Introduction to Data Science

> ***Data Science*** *is the* ***nontrivial*** *extraction of* ***implicit****, previously* ***unknown****, and potentially* ***useful information from data****.* (William J. Frawley, AI Magazine, 1995)

- Data Science is an application of **Machine Learning**, one of the major subfields of **Artificial Intelligence**;
- Similar terms: **analytics**, **data mining**, **KDD** (Knowledge Discovery in Databases), etc;
- **nontrivial** - we need more or less complex algorithms to extract the information;
- **implicit** - the information is not explicitly present in the data;
- **unknown** - we do not know the information beforehand;
- **useful** - the information is useful for some purpose.

Artificial Intelligence

> ***Artificial Intelligence*** *is branch of computer science concerned with making computers behave like humans.* (John McCarthy, 1956)
>
> AI Fields:
>
> - Search;
> - Reasoning;
> - Natural Language Processing;
> - **Machine Learning**;
> - Interaction.

The main subfields of of **Data Science** are:

- **Statistics** - tools for understanding data distribution, sparsity and correlation;
- **Machine Learning** - algorithms for discovering the models behind the data;
- **Databases** - efficiently deal with large amounts of data.

From Data Engineer to Data Scientist

- **Data Engineer** - deals with the data collection, storage and processing;
- **Data Analyst** - deals with the data exploration and preparation;
- **Data Scientist** - deals with the data modeling and evaluation;
- **?** - deals with prediction and decision making.

Basic Concepts

- **Records** - instances, data objects or observations;
  - Tuples of values described by a set of **variables**;
- **Variables** - attributes, fields, features, dimensions;
  - **Numeric** - real-valued, interval based, ratio;
  - **Symbolic** - binary, nominal, ordinal.
- **Tabular data** - data organized in a table/matrix `n x d`, where:
  - **n** - number of records;
  - **d** - number of variables - **dimensionality** of the data.

  Information *is the set of patterns or expectations that underlie the data.*

KDD (Knowledge Discovery in Databases) Process

1. Define the **goal**;
2. **Data Collection**;
3. **Data Profiling** - characterize the data under analysis with respect to its **distribution**, **sparsity**, **granularity** and **dimensionality**;
4. **Data Preparation** - **Selection** and **Transformation** (integration, cleansing and feature engineering) of the data to be used in the modeling phase;
5. **Modeling**;
   - **Supervised Learning** - classification, forecasting;

- **Unsupervised Learning** - clustering, pattern mining;

6. **Evaluation** - assess the quality of the model;
    - Evaluates **simplicity**, **utility** (**coverage** and **novelty**) and **certainty** of the model.

Some Definitions

- **Data Mining** - the process of **discovering patterns** in large data sets involving methods at the intersection of **machine learning**, **statistics**, and **database systems**;
- **Classification** - usage of a model that **describes and distinguishes data classes** or concepts;
- **Clustering** - analyzes data objects without consulting class labels (which may not be available). As such, it can be used to **generate class labels for a group of data**. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity;
- **Regression** - used to predict **missing or unavailable numerical data** values rather than discrete class labels.

Data

- **Data** - a collection of facts, such as numbers, words, measurements, observations or even just descriptions of things;
- **Number of rows** - number of records - size of the data - **n**;
- **Number of columns** - number of variables - **dimensionality** of the data - **d**;
- **Univariate analysis** - analysis of a single variable;
- **Multivariate analysis** - analysis of multiple variables;

Variables can be classified into two main types:

- **Numeric** - real-valued, interval based, ratio;
    - **Discrete** - take on a finite or countable number of values;
    - **Continuous** - take on an infinite number of possible values;

  - **Interval** - the difference between two values is meaningful;
  - **Ratio** - the ratio of two values is meaningful;
- **Symbolic/Categorical** - composed by a set of symbols;
  - **Nominal** - composed by a finite set of symbols;
  - **Ordinal** - composed by a finite set of symbols with an order.

  A variable is **binary** if it can take only two values, such as `true` and `false`, `yes` and `no`, `0` and `1`, etc.

Measures

- **Mean** - average value of a variable - $mean(D) = \frac{1}{n}\sum_{i=1}^{n} x_i$;
  - Only measure of central tendency that is affected by outliers;
- **Median** - middle value of a variable;
- **Mode** - most frequent value of a variable;
- **Standard Deviation** - measure of the amount of variation or dispersion of a set of values - $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$;
- **Variance** - measure of the amount of variation or dispersion of a set of values - $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$.

Visualization

- **Pixel-Oriented** - each pixel represents a data value; basically **heat maps**;
  - **Drawback** - cannot help to identify data distribution;
- **Geometric Projection** - each data value is represented by a geometric shape; **scatter plots**;
  - Useful to identify **correlation** between variables;
  - **Drawback** - not suitable for many dimensions;
- **Boxplots** - graphical representation of the **5-number summary** (`min, Q1, median, Q3, max`);
  - Useful to identify **outliers**;
- **Histograms** - graphical representation of the **frequency distribution** of a variable;
  - Useful to identify **data distribution**.