Chapter2

ML TYPES

What is Machine Learning?

Machine Learning is the art & science of programming computers to learn from data.

Why use Machine Learning?

When building non-learners, we usually follow these steps:

1. We make rules
2. We write an algorithm
3. If the algorithm performs well, we deploy. If not, we go back to step

Examples of Applications

- Image Classification: typically performed using convolutional neural networks.
- Semantic segmentation: the algorithm is trained to classify each pixel in an image, one example of this is brain tumor detection.
- Natural Language Processing (NLP): More specifically, text classification, which can be learned using RNNs, CNNs, or Transformers.
- Chatbots: Involve many NLP tasks such as Natural Language Understanding (NLU) and Question-Answering.
- Forecasting future revenue: a regression task that can be tackled using multiple algorithms such as:
    - Linear Regression
    - Polynomial Regression
    - SVM
    - Random Forest
    - Artificial Neural Networks
- Speech recognition: this problem can be tackled by recognizing the incoming audio signals using RNNs, CNNs or Transformers.
- Credit card fraud detection: detecting frauds can be solved using supervised (classification) or unsupervised (anomaly detection) learning.
- Clustering: segmenting clients based on their purchases so we can design targeted & more effective marketing campaigns.
- Dimensionality reduction: useful for high-dimensional data visualization and cluster analysis. It can be solved using algorithms such as PCA or T-SNE.
- Recommender systems: where we can feed in the sequence of client purchases (for example) to an artificial neural network to predict the next purchase.

Types of Machine Learning Systems

Supervised Learning

Supervised learning involves training a model on a labeled dataset, meaning that each training example is paired with an **output label**.

- **Classification**

  Predicts discrete class labels, such as whether an email is spam, by assigning new data to predefined categories.

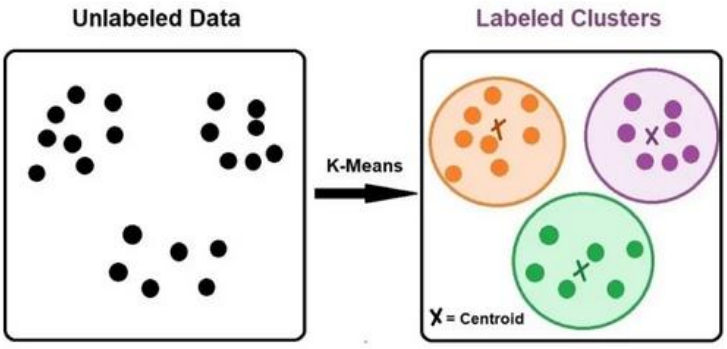
- **Regression**
  Predicts continuous quantities, such as a house's price, by providing an output variable based on one or more input variables.

Unsupervised Learning

In unsupervised learning, the data is unlabeled, the system is trying to learn without a teacher by finding internal structure within the dataset. Here are some unsupervised learning algorithms

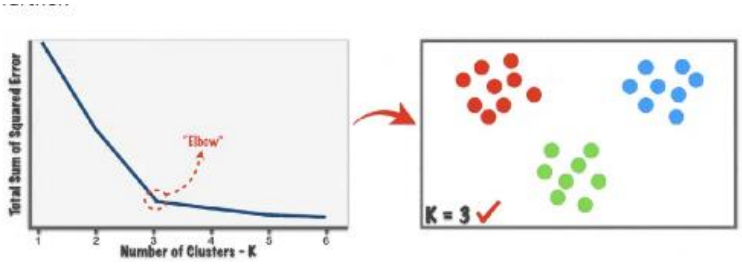
- **Clustering**

  - K-means: K-means aims to find groups in the data, with the number of groups represented by the variable K. Based on the provided features, the algorithm works iteratively to assign each data point to one of the “K” groups.



We can calculate variable K from Elbow method.

Elbow Method: calculate sum square error (clustering data and see the biggest change)



- Hierarchical Cluster Analysis (HCA): is a technique for optimal and compact connection of objects based on empirical similarity measures. The two most similar objects are assigned one after another until all objects are finally in one cluster. This then results in a tree-like structure.

