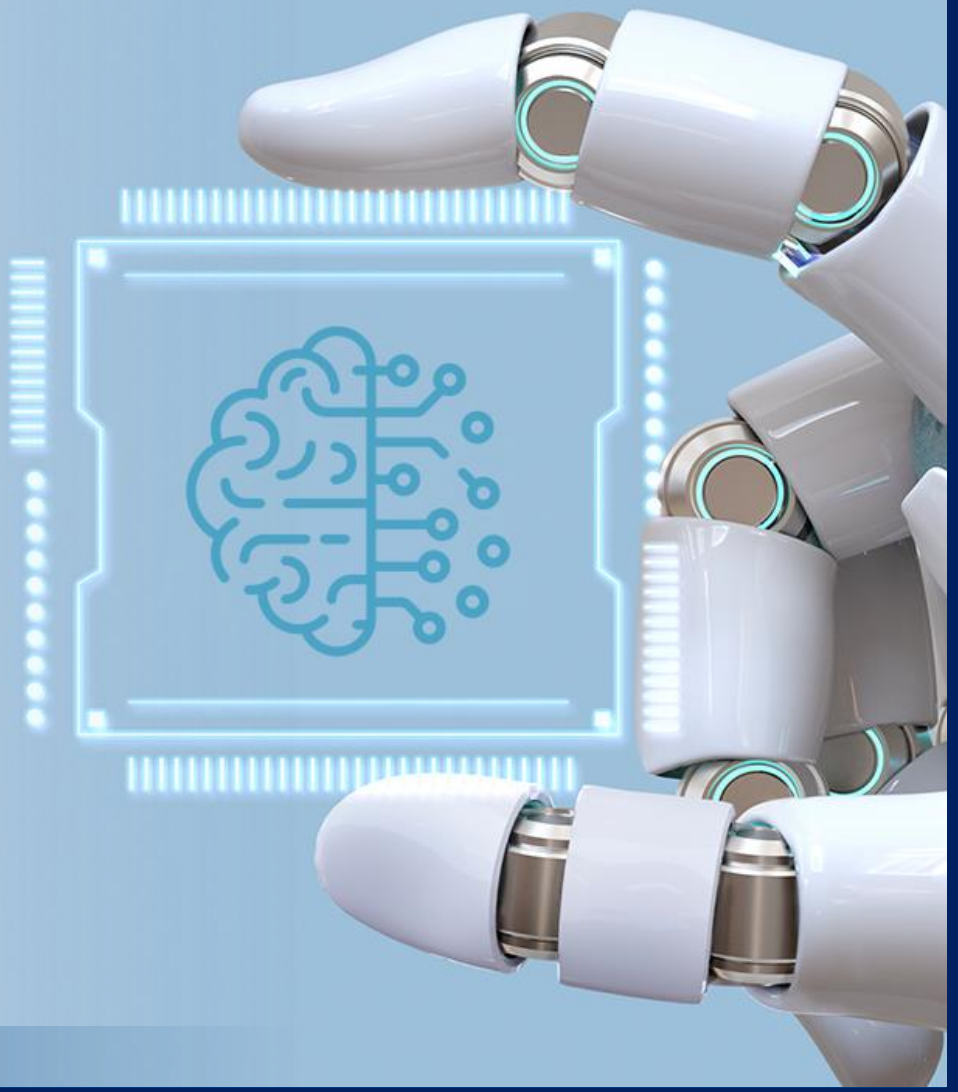
edunet foundation

Machine Learning Algorithms



Disclaimer

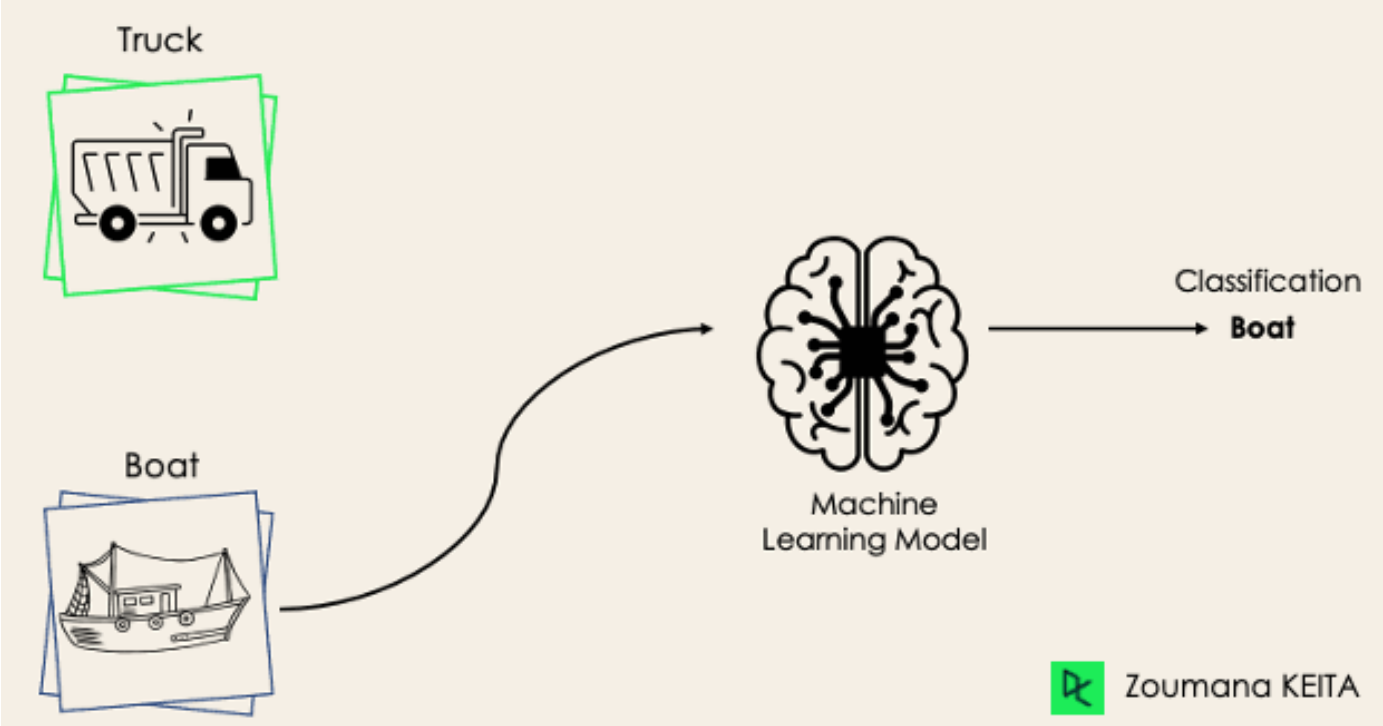
The content is curated from online/offline resources and used for educational purpose only

