Machine Learning and Data Mining

05. May 2021.

Unsupervised learning

Unsupervised learning

- the information used to train is unlabeled.
- discover “interesting structure” in the data
- knowledge discovery

The GOAL

- find the underlying structure of dataset
- group that data according to similarities
- represent that dataset in a compressed

Frame Title

Advantages of Unsupervised Learning

- Unsupervised learning is used for more complex tasks
- Unsupervised learning is preferable as it is easier to get unlabeled data than labeled data
- Unsupervised machine learning finds all kind of unknown patterns in data
- help you to find features which can be useful for categorization.

Disadvantages of Unsupervised Learning

- Unsupervised learning is more difficult than supervised
- algorithms do not know the exact output in advance $\implies$ the result might be less accurate

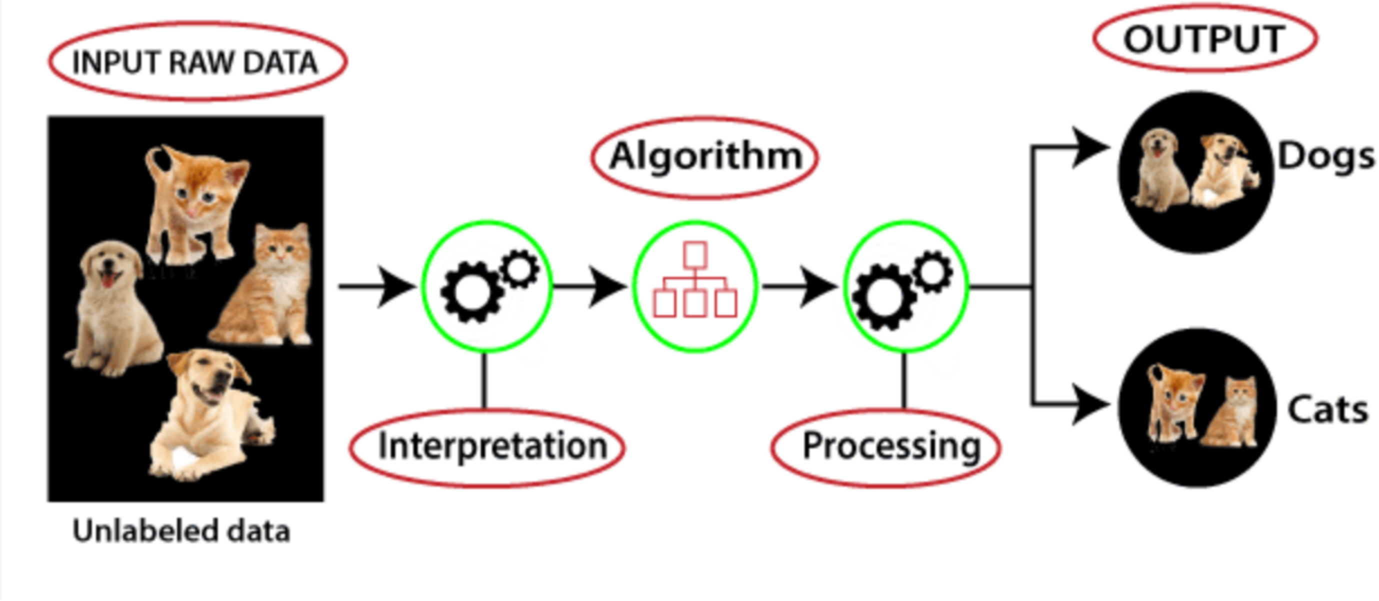


Figure 1: This Figure is from the https://www.javatpoint.com/unsupervised-machine-learning

Probabilistic methods

Types of Unsupervised Learning Algorithm

- Clustering (grouping the objects into clusters)
- Association (determines the set of items that occurs together in the dataset, e.g. Market Basket Analysis)

Unsupervised Learning Algorithms

- K-means clustering
- KNN (k-nearest neighbors)
- Hierarchal clustering
- Anomaly detection
- Neural Networks
- Principle Component Analysis
- Independent Component Analysis
- Singular value decomposition

