Machine Learning

CIberCATSS 2023

Cyberinfrastructure Comprehensive, Applied and Tangible Summer School

Instructor: Sammie Omranian

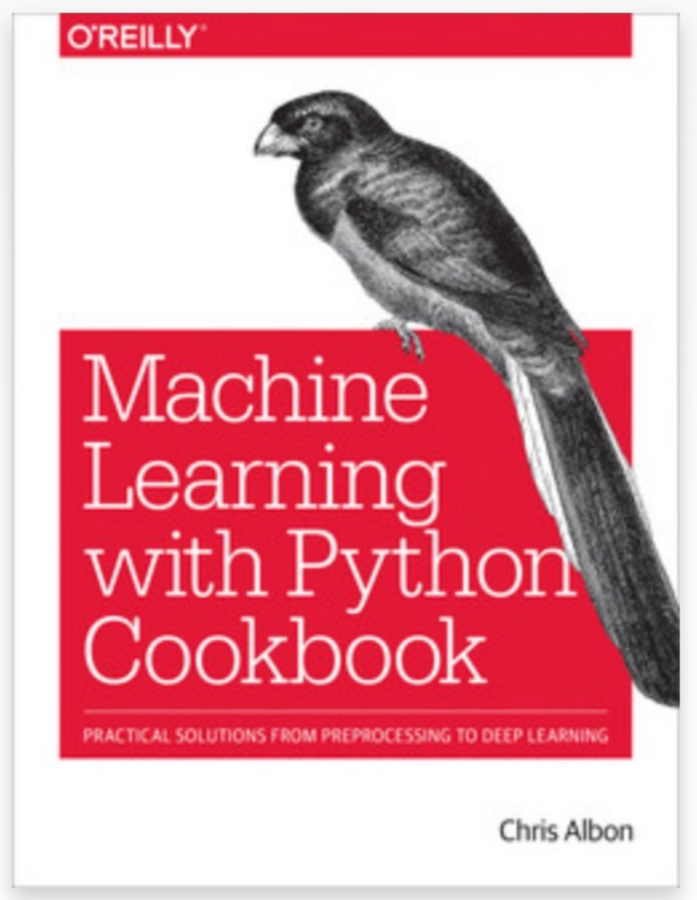
Book

Machine Learning with Python Cookbook

Book description from O'REILY website:

This practical guide provides nearly 200 self-contained recipes to help you solve machine learning challenges you may encounter in your daily work.

- Vectors, matrices, and arrays
- Handling numerical and categorical data, text, images, and dates and times
- Dimensionality reduction using feature extraction or feature selection
- Model evaluation and selection
- Linear and logical regression, trees and forests, and k-nearest neighbors
- Support vector machines (SVM), naïve Bayes, clustering, and neural networks
- Saving and loading trained models


O'REILLY
Machine Learning with Python Cookbook
PRACTICAL SOLUTIONS FROM PREPROCESSING TO DEEP LEARNING
Chris Albon

What is machine learning?

Arthur Samuel (1959), an American pioneer in the field of artificial intelligence, defined machine learning as:

The field of study that gives computers the ability to learn without being explicitly programmed.

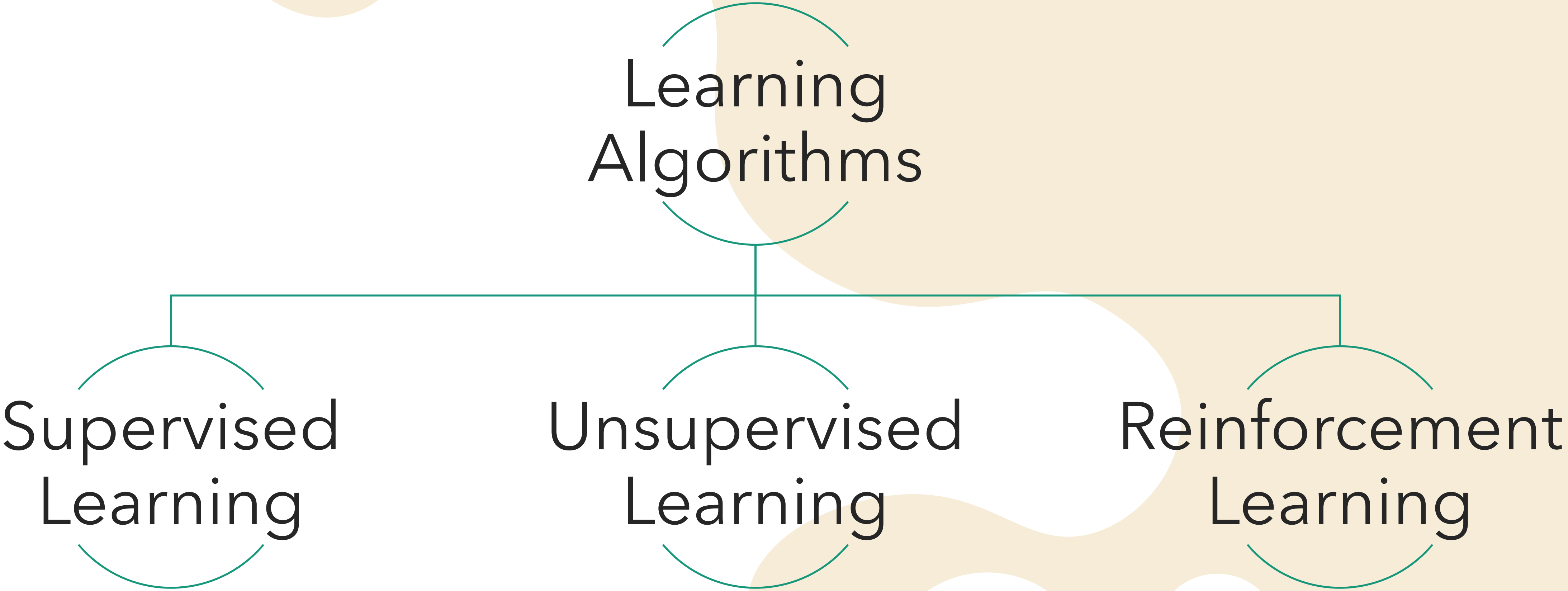
He developed the checkers-playing program, in the late 1950s, one of his most notable achievements in the field of artificial intelligence and machine learning.

What is machine learning?

Tom Mitchell, a renowned computer scientist, defined machine learning as:

"A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E."

