unsupervised learning algorithms

- Clustering
  - K-Means
  - DBSCAN
  - Hierarchical Cluster Analysis (HCA)
- Anomaly detection and novelty detection
  - One-class SVM
  - Isolation Forest

unsupervised learning algorithms

- Visualization and dimensionality reduction

  Principal Component Analysis (PCA)
  - Kernel PCA
  - Locally-Linear Embedding (LLE)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)
- Association rule learning
  - Apriori
  - Eclat

1) Traditional & Machine learning

- Traditional learning refers to process of acquiring knowledge and skills through methods like in-person teaching / reading books
- ML is subset of AI where computers use algorithm to learn from data & make decisions without programmed

2) Batch learning

- Trains model using entire dataset at once
- It suitable for where data is static and can't change frequently
- Process entire dataset

Online learning

- Well suited for scenarios with evolving data / when not feasible to stored & process entire dataset at once
- Updates model incrementally with new data

3) Instance Vs Model based

→ Instance based relies on similarity to existing data points for predictions, whereas model based learning uses mathematical model derived from training data to make predictions

4) Supervised learning

- Type of ML where model is trained on labeled dataset. Each data point has target labels / outcomes
- Used for tasks like classification & regression.

Unsupervised Learning

- Deals with unlabeled data, where model explore structures in data without guidance
- Aims to discover hidden relationships / groups within data

5) Variance Bias

- Refers to error introduced by approximating real world, may be complex.
- Results in model being overly generalized and not fitting training data well

Bias Variance

- Refers to error introduced due to model's sensitivity to small fluctuations.
- Results in model that is too complex & capture noise instead of true patterns

6) Machine learning:

→ subset of AI that automatically enables machine / system to learn and improve from experience

Overfitting

- Occurs when machine learning model is too complex & learns training data too well,
- Leads to poor generalization on new unseen data

Underfitting

- Happens when model is too simple and can't capture underlying patterns in training data
- Results in poor performance

Main Challenges of Machine Learning

- Insufficient Quantity of Training Data
- Non representative Training Data
- Poor-Quality Data
- Irrelevant Features
- Over fitting the Training Data
- Under fitting the Training Data

Supervised learning algorithms

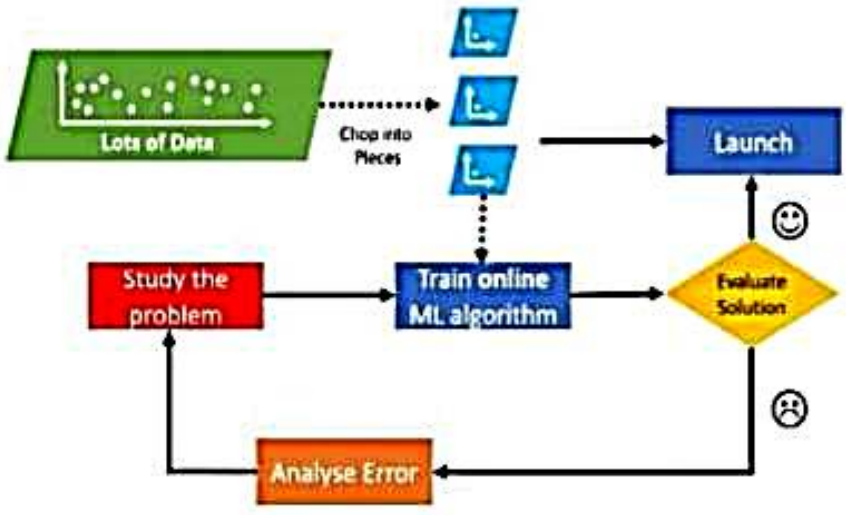
- k-Nearest Neighbors
- Linear Regression
- Logistic Regression
- Support Vector Machines (SVMs)
- Decision Trees and Random Forests
- Neural networks

- In online learning, you train the system incrementally by feeding it data instances sequentially, either individually or by small groups called mini-batches.

Why online learning vs batch or offline learning?