Agenda

- Classification
- Logistic Regression
- Decision Trees
- Splitting criteria(Gini Impurity, Entropy)
- Advantages and limitations of decision trees

Classification

- Classification involves learning a mapping from input variables (features) to discrete output labels.
- The output labels are categories or classes to which the input data points belong.
- The goal is to train a classifier that can accurately predict the class label of new, unseen data points.



Source: https://www.datacamp.com/blog/classification-machine-learning

Types of Classification

Binary Classification: Involves classifying data into two classes.

- Example: Spam detection (classifying emails as spam or not spam).

Multi-Class Classification: Involves classifying data into more than two classes.

- Example: Handwritten digit recognition (classifying digits from 0 to 9).

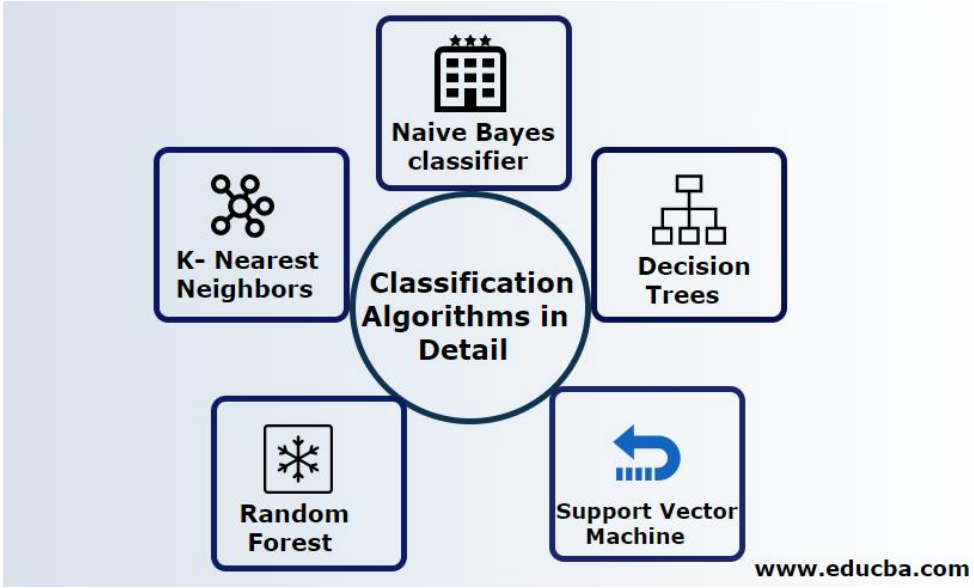
Multi-Label Classification: Involves assigning multiple labels to each instance.

- Example: Tagging documents with relevant topics (e.g., sports, politics, entertainment).

Source: https://www.zoumdatascience.com/different-methods-to-evaluation-binary-classification-models/

Classification Algorithms

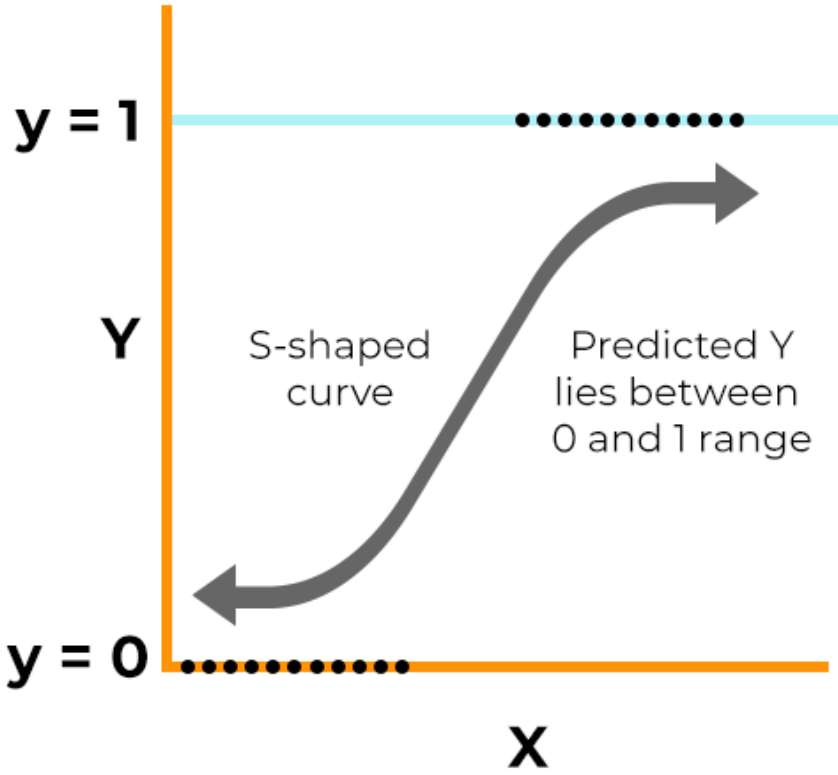
- **Logistic Regression:** Despite its name, logistic regression is a classification algorithm used for binary classification.
- **Support Vector Machines (SVM):** Effective for both binary and multi-class classification by finding the optimal hyperplane that separates classes.
- **Decision Trees:** Hierarchical tree-like structures that recursively split the feature space based on feature values.
- **Random Forest:** Ensemble of decision trees where predictions are averaged over multiple trees, suitable for both binary and multi-class classification.
- **K-Nearest Neighbors (KNN):** Classifies data points based on the majority class among their nearest neighbors.
- **Naive Bayes:** Probabilistic classifier based on Bayes' theorem, often used in text classification and spam filtering.