Machine Learning Algorithms



Learning Algorithms

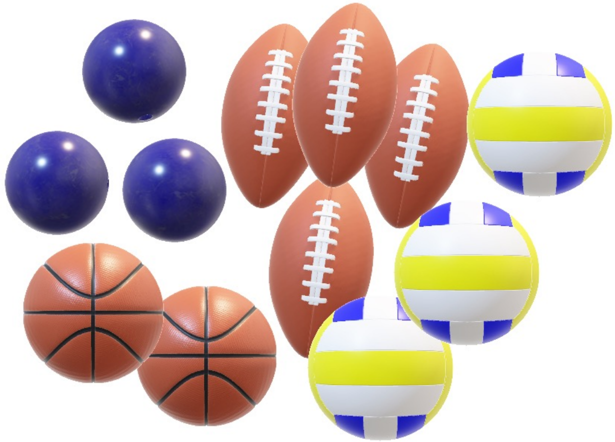
Supervised Learning

Unsupervised Learning

Reinforcement Learning

UNIVERSITY of WISCONSIN UWM MILWAUKEE

Supervised Learning



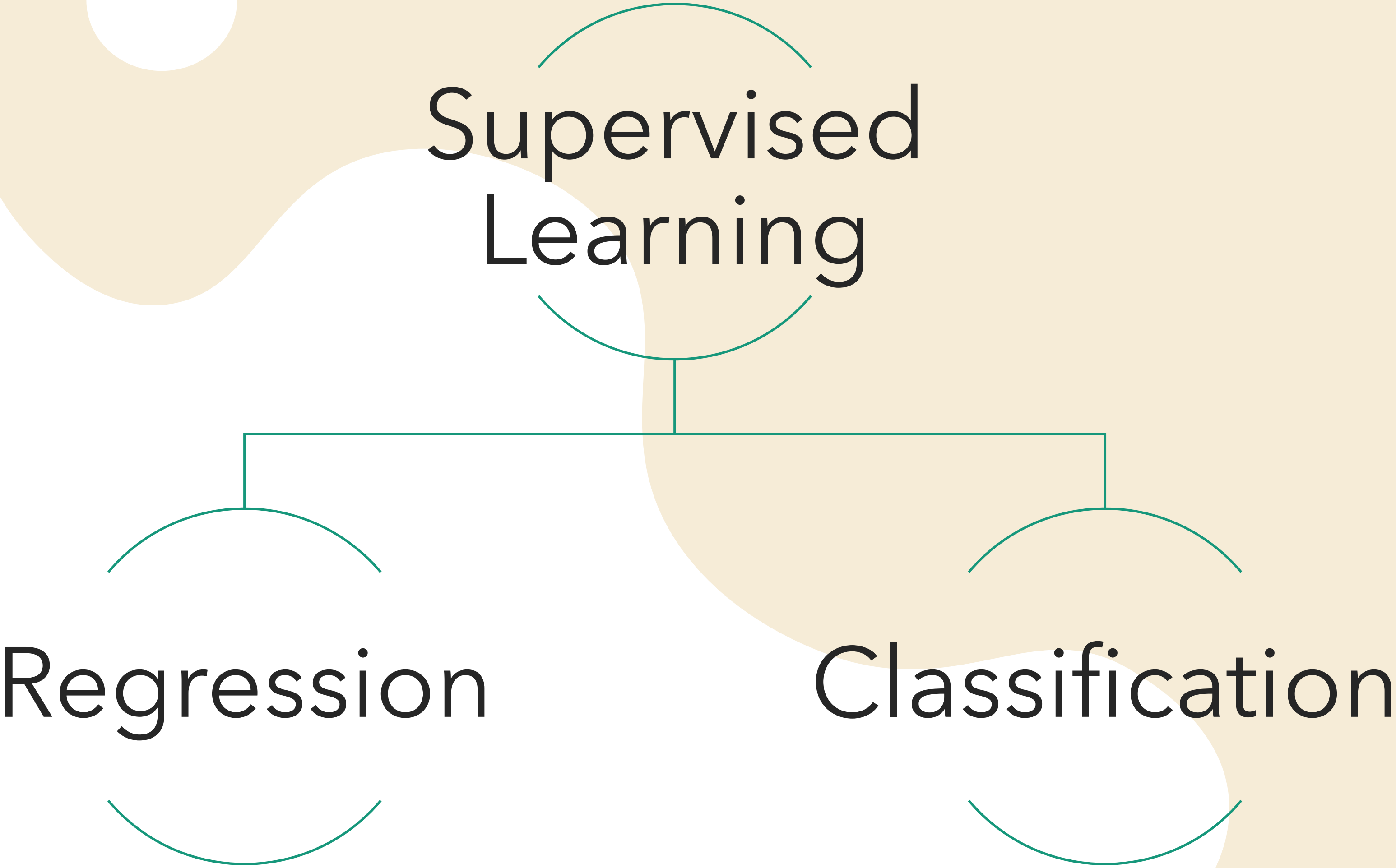
Labeled Data

Labels

Model Training

Prediction

Test Data



Supervised Learning

Regression

Classification

Regression

The task of predicting a continuous numerical value or a set of values based on input features.

The main objective in regression is to find the best-fit line or curve that minimizes the difference between the predicted values and the actual values.

Example: Predicting House Prices

Classification

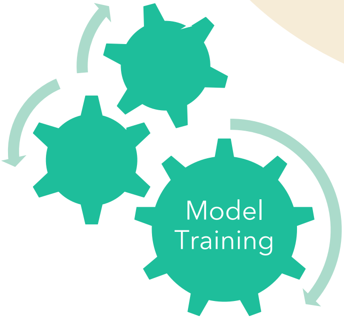
The task of categorizing data into predefined classes or categories based on their features.