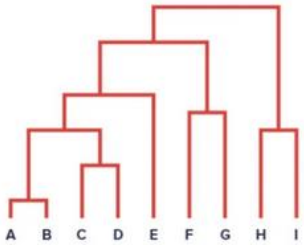
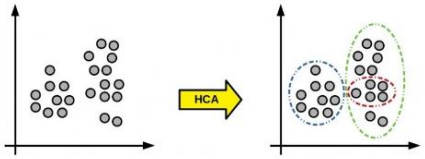
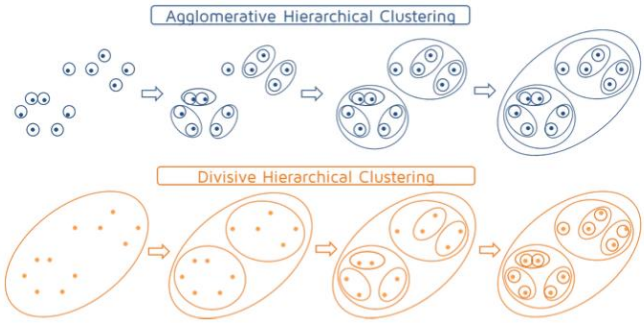


Figure 2: Visual from Data Viz Project

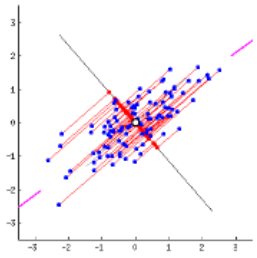


  - Bottem up
  - Devision

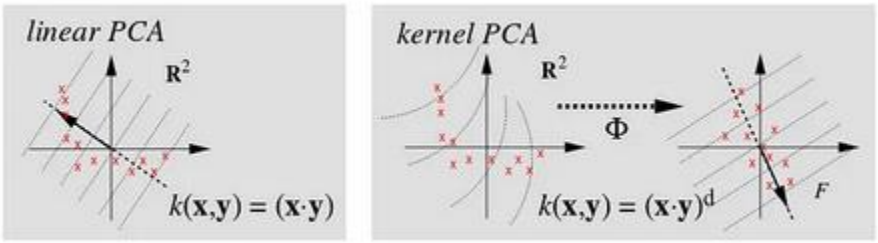


- Expectation Maximization(EM): is an iterative optimization method that combines different unsupervised machine learning algorithms to find maximum likelihood or maximum posterior estimates of parameters in statistical models that involve unobserved latent variables. The EM algorithm is commonly used for latent variable models and can handle missing data. It consists of an estimation step (E-step) and a maximization step (M-step), forming an iterative process to improve model fit.

- **Visualization and dimensionality reduction**
  - **Principal Component Analysis (PCA):** Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction while preserving as much variability in the data as possible.(it's a linear model that can only be applied to datasets which are linearly separable.)

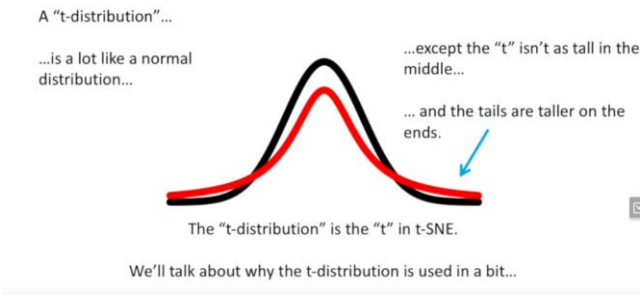


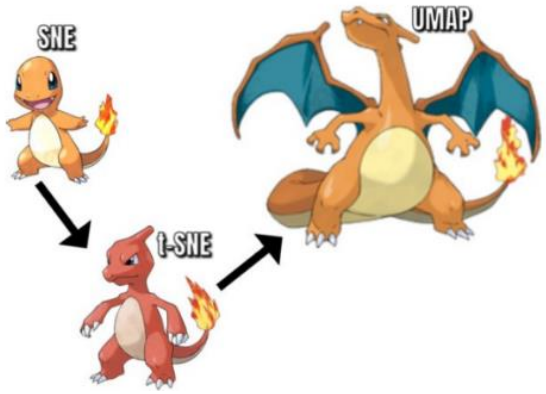
  - **Kernel PCA:** is a technique used in machine learning for nonlinear dimensionality reduction.



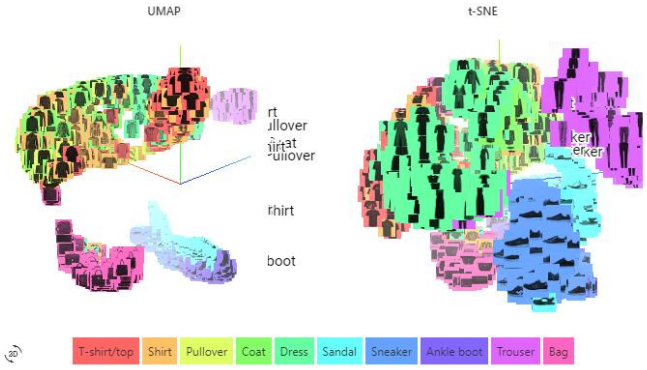
  - **Locally-Linear Embedding (LLE):** Unlike PCA, which is a linear method, LLE can capture the nonlinear structure of the data.