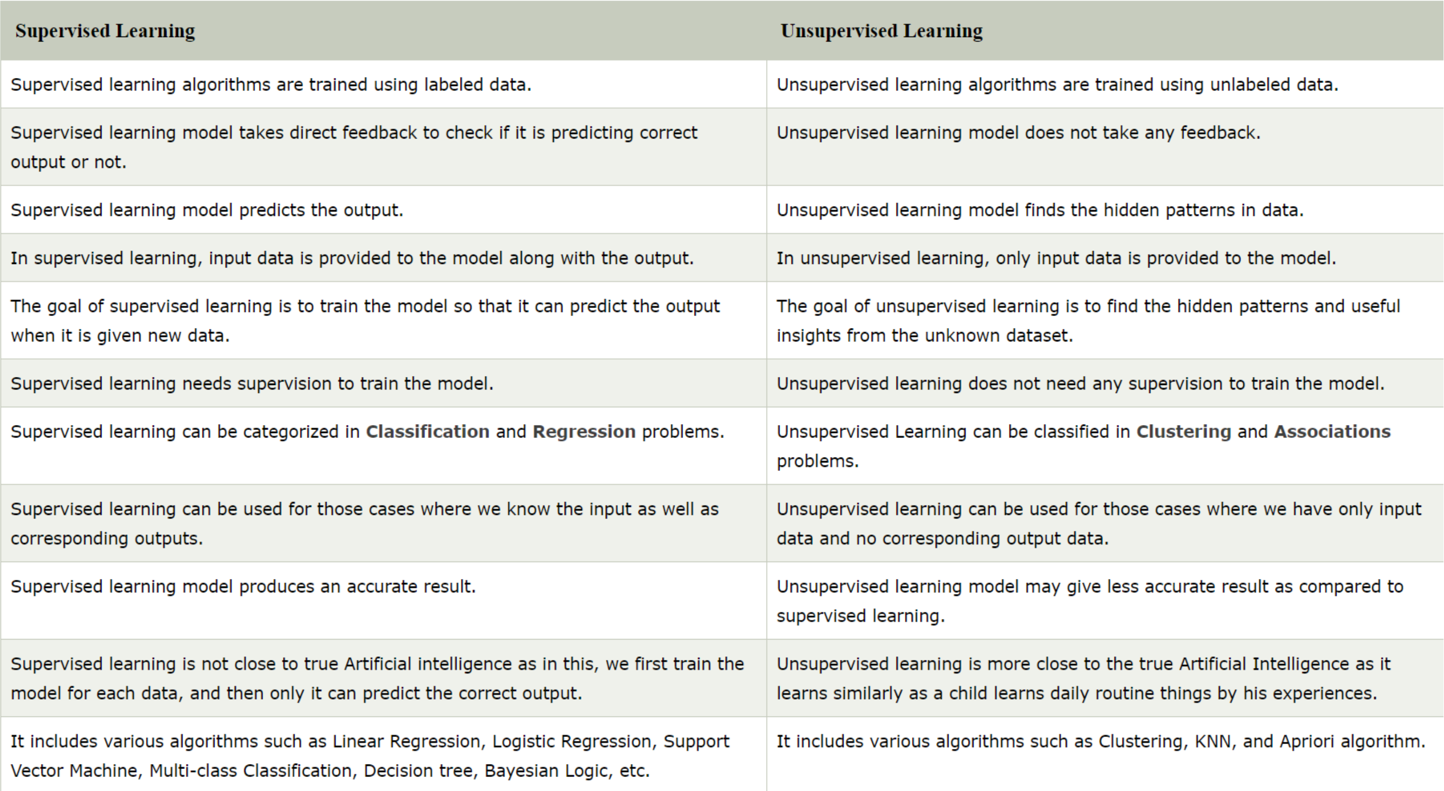
Clustering

different types of clustering

- Exclusive (partitioning) (Example: K-means)
- Agglomerative (Example: Hierarchical clustering)
- Overlapping (Example: Fuzzy C-Means)
- Probabilistic (Example: Following keywords: ”man’s shoe.” ”women’s shoe.” ”women’s glove.” ”man’s glove.” can be clustered into two categories ”shoe” and ”glove” or ”man” and ”women.”)

Clustering Types

- K-means clustering
- K-NN (k nearest neighbors)
- Hierarchical clustering
- Principal Component Analysis (PCA)
- Independent Component Analysis



