Classification

Pre-requisites

Hope you have gone through the self-learning content for this session on the PRISM portal.

Recap

By the End of this Session, You Will:

- Learn classification and different classification algorithms
- Explore evaluation and performance metrics for classification models
- Understand logistic regression and logit function
- Interpret of Beta coefficients in Logistic Regression
- Build a logistic regression model

What’s in It for Me?

Understand the concept of classification in machine learning and gain insights into when to use different classification algorithms based on real world data characteristics.

Gain a deep understanding of various evaluation metrics used to measure the performance of Logistic Regression models.

Learn about the logit function and grasp the mathematical details of logistic regression for intuitive understanding.

Discover practical implications of Beta coefficient interpretation in real-world scenarios and how to build a basic logistic regression model.

Poll Time

Q. What is the primary goal of classification in machine learning?

a. To predict a continuous output value

b. To predict a categorical label or class

c. To minimize the mean squared error

d. To cluster similar data points together

Poll Time

Q. What is the primary goal of classification in machine learning?

a. To predict a continuous output value

b. To predict a categorical label or class

c. To minimize the mean squared error

d. To cluster similar data points together

Introduction to Classification

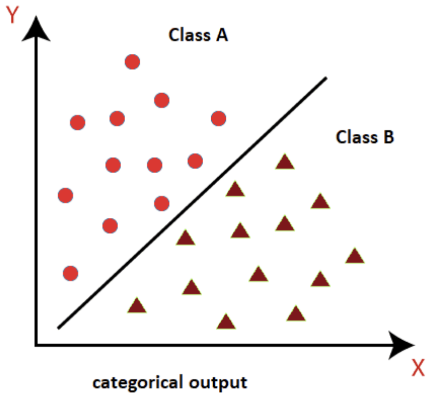
Introduction to Classification

Classification is all about categorizing data points into predefined classes or categories based on independent features.

Classes or Categories: In classification tasks, you have a set of classes or categories that you want to assign data points to.

Features: Features are the characteristics or attributes of the data points that the algorithm uses to make predictions.

Classification is a powerful tool with a wide range of **applications,** including image recognition, natural language processing, fraud detection, and more.



