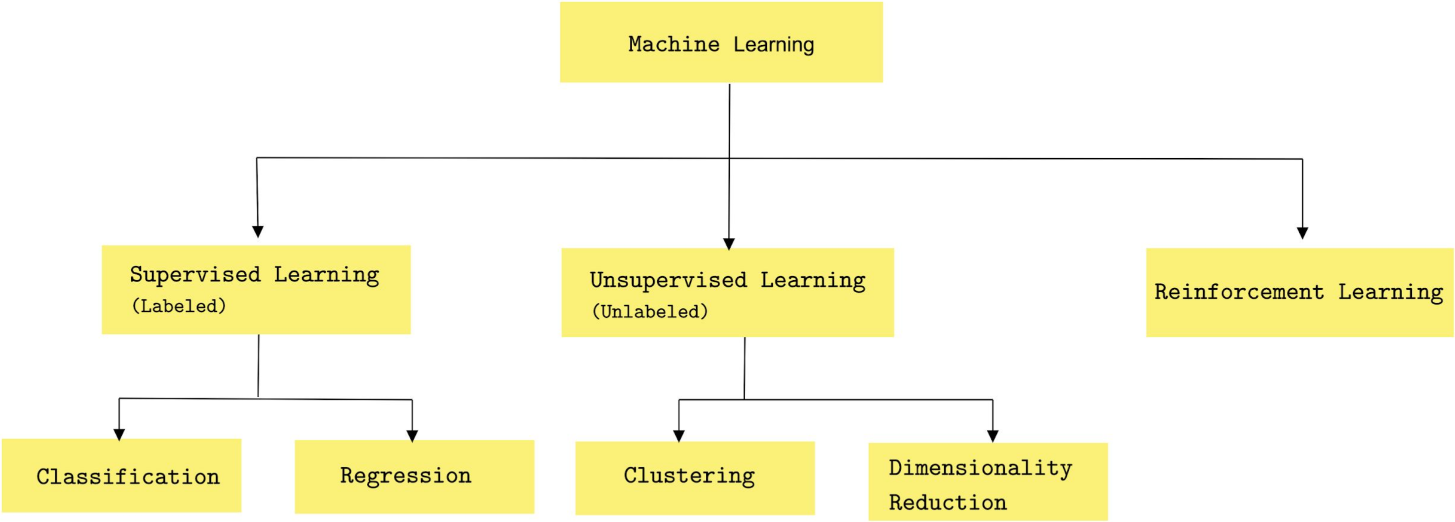
Model Development & Evaluation

DEC26, 2020

From EDA to prediction and analysis

- The reason we performed EDA was to prepare our dataset and make sense of it so that it can be used for predictive and analytical purposes.
- By predictive and analytical, we mean to create and evaluate Machine Learning (ML) models.
- Data scientist need to understand the different types of machine learning
    - supervised learning
    - unsupervised learning
    - reinforcement learning

Types of machine learning



Machine learning (ML) is a field of computer science that deals with the creation of algorithms that can discover patterns by themselves without being explicitly programmed

Supervised learning

- The primary objective of supervised learning is to generalize a model from **labeled training data.**
- Once a model has been trained, it allows users to make predictions about unseen future data.
- **Labeled training data** mean the training examples know the associated output labels (supervised learning).
- The learning process can be thought of as a teacher supervising the entire process. In such a learning process, we know the correct answer initially, and the students learn enough iteratively over time and try to answer unseen questions. The errors in the answers are corrected by the teacher. The process of learning stops when we can ensure the performance of the student has reached an acceptable level.
- In supervised learning, we have input variables (x_i) and output variables (Y_i). With this, we can learn a function, f, as shown by the following equation:

$$Y_i = f(x_i)$$

Supervised learning $Y_i = f(x_i)$

- The objective is to learn a general mapping function, *f,* so that the function can predict the output variable, *Y*, for any new input data, *x*. Supervised learning algorithms can be categorized into two groups:
  - **Regression** : A regression problem has an output variable or dependent variable that is a real value, such as weight, age, or any other real numbers. It include different types of regression
    - simple linear regression
    - multiple linear regression
    - non-linear regression
  - **Classification:** A classification problem has the output variable in the form of a category value; for example, red or white; young, adult, or old. For classification problems, there are different types of classification algorithms.
    - Linear classifier: Naive Bayes classifier, logistic regression, linear SVM
    - Decision tree classifier
    - Support vector machines
    - Random Forest classifier