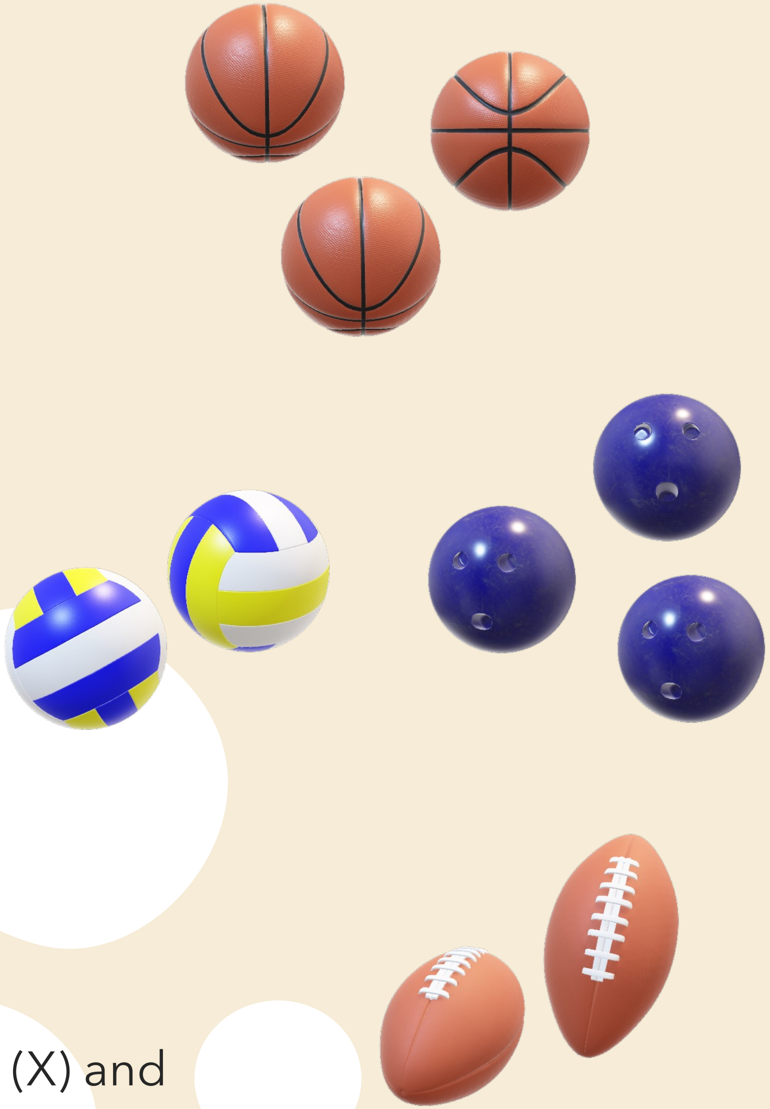
The goal is to build a model that can learn from labeled training data and then predict the class labels of new, unseen instances.

Example: Spam Email Detection

Unsupervised Learning

No labeled data! No Y!





Unsupervised learning is the concept of using unlabeled data (X) and finding interesting things about it.

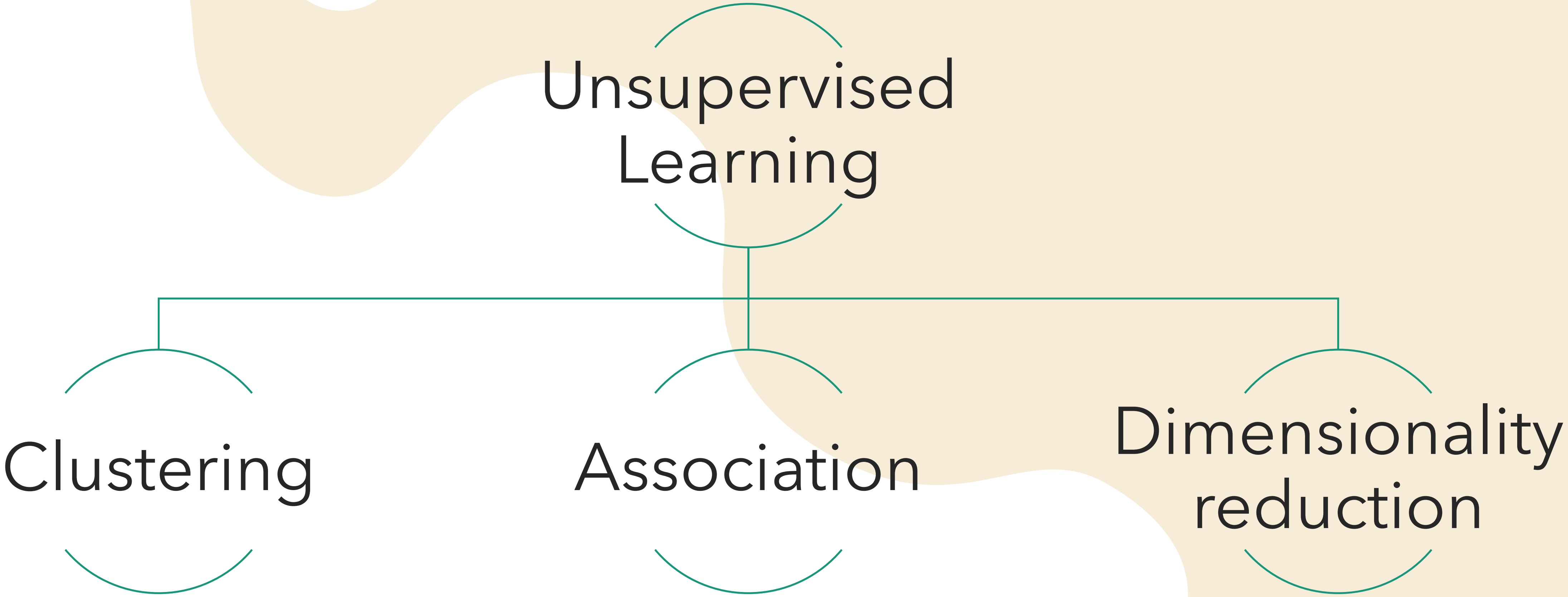
Unsupervised Learning

Example: Finding out which customers made similar product purchases

Clustering algorithms can group customers based on their purchasing patterns or demographics to identify different customer segments.

Example: Grouping news articles

Clustering news articles involves grouping them based on their similarities in terms of content, topics, or themes.



Unsupervised Learning

Clustering

Association

Dimensionality reduction

market segmentation

"Customers Who Bought This Item Also Bought”

To eliminate visual noise and enhance image quality

Reinforcement Learning

Reinforcement Learning (RL) is the science of decision making. In essence how an intelligent agent learn to make a good sequence of decisions.

In reinforcement learning, the agent receives feedback in the form of rewards or punishments based on its actions. The goal of the agent is to maximize the cumulative reward over time by learning an optimal policy.

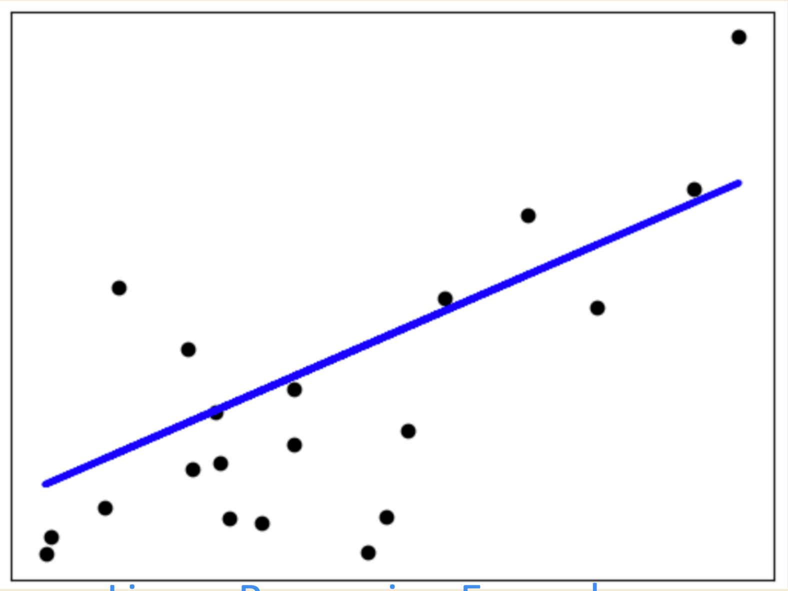
Examples: clinical decision-making, autonomous driving

