AIM Lab Introduction: Data Analysis

Mathematics and Statistics packages

In research we will inevitably have to do some statistics. It is good to know how we can do mathematics and statistics operations with Python. As we are using Google Colab which already installed most required packages, so you do not need to install additional packages here.

NumPy

NumPy is a python library to help dealing with mathematics, especially arrays. As it can do so many things, the examples listed below may not be able to cover all.

Pandas

Pandas is an easy-to-use python library for data structures and data analysis. pandas implements a number of statistical functions that can be used in research.

Matplotlib

Matplotlib is a popular Python 2D plotting library. You can do most of the data visualization using Matplotlib. pandas plotting methods also use Matplotlib as a based tool.

Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Sklearn

Sklearn is a machine learning module which is simple but very useful. It provides numerous tools for model fitting, data processing, evaluation and many other utilities.

Classical Machine Learning

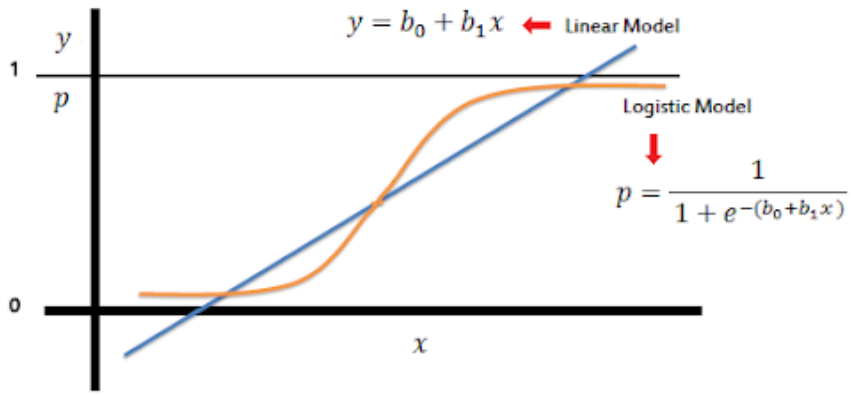
Simple Linear Regression

y = X * b

A linear function has one independent variable and one dependent variable. The independent variable is x and the dependent variable is y.

- y is the output value that we are interested in.
- X is the input values that we have.
- b is the constant term or the model. It describes how the system works.

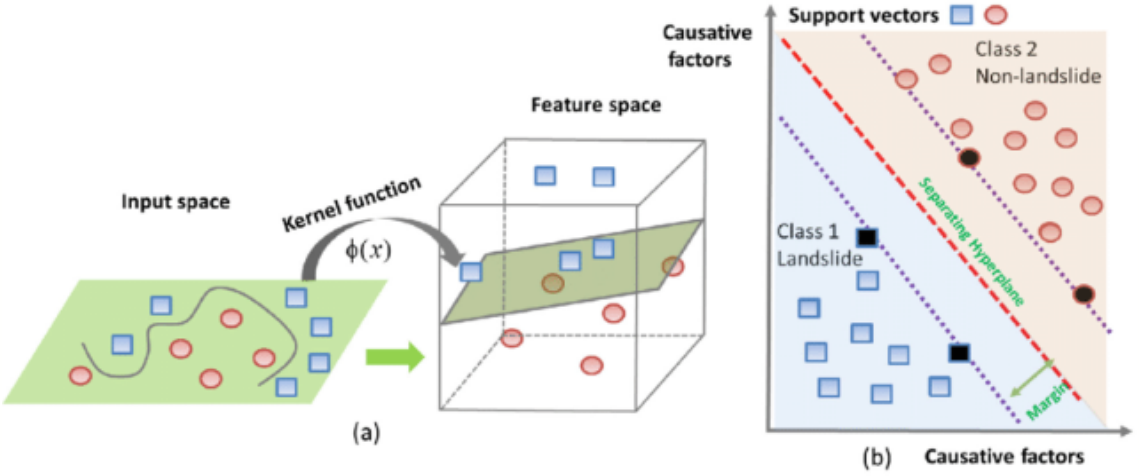
Logistic Regression



Logistic regression is named for the function used at the core of the method, the logistic function.

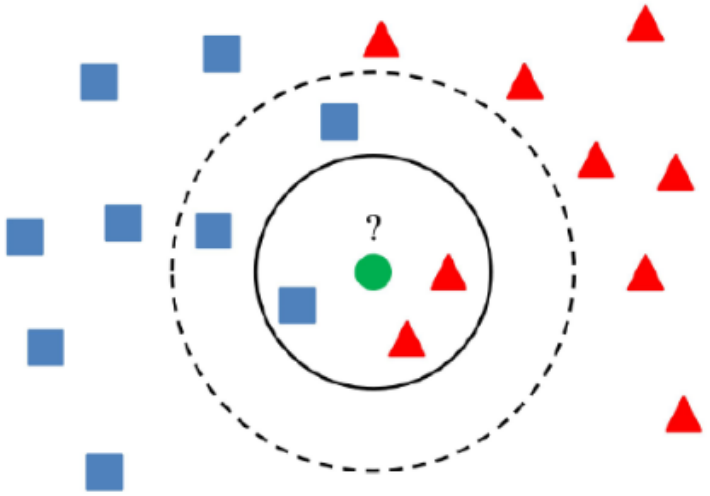
The logistic function, also called the sigmoid function, was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