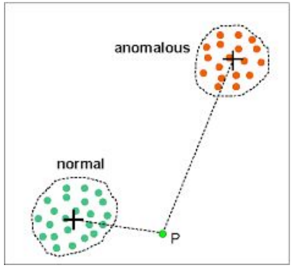
Unsupervised learning

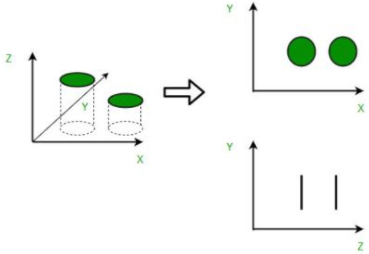
- Unsupervised machine learning deals with unlabeled data to discover all kinds of unknown patterns in the data and can facilitate useful categorization.
- There are several types of unsupervised learning algorithms but the two major unsupervised learning tasks are:
  - **Clustering** : Categorize the dataset into several similar groups, referred to as a cluster. Each cluster represents a group of similar points.
  - **Association mining**: Find frequently occurring items in our dataset.
  - **Anomaly detection**: Determine unusual data points in any existing dataset.
  - **Dimensionality reduction**: Used in data processing in order to reduce the number of features in a dataset. This is one of the most important tasks to perform in unsupervised learning.

sample

Cluster/group

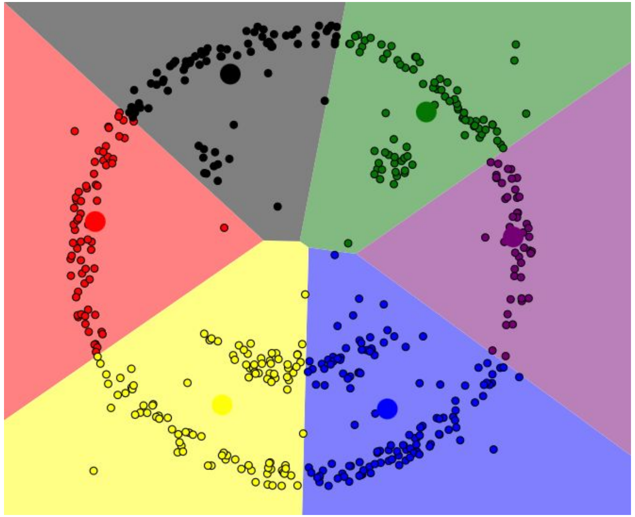
Milk, eggs, sugar, bread
Milk, eggs, cereal, bread
Eggs, sugar
Customer1
Customer2
Customer3





Clustering Techniques : K-means

The k-means algorithm captures the insight that each point in a cluster should be near to the center of that cluster.

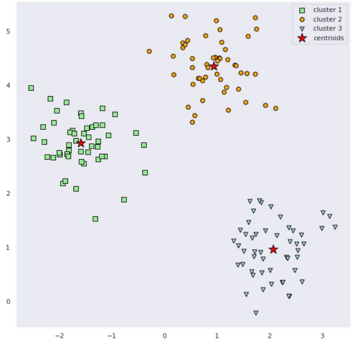


- ***Choose k***, the number of clusters we want to find in the data.
- **Centers** of those k clusters, called centroids, are initialized
- ***Reassign*** every point in the data to the cluster whose centroid is nearest to it.
- ***Update Centroids:*** recalculate each centroid's location as the mean (center) of all the points assigned to its cluster.
- ***Iterate*** until the centroids stop moving, or equivalently until the points stop switching clusters.
- Demo --> https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Clustering Techniques : K-means

```
from sklearn.cluster import KMeans

km = KMeans(
    n_clusters=3, init='random',
    n_init=10, max_iter=300,
    tol=1e-04, random_state=0
)
y_km = km.fit_predict(X)
```


cluster 1
cluster 2
cluster 3
centroids