Source: https://cloud2data.com/need-for-different-types-of-classifiers/

Logistic Regression

- Logistic regression is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm which analyze the relationship between two data factors.
- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value.
- It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- In Logistic regression, instead of fitting a regression line, we fit an “S” shaped logistic function, which predicts two maximum values (0 or 1).

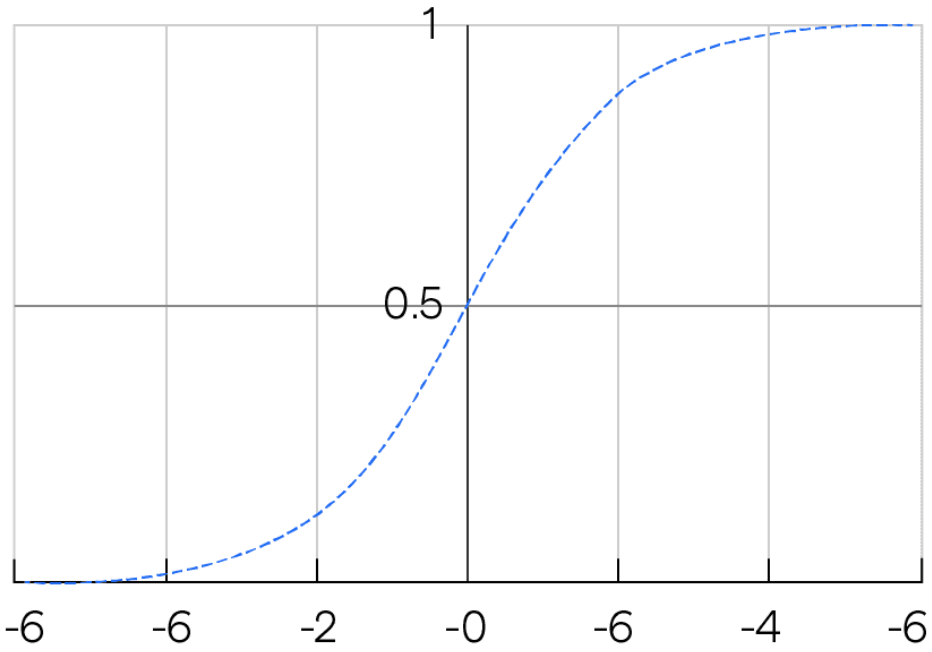


Source: https://www.linkedin.com/pulse/decoding-liver-health-using-logistic-regression-predict-dr-jay-mogare-homlf

Logistic Function – Sigmoid Function

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the “S” form.
- The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

$$f(x) = \frac{1}{1 + e^{-fx}}$$



Source: https://www.researchgate.net/figure/Standard-logistic-sigmoid-function_fig11_336216799

Types of Logistic Regression

- Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
- Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as “cat”, “dogs”, or “sheep”
- Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as “low”, “Medium”, or “High”.

	Binomial Logistic Regression	Multinomial Logistic Regression	Ordinal Logistic Regression
Number of Categories for Response Variable	2	3 or more	3 or more
Does Order of Categories Matter?	No	No	Yes

Source: https://www.almabetter.com/bytes/tutorials/data-science/logistic-regression

Assumptions of Logistic Regression

Independent observations: Each observation is independent of the other. meaning there is no correlation between any input variables.

Binary dependent variables: It takes the assumption that the dependent variable must be binary or dichotomous, meaning it can take only two values. For more than two categories SoftMax functions are used.

Linearity relationship between independent variables and log odds: The relationship between the independent variables and the log odds of the dependent variable should be linear.

No outliers: There should be no outliers in the dataset.

Large sample size: The sample size is sufficiently large

Terminologies involved in Logistic Regression

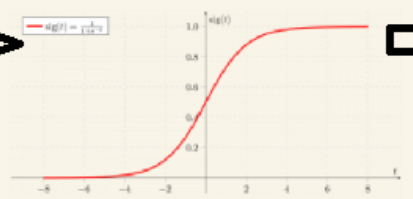
- **Independent variables:** The input characteristics or predictor factors applied to the dependent variable's predictions.
- **Dependent variable:** The target variable in a logistic regression model, which we are trying to predict.
- **Logistic function:** The formula used to represent how the independent and dependent variables relate to one another. The logistic function transforms the input variables into a probability value between 0 and 1, which represents the likelihood of the dependent variable being 1 or 0.
- **Odds:** It is the ratio of something occurring to something not occurring. it is different from probability as the probability is the ratio of something occurring to everything that could possibly occur.
- **Log-odds:** The log-odds, also known as the logit function, is the natural logarithm of the odds. In logistic regression, the log odds of the dependent variable are modeled as a linear combination of the independent variables and the intercept.
- **Coefficient:** The logistic regression model's estimated parameters, show how the independent and dependent variables relate to one another.
- **Intercept:** A constant term in the logistic regression model, which represents the log odds when all independent variables are equal to zero.
- **Maximum likelihood estimation:** The method used to estimate the coefficients of the logistic regression model, which maximizes the likelihood of observing the data given the model.

How does Logistic Regression work?