Supervised Learning	Unsupervised Learning
Supervised learning algorithms are trained using labeled data.	Unsupervised learning algorithms are trained using unlabeled data.
Supervised learning model takes direct feedback to check if it is predicting correct output or not.	Unsupervised learning model does not take any feedback.
Supervised learning model predicts the output.	Unsupervised learning model finds the hidden patterns in data.
In supervised learning, input data is provided to the model along with the output.	In unsupervised learning, only input data is provided to the model.
The goal of supervised learning is to train the model so that it can predict the output when it is given new data.	The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.
Supervised learning needs supervision to train the model.	Unsupervised learning does not need any supervision to train the model.
Supervised learning can be categorized in **Classification** and **Regression** problems.	Unsupervised Learning can be classified in **Clustering** and **Associations** problems.
Supervised learning can be used for those cases where we know the input as well as corresponding outputs.	Unsupervised learning can be used for those cases where we have only input data and no corresponding output data.
Supervised learning model produces an accurate result.	Unsupervised learning model may give less accurate result as compared to supervised learning.
Supervised learning is not close to true Artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output.	Unsupervised learning is more close to the true Artificial Intelligence as it learns similarly as a child learns daily routine things by his experiences.
It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc.	It includes various algorithms such as Clustering, KNN, and Apriori algorithm.

Figure 2: This Table is from the https://www.javatpoint.com/difference-between-supervised-and-unsupervised-learning

K-means

- aims K-means aims to partition n observations $X = \{x_1, ..., x_n\}$ into K clusters $S = \{S_1, S_2, ..., S_k\}$ in which each observation belongs to the cluster with the nearest mean
- μ_i is the mean of the points in the cluster S_i
- each cluster is associated with a centroid.

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2 = \underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

Standard K-means algorithm

Standard algorithm

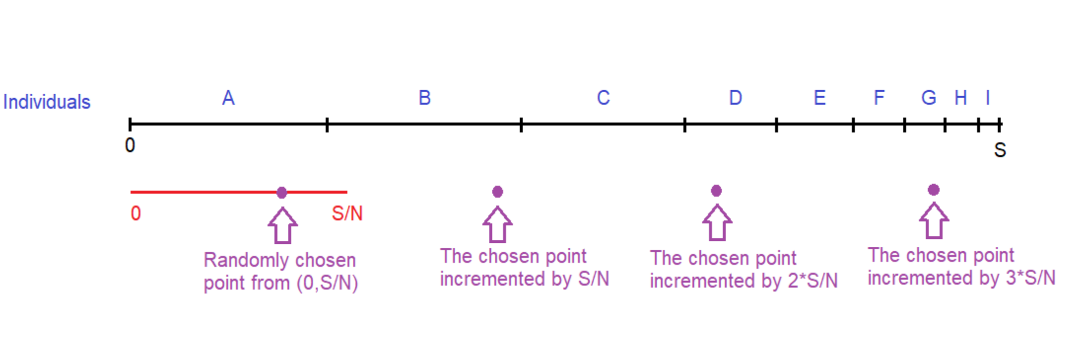
- Given an initial set of K means $\mu_1, ..., \mu_k$ the algorithm proceeds by alternating between two steps
- Assignment step: Assign each observation to the cluster with the nearest mean
- Update step: Recalculate means (centroids) for observations assigned to each cluster.
- The algorithm has converged when the assignments no longer change.

K-means

The steps of the K-Means algorithm

- Step-1: Select the number the number of clusters K.
- Step-2: Select K random points (centroids).
- Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.
- Step-4: Calculate the variance and place a new centroid of each cluster.
- Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.
- Step-6: If any reassignment occurs, then go to step-4.

Stochastic universal sampling (SUS)


Individuals
A
B
C
D
E
F
G
H
I
0
S
0
S/N
Randomly chosen point from (0,S/N)
The chosen point incremented by S/N
The chosen point incremented by 2*S/N
The chosen point incremented by 3*S/N

This Figure is from the dissertation entitled ”Stochastic modeling and statistical properties of the Limit Order Book”, Dragana Radojicic

The SUS algorithm aims to choose $N = 4$ individuals from the population $\{A, B, \ldots, I\}$. Assume that the individuals are sorted descending with respect to their fitness values and each individual is placed on the $(0, S)$ line taking exactly the same length as its fitness value. One number is randomly chosen from $(0, S/N)$ and the first selected individual is A. After incrementation of that point by S/N three times, each time one new individual is selected (i.e. in particular B, D, G in this Figure).

The world of automation

- article from Washington Post: ”The robots-vs.-robots trading that has hijacked the stock market”, roughly 50% of all trading volume is executed by the robots.
- Stock markets are nowadays producing vast portions of data.
- The financial markets hold memory properties.

Introduction: Limit Order Book (LOB)

List of all the waiting buy and sell orders.

- The LOB records all unexecuted limit orders.
- For a given price, orders are arranged in a FIFO stack.
- Tick is a minimal distance between two price levels (points in a discrete price grid).
- The spread is difference between the best ask and the best bid price.