Machine Learning Algorithms

• Linear Regression	to predict continuous numbers
• Logistic regression	to predict classes/categories
• Decision Trees	to predict both categorical and continuous variables
• Random Forest	classification and regression
• Support Vector Machine	classification
• Naïve Bayes	classification
• K-Nearest Neighbor	classification
• K-Mean Clustering	unsupervised learning

Machine Learning Algorithms

- Linear Regression



Linear Regression Example

Fitting a line

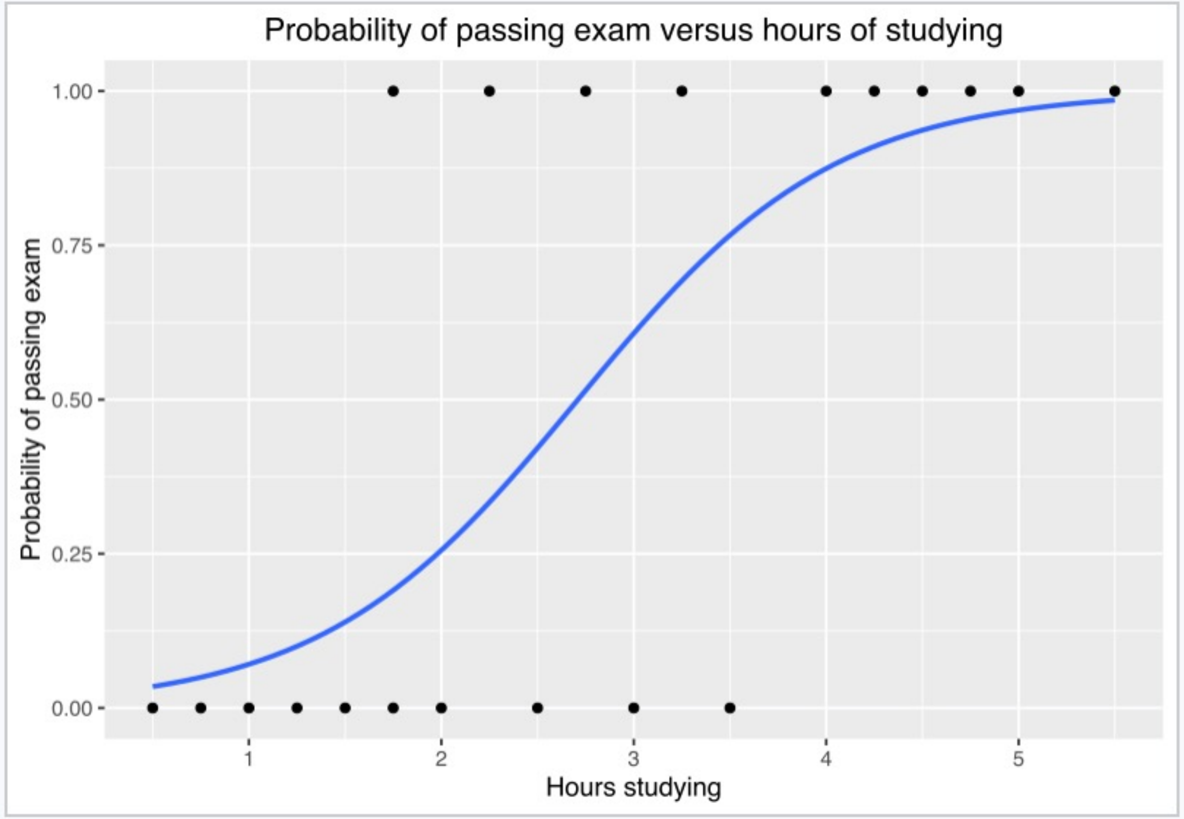
Linear regression, along with its variations and extensions, remains a widely used and valuable approach for making predictions in scenarios where the target variable is a numerical value, such as predicting home prices or estimating ages.

We will have a hands-on exercise that uses scikit-learn *Linear Regression* model.

Machine Learning Algorithms

- Logistic regression

Logistic regression statistical method used for binary classification, where the outcome variable is categorical with two possible values (e.g., yes/no, true/false).



Graph of a logistic regression curve fitted to the (x_m, y_m) data. The curve shows the probability of passing an exam versus hours studying.

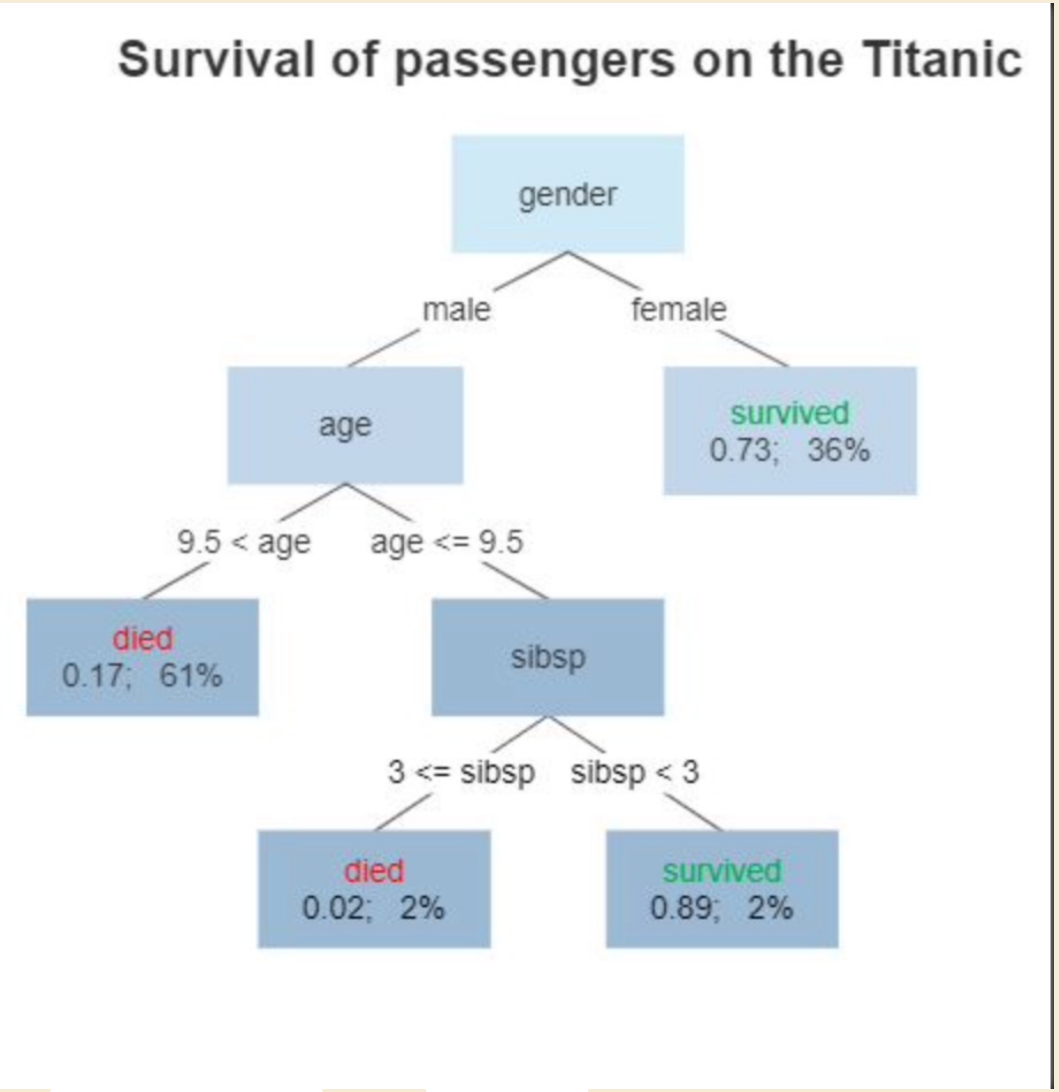
https://en.wikipedia.org/wiki/Logistic_regression#Example