- The logistic regression model transforms the linear regression function continuous value output into categorical value output using a sigmoid function, which maps any real-valued set of independent variables input into a value between 0 and 1. This function is known as the logistic function.
- Let the independent input features be X, dependent variable be Y and the multi-linear function be z. Once the multi-linear function is calculated, the sigmoid function will be applied on it to get the probability for the classes.

$$X = \begin{bmatrix} x_{11} & \dots & x_{1m} \\ x_{21} & \dots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nm} \end{bmatrix} \Rightarrow z = \left(\sum_{i=1}^{n} w_i x_i \right) + bz$$

$$z = w \cdot X + bz{=}w{\cdot}X{+}b \Rightarrow \sigma(z) = \frac{1}{1 - e^{-z}} \sigma(z){=}1{-}e{-}z1 \Rightarrow Y = \begin{cases} 0 & \text{if } Class1 \\ 1 & \text{if } Class2 \end{cases}$$

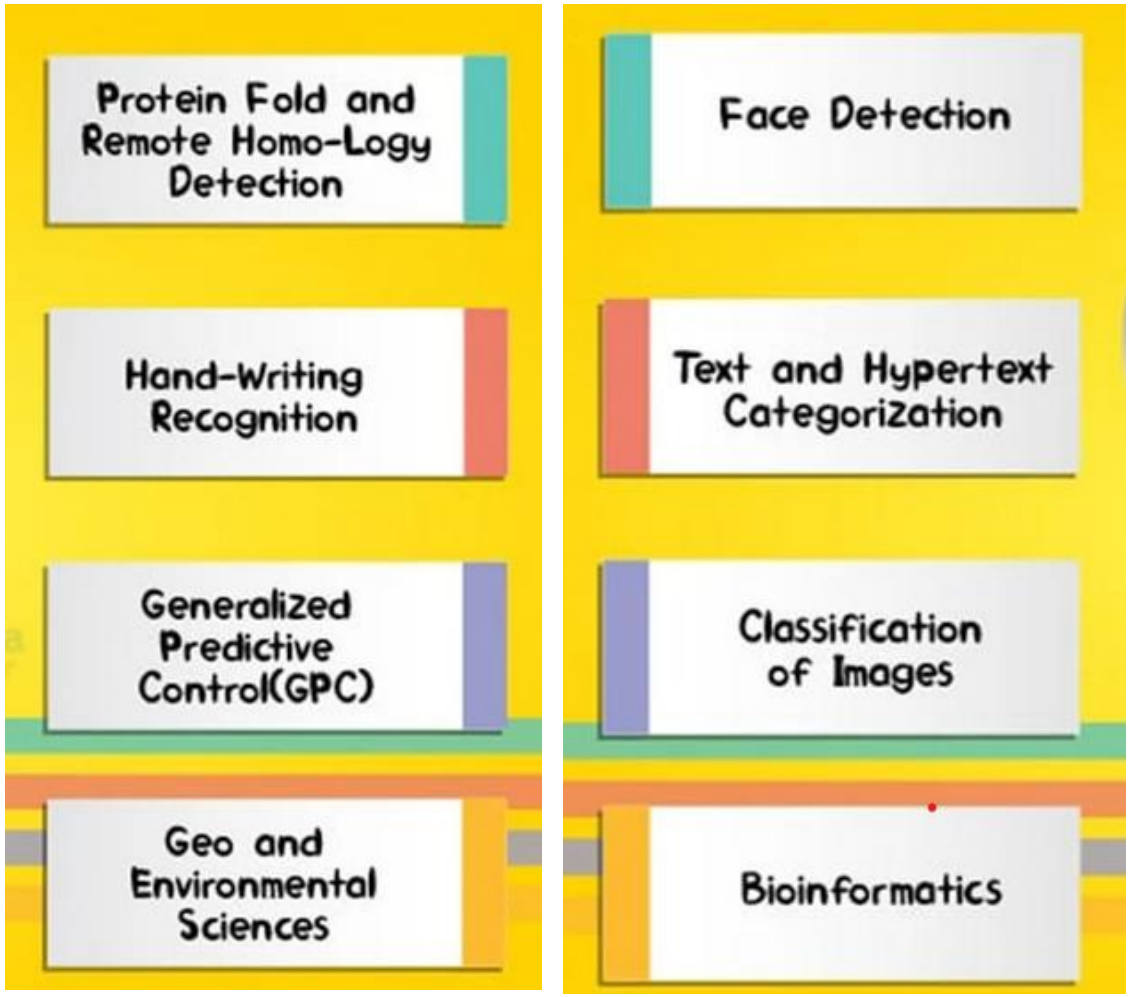


Sigmoid function

Source:

Applications of Classification

- Text and hypertext categorization
- Classification of images
- Bioinformatics
- Protein fold and remote homology detection
- Handwriting recognition
- Generalized predictive control(GPC)
- Face detection



Source: https://pianalytix.com/an-intro-to-support-vector-machine-svm/

Evaluation Metrics for Classification

Evaluation metrics are used to assess the performance of machine learning models. They provide quantitative measures of how well the model is performing on the task at hand. Here are some common evaluation metrics used for classification tasks:

- **Accuracy :** Accuracy measures the proportion of correctly classified instances out of all instances.
- **Precision :** Precision measures the proportion of true positive predictions among all positive predictions made by the model.
- **Recall (Sensitivity) :** Recall measures the proportion of true positive predictions among all actual positive instances in the dataset.
- **F1-score :** F1-score is the harmonic mean of precision and recall, providing a balanced measure between the two.
- **Specificity (True Negative Rate) :** Specificity measures the proportion of true negative predictions among all actual negative instances in the dataset.
- **ROC Curve and AUC :** Receiver Operating Characteristic (ROC) curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- **Confusion Matrix :** A confusion matrix is a table that summarizes the performance of a classification algorithm.

