**Glossary**

Here you will find the description of some words and phrases that pertain to the concepts in this course, and also words and phrases used in the videos and readings. If you want to add another word or expression that we missed, please suggest in it in the forums.

**A**

**Agglomerative** is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

**Apriori Algorithm** refers to an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

**B**

**Bayes Rule**, in probability and statistics, refers to a theorem that  describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

**Bayesian Model** refers to a statistical model where probability is used to represent all uncertainty within the model, both the uncertainty regarding the output and the uncertainty regarding the input to the model.

**C**

**Clustering** is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group.

**Confusion Matrix** refers to a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

**Complexity Parameter (CP):** is used to control the size of the decision tree and to select the optimal tree size.

**Cross-validation:** sometimes called rotation estimation, or out-of-sample testing is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set.

**D**

**Dendrogram** refers to a tree diagram, especially one showing taxonomic relationships.

**Decision Boundary**: is the region of a problem space in which the output label of a classifier is ambiguous. In a statistical-classification problem with two classes, a decision boundary or decision surface is a hypersurface that partitions the underlying vector space into two sets, one for each class.

**Decision Tree**: is a [decision support](https://en.wikipedia.org/wiki/Decision_support_system) tool that uses a [tree-like](https://en.wikipedia.org/wiki/Tree_(graph_theory)) [model](https://en.wikipedia.org/wiki/Causal_model) of decisions and their possible consequences, including [chance](https://en.wikipedia.org/wiki/Probability) event outcomes, resource costs, and [utility](https://en.wikipedia.org/wiki/Utility). It is one way to display an [algorithm](https://en.wikipedia.org/wiki/Algorithm) that only contains conditional control statements. It is usually used in operations research.

**Discrimination** refers to a function of the coefficients of a polynomial equation whose value gives information about the roots of the polynomial.

**Divisive** is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

**E**

**Entropy**: A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). ID3 algorithm uses entropy to calculate the homogeneity of a sample.

**Error rate** refers to the frequency of errors can have.

**Euclidean distance** between two points in either the plane or 3-dimensional space measures the length of a segment connecting the two points. It is the most obvious way of representing distance between two points.

**G**

**Gini Index:** is also known as the Gini coefficient, is a measure of statistical dispersion intended to represent the income or wealth distribution of a nation's residents, and is the most commonly used measurement of inequality.

**H**

**Hierarchical clustering** is a set of nested clusters that are organized as a tree.

**I**

**Independence assumption**, refers to the assumption that attribute values in a dataset are independent of each other. This is a very important baseline assumption in the application of the Naïve Bayes algorithm.

**K**

**K-means Clustering** aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

**K-Nearest Neighbor** refers to a non-parametric method used for classification and regression.

**L**

**Lift Chart** refers to a chart that can. Visualize the effectiveness of a predictive model calculated as the ration between the results obtained and without the predictive model.

**Leaf Node**: is any node that does not have child nodes in a decision tree.

**M**

**Misclassification Error**: is one of the splitting criteria, also known as impurity node.

**Model complexity**often refers to the number of features or terms included in a given predictive model, as well as whether the chosen model is linear, nonlinear, and so on. It can also refer to the algorithmic learning complexity or computational complexity.

**N**

**Naïve Bayes Algorithm**, refers to a machine learning model, which is based on the Bayes’ probability theorem, and is used for the classification of data.

**Nearest neighbor search** (**NNS**): as a form of proximity search, is the [optimization problem](https://en.wikipedia.org/wiki/Optimization_problem) of finding the point in a given set that is closest (or most similar) to a given point.

**Neural Networks** refers to a computer network modeled on the human brain and nervous system.

**O**

**Object function** refers to the function that it is desired to maximize or minimize.

**Over-fitting** refers to a modelling error which occurs when a function is too closely fit to a limited set. Of data points. Overfitting a model generally takes the form of making an overly complex model to explain idiosyncrasies in the data under stufy.

**P**

**Parameters** refers to a numerical characteristic of a population, as distinct from a statistic of a sample.

**Partition Clustering** is simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.

**Parzen window density estimation**: is another name for *kernel density estimation*. It is a nonparametric method for estimating continuous density function from the data.

**Principal component analysis (PCA)** is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.

**Pruning**: When we reduce the size of decision trees by removing nodes (opposite of Splitting), the process is called pruning**.**

**Pseudo R-Squared**: is a logical analog to OLS R2 measures, usually used for Goodness-Of-Fit test of logistic regression (to investigate binary or multinomial outcomes).

**R**

**Root Node Error:**  is the percent of correctly sorted records at the first (root) splitting node.

**S**

**Sensitivity in statistics** refers to the proportion of actual positives that are correctily identified as such.

**Specificity in statistics** refers to the proportion of actual negatives that are correctly identified as such.

**Supervised Learning** refers to the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.

**Support Vector Machines (SVMs)**: are a set of supervised learning methods used for [classification](https://scikit-learn.org/stable/modules/svm.html#svm-classification), [regression](https://scikit-learn.org/stable/modules/svm.html#svm-regression) and [outliers detection](https://scikit-learn.org/stable/modules/svm.html#svm-outlier-detection).

**SVM or Support Vector Machine** refers to a descriptive classifier formally defined by a serparating hyperplane. In other words, given labeled training data, the algorithm outputs an optimal hyperplane which categorizes new examples.

**T**

**Tuning parameter**: sometimes called a penalty parameter, controls the strength of the penalty term in ridge regression and lasso regression. It is basically the amount of shrinkage, where data values are shrunk towards a central point, like the mean.

**U**

**Unsupervised Learning** refers to a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data.