



电子科技大学
格拉斯哥学院
Glasgow College, UESTC

Regression Analysis & Naïve Bayes

Supervised Learning

- Various types of algorithms and computation methods are used in the supervised learning process.
 - Regression
 - Naïve Bayes
 - K-Nearest Neighbors
 - Support Vector Machines
 - Decision Tree
 - Neural Networks

Regression Analysis

Definition

- Regression analysis is the process of **estimating the relationship** between a **dependent variable** and **independent variables**.
- It means **fitting a function** from a **selected family of functions** to the sampled data under some error function.
- Regression analysis is one of the **most basic tools** in the area of machine learning used for prediction, in which an algorithm is used to predict continuous outcomes.

Objective

- Solving regression problems is one of the **most common applications** for ML models, especially in supervised ML.
- Using regression, we **fit a function on the available data** and try to **predict the outcome** for the **future or hold-out data points**.
- This fitting of function serves two purposes:
 - 1) You can **estimate missing data** within your data range (Interpolation).
 - 2) You can **estimate future data** outside your data range (Extrapolation).

Applications of Regression Models

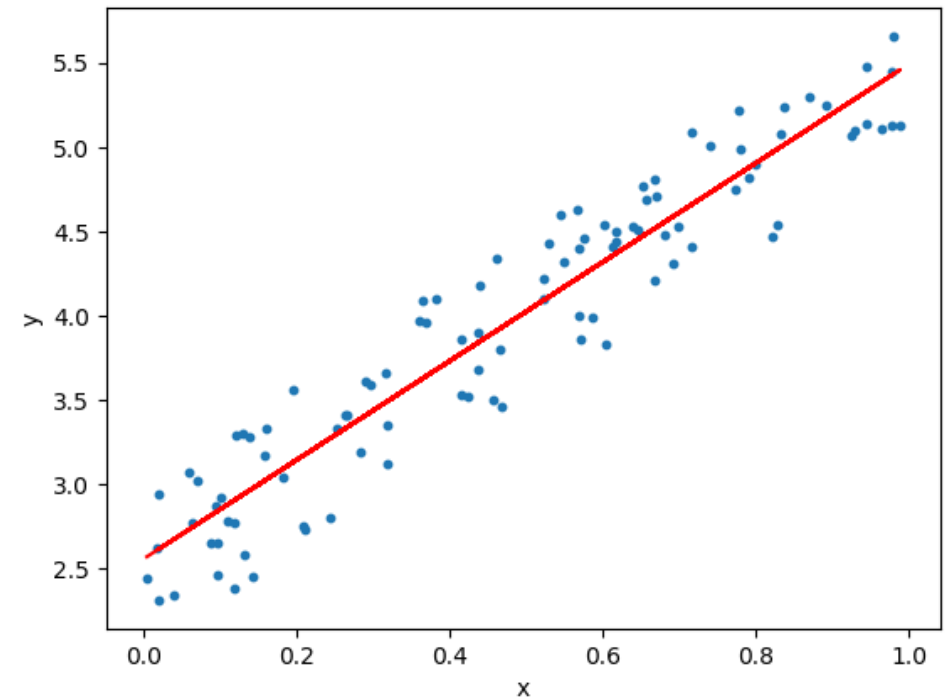
- Common uses for regression models include:
 - Forecasting continuous outcomes like house prices, stock prices, or sales.
 - Predicting the success of future retail sales or marketing campaigns to ensure resources are used effectively.
 - Predicting customer or user trends, such as streaming services or e-commerce websites.
 - Analysing datasets to establish the relationships between variables and output.
 - Predicting interest rates or stock prices from a variety of factors.
 - Creating time series visualisations.

Types of Regression Analysis

- Based on the family-of-functions (f_{β}), and the loss function (l) used, we can categorize regression into the following categories..
 - Linear Regression
 - Polynomial Regression
 - Ridge Regression
 - LASSO regression
 - Bayesian Regression
 - Logistic Regression

Linear Regression

- In linear regression (LR), the objective is to **fit a hyperplane** (a line for 2D data points) by **minimizing the sum of the mean-squared error** for each data point.
- LR finds the **linear relationship** between the dependent variable and one or more independent variables using a **best-fit straight-line**.