Machine Learning Algorithms

- Decision Trees

Tree-based learning algorithms are widely used supervised methods for classification and regression tasks. They rely on decision trees, which use decision rules to make predictions.

DecisionTreeClassifier is capable of both binary (where the labels are [-1, 1]) classification and multiclass (where the labels are [0, …, K-1]) classification.

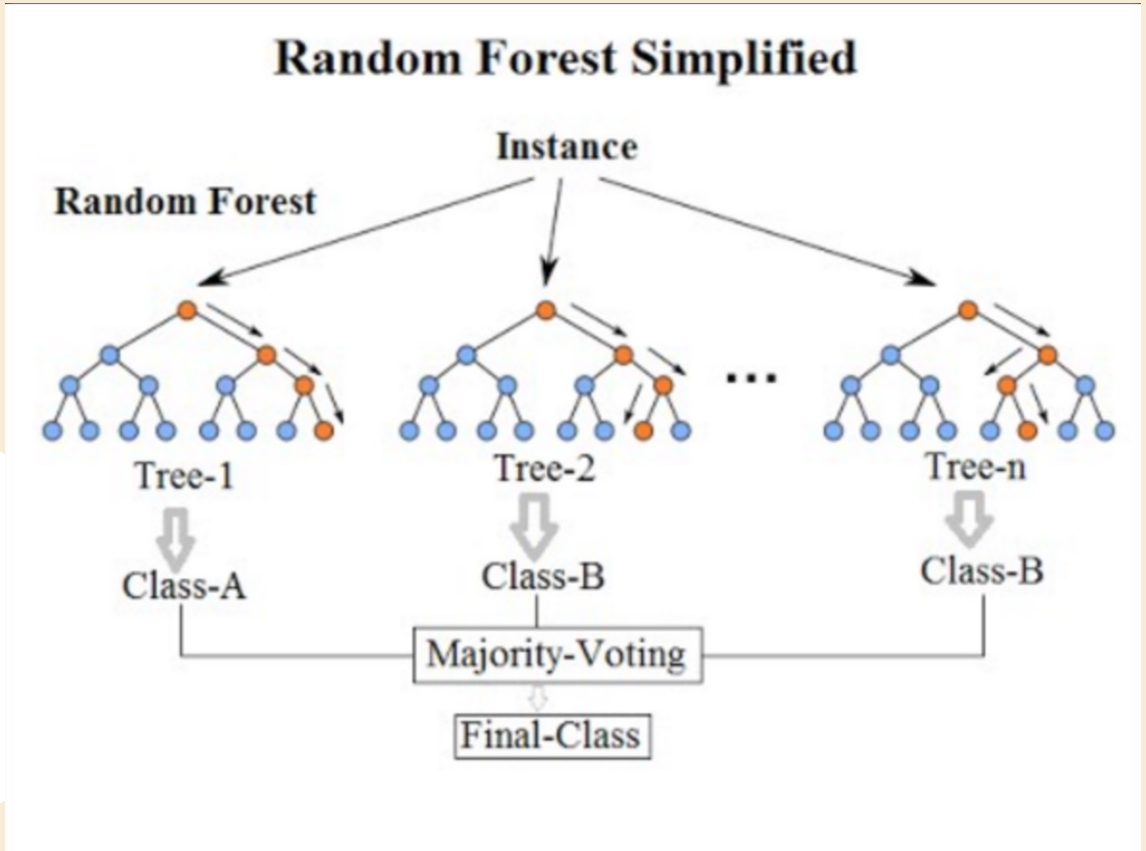


https://en.wikipedia.org/wiki/Decision_tree_learning#/media/File:Decision_Tree.jpg

Machine Learning Algorithms

- Random Forest

The Random Forest algorithm is a popular ensemble learning method for both classification and regression tasks. It is based on the idea of creating multiple decision trees and combining their predictions to make a final prediction.

