Freie Universitat Berlin
Erasmus Program

10 ECTS

Machine learning for data science

Professor:
Prof. G. Montavon

Autor:
Filippo Ghirardini

Winter Semester 2024-2025

Contents

Machine learning for data science

Autor: Ghirardini Filippo

Winter Semester 2024-2025

---

1 Data science

Definition 1.0.1 (Study of data)**.** *Data science is a study of data. It includes:*

- *Processing data*
- *Representation of data (compressed, understandable)*
- *Value extraction from data*
- *Knowledge extraction from data*

Definition 1.0.2 (Data-Driven science)**.** *Data science as a tool scientific discovery, complement to hypothesis-drive science.*

1.1 Comparison

1.1.1 Hypothesis-Driven

In the **hypothesis-driven** experimentation, we first form an *hypothesis*, then we perform an *experiment* and in the end we *check* the results.

1.1.2 Data Science

Data-Driven science instead starts with taking *existing data*, performing an *analysis* on them and *observe* the result.

1.2 History

In the past, experimentation and data collection existed with separate purposes. When the **cost** of collection and accessing the data dropped dramatically (thanks to the internet and all the sensors), they started being implemented in science, since on the other hand the cost of an experiment was the same. Furthermore, computers are now able to process large amount of data.

Observation 1.2.1. The first and most important thing in data science is collecting the data and only later working on it.

The first example of data science can be found in a map made by John Snow in 1854 showing the clusters of cholera cases in London. With the map he noticed that there was a center place with a fountain/well and understood that the disease was transmitted via water and not air.

1.3 Sources of data

There are multiple sources of data:

- Physics equations: data generated artificially via simulation running multiple times with different parameters
- User content
- User activity: for example public GPS traces of OpenStreetMap, that enables to gain insights on mobility patters (e.g. relation between location, mode of transportation, date and user).
- Historical books or artifacts: for example *De Sphaera Corpus*, the corpus of multiple editions of the cited textbook made available in digital form. It enables similar dataset-wide analysis, relating content, year and city of publications.
- Earth data
- Biomedical data: for example the National Cancer Institute in the USA and the UK Biobank data

Note 1.3.0.1. Data is often not intended for scientific use. It is often a by-product of an activity with a different purpose (e.g. clinical practice, accounting).

1.4 Usages

The outcome of a data science analysis can:

- be **insightful** and of interest by itself (and be published)
- inform on further experiments or investigations to be conducted in order to **verify a hypothesis**
- inform on whether it is **feasible** to build a system that accurately predicts the data

1.5 Examples

1.5.1 T-SNE analysis of TCGA high-dimensional omics data

We start by defining:

- x_i is the representation of i in the original measurement space (e.g. $x_i \in \mathbf{R}^{1000}$)
- d_{ij} is the distance between instance i and j in measurement space

$$d_{ij} = ||x_i - x_j||$$

- y_i is the representation of instance i in low-dimensional space, usually $\mathbf{R}^2$
-

1.5.2 Planetary science

Can internal properties of a planet be predicted from a few observables? Yes, starting from predetermined parameters we then run tons of simulations via convection models and check the observables. In the end we try to build a model that can predicts them. This is an example of a **correlational** research.

1.5.3 Digital humanities

The idea was to extract correlations between visual patterns and categorization of images.

1.6 Comparisons

1.6.1 Statistics

Statistics makes particular assumptions about the nature of the data whereas **data science** addresses the data as it occurs in real-world applications. Data science also addresses the technical challenge of processing the data.

1.6.2 Decision making

Data science focuses on extracting insightful structures and relations from the data and presenting them in an interpretable manner, while **Decision Making** focuses on learning good autonomous decisions so that repetitive tasks can be performed without human intervention.

2 Data

2.1 Structures

2.1.1 Classic

A classical dataset consists of a collection of N **instances** (data points, examples) where each instance can be represented as a vector of d **features** (attributes, measurements).
They can be stored in a two-dimensional array structure of $N \times d$ and they usually come with **metadata** that explains the data.

2.1.2 Images, text, sound

Sometimes the data is **not tabular** but, for example, is an image, text or a sound. In these cases the most common approach is to provide the dataset as a folder that contains the file. Sub folders may be used to organize the data according to metadata.

2.1.3 Network

In this case the data consists of a network of N instances with connections (directed or not, weighed or not) between pairs of related instances. It can be represented as an **adjacency matrix** of size $N \times N$. Since it is typically sparse (one node connected to few), it may be better to use a **sparse representation** such as a table of size $\#edges \times 2$ storing the links.

2.1.4 Relational databases

It's a collection of tables of two different types:

- The first one has a row for each entity and a column for each attribute
- The second one stores relations between instances of two different tables

The analysis may proceed either by:

- Focusing on data from a single table
- Joining tables
- Operating on the relational structure using advanced techniques

2.1.5 Fusion of datasets

Aggregation of multiple small datasets can enable the learning of more general and accurate models. For them to be valuable, data coming from the multiple sources needs to be **homogenized**. Furthermore, **implicit information** in the original datasets needs to be included in the aggregated one, ideally as additional features or metadata.

2.1.6 Unstructured data

Data may have a level of heterogeneity such that there is no obvious data model that can be used. In that case, the data model must be rebuilt from scratch using expert knowledge from the field.

2.1.7 Large datasets

Datasets whose size is too large to be processed with classical techniques (e.g. high throughput devices like FMRI or complex simulations). In this situation advanced approaches are needed like data parallelism and model synchronization between multiple machines.

2.1.8 Streaming data

In this case data arrives continuously at a high rate. Insights need to be delivered in a timely fashion since there is no time to collect a full batch before the analysis.

2.1.9 Data subject to regulations

User data or medical data is subject to regulations fro privacy reasons that determine who can access the data. There could be the need for a two-level data analysis: the first one may be performed only on non sensitive data.

2.2 Preprocessing

Tabular data can be converted into an array via

```
numpy.getfromtxt
```

while non numerical may be discarded or converted to a numerical value.
Images can be loaded in python via **PIL** or **cv2**. Otherwise, one can use raw pixel values to compute low level features or feed the image to a pretrained neural network feature extractor.
Sound data are usually converted in spectrograms showing the frequency information at coarser time steps.
Text data can be converted to numbers with encodings. You can also remove non important words such as "the" or "and".

2.2.1 Missing values

When there are missing values (e.g. faulty sensor) we can:

- Replace missing values with standard ones
- Replace with the most likely given the others
- Encode each value as a two-dimensional vector, e.g.

$$x \mapsto (x, I\{\text{missing}\}) \qquad x \mapsto (x, 1-x) \cdot (1 - I\{\text{missing}\})$$