Linear Regression

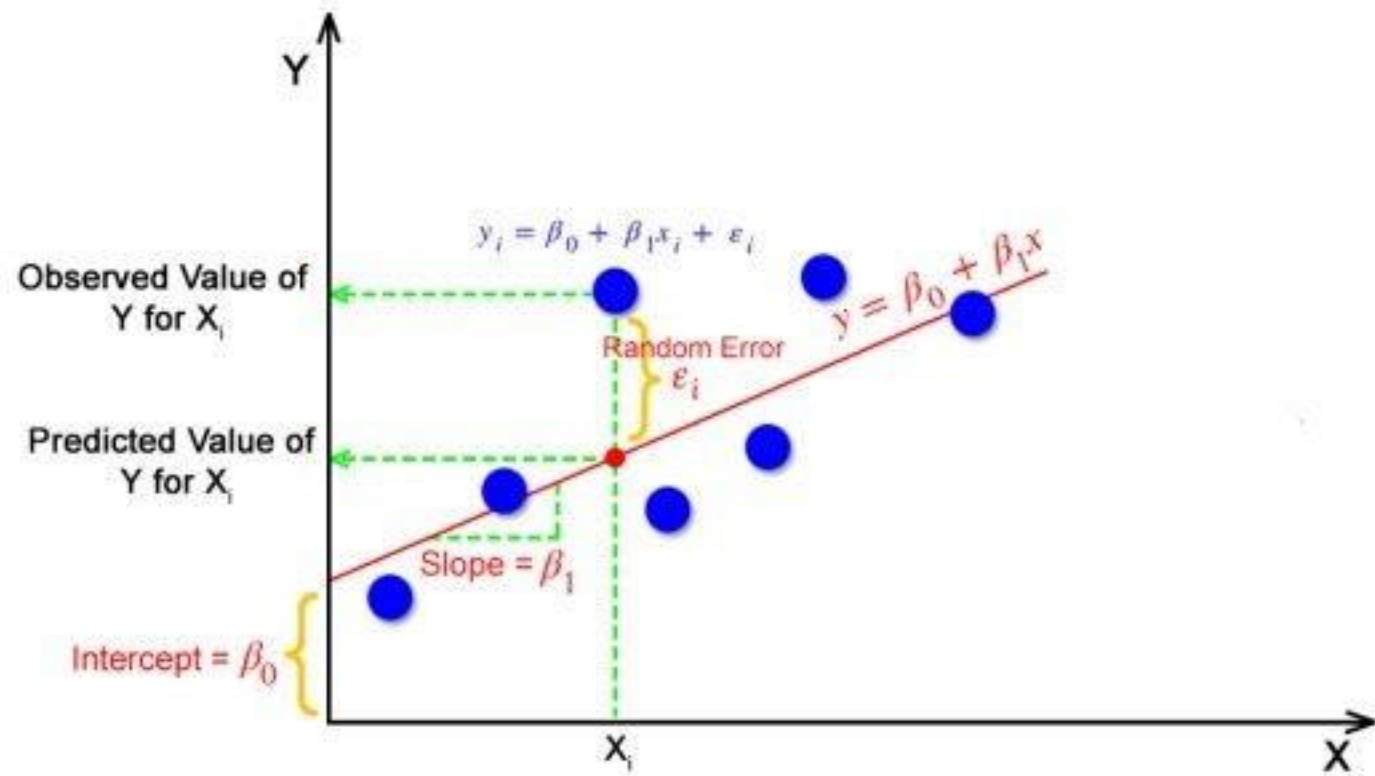
- A linear model makes a prediction by simply computing a **weighted sum of the input features**, plus a **constant called the bias term** (also called the intercept term).
- In linear regression technique,
 - the dependent variable is **continuous**,
 - the independent variable(s) can be **continuous or discrete**, and
 - the nature of the regression line is linear.

