hadoop allows you to divide the dataset into small parts, process each part in different computers, and then combine the results. This renders big data processing a lot faster and efficient. It uses the *MapReduce* programming model, which consists of two main steps: the *map* step, which processes the data and produces intermediate results, and the *reduce* step, which combines the intermediate results to produce the final output. As a data scientist, you will often work with Hadoop to process large datasets.

Data mining is the process of discovering patterns in large datasets. For it to be successful, the following steps are recommended:

1. Select the data: Identify useful data sources and collect the data. you may even need to plan new data collection initiatives

2. Preprocess the data: Identify relevant attributes of data, remove the irrelevant ones, and spot and correct errors in the data. It is also important to develop a way in which to deal with missing data, especially when this happens systematically—since this may lead to a bias in the analysis.

3. Transform the data: Determine a format in which to store the data and reduce the amount of attributes to the furthest extent possible. This can often be achieved by using **data reduction algorithms**, such as *Principal Component Analysis* (PCA) or *Singular Value Decomposition* (SVD). A common way in which data is transformed is by converting continuous data values into discrete ones, which is called *discretization*.

4. Store the data: The data must be readily available for the data scientist to manipulate. Furthermore, it must be easy for the data scientist to edit the data and to store the results of the analysis. The servers/storage media containing the data must be secure, especially in cases where the data being handled contains sensitive personal information.

5. Mine the data: This is the step where the data scientist uses algorithms to extract patterns from the data. This can be done using a variety of techniques, such as clustering, classification, and regression

6. Evaluate the results: The data scientist must evaluate the results of the data mining process to determine if the patterns that were extracted are useful. This can be done by comparing the results to existing knowledge or by using statistical tests to determine if the patterns are statistically significant.