- Volume: The data comes in large volumes. This would thus require IT infrastructures, software systems, and appropriate expertise and experience to do the data processing.
- Velocity: As like in the case of the high volume of data, the data coming at high speed (for example, tweets) can also become key criteria
- Variety: Similar to volume and variety, the data can become of a different variety. For example, data for aggregator services such as Uber

Online Learning



Online learning algorithms can also be used to train systems on huge datasets that cannot fit in one machine's main memory (this is called out-of-core learning).

A big challenge with online learning is that if bad data is fed to the system, the system performances will gradually decline.

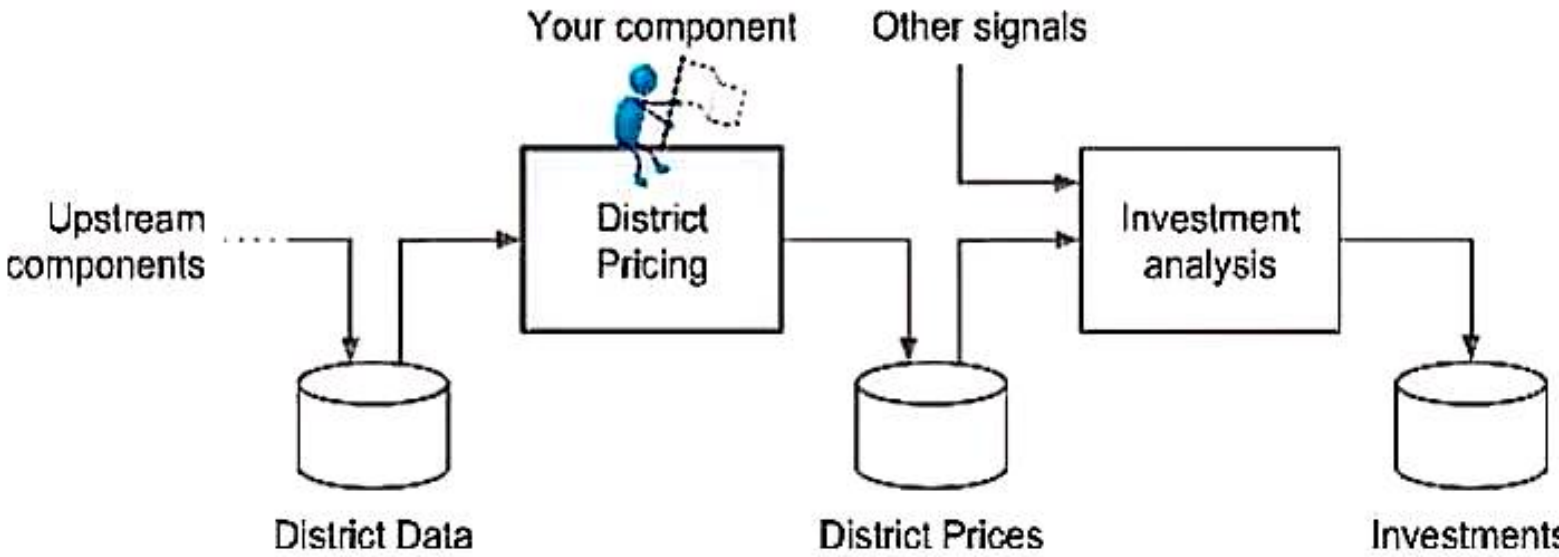
To reduce this risk, you need to monitor the systems closely and promptly switch learning off and possibly you want to revert to a previous working state if you detect a drop-in performance.

Pipeline

Pipeline:

A sequence of data processing *components* is called a data *pipeline*. Pipelines are very common in Machine Learning systems, since there is a lot of data to manipulate and many data transformations to apply.

