**1. Difference between Simple Linear Regression and Multiple Linear Regression**

**- Simple Linear Regression:** Simple linear regression involves predicting a target variable using a single predictor variable. The relationship between the predictor and the target is assumed to be linear, and the model attempts to fit a straight line to the data.

**- Multiple Linear Regression :** Multiple linear regression involves predicting a target variable using multiple predictor variables. The model assumes a linear relationship between the predictors and the target, and it fits a hyperplane to the data in a higher-dimensional space.

**2.Explain the concept of the cost function in linear regression.**

- The cost function, also known as the loss function or objective function, measures the difference between the actual target values and the predicted values by the linear regression model.

- In linear regression, the most commonly used cost function is the Mean Squared Error (MSE) or the Mean Absolute Error (MAE). MSE is calculated by averaging the squared differences between the actual and predicted values, while MAE is calculated by averaging the absolute differences.

- The goal of linear regression is to minimize the cost function, which means finding the set of coefficients (parameters) that result in the smallest error between the predicted and actual values.

**3.How do you interpret the coefficients in a linear regression model?**

- In a linear regression model, each coefficient represents the change in the target variable for a one-unit change in the corresponding predictor variable, holding all other predictors constant.

- For example, in the equation (y = beta\_0 + beta\_1 x1 + beta\_2 x2 + ……+ beta\_p n), (beta\_1) represents the change in the target variable for a one-unit change in (x1), while holding all other predictors constant.

- The sign of the coefficient indicates the direction of the relationship between the predictor and the target. A positive coefficient indicates a positive relationship, while a negative coefficient indicates a negative relationship.

**4.What are the assumptions of linear regression?**

**- Linearity:** The relationship between the predictors and the target variable is assumed to be linear.

**- Independence :** The observations in the dataset are assumed to be independent of each other.

**- Homoscedasticity** : The variance of the residuals (the differences between the actual and predicted values) is constant across all levels of the predictor variables.

**- Normality of Residuals** : The residuals are assumed to be normally distributed.

**- No Multicollinearity :** There should be no significant multicollinearity among the predictor variables, meaning that the predictors should not be highly correlated with each other.

**1.How does logistic regression differ from linear regression?**

- **Logistic Regression**: Logistic regression is used for binary classification tasks where the target variable is categorical and has two classes (e.g., 0 or 1, yes or no). It models the probability that an instance belongs to a particular class. Logistic regression uses a logistic or sigmoid function to map the output of a linear combination of predictors to a probability between 0 and 1.

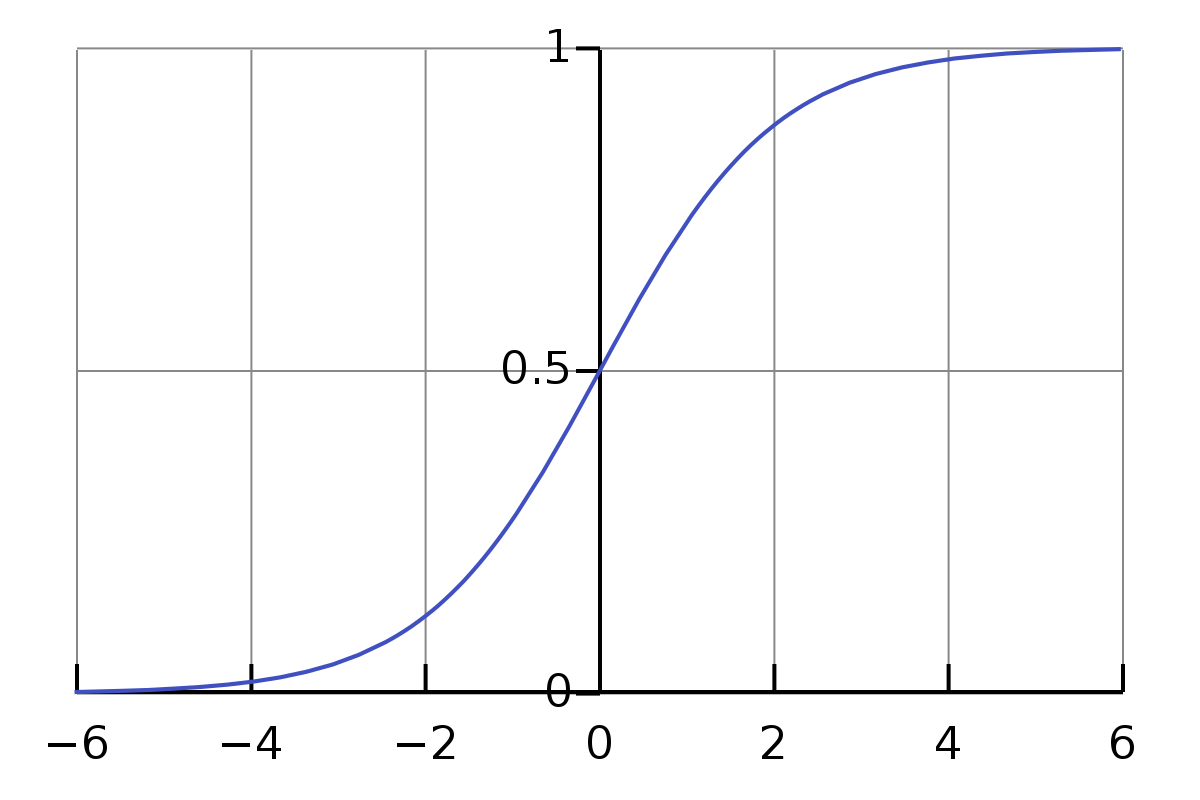
**- Linear Regression**: Linear regression is used for regression tasks where the target variable is continuous. It models the relationship between the predictors and the target as a linear equation. Linear regression p2.