Finding/Drawing the Best-fit Line

- Let's see how linear regression **adjusts** the line between the data for **accurate predictions**.
- Imagine, you're given a set of data, and your **goal is to draw the best-fit line** which passes through the data. This is the step-by-step process you proceed with:
 - 1) Consider your linear equation to be $y = \theta_1 x + \theta_0$, where y is the dependent data and x is the independent data given in your dataset
 - 2) Adjust the line by varying the values of θ_1 and θ_0 , i.e., the **coefficient** and the **bias**
 - 3) Come up with some random values for the coefficient and bias initially and plot the line
 - 4) Since the line won't fit well, change the values of ' θ_1 ' and ' θ_0 .' This can be done using the '**gradient descent algorithm**' or '**least squares method**'

Finding/Drawing the Best-fit Line

- The goal of the linear regression algorithm is to **get the best values for B_0 and B_1** to find the **best fit line**.
- The best fit line is a line that has the **least error** which means the error between predicted values and actual values should be minimum.



Mathematical Representation

- Mathematically speaking, linear regression solves the following problem

Given P number of data points (x_i, y_i) where $x_i, y_i \in \mathbb{R} \forall i \in \{0, 1, \dots, P-1\}$,
fit a linear function

$$\hat{y} = f_{\beta}(x) = \beta_0 + \beta_1 x$$

by minimizing

$$\min_{\beta} \sum_p \|y^p - f(x^p)\|^2$$

- Hence, we need to find 2 variables denoted by beta that parameterize the linear function $f(\cdot)$

Polynomial Regression

- Linear regression assumes that the **relationship** between the dependent (y) and independent (x) variables are **linear**.
- It fails to fit the data points where the **relationship is not linear**.
- Polynomial regression expands the fitting capabilities of linear regression by fitting a **polynomial of degree m** to the data points instead.
- The richer the function under consideration, the better (in general) it's fitting capabilities.

Mathematical Representation

- Polynomial regression solves the following problem

Given P number of data points (x_i, y_i) where $x_i, y_i \in \mathbb{R} \forall i \in \{0, 1, \dots, P-1\}$,
fit a polynomial function of degree m

$$\hat{y} = f_{\beta}(x) = \beta_0 + \sum_{j=1}^m \beta_j x^j$$

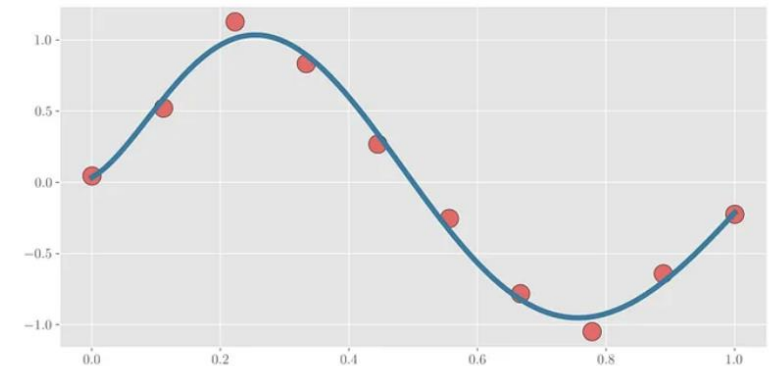
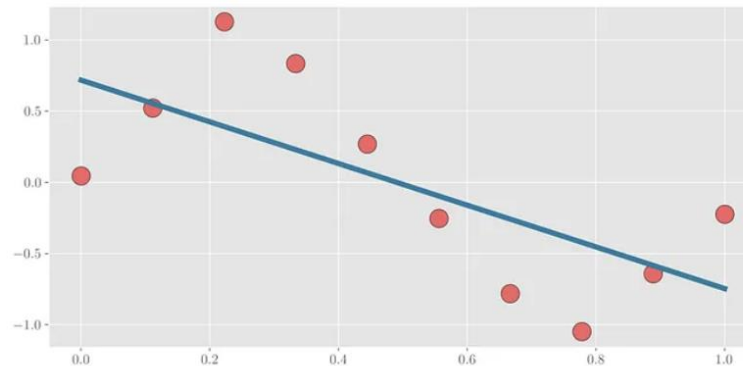
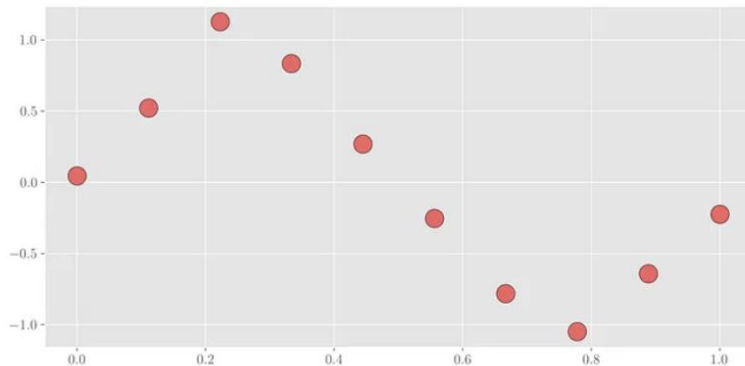
by minimizing