Lab Exercise - Code Implementation for Binomial Logistic Regression

Hands On

Refer: Lab 1

- Target variable can have only 2 possible types: “0” or “1” which may represent “win” vs “loss”, “pass” vs “fail”, “dead” vs “alive”, etc., in this case, sigmoid functions are used.
- Use Breast Cancer dataset from Sklearn dataset to train a binomial logistic regression model. (Refer: Lab-1)

Lab Exercise - Code Implementation for Multinomial Logistic Regression

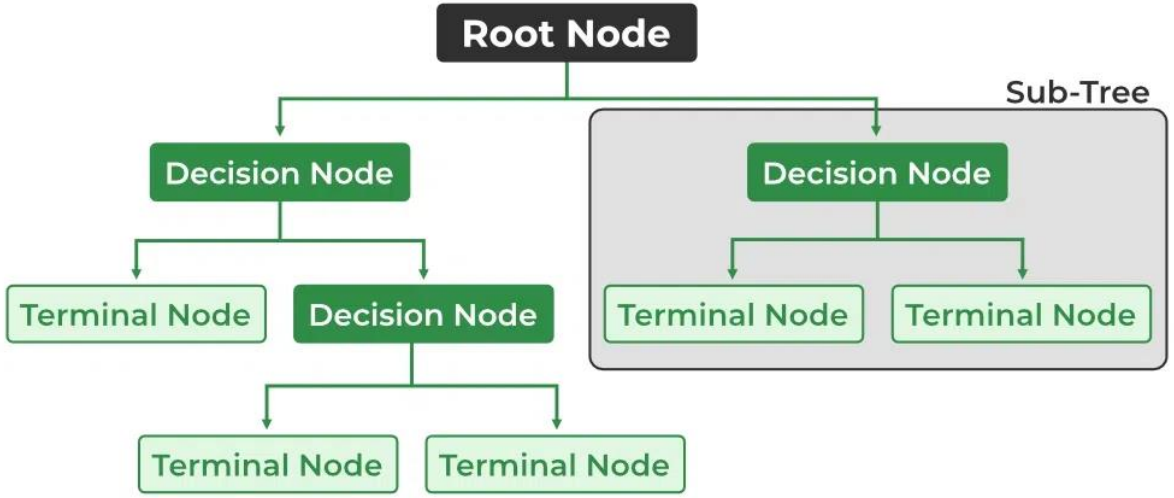
Hands On

Refer: Lab 2

- Target variable can have 3 or more possible types which are not ordered (i.e. types have no quantitative significance) like “disease A” vs “disease B” vs “disease C”.
- In this case, the softmax function is used in place of the sigmoid function.
- Use digits dataset from Sklearn dataset to train a multinomial logistic regression model. (Refer: Lab-2)

Decision Tree

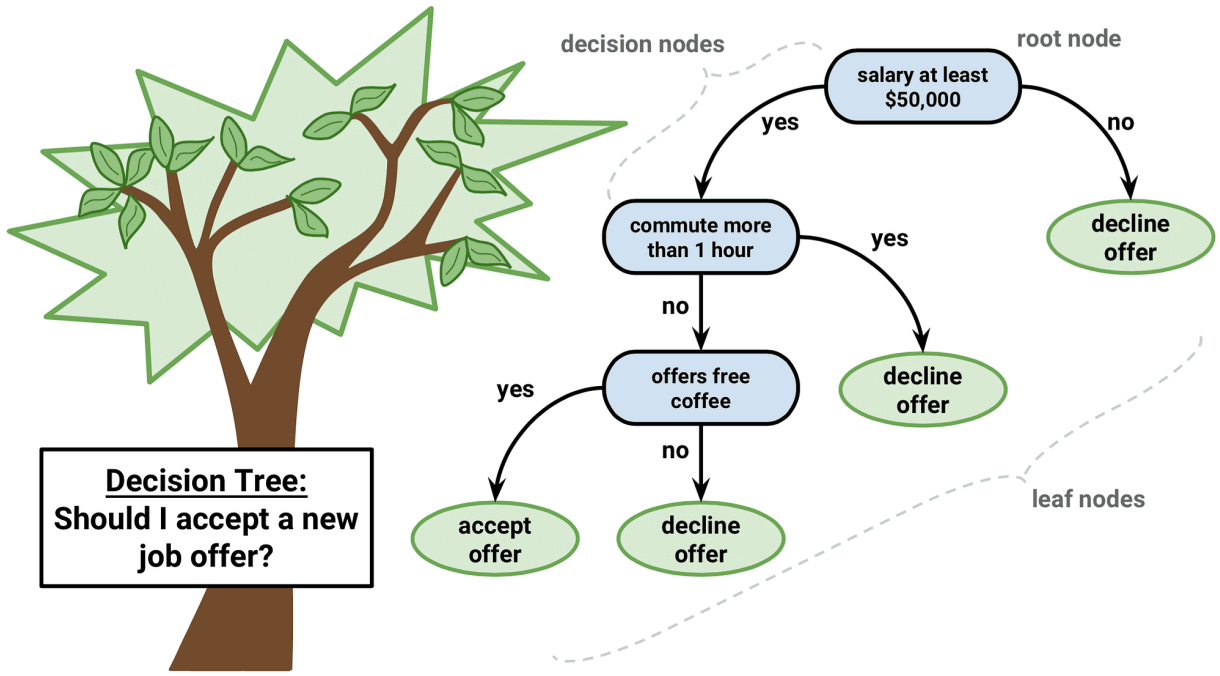
- Decision trees are a popular type of machine learning algorithm used for both classification and regression tasks.
- A decision tree is a hierarchical tree-like structure that represents a sequence of decisions and their possible consequences.
- It partitions the feature space into a set of rectangular regions, each associated with a class label (for classification) or a predicted value (for regression).
- Decision trees are intuitive and easy to interpret, making them particularly useful for understanding the decision-making process of a model.