**2.Explain the sigmoid function and its role in logistic regression.**

- The sigmoid function, also known as the logistic function, is defined as ( sigma(z) = 1/(1 + e^{-z} ), where (z) is the linear combination of predictor variables and their coefficients.

The sigmoid function maps any real-valued number (z) to the range [0, 1]. This allows logistic regression to output probabilities, making it suitable for binary classification tasks.

In logistic regression, the output of the linear combination of predictors is passed through the sigmoid function to obtain the probability that an instance belongs to the positive class.



**3.What are the key performance metrics used to evaluate a logistic regression model?**

- Accuracy : The proportion of correctly classified instances out of the total number of instances.

- Precision : The proportion of true positive predictions out of all positive predictions made by the model.

- Recall (Sensitivity) : The proportion of true positive predictions out of all actual positive instances in the dataset.

**4.How do you handle multicollinearity in logistic regression?**

- Feature Selection : Remove highly correlated predictors from the model.

- Regularization : Apply techniques like Lasso or Ridge regression, which penalize large coefficients and can reduce the impact of multicollinearity.

**1.What is the Naive Bayes algorithm based on?**

- The Naive Bayes algorithm is based on Bayes' theorem, which is a fundamental theorem in probability theory. It calculates the probability of a hypothesis (class label) given the observed evidence (predictor variables) by incorporating prior knowledge and new evidence.

Naive Bayes assumes that the features are conditionally independent given the class label, which means that the presence of a particular feature in a class is independent of the presence of other features. This is a simplifying assumption that allows the algorithm to make predictions efficiently.

**2.Explain the concept of conditional probability in the context of Naive Bayes.**

- In the context of Naive Bayes, conditional probability refers to the probability of observing a particular feature value given a specific class label.

The algorithm calculates the conditional probability of each feature given each class using the training data. These probabilities are then used to compute the posterior probability of each class given a new instance's features.

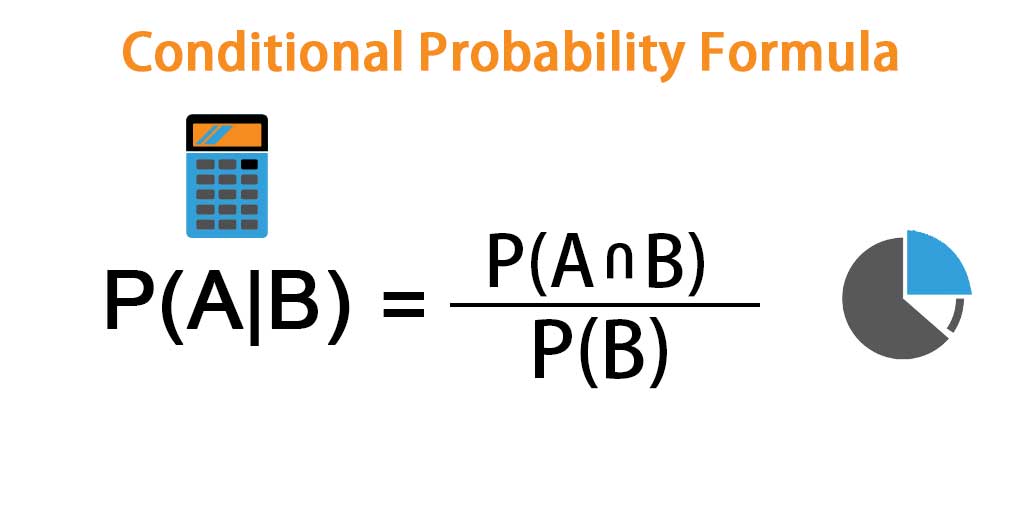
Consider a random experiment of tossing 2 coins. The sample space here will be:

S = {HH, HT, TH, TT}

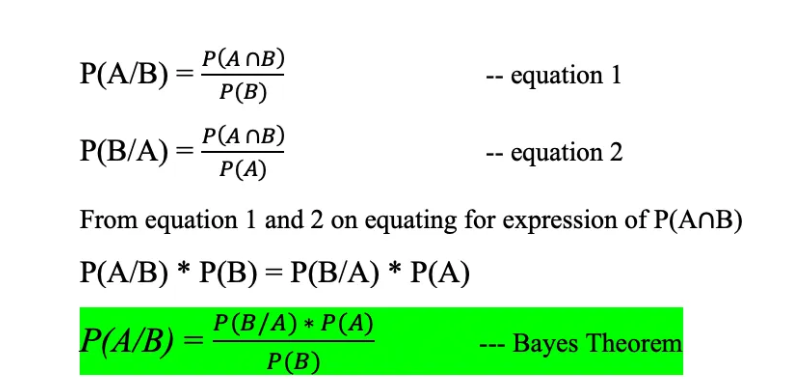
If a person is asked to find the probability of getting a tail his answer would be 3/4 = 0.75

Now suppose this same experiment is performed by another person but now we give him the condition that both the coins should have heads. This means if event A: ‘Both the coins should have heads’, has happened then the elementary outcomes {HT, TH, TT} could not have happened. Hence in this situation, the probability of getting heads on both the coins will be 1/4 = 0.25

From the above examples, we observe that the probability may change if some additional information is given to us. This is exactly the case while building any machine learning model, we need to find the output given some features.



Bayes theorem is derived through conditional probability equation by equating P(A and B) of below mentioned equation 1 and equation 2.



**3.What are the advantages and disadvantages of Naive Bayes?**

**- Advantages:**

- Naive Bayes is computationally efficient and scales well to large datasets.

- It performs well in practice, especially for text classification and spam filtering tasks.

- Naive Bayes can handle high-dimensional data well, making it suitable for datasets with many features.

- It requires relatively few parameters and is less prone to overfitting, especially when there is limited training data.

**- Disadvantages:**

- The "naive" assumption of feature independence may not hold true in real-world datasets, leading to suboptimal performance.

- It tends to perform poorly when there are strong dependencies among features.

- Naive Bayes is sensitive to the presence of irrelevant or redundant features, which can degrade performance.

- It estimates probabilities using the observed frequencies in the training data, which may lead to inaccurate probability estimates for rare or unseen feature combinations.

**4.How does Naive Bayes handle missing values and categorical features?**

**- Missing Values :** Naive Bayes can handle missing values by ignoring instances with missing values during training and classification.Naive Bayes Imputation (NBI) is used to fill in missing values by replacing the attribute information according to the probability estimate. The NBI process divides the whole data into two sub-sets is the complete data and data containing missing data. Complete data is used for the imputation process at the lost value. The process is repeated for each missing attribute to generate complete data for classification. This research applies NBI for imputation and preprocessing as preparation of classification process.

**- Categorical Features :** Naive Bayes naturally handles categorical features by calculating the conditional probabilities of each feature value given each class label. It does not require one-hot encoding or other transformations for categorical variables. However, if a categorical feature has a large number of levels, it may lead to sparse data, which can affect performance.

**1.How does a decision tree make decisions?**

- Decision trees make decisions by recursively partitioning the feature space into smaller regions or subsets based on the values of predictor variables.

- At each node of the tree, the algorithm selects the best split based on a chosen criterion (e.g., Gini impurity, entropy) to maximize the homogeneity (purity) of the resulting subsets with respect to the target variable.

- The process continues until a stopping criterion is met, such as reaching a maximum tree depth, having a minimum number of samples in a node, or when further splitting does not improve the model's performance.

**2.What are the main criteria for splitting nodes in a decision tree?**

**- Gini Impurity:** Measures the probability of incorrectly classifying a randomly chosen element if it were randomly labeled according to the distribution of classes in the node. The split that minimizes the Gini impurity is selected.

**- Entropy:** Measures the level of impurity or disorder in a set of data. It is calculated as the sum of the probability of each class multiplied by the logarithm of the probability. The split that maximizes the information gain (reduction in entropy) is chosen.

**- Classification Error :** Measures the proportion of instances in a node that do not belong to the most common class. The split that minimizes the classification error is selected.

**3.How do decision trees handle categorical variables?**

- Decision trees can handle categorical variables naturally by considering each level or category as a distinct branch in the tree.