$$\min_{\beta} \sum_i \|y_i - f(x_i)\|^2$$

- We need to find $(m+1)$ variables denoted by $\beta_0, \beta_1, \dots, \beta_m$.
- It can be seen that [linear regression](#) is a special case of [polynomial regression with degree 1 and two variables](#).

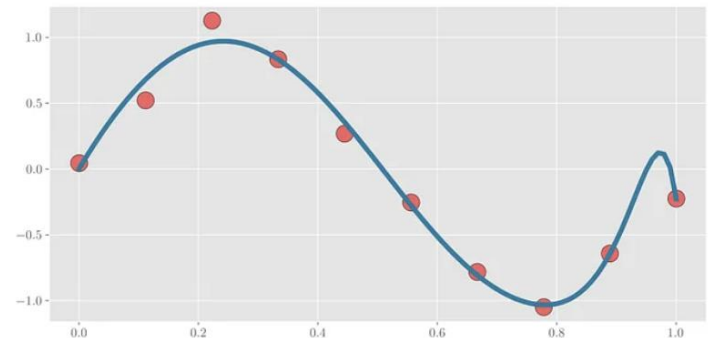
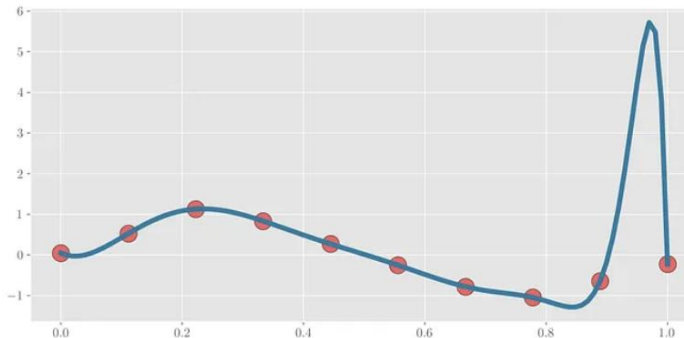
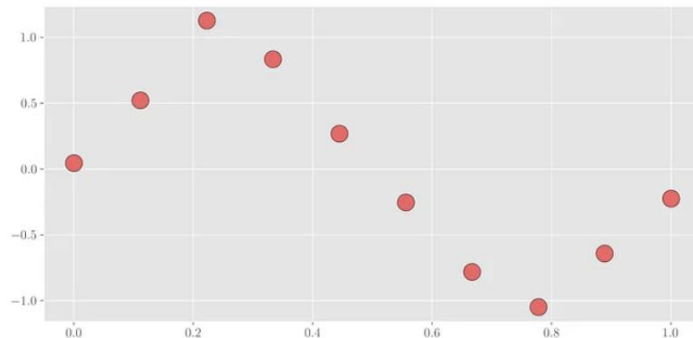
Finding the Better Fit

- Consider the following set of data points plotted as a scatter plot.
- If we use linear regression, we get a fit that clearly **fails to estimate the data points**.
- But if we use polynomial regression with **degree 6**, we get a much better fit.