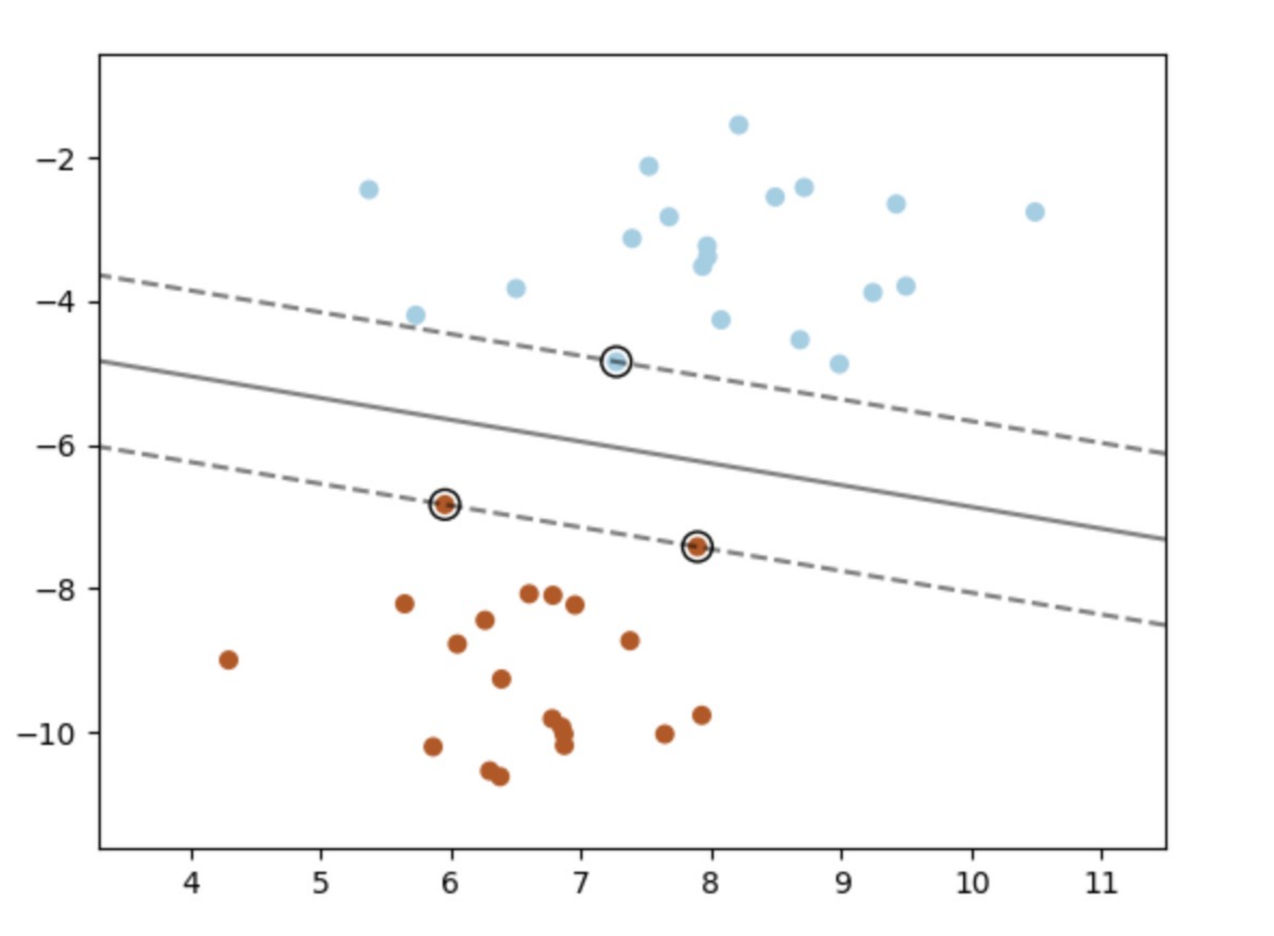


Random Forest Classifier

Machine Learning Algorithms

- Support Vector Machine

Support Vector Machines (SVMs) classify data by identifying the hyperplane that maximizes the margin between the classes in the training data. In a simple scenario with two classes represented in a two-dimensional space, we can visualize the hyperplane as the widest straight "band" or line that separates the two classes with clear margins on each side.



SVM: Maximum margin separating hyperplane

Machine Learning Algorithms

- Naïve Bayes

Naive Bayes methods are a collection of supervised learning algorithms that utilize Bayes' theorem under the assumption of conditional independence between each pair of features, given the class variable. This "naive" assumption simplifies the computation and allows for efficient training and prediction.

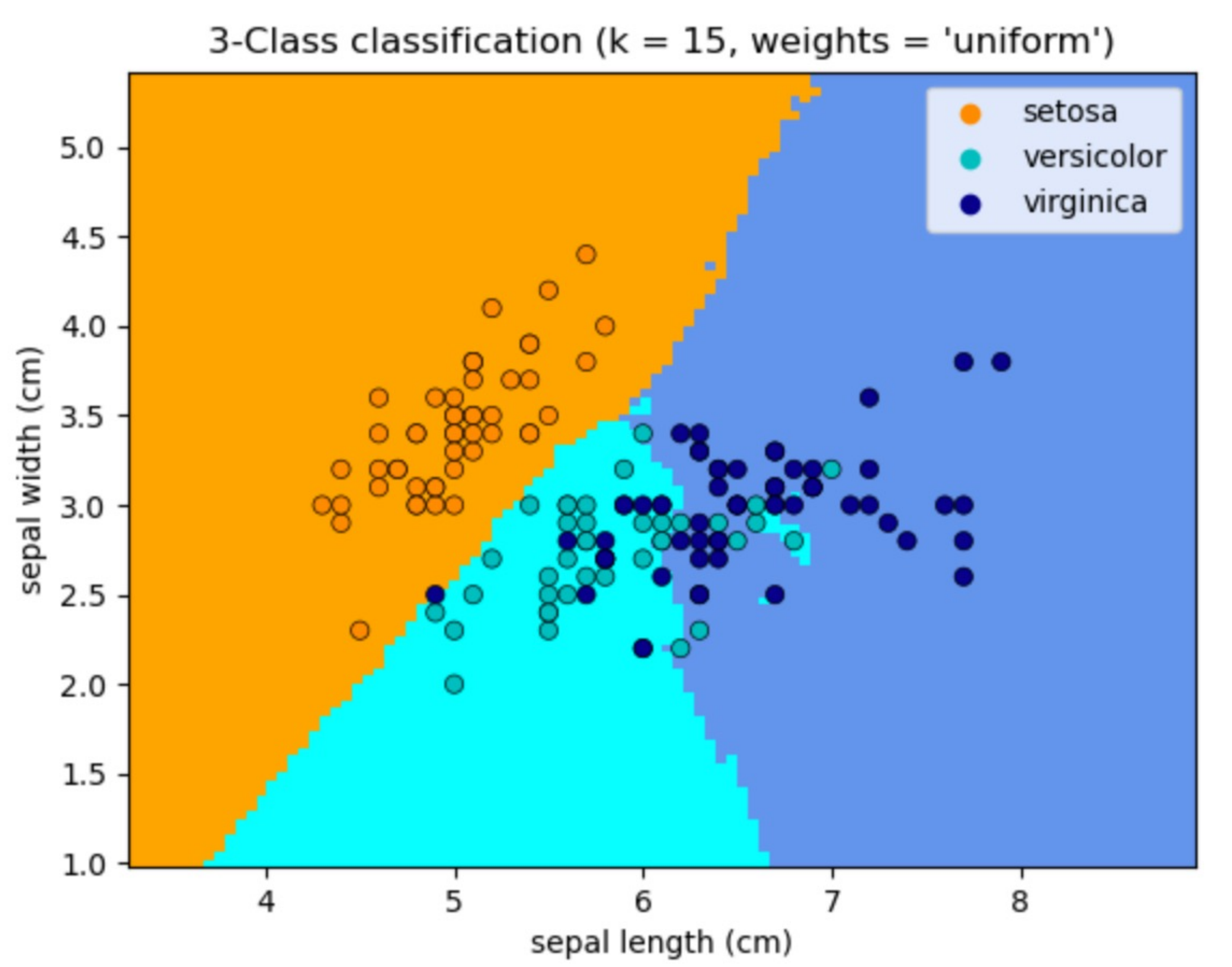
Bayes' theorem:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Machine Learning Algorithms

- K-Nearest Neighbor

The K-Nearest Neighbors classifier (KNN) is a popular and straightforward supervised machine learning algorithm. It is often categorized as a lazy learner because it doesn't explicitly train a model. Instead, it classifies an observation based on the majority class among its k nearest neighbors in the training data.