Support Vector Classifier



The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

k-Nearest Neighbors



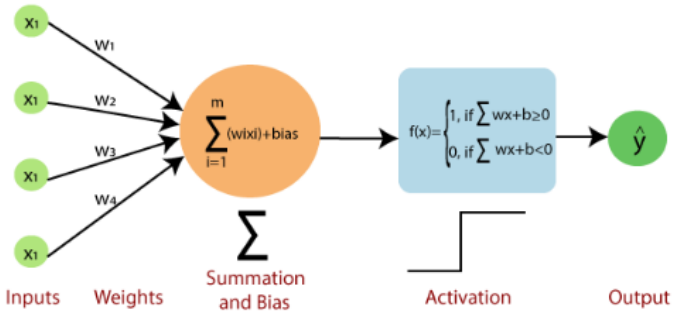
kNN is one of the simplest yet powerful supervised ML algorithms. It is widely used for classification problems as well as can be used for regression problems. The data-point is classified on the basis of its k Nearest Neighbors, followed by the majority vote of those nearest neighbors; a query point is assigned the data class which has the most representatives within the nearest neighbors of the point.

Gaussian Naive Bayes

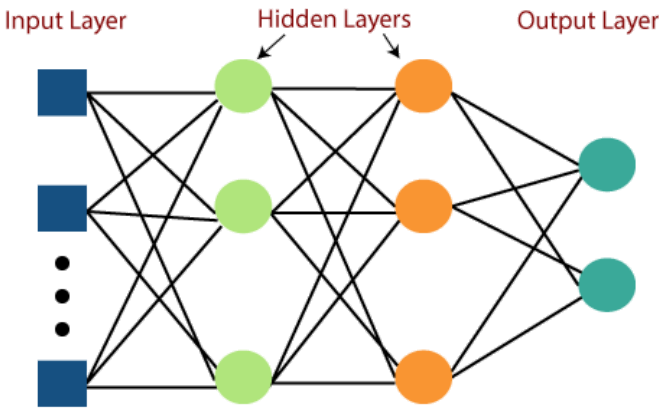
Naive Bayes is a probabilistic machine learning algorithm based on the **Bayes Theorem**. It was used in a wide variety of classification tasks, which is carried out by finding the probability of given feature being associated with a label and assigning the label with the highest probability. It was named **naive** because it assumes the features that go into the model are independent of each other, which is rarely the case in real life.

Gaussian Naive Bayes is the Naive Bayes algorithm that assumes the likelihood of the features is Gaussian.

Multilayer Perceptron

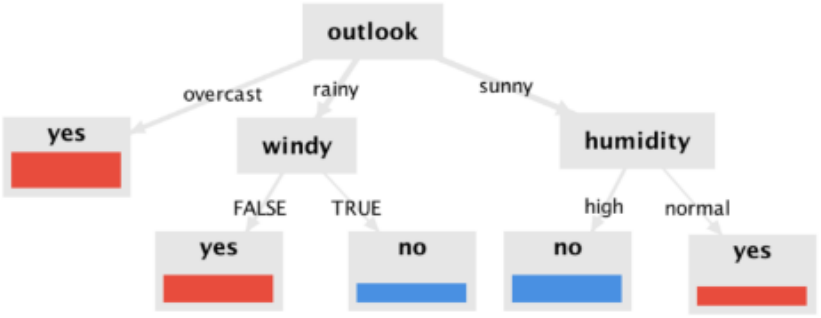


Perceptron is an artificial neuron. It is working with 5 connecting parts. They are input nodes, weighting, summation, activation function, and the output node.



Multilayer Perceptron (MLP) is a deep, artificial neural network. It is formed from the connections of perceptrons. It has 3 types of layer which are, an input layer that receives signals, hidden layers that extracts features, and an output layer that gives response.

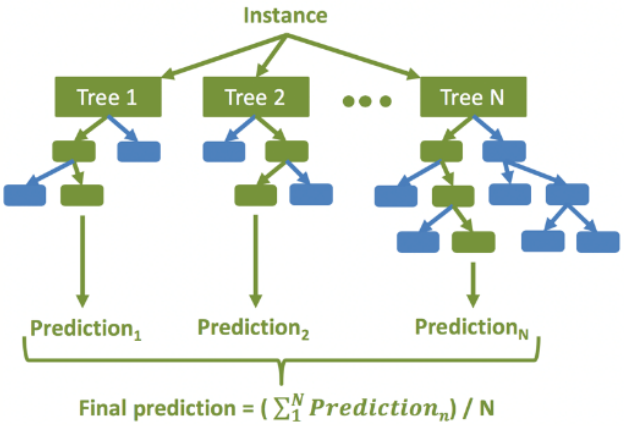
Decision Tree



Decision Tree is a supervised learning method used for classification and regression. 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.

Decision Tree is famous because it can generate a tree-diagram that helps us understand a course of action with statistical probability. it consists of 3 elements, which are node (the decision), branch (the decision probability), and leaf (the outcome).

Random Forest Classifier



The random forest is a classification algorithm consisting of many decision trees. It uses bagging and features randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.