Ridge Regression

- Ridge regression addresses the issue of **overfitting in regression analysis**.
- When a polynomial of degree 25 is fit on the data with 10 training points, it can be seen that it fits the red data points perfectly (center figure below). But in doing so, it **compromises other points** in between (spike between last two data points).
- Ridge regression tries to address this issue. It tries to **minimize the generalization error** by compromising the fit on the training points



Mathematical Representation

- Ridge regression solves the following problem by modifying the loss function.

Given P number of data points (x_i, y_i) where $x_i, y_i \in \mathbb{R} \forall i \in \{0, 1, \dots, P-1\}$, fit a function f_β (parameterized by β) by minimizing

$$\min_{\beta} \sum_i \|y_i - f(x_i)\|^2 + \alpha \|\beta\|^2$$

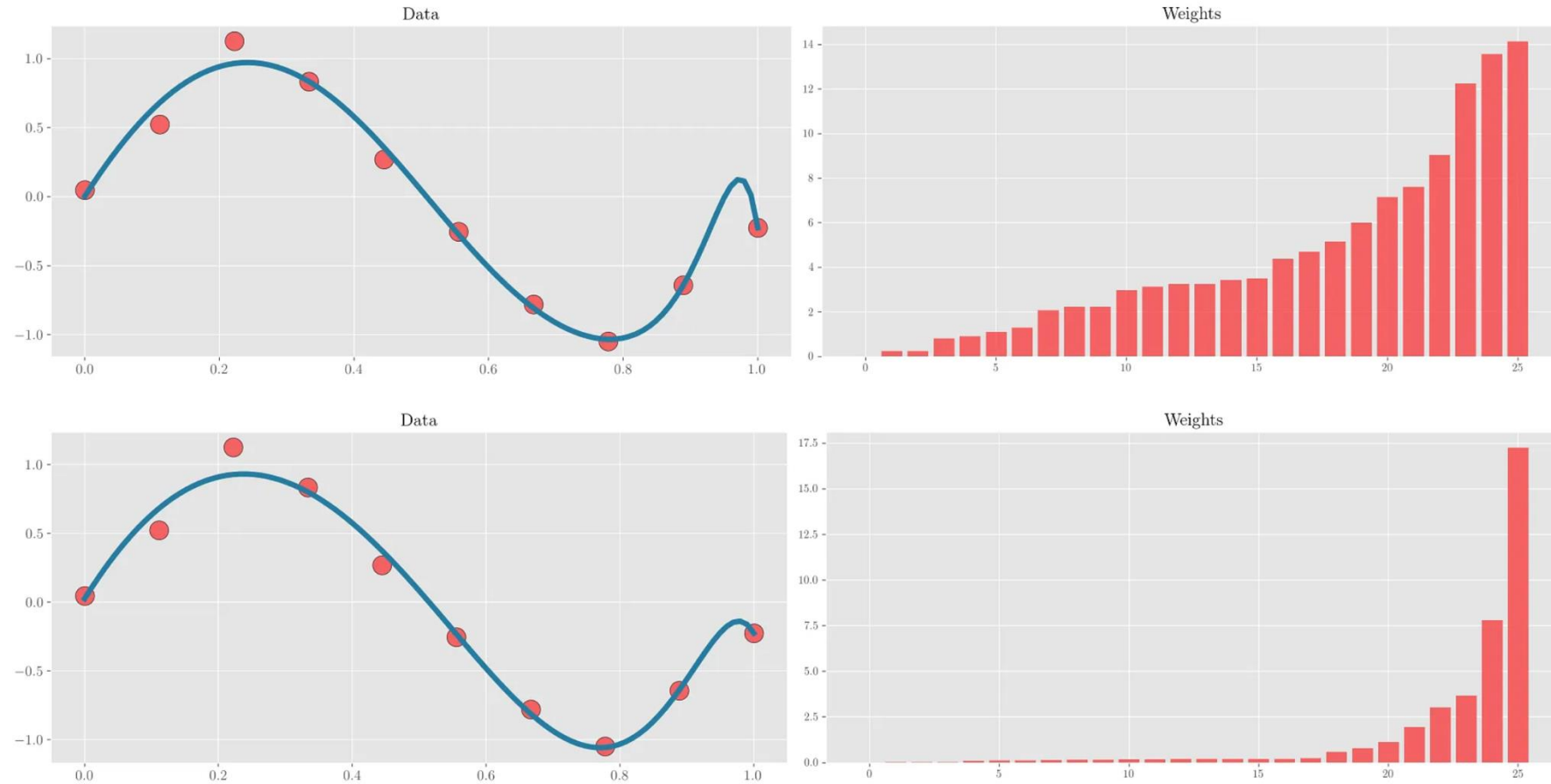
where $\alpha \in \mathbb{R}$ is a scaling factor.