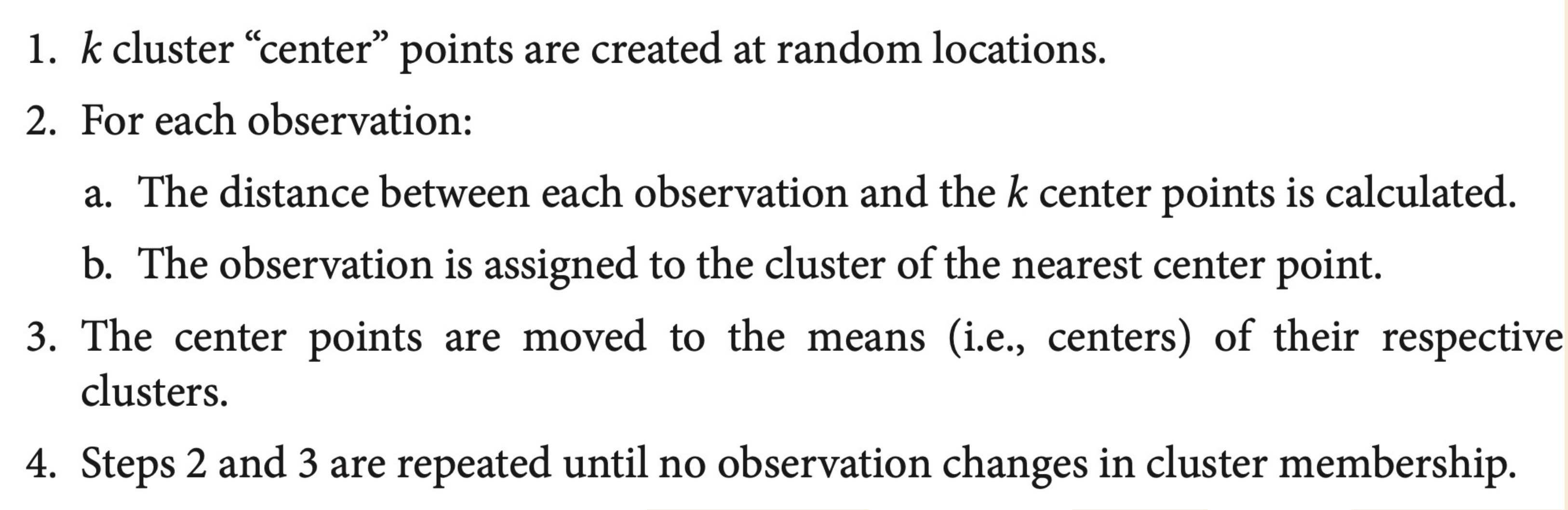
K-Nearest Neighbor Classification

Machine Learning Algorithms

- K-Mean Clustering

K-means clustering is a widely used clustering technique in machine learning. In k-means clustering, the algorithm aims to partition observations into k clusters, where each cluster has similar variance. The value of k, representing the number of clusters, is chosen by the user as a hyperparameter.

Specifically, in k-means:



1. *k* cluster "center" points are created at random locations.
2. For each observation:
   a. The distance between each observation and the *k* center points is calculated.
   b. The observation is assigned to the cluster of the nearest center point.
3. The center points are moved to the means (i.e., centers) of their respective clusters.
4. Steps 2 and 3 are repeated until no observation changes in cluster membership.

Model Evaluation

- Train set

  The train set is a subset of the available labeled data that is used to train or fit the machine learning model.

- Test set

  The test set, on the other hand, is a separate subset of the labeled data that is not used during the training phase.

- train-test split

  The train-test split is typically done randomly to ensure that the train and test sets represent the underlying data distribution. The common practice is to allocate a majority portion of the data to the train set and reserve the remaining portion for the test set.

- It's important to note that the train-test split is just one approach to assess model performance.

Model Evaluation

- Cross-validation technique

One approach for evaluating a supervised learning model is to split the data into a training set and a test set, reserving a portion for evaluation. However, this validation method has limitations: the model's performance can be influenced by the specific observations in the test set, and it doesn't utilize all available data for training and evaluation.

- K-fold cross validation

  The most common type of cross-validation is k-fold cross-validation, where the data is divided into k subsets/folds. Example.

Confusion Matrix

	Actual Values		
Predicted Values		Positive	Negative
	Positive	#True Positive	#False Positive
	Negative	#False Negative	#True Negative

Confusion Matrix

- If classifying 100 news articles with:
  - 70 real
  - 43 fake_news

		Actual Values		
Predicted Values		real	fake_news	Total
	real	42	9	51
	fake_news	28	21	49
	Total	70	30	100

Model Performance Metrics

Model performance metrics are quantitative measures used to evaluate the performance of a machine learning model.

Some commonly used model performance metrics include

- Precision

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- Recall

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Model Performance Metrics

- F1 score

$$F_1 = 2 * \frac{1}{\frac{1}{recall} + \frac{1}{precision}}$$

- Accuracy

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Model Performance Metrics

- ROC-AUC

An **ROC curve (receiver operating characteristic curve)** is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

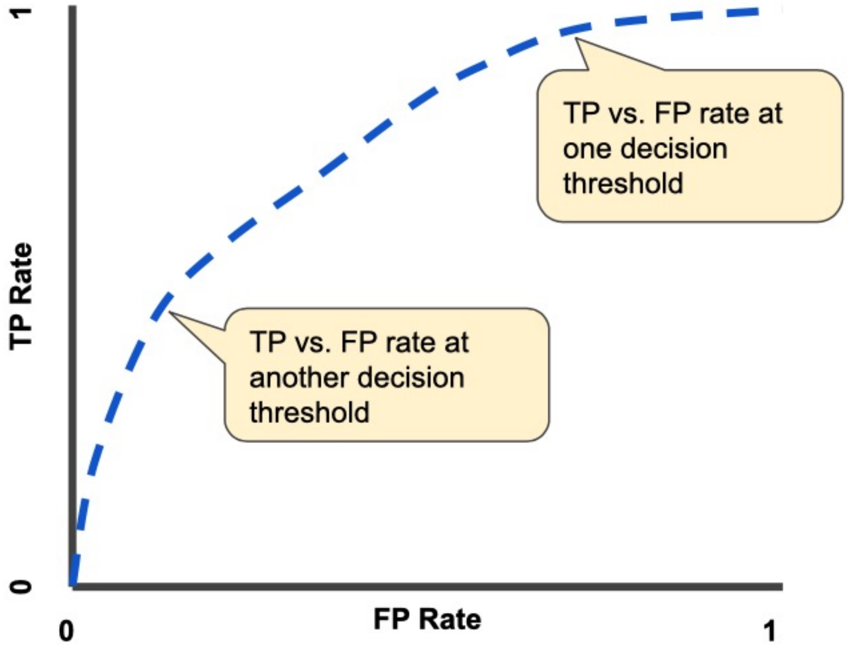
- True Positive Rate (TPR) is a synonym for recall