  - **t-distributed Stochastic Neighbor Embedding (t-SNE):** Unlike PCA and LLE, t-SNE is specifically designed to handle nonlinear data and is particularly effective for creating two- or three-dimensional maps from data with many dimensions.





UMAP: is a new technique by McInnes et al. that offers a number of advantages over t-SNE, most notably increased speed and better preservation of the data's global structure



- **Association rule learning algorithms find interesting relations between attributes**

  a rule-based machine learning method for discovering interesting relations between variables in large databases

  - Apriori: This algorithm uses frequent datasets to generate association rules. We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets. This algorithm uses a ***Breadth-First Search*** algorithm and ***Hash-Tree*** to calculate the itemset efficiently.

    Apriori uses a "bottom-up" approach, where it starts by finding frequent individual items and then builds up to larger and larger itemsets, as long as they appear together often enough in the data.

    Example:

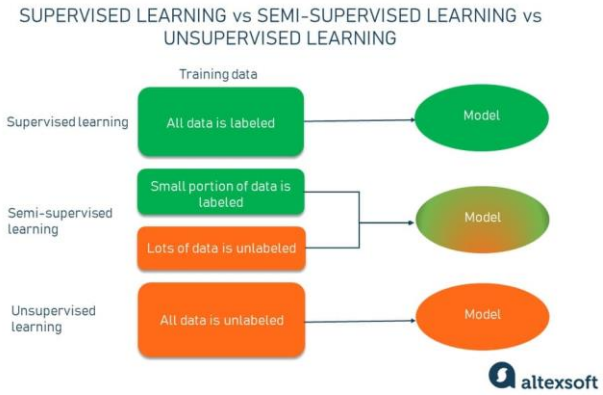
    https://kaumadiechamalka100.medium.com/apriori-algorithm-examples-be8915b01cf2

  - Eclat: It is a more efficient and scalable version of the Apriori algorithm. While the Apriori algorithm works in a horizontal sense imitating the Breadth-First Search of a graph, the ECLAT algorithm works in a vertical manner just like the Depth-First Search of

a graph. This vertical approach of the ECLAT algorithm makes it a faster algorithm than the Apriori algorithm.

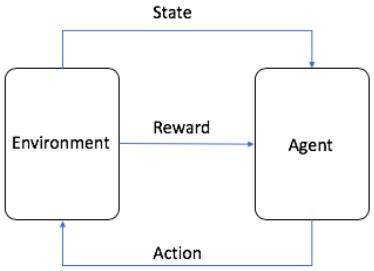
Semi-supervised Learning

In semi-supervised learning, we have partially-labeled data. The goal is to to use unlabeled data around the labeled data as helpers to solve the task. Most semi-supervised learning algorithms are a combination of unsupervised and supervised learning algorithms.



Reiforcement Learning

Reinforcement learning (RL) is a machine learning (ML) technique that trains software to make decisions to achieve the most optimal results. It mimics the trial-and-error learning process that humans use to achieve their goals.

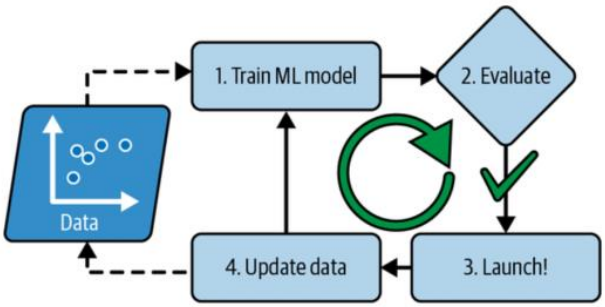


An agent observes the environment, selects an action, gets a reward, and updates its policy.

We can also categorize ML systems to **batch** or **online** algorithms. The question is is whether the algorithm will learn from an incoming stream of data or not.

Batch vs. Online ML Algorithms

In batch learning, the model is incapable of incremental learning, it starts by learning from all of the available data offline, and then gets deployed to produce predictions without feeding it any new data points. Another name of batch learning is Offline Learning.