Source: https://medium.com/@francescofranco_39234/decision-trees-in-machine-learning-5e21a8446749

Construction of Decision Tree

- **Tree Induction:** The process of building a decision tree from the training data.
- **Top-Down Approach:** Decision trees are constructed recursively from the root node to the leaf nodes.
- **Splitting Criteria:** At each node, the algorithm selects the best feature to split the data based on certain criteria (e.g., information gain, Gini impurity, or entropy).
- **Stopping Criteria:** The tree-growing process stops when one of the stopping criteria is met, such as reaching a maximum depth, minimum number of samples at a node, or when no further improvement can be made.

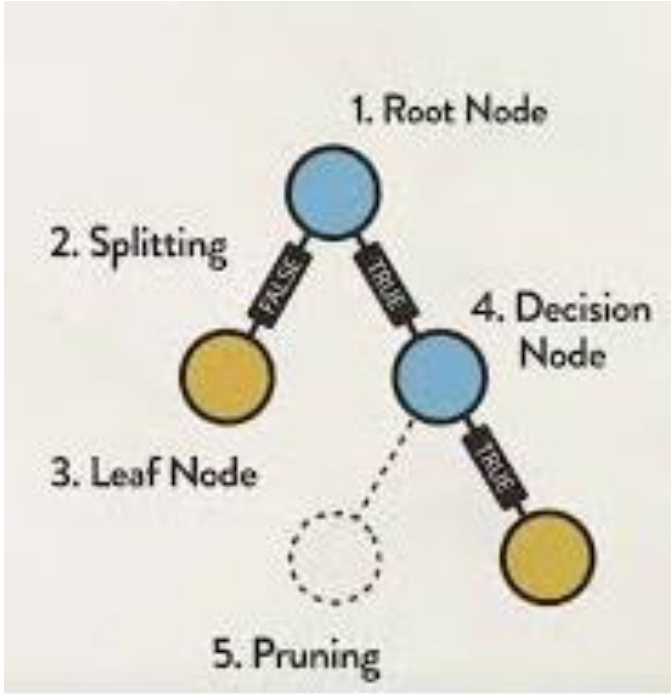


Source: https://medium.com/@sunils0506/decision-tree-types-metrics-dc9b97150f8d

Decision Tree Terminologies

In decision tree, each internal node denotes the feature, branches denote the rules and the leaf nodes denote the result of the algorithm.

- **Root Node:** It is the topmost node in the tree, which represents the complete dataset. It is the starting point of the decision-making process.
- **Decision/Internal Node:** A node that symbolizes a choice regarding an input feature. Branching off of internal nodes connects them to leaf nodes or other internal nodes.
- **Leaf/Terminal Node:** A node without any child nodes that indicates a class label or a numerical value.
- **Splitting:** The process of splitting a node into two or more sub-nodes using a split criterion and a selected feature.
- **Branch/Sub-Tree:** A subsection of the decision tree starts at an internal node and ends at the leaf nodes.
- **Parent Node:** The node that divides into one or more child nodes.



Source: https://medium.com/@jdseo/to-a-5-year-old-decision-trees-explained-without-math-2ae1b489c7a9

Decision Tree Terminologies

- **Child Node:** The nodes that emerge when a parent node is split.
- **Impurity:** A measurement of the target variable's homogeneity in a subset of data. It refers to the degree of randomness or uncertainty in a set of examples. The Gini index and entropy are two commonly used impurity measurements in decision trees for classifications task
- **Variance:** Variance measures how much the predicted and the target variables vary in different samples of a dataset. It is used for regression problems in decision trees. Mean squared error, Mean Absolute Error, friedman_mse, or Half Poisson deviance are used to measure the variance for the regression tasks in the decision tree.
- **Information Gain:** Information gain is a measure of the reduction in impurity achieved by splitting a dataset on a particular feature in a decision tree. The splitting criterion is determined by the feature that offers the greatest information gain, It is used to determine the most informative feature to split on at each node of the tree, with the goal of creating pure subsets
- **Pruning:** The process of removing branches from the tree that do not provide any additional information or lead to overfitting.

How Decision Tree Works?

Decision trees work by recursively partitioning the feature space into regions, with each partition corresponding to a decision or a split based on feature values.

Starting Point:

- Begin with the entire dataset, where each data point represents an observation with a set of features and a corresponding target variable (class label for classification or numerical value for regression).

Feature Selection:

- Choose the best feature to split the dataset at the current node.
- The selection criteria can be based on measures like information gain, Gini impurity, or entropy.
- The goal is to find the feature that best separates the data into distinct classes or reduces the impurity of the data.

Splitting:

- Once the best feature is selected, the dataset is partitioned into subsets based on the possible values of this feature.
- Each subset corresponds to a branch or child node in the decision tree.
- This process is repeated recursively for each child node until one of the stopping criteria is met.

How Decision Tree Works?

Stopping Criteria:

- The recursive splitting process stops when one of the following criteria is met:
- Maximum tree depth is reached.
- Minimum number of samples required to split a node is reached.
- No further improvement can be made in terms of reducing impurity or increasing information gain.

Leaf Node Assignment:

- Once a stopping criterion is met, the node becomes a leaf node, and a decision is made based on the majority class (for classification) or the average value (for regression) of the target variable in that node.
- If the leaf node contains instances of multiple classes, the majority class determines the predicted class label.