The function $f(x)$ can either be linear or polynomial. In the absence of ridge regression, when the function overfits the data points, the **weights learned to tend to be pretty high**. Ridge regression avoids over-fitting **by limiting the norm of the weights** being learned by introducing the **scaled L2 norm of the weights (beta) in the loss function**.

LASSO Regression

- Both LASSO and Ridge regression use regularizers against overfitting on the training data points, but only LASSO enforces sparsity on the learned weights.
- Ridge regression enforces the norm of the learned weights to be small, yielding a set of weights where the total norm is reduced. Most of the weights (if not all) will be non-zero.
- LASSO tries to find a set of weights by making most of them really close to zero.
- This yields a sparse weight matrix whose implementation can be much more energy-efficient than a non-sparse weight matrix while maintaining similar accuracy in terms of fitting to the data points.

Ridge vs Lasso Regression



Mathematical Representation

- LASSO regression solves the following problem by modifying the loss function.

Given P number of data points (x_i, y_i) where $x_i, y_i \in \mathbb{R} \forall i \in \{0, 1, \dots, P-1\}$, fit a function f_β (parameterized by β) by minimizing

$$\min_{\beta} \sum_i \|y_i - f(x_i)\|^2 + \alpha \|\beta\|_1$$

where $\alpha \in \mathbb{R}$ is a scaling factor and $\|\beta\|_1 = \sum_k |\beta_k|$.

The difference between LASSO and Ridge regression is that LASSO uses the L1 norm of the weights instead of the L2 norm. This L1 norm in the loss function tends to increase sparsity in the learned weights.

Naïve Bayes Classifiers

Prior Probability

- Prior probability is the probability of an event before new data is collected.
- This is the best rational assessment of the probability of an outcome based on the current knowledge before an experiment is performed.

Posterior Probability

- The posterior probability is the probability of event A occurring given that event B has occurred.
- The prior probability of an event will be revised as new data or information becomes available, to produce a more accurate measure of a potential outcome.
- That revised probability becomes the posterior probability and is calculated using Bayes' theorem.

Prior vs Posterior Probability

- **Prior probability** represents **what is originally believed** before new evidence is introduced, and
- **Posterior probability** takes this **new information** into account.

Definition - Naïve Bayes Algorithm

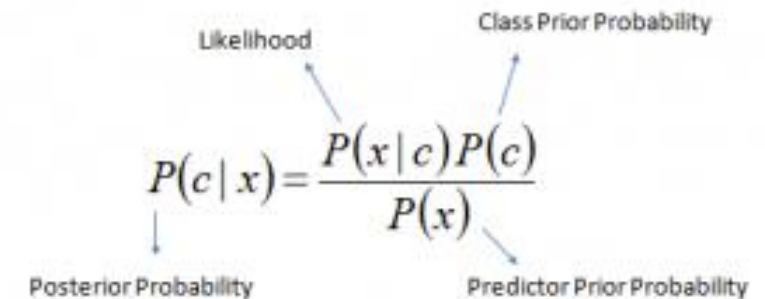
- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- This is one of the simplest and most effective classification algorithms which helps build fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts based on the probability of an object.
- Some popular examples of Naïve Bayes Algorithm are spam filtration, sentimental analysis, and classifying articles.

