# 1 Quick notes

This document will serve to recite, remember and maybe even quote the different things learned from machine learning. The first chapter is the different chapters in the book [1]

# 2 Introduction to Machine Learning and Data Mining

## 2.1 Introduction

Machine learning is all around, and very usefull theses days. It is a subarea of artificial intelligense, where most progress in is arguably being made. There are different ways of training a model, below are some.

### 2.1.1 Supervised learning

In supervised machine learning the task is to predict a quantity based on other quantities. It is useful to distinguish between classification and regression. Machine learning problems where we have both observed observations and observed target values is known as supervised learning problems or simply supervised learning.

- **Classification**: In classification we are given observed values x and have to predict a discrete response y. I.e., we are given discrete observations of some object and have to determine what class the object belongs to. Some examples are: given hand-written digits determine what number is contained in an image (multi-class classification), given hospital records decided whether a patint will survive for another year.
- **Regression**: In regression problems we have given ovserved values x and have to predict a continious response y, e.g: given a persons height predict their weight, given weather information predict whether it will rain.

### 2.1.2 Unsupervised learning

The oppositive of supervised learning is unsupervised learning. Consider an example dataset consisting of images of animals. We may immediately consider building a machine learning method which tries to discover what animal is in the picture (a duck, a lion, an elephant, etc.). However, for most images on the internet we do not know what animal is actually in the image. Surely, we can sit down and label a few thousand of the images ourselves, train a supervised method (a classification method in this case) on the labelled images, and then use this method to determine the labels of all other animal pictures on the internet – but this is very boring and not reflective of how humans actually learn. Unsupervised learning tries to solve this, and similar problems, without access to any "ground-truth" label information, but try to discover the labelling from the data alone.

- **Clustering**: Dividing unlabeled data into different clusters, hopefully representing some features.
- **Density estimation**: An attempt to quantify the probability of a given future observation given the probability distribution of past observations.
- **Anomaly detection**: Find outliers in data
- **Association mining (rule-induction)**: Setting up rules based on a big dataset. This could be in relation to what people often buy together, gin and tonic, and then make a rule like $gin \leftrightarrow tonic$. This resemlbes discrete math a lot.
- **Dimensionality reduction**: PCR. Trying to boild down a lot of dimensions into few dimensions.

## 2.2 Data and attribute types

### 2.2.1 Data

- Temporal: Sequential over time (stock market)
- Varying number of features: E

### 2.2.2 Attributes

A little vocabulary to distinguish between attributes:

- **Continous**: A continuous attribute is one which, if it can take values a and b, then it can also take any value between a and b.
- **Discrete**: A discrete feature is one where if it can take a value a, then there is a positive minimum distance to the nearest other value b it can take
- **Binary**: A binary feature is a discrete feature which can only take two values. Usually these are denoted 0 (false) or 1 (true).

### 2.2.3 Attribute types

- **Nominal**: If the variable is not ordered and only uniqueness matters. An example is the country of origin or the id.
- **Ordinal**: If the variable is ordered (smaller, larger). An example is the safety rating.
- **Interval**: If the variable is ordered and the relative magnitude of the variable has a physical meaning. The year is interval since the difference between say 82 and 85 (three years) has the same meaning as the difference between years 77 and 80, however the safety rating is not interval since a difference in safety rating of 2 and 4 and 3 and 5 does not have a physical meaning.
- **Ratio**: If the value 0 of the variable has a specific, physical meaning. I.e. it makes sense to say one value of the variable is "twice as large" as another. The year is not ratio since it has no particular meaning to say that 62 is twice as large as 31, whereas the volume of the engine does have a particular physical meaning since a value of 0 means the combustion chamber has volume 0 cm3.

### 2.2.4 Data issues

Data can have different issues, that in turn makes the dataset incomplete. Some machine learning methods, can handle incomplete datasets; most cannot. Below are solutions for such occasitions.

- If the missing data is contained within one attribute, that is deemed non-essential, the attribute can be discarded.
- A lot of observations where a few has missing values, we can discard them.
- If above not possible, the values can be imputed/guessed with some kind of neutral guess.

Below are some examples of what to look for, when searching for bad data.

- Irrelevant or spurious attributes: An example of this would be the ID of an item in the dataset, as this is not a property of the data.
- Outliers: A value that seems impropable. This could be a discrete value in a ratio in an interval of 1-5, where the outlier had a value of 100.
- Missing data: Nan.

# References

[1] Mikkel N. Schmidt Tue Herlau and Morten Mørup. *Introduction to Machine Learning and Data Mining*. DTU, 2019 Fall.