Unsupervised Learning algorithms cheat sheet

A complete cheat sheet for all unsupervised machine learning algorithms you should know

Unsupervised learning tasks. Image by Author

This article provides cheat sheets for different unsupervised learning machine learning concepts and algorithms. This is not a tutorial, but it can help you to better understand the structure of machine learning or to refresh your memory.

To know more about a particular algorithm, just Google it or check for it in *sklearn documentation*.

Next, the following tasks will be explored:

- *Dimensionality Reduction;*
- *Anomaly Detection;*
- *Clustering;*
- and other unsupervised learning algorithms (*Density Estimation* and *Association Rule Learning*)

Since these topics are extensive, *Dimensionality Reduction*, *Anomaly Detection*, and *Clustering* sections are separate articles. I've been working on them for a long time, but I still want to put them in one place.

If we perceive this article as a concatenation of these three, it is quite voluminous, so I don't recommend you to read it all at one time. Add this article to the reading list to come back later, read the chapters through GitLab or download the pdf version of this article and print it (available in the same place).

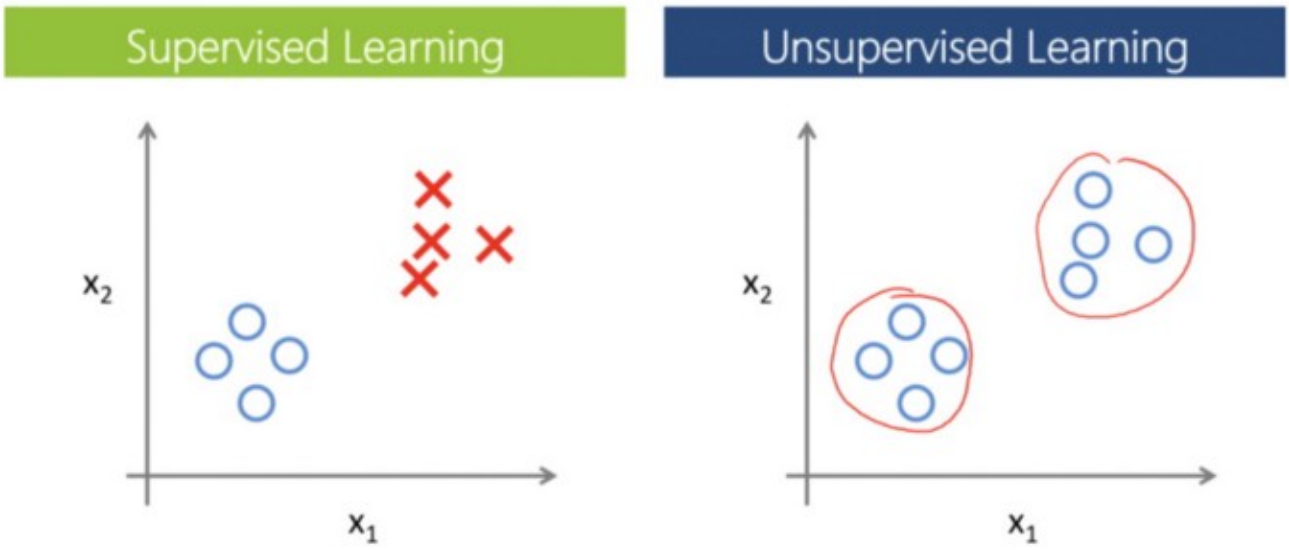
Dmytro Nikolaiev / unsupervised_algorithms

Unsupervised Learning algorithms cheat sheet

gitlab.com

Introduction

Unsupervised learning is a machine learning technique in which developers don't need to supervise the model. Instead, this type of learning allows the model to work independently *without any supervision* to discover hidden patterns and information that was previously undetected. It mainly deals with the *unlabeled data*, while supervised learning, as we remember, deals with labeled data.



Supervised vs Unsupervised Learning. Public Domain

Three of the most popular unsupervised learning tasks are:

- **Dimensionality Reduction**— the task of reducing the number of input features in a dataset,

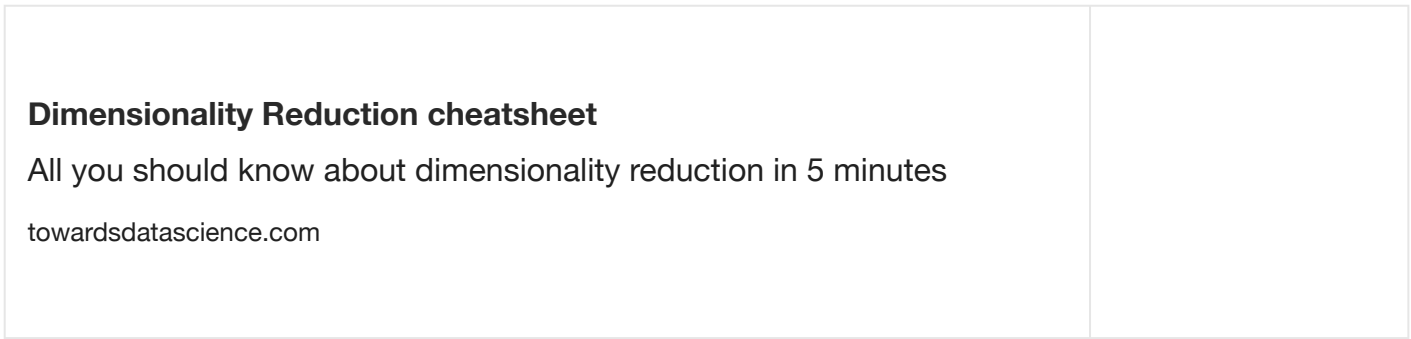
- **Anomaly Detection**— the task of detecting instances that are very different from the norm, and
- **Clustering** — the task of grouping similar instances into clusters.

