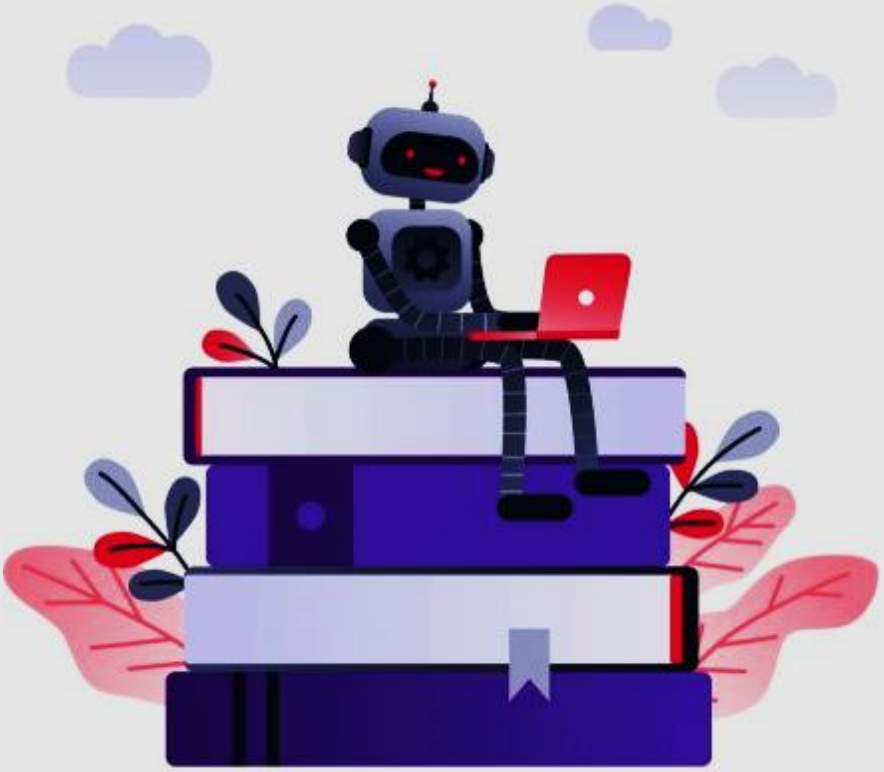
Machine Learning Algorithms Explained in Less Than 1 Minute Each



Linear Regression

One of the simplest Machine learning algorithms out there, Linear Regression is used to make predictions on continuous dependent variables with knowledge from independent variables. A dependent variable is the effect, in which its value depends on changes in the independent variable.

You may remember the line of best fit from school - this is what Linear Regression produces. A simple example is predicting one's weight depending on their height.

Logistic Regression

Logistic Regression, similar to Linear Regression, is used to make predictions on categorical dependent variables with knowledge of independent variables. A categorical variable has two or more categories. Logistic Regression classifies outputs that can only be between 0 and 1.

For example, you can use Logistic Regression to determine whether a student will be admitted or not to a particular college depending on their grades - either Yes or No, or 0 or 1.

Decision Trees

Decision Trees (DTs) is a probability tree-like structure model that continuously splits data to categorize or make predictions based on the previous set of questions that were answered. The model learns the features of the data and answers questions to help you make better decisions.

For example, you can use a decision tree using the answers Yes or No to determine a specific species of bird using data features such as feathers, ability to fly or swim, beak type, etc.

Random Forest

Similar to Decision Trees, Random Forest is also a tree-based algorithm. Where Decision Tree consists of one tree, Random forest uses multiple decision trees for making decisions - a forest of trees.

It combines multiple models to make predictions and can be used in Classification and Regression tasks.

K-Nearest Neighbors

K-Nearest Neighbors uses the statistical knowledge of how close a data point is to another data point and determines if these data points can be grouped together. The closeness in the data points reflects the similarities in one another.

For example, if we had a graph which had a group of data points that were close to one another called Group A and another group of data points that were in close proximity to one another called Group B. When we input a new data point, depending which group the new data point is nearer to - that will be their new classified group.

Support Vector Machines

Similar to Nearest Neighbor, Support Vector Machines performs classification, regression and outlier detection tasks. It does this by drawing a hyperplane (a straight line) to separate the classes. The data points that are located on one side of the line will be labeled as Group A, whilst the points on the other side will be labeled as Group B.

For example, when a new data point is inputted, depending on which side of the hyperplane and its location within the margin it is - this will determine which group the data point belongs to.

Naive Bayes

Naive Bayes is based on Bayes' Theorem which is a mathematical formula used for calculating conditional probabilities. Conditional probability is the chance of an outcome occurring given that another event has also occurred.

It predicts that the probabilities for each class belongs to a particular class and that the class with the highest probability is considered the most likely class.

k-means Clustering

K-means clustering, similar to nearest neighbors but uses the method of clustering to group similar items/data points in clusters. The number of groups is referred to as K. You do this by selecting the k value, initializing the centroids and then selecting the group and finding the average.

For example, if there are 3 clusters present and a new data point is inputted, depending on which cluster it falls in - that is the cluster they belong to.

Bagging

Bagging is also known as Bootstrap aggregating and is an ensemble learning technique. Bagging is used in both regression and classification models and aims to avoid overfitting of data and reduce the variance in the predictions.

Overfitting is when a model fits exactly against its training data - basically not teaching us anything and can be due to various reasons. Random Forest is an example of Bagging.

Boosting

The overall aim of Boosting is to convert weak learners to strong learners. Weak learners are found by applying base learning algorithms which then generates a new weak prediction rule. A random sample of data is inputted in a model and then trained sequentially, aiming to train the weak learners and trying to correct its predecessor

XGBoost, which stands for Extreme Gradient Boosting, is used in Boosting.

Dimensionality Reduction

Dimensionality reduction is used to reduce the number of input variables in the training data, by reducing the dimension of your feature set. When a model has a high number of features, it is naturally more complex leading to a higher chance of overfitting and decrease in accuracy.

Dimensionality Reduction

For example, if you had a dataset with a hundred columns, dimensionality reduction will reduce the number of columns down to twenty. However, you will need Feature Selection to select relevant features and Feature Engineering to generate new features from existing features.

The Principal Component Analysis (PCA) technique is a type of Dimensionality Reduction.

Stay in touch !

@DataCleanic

www.DataCleanic.ml