In online learning, we train the data incrementally by continuously feeding it data instances as they come, either individually or in small groups of instances called *mini-batches*. Each learning step is fast and cheap, so the system can learn as data comes, on the fly. Online learning is great for systems that receive data in a continuous flow.

the model is trained continuously as new data becomes available. Unlike traditional batch learning, which trains the model on a fixed dataset all at once, online learning updates the model incrementally, making it well-suited for dynamic and streaming data environments.

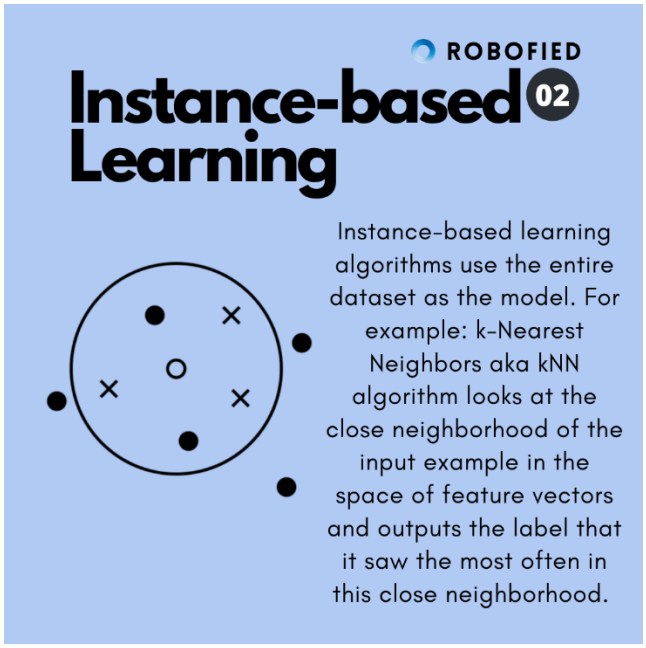
One important aspect of online learning is how fast the learning algorithm should adapt to new data points or to changes to the overall data distribution. With a big learning rate, the model tends to forget past data and lean heavily towards new data points. With a small learning rate, the model tends to slightly adapt to new data points but keeps its knowledge on old data points mostly intact.

A big challenge with online learning algorithms is that they can be damaged with bad incoming data points and clients will notice that on the fly to mitigate this, we can closely monitor the system through performance metrics and turn off online learning or revert back to a previous model state. We have to also make sure we clean the data before feeding it to the model by conducting anomaly/outlier detection.

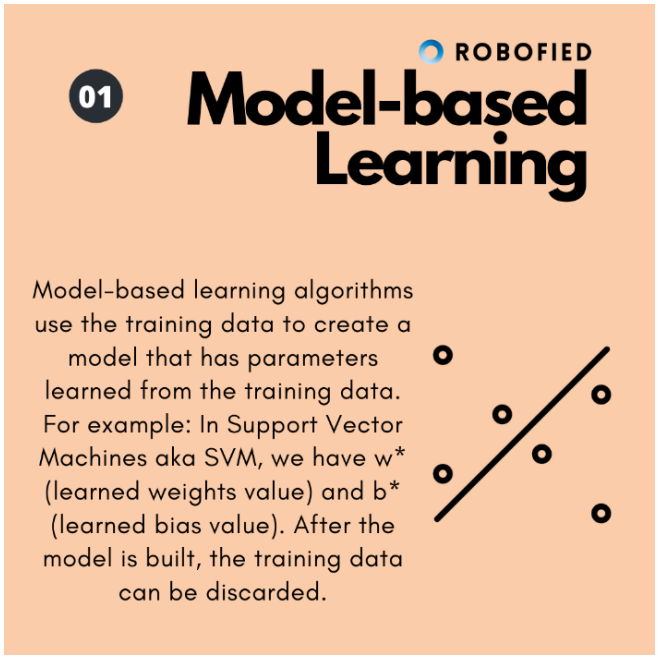
Instance-based versus Model-based Learning

One other way to categorize machine learning algorithms is how they generalize. There are two approaches to generalization: **instance-based approaches** and **model-based approaches.**

With instance-based Learning, we perform similarity-based comparisons, a new data point would be classified based on its similarity to the target group in the training set, this would require a measure of similarity.



In model-based learning we build a model for each class of data points and then use the model to classify a new data point (from the validation/test/production environment).



Compare T_SNE and UMAP:

https://pair-code.github.io/understanding-umap/