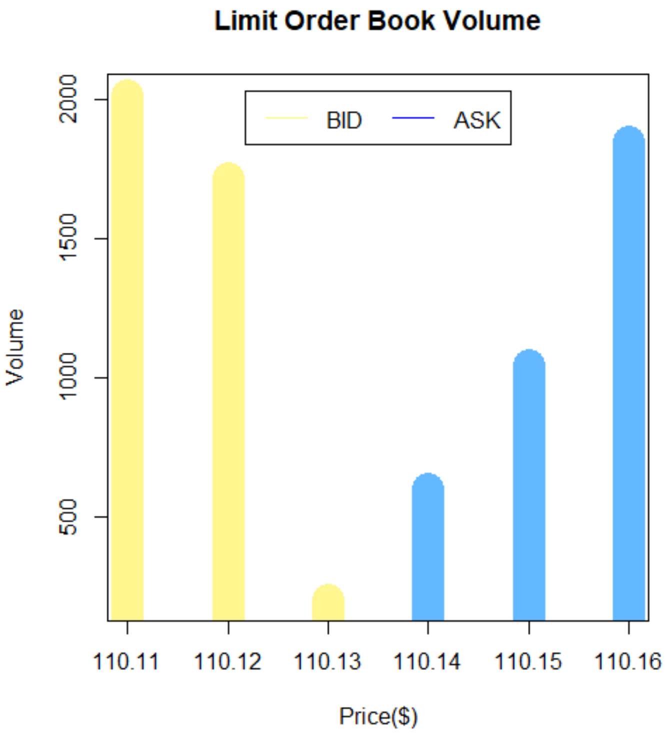


Figure 2: LOB

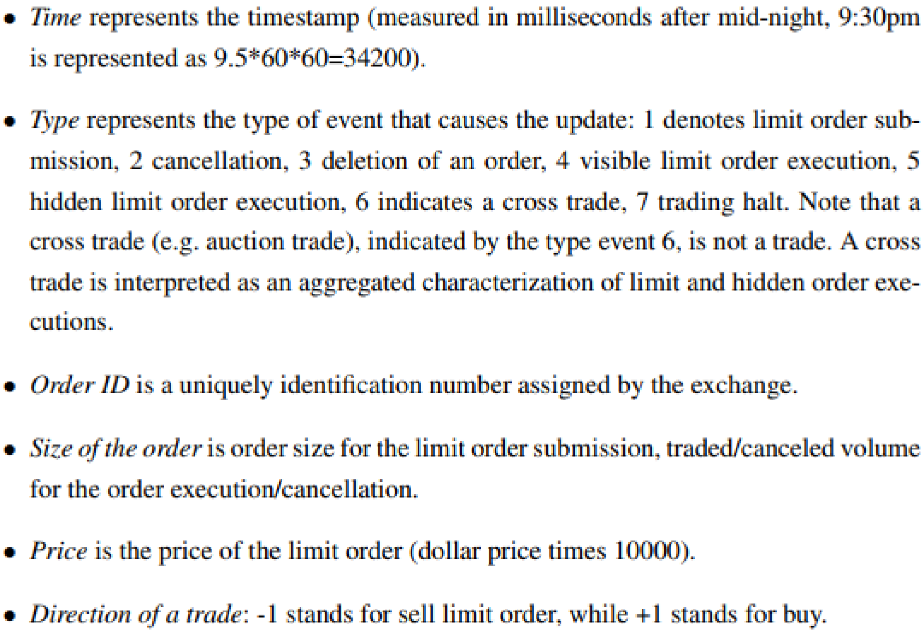
The qualitative data analysis

NASDAQ (second largest exchange in the world)

- Our research is based on high-quality online limit order book data tool LOBSTER (www.lobsterdata.com).
- LOBSTER has information for the entire NASDAQ stock exchange from the 27th of June 2007 up to the two days ago from the current day.
- 'orderbook' file - keep track of evolution of the limit order book
- 'message' file - contains information of the kind of event which update the limit order book (i.e. Time, Type, Order ID, Size of the order, Price, Direction of a trade)