A Machine Learning pipeline for real estate investments



Mean Absolute Error

$$\text{MAE}(\mathbf{X}, h) = \frac{1}{m} \sum_{i=1}^{m} \left| h\left(\mathbf{x}^{(i)}\right) - y^{(i)} \right|$$

For example, suppose that there are many outlier districts. In that case, you may consider using the *Mean Absolute Error*

Scientific Python

Python modules:

- ✓ Jupyter
- ✓ NumPy
- ✓ Pandas,
- ✓ Matplotlib
- ✓ Scikit-Learn

$$\mathrm{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2}$$

- m is the number of instances in the dataset
- $\mathbf{x}(i)$ is a vector of all the feature values (excluding the label) of the i th instance in the dataset, and
- $y(i)$ is its label (the desired output value for that instance).
- $\mathbf{X}$ is a matrix containing all the feature values (excluding labels) of all instances in the dataset
- h is hypothesis

At a high level, ML problem framing consists of two distinct steps,

1. Determining whether ML is the right approach for solving a problem.
2. Framing the problem in ML terms

HISTOGRAM

A histogram shows the number of instances (on the vertical axis) that have a given value range (on the horizontal axis).

You can either plot this one attribute at a time, or you can call the hist() method on the whole dataset, and it will plot a histogram for each numerical attribute

The info() method is useful to get a quick description of the data, in particular the total number of rows, and each attribute's type and number of non-null values

- Used for predicting the continuous dependent variable with the help of independent variables
- To find the that can accurately predict the output

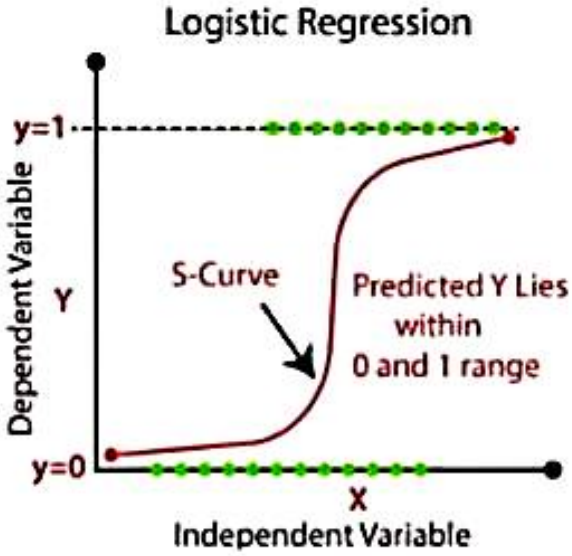
Difference b/w Linear and Logistic Regression

Linear Regression	Logistic Regression
Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.
Linear Regression is used for solving Regression problem.	Logistic regression is used for solving Classification problems.
In Linear regression, we predict the value of continuous variables.	In logistic Regression, we predict the values of categorical variables.
In linear regression, we find the best fit line, by which we can easily predict the output.	In Logistic Regression, we find the S-curve by which we can classify the samples.

Difference b/w Linear and Logistic Regression

Linear Regression	Logistic Regression
Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for estimation of accuracy.
The output for Linear Regression must be a continuous value, such as price, age, etc.	The output of Logistic Regression must be a Categorical value such as 0 or 1, Yes or No, etc.
In Linear regression, it is required that relationship between dependent variable and independent variable must be linear.	In Logistic regression, it is not required to have the linear relationship between the dependent and independent variable.
In linear regression, there may be collinearity between the independent variables.	In logistic regression, there should not be collinearity between the independent variable.

- Used for Classification as well as for Regression problems



- used to predict the categorical dependent variable with the help of independent variables
- output of Logistic Regression problem can be only between the 0 and 1
- used where the probabilities between two classes is required

Bayes' Theorem

- used to determine the probability of a hypothesis with prior knowledge

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Minimum Description Length Principle

- 'MDL' is a method for inductive inference...
- machine learning
- pattern recognition
- statistics
- ...based on ideas from data compression (information theory)
- In contrast to most other methods, MDL automatically deals with overfitting, arguably the central problem in machine learning and statistics

Minimum Description Length Principle

- MDL is based on the correspondence between 'regularity' and 'compression'
- The more you are able to compress a sequence of data, the more regularity you have detected in the data
- Example: 001001001001001001001001001001: :::001 01011011100100111010001010 1:::: 010

The eigenvectors **x** and eigenvalues λ of a matrix A satisfy

$$A\mathbf{x} = \lambda\mathbf{x}$$

If A is an n x n matrix, then **x** is an n x 1 vector, and λ is a constant.

The equation can be rewritten as $(A - \lambda I)\,\mathbf{x} = 0$, where I is the n x n identity matrix.