Background

- The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:
- **Naïve:** It is called Naïve because it assumes that the **occurrence of a certain feature is independent** of the occurrence of other features. For example, if the fruit is identified on the bases of **color**, **shape**, and **taste**, then **red**, **spherical**, and **sweet** fruit is recognized as an apple. Hence, each feature individually contributes to identifying that it is an apple without depending on each other.
- **Bayes:** It is called Bayes because it depends on the principle of **Bayes' Theorem**.

Bayes' Rule

- Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
- The formula for Bayes' theorem is given as
 - $P(c/x)$ is the **posterior probability** of **class (c, target)** given **predictor (x, attributes)**.
 - $P(c)$ is the **prior probability** of **class**.
 - $P(x/c)$ is the **likelihood** which is the probability of the **predictor** given **class**.
 - $P(x)$ is the **prior probability** of the **predictor**.



The diagram shows the formula $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$ with arrows pointing from labels to the terms in the formula. 'Likelihood' points to $P(x|c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c|x)$, and 'Predictor Prior Probability' points to $P(x)$.

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

Example - Bayes' Rule

- Suppose that an individual is extracted at random from a population of men.
- We know the following information:
 - the probability of extracting a married individual is 50%;
 - the probability of extracting a childless individual is 40%;
 - the conditional probability that an individual is childless given that he is married is equal to 20%.
- If the individual extracted at random from the population turns out to be childless, what is the conditional probability that he is married?

Example - Solution

- This conditional probability is called **posterior probability** and can be computed using **Bayes' rule**.
- The quantities involved in the computation are:

$$P(\text{married}) = 1/2$$

$$P(\text{childless}) = 2/5$$

$$P(\text{childless}|\text{married}) = 1/5$$

- The **posterior probability** is:

$$\begin{aligned} P(\text{married}|\text{childless}) &= \frac{P(\text{childless}|\text{married})P(\text{married})}{P(\text{childless})} \\ &= \frac{\frac{1}{5} \cdot \frac{1}{2}}{\frac{2}{5}} = \frac{1}{10} \cdot \frac{5}{2} = \frac{1}{4} \end{aligned}$$

Advantages of Naïve Bayes

- **Less complex:** Compared to other classifiers, Naïve Bayes is considered a simpler classifier since the parameters are easier to estimate. As a result, it's one of the first algorithms learned within data science and machine learning courses.
- **Scales well:** Compared to logistic regression, Naïve Bayes is considered a fast and efficient classifier that is fairly accurate when the **conditional independence assumption holds**. It also has low storage requirements.
- **Can handle high-dimensional data:** Use cases, such as document classification, can have a high number of dimensions, which can be difficult for other classifiers to manage.

Disadvantages of Naïve Bayes

- **Subject to Zero frequency:** Zero frequency occurs when a categorical variable does not exist within the training set.

For example, imagine that we're trying to find the maximum likelihood estimator for the word, "sir" given class "spam", but the word, "sir" doesn't exist in the training data. The probability in this case would be zero, and since this classifier multiplies all the conditional probabilities together, this also means that the posterior probability will be zero. To avoid this issue, **Laplace smoothing** can be leveraged.

- **Unrealistic core assumption:** While the conditional independence assumption overall performs well, the assumption does not always hold, leading to incorrect classifications.

Applications

- **Real-time Prediction:** Naïve Bayesian classifier is an eager learning classifier, and it is superfast. Thus, it could be used for making predictions in real-time.
- **Multi-class Prediction:** This algorithm is also well known for its multi-class prediction feature. Here, we can predict the probability of multiple classes of target variables.
- **Text classification / Spam Filtering / Sentiment Analysis:** Naïve Bayesian classifiers mostly used in text classification (due to better results in multi-class problems and independence rule) have a higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)
- **Recommendation System:** Naïve Bayes Classifier and **Collaborative Filtering** together build a **Recommendation System** that uses machine learning and data mining techniques to filter unseen information and predict whether a user would like a given resource or not.