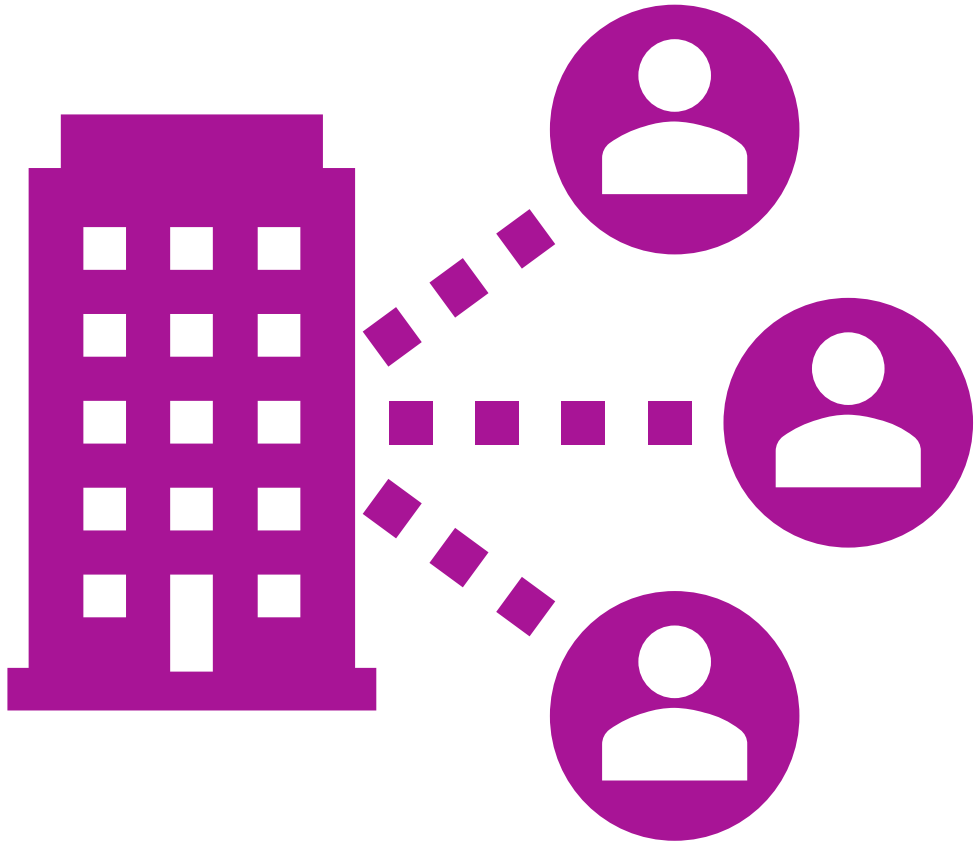
Pruning (Optional):

- After the tree is fully grown, it may be pruned to prevent overfitting and improve generalization.
- Pruning involves removing unnecessary branches from the tree while preserving the overall structure.
- Pre-pruning and post-pruning are two common strategies for controlling the size and complexity of decision trees

How Decision Tree Works?

Prediction

- To make predictions for new data, traverse the decision tree from the root node to a leaf node based on the values of its features.
- At each node, follow the corresponding branch based on the feature value until reaching a leaf node.
- The predicted class label (for classification) or value (for regression) associated with the leaf node is the final prediction.



Splitting Criteria(Gini Impurity, Entropy)

Splitting criteria, such as Gini impurity and entropy, are used in decision trees to measure the impurity or uncertainty of a dataset before and after a split. These criteria help the decision tree algorithm decide which feature and threshold to use for splitting the data.

- Gini Impurity
- Entropy
- Information gain

Gini Impurity

- Gini Impurity is a score that evaluates how accurate a split is among the classified groups.
- The Gini Impurity evaluates a score in the range between 0 and 1, where 0 is when all observations belong to one class, and 1 is a random distribution of the elements within classes.
- In this case, we want to have a Gini index score as low as possible. Gini Index is the evaluation metric we shall use to evaluate our Decision Tree Model.
- Gini impurity measures the degree of impurity or disorder in a dataset.
- For a given dataset with N data points and K classes, the Gini impurity G is calculated as:

$$G = 1 - \sum_{i=1}^{K} (p_i)^2$$

- Where p_i is the probability of randomly selecting a data point of class i from the dataset.
- Gini impurity is minimized when all data points belong to the same class (i.e., G=0).

Entropy

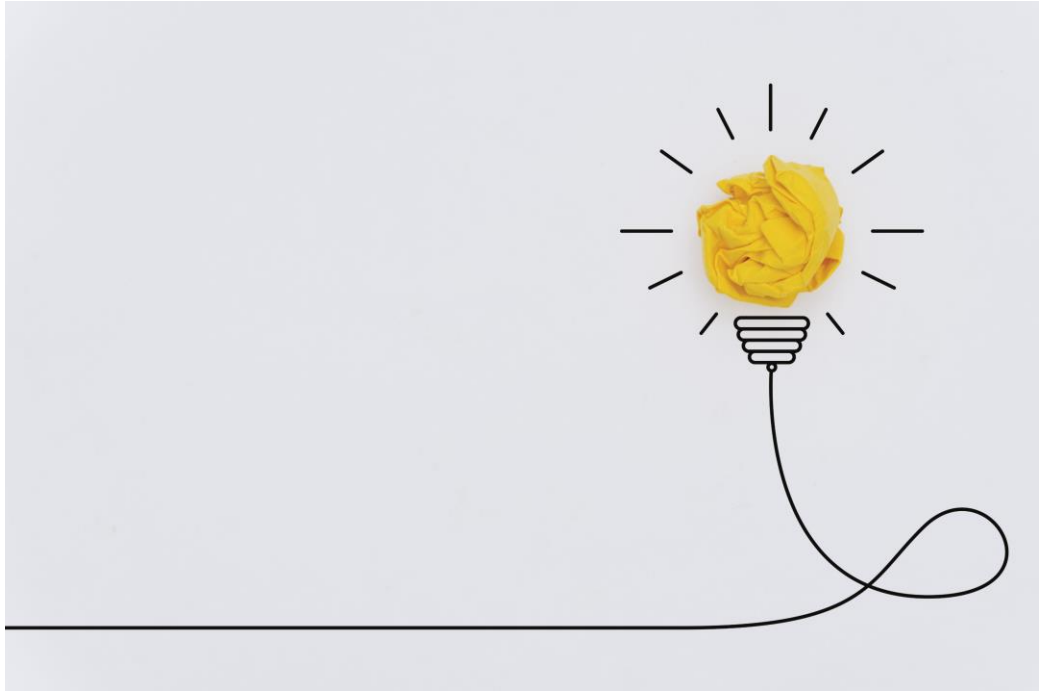
- Entropy is the measure of the degree of randomness or uncertainty in the dataset. In the case of classifications, It measures the randomness based on the distribution of class labels in the dataset.
- Entropy is another measure of impurity or uncertainty in a dataset.
- For a given dataset with N data points and K classes, the entropy H is calculated as:

$$H = -\sum_{i=1}^{K} p_i \log_2(p_i)$$

- Where p_i is the probability of randomly selecting a data point of class i from the dataset.
- Entropy is minimized when all data points belong to the same class (i.e., H=0).