- When splitting a node with a categorical variable, the algorithm evaluates each category and selects the one that maximizes the purity of the resulting subsets (based on the chosen impurity criterion).

- Some algorithms, like CART (Classification and Regression Trees), may use binary splits for categorical variables, creating separate branches for each category.

**4.What are some common techniques to prevent overfitting in decision trees?**

**- Pruning :** Pruning involves removing branches or nodes of the tree that do not significantly improve predictive performance on a validation set. This helps prevent the model from capturing noise in the training data and improves generalization to unseen data.

**- Minimum Samples per Leaf:** Setting a minimum number of samples required to be present in a leaf node before a split is considered. This prevents the creation of leaf nodes with very few samples, which can lead to overfitting.

**- Maximum Depth:** Limiting the maximum depth of the tree can prevent overly complex trees that memorize the training data but perform poorly on new data.

**- Minimum Samples per Split :** Requiring a minimum number of samples in a node to be eligible for splitting can prevent the algorithm from splitting nodes with too few samples, which may lead to overfitting.

**1.What is the basic idea behind SVM?**

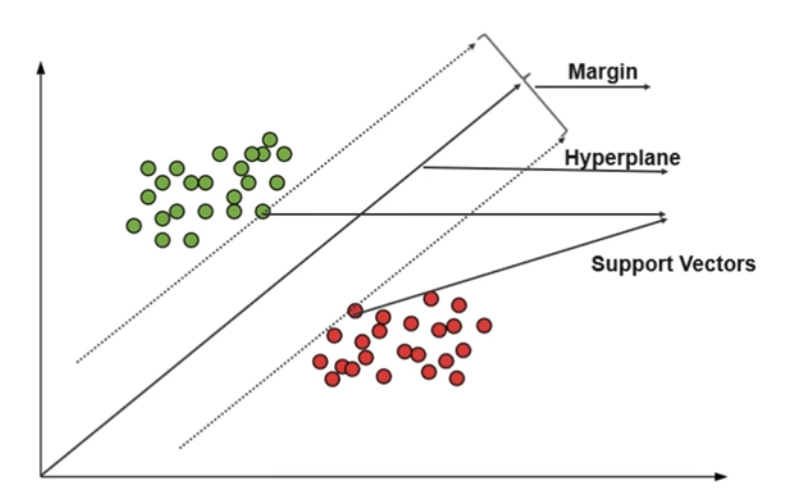
- The basic idea behind Support Vector Machines (SVM) is to find the hyperplane that best separates the data into different classes while maximizing the margin between the classes.

- SVM aims to find the optimal decision boundary that maximizes the margin, which is the distance between the hyperplane and the closest data points from each class (support vectors).

**2.Explain the concepts of margin and support vectors in SVM.**

- Margin : The margin is the distance between the decision boundary (hyperplane) and the nearest data points from each class. The optimal hyperplane in SVM is the one that maximizes this margin.

- Support Vectors : Support vectors are the data points that lie closest to the decision boundary and define the margin. These are the critical points that determine the position and orientation of the hyperplane. Only support vectors influence the placement of the decision boundary; other data points are irrelevant.



**3.What are the different kernel functions used in SVM, and when would you use each?**

**- Linear Kernel :** The linear kernel is the simplest kernel function, and it represents a linear decision boundary. It works well when the data is linearly separable or when the number of features is very high compared to the number of samples.

**- Polynomial Kernel :** The polynomial kernel computes the dot product of the feature vectors in a higher-dimensional space. It allows for non-linear decision boundaries and is suitable for data that cannot be separated linearly but may have some degree of polynomial separation.

**- Radial Basis Function (RBF) Kernel :** The RBF kernel measures the similarity between data points in an infinite-dimensional space. It is versatile and can handle non-linear decision boundaries effectively. The RBF kernel is commonly used when the data is not linearly separable or when the relationship between features and the target variable is complex.

- Sigmoid Kernel : The sigmoid kernel calculates the similarity between data points based on the hyperbolic tangent function. It can be useful for neural network-like behaviors, but it is less commonly used compared to other kernels.

**4.How does SVM handle outliers?**

- SVM is inherently robust to outliers due to its reliance on the support vectors, which are the data points closest to the decision boundary.The SVM algorithm has a feature to ignore outliers and find the hyper-plane that has the maximum margin. Outliers that are far from the decision boundary have little effect on the placement of the boundary.

- However, if outliers significantly impact the decision boundary or reduce the margin, one approach is to adjust the penalty parameter (C) in the SVM optimization problem. Increasing the value of C allows for more flexibility in the decision boundary, potentially accommodating outliers. Additionally, using a robust kernel like the RBF kernel can help make SVM more resistant to outliers by modeling complex decision boundaries.