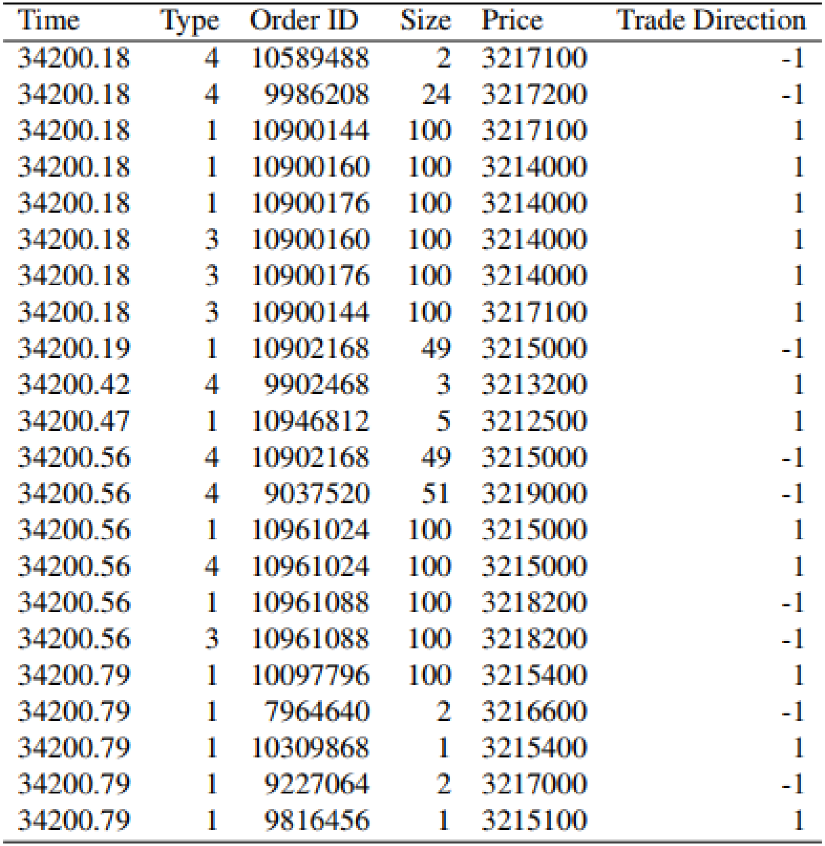
GOAL

Goal is to develop a foundation which allows to easily match similar points together via unsupervised learning as well as to classify elements into groups via supervised learning (more precisely classification).



- *Time* represents the timestamp (measured in milliseconds after mid-night, 9:30pm is represented as 9.5*60*60=34200).
- *Type* represents the type of event that causes the update: 1 denotes limit order submission, 2 cancellation, 3 deletion of an order, 4 visible limit order execution, 5 hidden limit order execution, 6 indicates a cross trade, 7 trading halt. Note that a cross trade (e.g. auction trade), indicated by the type event 6, is not a trade. A cross trade is interpreted as an aggregated characterization of limit and hidden order executions.
- *Order ID* is a uniquely identification number assigned by the exchange.
- *Size of the order* is order size for the limit order submission, traded/canceled volume for the order execution/cancellation.
- *Price* is the price of the limit order (dollar price times 10000).
- *Direction of a trade*: -1 stands for sell limit order, while +1 stands for buy.

This explanation is from the dissertation entitled ”Stochastic modeling and statistical properties of the Limit Order Book”, Dragana Radojicic



Time	Type	Order ID	Size	Price	Trade Direction
34200.18	4	10589488	2	3217100	-1
34200.18	4	9986208	24	3217200	-1
34200.18	1	10900144	100	3217100	1
34200.18	1	10900160	100	3214000	1
34200.18	1	10900176	100	3214000	1
34200.18	3	10900160	100	3214000	1
34200.18	3	10900176	100	3214000	1
34200.18	3	10900144	100	3217100	1
34200.19	1	10902168	49	3215000	-1
34200.42	4	9902468	3	3213200	1
34200.47	1	10946812	5	3212500	1
34200.56	4	10902168	49	3215000	-1
34200.56	4	9037520	51	3219000	-1
34200.56	1	10961024	100	3215000	1
34200.56	4	10961024	100	3215000	1
34200.56	1	10961088	100	3218200	-1
34200.56	3	10961088	100	3218200	-1
34200.79	1	10097796	100	3215400	1
34200.79	1	7964640	2	3216600	-1
34200.79	1	10309868	1	3215400	1
34200.79	1	9227064	2	3217000	-1
34200.79	1	9816456	1	3215100	1

The 'message' file sample of the LOBSTER data for TSLA ticker (Tesla, Inc.) company on January 7th 2019.

The data analysis

Label data

- The market data at a given time point t can be formally defined as a vector x_t, which will consist of market data information and various technical analysis markers
- The idea is to express trader as a function with an input vector x_t such that output is one of the values from the set $\{S = idle, sell, buy\}$.

The classification, regression predictions and the latest research in the field of Artificial Intelligence can be applied in order to successfully classify a time series of market data.