Each of these three tasks and the algorithms for solving them will be discussed in more detail later in the corresponding sections. However, note that the **Other Unsupervised Learning Tasks** section lists other less popular tasks that can also be attributed to unsupervised learning.

Dimensionality Reduction

The following algorithms are mentioned for dimensionality reduction:

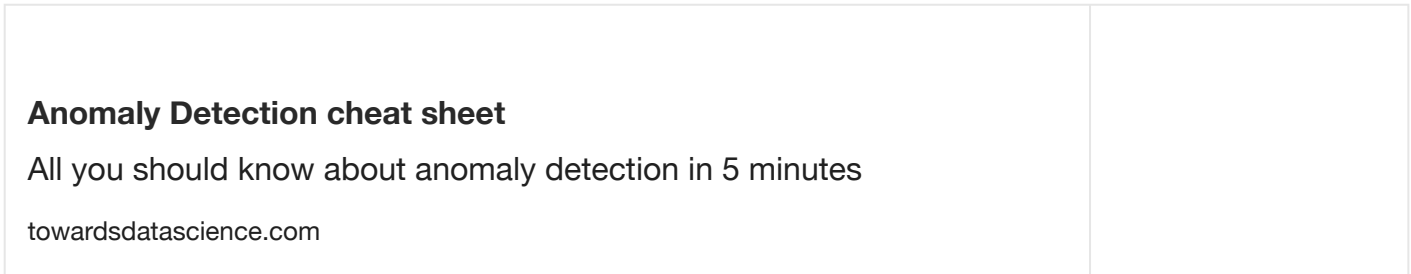
- *Principal Component Analysis;*
- *Manifold Learning — LLE, Isomap, t-SNE;*
- *Autoencoders* and others.



Anomaly Detection

The following algorithms are mentioned for anomaly detection:

- *Isolation Forest;*
- *Local Outlier Factor;*
- *Minimum Covariance Determinant* and other algorithms from dimensionality reduction or supervised learning.



Clustering

The following algorithms are mentioned for clustering:

- *K-Means;*
- *Hierarchical Clustering* and *Spectral Clustering;*
- *DBSCAN* and *OPTICS;*
- *Affinity Propagation;*
- *Mean Shift* and *BIRCH;*
- *Gaussian Mixture Models.*

Clustering cheat sheet
All you should know about clustering in 11 minutes
towardsdatascience.com

Other Unsupervised Learning Tasks

Although dimensionality reduction, anomaly detection, and clustering are the main and the most popular unsupervised learning tasks, there are others.

Since the definition is blurry, any algorithm that deals with an unlabeled dataset can be considered solving some unsupervised learning task (for example calculating the mean or applying *Student's t-test*). However, researchers often identify two other tasks among others: *Density Estimation* and *Association Rule Learning*.

Density Estimation

I have already briefly mentioned density estimation in the anomaly detection section.

Density Estimation is the task of estimating the density of the distribution of data points. More formally, it estimates the *probability density function* (PDF) of the random process that is generated by the given dataset. This task historically came from statistics, when it

was necessary to estimate the PDF of some random variable and can be solved using statistical approaches.

In the modern era, it is used mostly for *data analysis* and as an auxiliary tool for *anomaly detection* — data points located in regions of low density are more likely to be anomalies or outliers. Now it is usually solved with *density-based clustering* algorithms such as **DBSCAN** or **Mean Shift**, and using *Expectation-Maximization* algorithm into **Gaussian Mixture Models**.

Association Rule Learning

Association Rule Learning (also called *Association Rules* or simply *Association*) is another unsupervised learning task. It is most often used in business analysis to maximize profits.

It aims to detect unobvious relationships between variables in a dataset, so also can be considered as a *data analysis* tool. There are many complex algorithms to solve it, but the most popular are:

- **Apriori** — based on breadth-first search;
- **Eclat** (*Equivalence Class Transformation*) — based on depth-first search; and
- **FP-Growth**— designed to detect frequently occurring patterns in the data.

A common example of such a task is *product placement*. For example, knowing that people often buy onions together with potatoes in supermarkets, it makes sense to place them side by side to increase sales. Therefore, associative rules are used in promotional pricing, marketing, continuous production, etc.