## What is Linear Regression?

Linear regression is a **linear model**. It is a model that assumes a linear relationship between the input variables (**x**) and the single output variable (y). When there is a single input variable (x), the method is referred to as **simple linear regression**. The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B).

***Y = a + bX***, where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0). The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line

**linear Regression** is a machine **learning** algorithm based on **supervised learning**.

Linear regression provides a powerful statistical method to find the relationship between variables. It hardly needs further tuning. However, it’s only limited to linear relationships.

Linear regression produces the best predictive accuracy for linear relationship whereas its little sensitive to outliers and only looks at the mean of the dependent variable

**Logistic Regression**

Logistic regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems, where the dependent variable(target) is categorical. Instead of fitting a straight line or hyperplane, the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1. In machine learning, **Logistic regression** is also **supervised**

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Example of logistic regression questions

1. *Do body weight, calorie intake, fat intake, and age have an influence on the probability of having a heart attack (yes vs. no)?*
2. Binary (eg. Tumor Malignant or Benign)

*Whilst linear regression* used the following formula: *hΘ(x) = β₀ + β₁X*

logistic regression is modified a little bit i.e. *σ(Z) = σ(β₀ + β₁X)*

**Decision Tree**

**Decision Trees** are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. The deeper the tree, the more complex the decision rules and the fitter the model

A **decision tree** is a map of the possible outcomes of a series of related choices. It allows an individual or organization to weigh possible actions against one another based on their costs, probabilities, and benefits. They can be used either to drive informal discussion or to map out an algorithm that predicts the best choice mathematically.

A **decision tree** can also be used to help build automated predictive models, which have applications in machine learning, data mining, and statistics. Known as decision tree learning, this method considers observations about an item to predict that item’s value.

There are a variety of trees including classification tree, regression tree, random trees. In these decision trees, nodes represent data rather than decisions. Some trees are used for classification and some for regression, both share similarities but also some differences, such as the procedure used to determine where to split.

**SVM (Support Vector Machine)**

**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). “Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges.

The advantages of support vector machines are:

* Effective in high dimensional spaces.
* Still effective in cases where number of dimensions is greater than the number of samples.
* Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
* Versatile: different [Kernel functions](https://scikit-learn.org/stable/modules/svm.html#svm-kernels) can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

* It doesn’t perform well when we have large data set because the required training time is higher
* SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see [Scores and probabilities](https://scikit-learn.org/stable/modules/svm.html#scores-probabilities), below).

## What is Naive Bayes algorithm?

In machine learning, the Naive Bayes is a classification algorithm based on the concept of Bayes Theorem. Bayes theorem is one of the fundamental theorems in probability. It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. The Bayes Theorem forms the backbone of the Naive Bayes algorithm. We use conditional probability to classify the data - thus, the Naive Bayes algorithm basically gives us the probability of a record being in a particular class, given the values of the features.

Naive Bayes is an eager learning classifier and it is sure fast. Thus, it could be used for making predictions in real time. Though classification algorithms like Logistic Regression, random forest, etc are quite popular, Naive Bayes holds its own among them. It is faster than Random forest, since it can adapt to changing data pretty quickly. If the assumptions of Naive Bayes hold true, then it is much faster than logistic regression as well.

What are the different types of Naive Bayes algorithms?

There are 3 main types of Naive Bayes algorithms:

• Gaussian Naive Bayes

• Multinomial Naive Bayes

• Bernoulli Naive Bayes

**Advantage Naïve’s Bayes**

• It is easy and fast to predict class of test data set. It also performs well in multi class prediction

• When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.

• It performs well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

**Disadvantage of Naïve’s Bayes**

• If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.

• On the other side naive Bayes is also known as a bad estimator

• Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

KNN (K-Nearest Neighbors)

What is **KNN**? **KNN** is a model that classifies data points based on the points that are most similar to it. It uses test data to make an “educated guess” on what an unclassified point should be classified as.

The k-nearest neighbours (**KNN**) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

KNN algorithm is considered both non-parametric and an example of lazy learning.

**• Non-parametric** = makes no assumptions. The model is made up entirely from the data given to it rather than assuming its structure is normal.

• **Lazy learning** means that the algorithm makes no generalizations. This means that there is little training involved when using this method. Because of this, all of the training data is also used in testing when using KNN.

**Advantage of KNN**

* Easy to use.
* Quick calculation time.
* Does not make assumptions about the data.

**Disadvantages of KNN**

* Accuracy depends on the quality of the data.
* Must find an optimal k value (number of nearest neighbours).
* Poor at classifying data points in a boundary where they can be classified one way or another.

K-Means

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. The objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.

A cluster refers to a collection of data points aggregated together because of certain similarities. To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

**It halts creating and optimizing clusters when either:**

* The centroids have stabilized — there is no change in their values because the clustering has been successful.
* The defined number of iterations has been achieved.

Random Forest

Random forest is a **supervised learning algorithm.** The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. **Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.**

One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

The Random forest is a supervised learning algorithm. This can be used for regression and classification tasks both. Random forest is one of the most used algorithms because of its simplicity and stability. Random forests are more stable and reliable than just a decision tree. This is just saying like- it’s better to take a vote from all cabinet ministers rather than just accepting the decision given by the PM.