- False Positive Rate

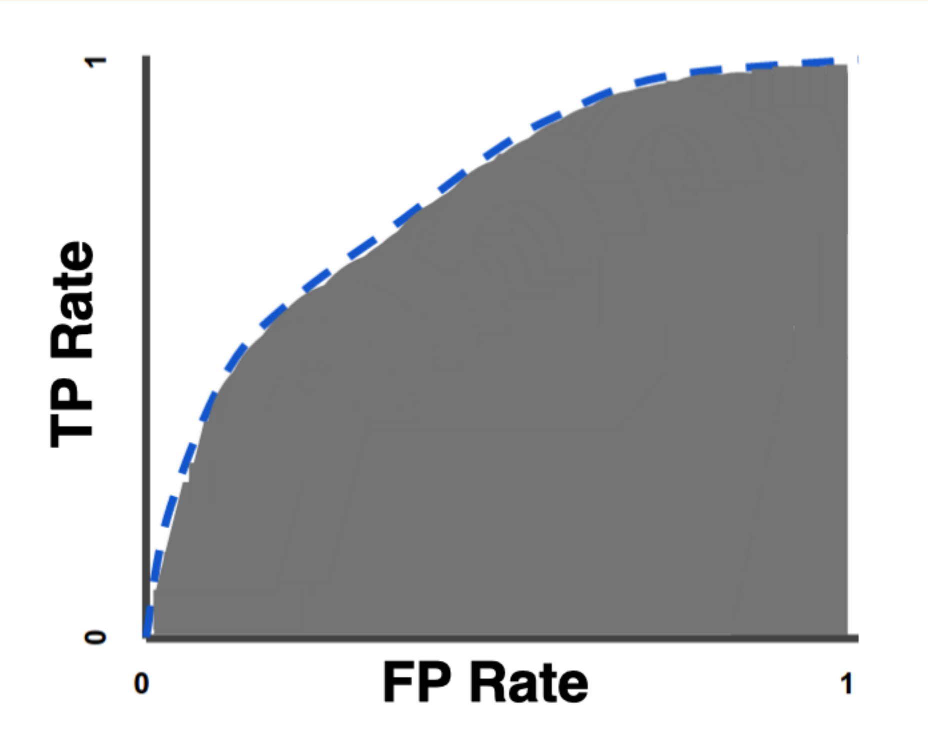
$$FPR = \frac{FP}{FP + TN}$$

Model Performance Metrics

- ROC-AUC



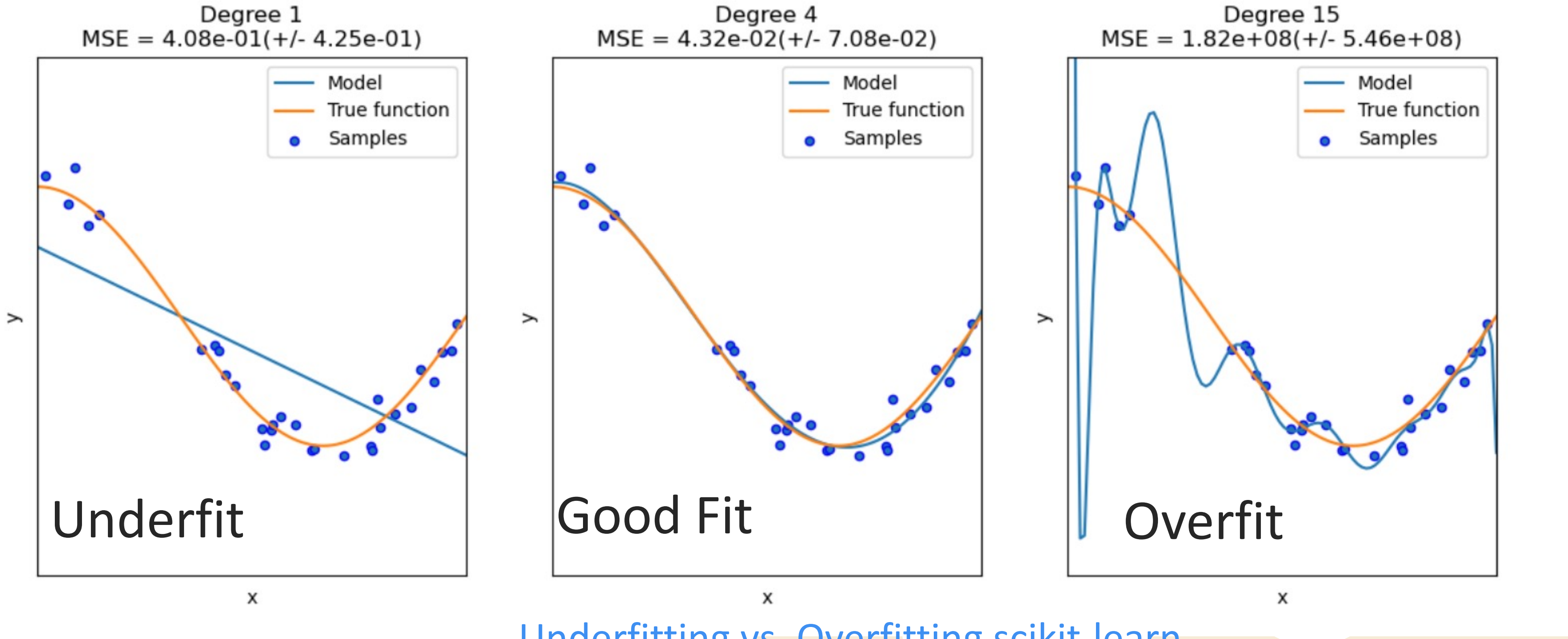
ROC



ROC-AUC

Overfitting

Overfitting occurs when a machine learning model performs very well on the training data but fails to generalize to new, unseen data. In other words, the model has “memorized” the training data, rather than learning the pattern of the data.



[Underfitting vs. Overfitting scikit-learn]()