Important Points Related To Entropy

- When every instance in a dataset belongs to the same class, the entropy is 0, indicating complete homogeneity. This represents the lowest uncertainty in the dataset.
- Conversely, when the dataset is evenly split among multiple classes, the entropy reaches its maximum value. Thus, entropy is highest when class labels are distributed equally, signifying maximum uncertainty in the dataset.
- Entropy is employed to assess the effectiveness of a split. Its objective is to identify the attribute that minimizes the entropy of resulting subsets by creating more homogeneous subsets based on class labels.
- The attribute with the highest information gain, which represents the reduction in entropy after splitting on that attribute, is chosen as the splitting criterion. This process is repeated recursively to construct the decision tree.



Information Gain

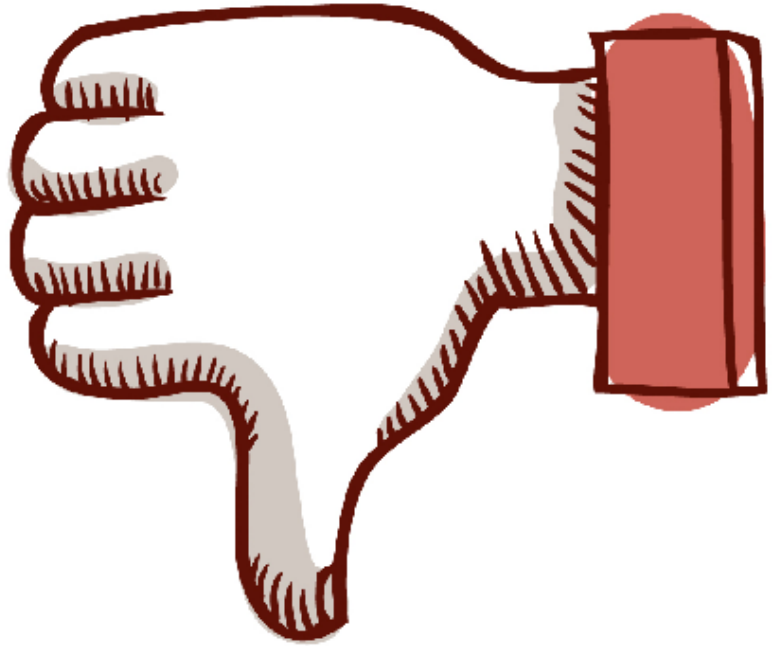
- Information gain measures the reduction in entropy or variance that results from splitting a dataset based on a specific property. It is used in decision tree algorithms to determine the usefulness of a feature by partitioning the dataset into more homogeneous subsets with respect to the class labels or target variable. The higher the information gain, the more valuable the feature is in predicting the target variable.

Advantages of the Decision Tree

- It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
- It can be very useful for solving decision-related problems.
- It helps to think about all the possible outcomes for a problem.
- There is less requirement of data cleaning compared to other algorithms.

Disadvantages of the Decision Tree

- The decision tree contains lots of layers, which makes it complex.
- It may have an overfitting issue, which can be resolved using the Random Forest algorithm.
- For more class labels, the computational complexity of the decision tree may increase.



Lab Exercise - Code Implementation For Decision Tree Classifier

Hands On

Refer: Lab 3

- Use Iris dataset from Sklearn dataset to train a decision tree classifier model. (Refer: Lab-3)

Conclusion

Congratulations! You have completed this course and now possess a solid understanding of several critical concepts in machine learning classification. You have learned about classification techniques, including logistic regression and decision trees, and the criteria used for splitting in decision trees, such as Gini impurity and entropy. Additionally, you have explored the advantages and limitations of decision trees.

Quiz

1. **What is the main purpose of logistic regression?**

A) To perform regression analysis on continuous data

B) To classify data into binary outcomes

C) To cluster data points

D) To reduce data dimensionality

Answer: B
To classify data into binary outcomes

Quiz

2. Which function does logistic regression use to map predictions to probabilities?

A) Linear function

B) ReLU function

C) Sigmoid function

D) Tanh function

Answer: C
Sigmoid function

Quiz

3. What is a decision tree?

A. A supervised learning algorithm used for both classification and regression tasks

B. An unsupervised learning algorithm used for clustering

C. A method to perform linear regression

D. A technique for dimensionality reduction

Answer: A
A supervised learning algorithm used for both classification and regression tasks

Quiz

4. What is entropy used for in a decision tree?

A) To determine the complexity of the tree

B) To measure the purity of a node

C) To calculate the learning rate

D) To evaluate the performance of the tree

Answer: B
To measure the purity of a node

Quiz

5. Which of the following is an advantage of decision trees?

A) They require a lot of data preprocessing

B) They can handle both numerical and categorical data

C) They are prone to overfitting

D) They are difficult to interpret

Answer: B
They can handle both numerical and categorical data

References

- https://www.geeksforgeeks.org/getting-started-with-classification/
- https://www.geeksforgeeks.org/understanding-logistic-regression/
- https://www.ibm.com/topics/logistic-regression
- https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/
- https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
- https://medium.com/@arpita.k20/gini-impurity-and-entropy-for-decision-tree-68eb139274d1#:~:text=It%20aims%20to%20reduce%20the,criterion%20measure%20similar%20performance%20metrics.

edunet
foundation

Thank You!