Real World Use Cases of Classification

1 Spam Email Detection

2 Medical Diagnosis

3 Fraud Detection

4 Credit Scoring

5 Sentiment Analysis

6 Image Recognition

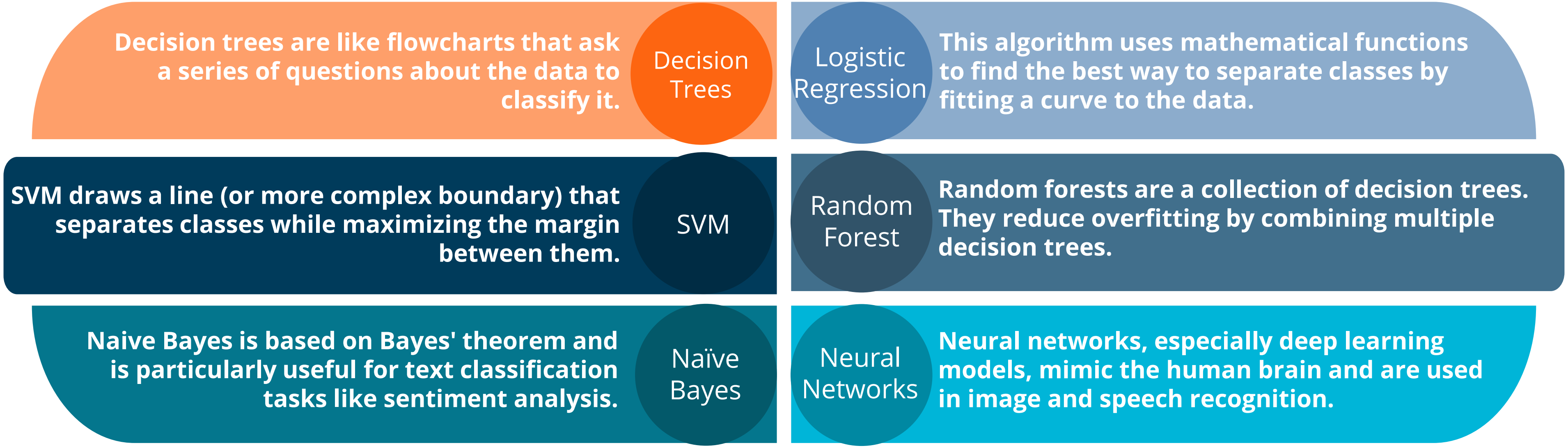
7 Customer Churn Prediction

8 Disease Outbreak Prediction

Different Classification Algorithms

Definition

Classification algorithms are like teaching a computer to recognize patterns and make predictions about which class or category new data belongs to.



Decision Trees

Decision trees are like flowcharts that ask a series of questions about the data to classify it.

Logistic Regression

This algorithm uses mathematical functions to find the best way to separate classes by fitting a curve to the data.

SVM

SVM draws a line (or more complex boundary) that separates classes while maximizing the margin between them.

Random Forest

Random forests are a collection of decision trees. They reduce overfitting by combining multiple decision trees.

Naïve Bayes

Naive Bayes is based on Bayes' theorem and is particularly useful for text classification tasks like sentiment analysis.

Neural Networks

Neural networks, especially deep learning models, mimic the human brain and are used in image and speech recognition.

Poll Time

Q. What is the primary goal of credit scoring using classification algorithms?

a. Predicting the stock market

b. Assessing credit risk for loan applicants

c. Identifying plant species

d. Detecting fraudulent credit card transactions

Poll Time

Q. What is the primary goal of credit scoring using classification algorithms?

a. Predicting the stock market

▌**b. Assessing credit risk for loan applicants**

c. Identifying plant species

d. Detecting fraudulent credit card transactions

Evaluation Metrics for Classification Models

Evaluation metrics provide valuable insights into how well a model is performing, helping make informed decisions.

Some of the most common evaluation metrics for classification models:

1. Confusion Matrix
2. Accuracy
3. Precision
4. Recall (Sensitivity or True Positive Rate)
5. F1-Score
6. Specificity (True Negative Rate)
7. Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC)

- Different metrics highlight different aspects of model performance, so choosing the right ones ensures a comprehensive assessment of your classification model.

Confusion Matrix

Confusion matrix provides a clear and detailed breakdown of the model's predictions and actual outcomes.

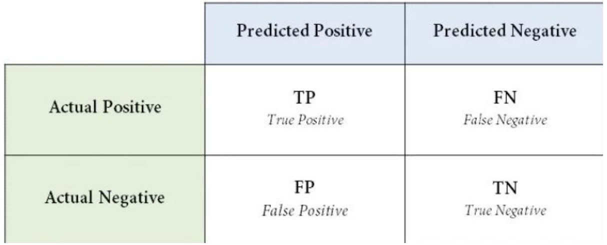
Components of a Confusion Matrix:
A confusion matrix consists of four key components:

1. **True Positives (TP)**: Model predicted 'yes,' and the actual outcome was also 'yes.'

2. **True Negatives (TN)**: Model predicted 'no,' and the actual outcome was 'no.'

3. **False Positives (FP)**: It is also known as a Type I error.
The model predicted 'yes,' but the actual outcome was 'no.'

4. **False Negatives (FN)**: This is a Type II error.
The model predicted 'no,' but the actual outcome was 'yes.'



	Predicted Positive	Predicted Negative
Actual Positive	TP *True Positive*	FN *False Negative*
Actual Negative	FP *False Positive*	TN *True Negative*

Confusion matrices are particularly useful when you want to understand the types of errors your model is making. For example, in medical
diagnosis, knowing the number of false positives and false negatives can be critical.

Demo : Components of Confusion Matrix

Metrics Derived from Confusion Matrix

Accuracy

It is the ratio of correctly predicted instances (TP + TN) to the total number of instances. Accuracy measures the overall correctness of predictions.

Precision

Precision is calculated as TP / (TP + FP). It measures the proportion of positive predictions that were correct.

Recall (Sensitivity or True Positive Rate)

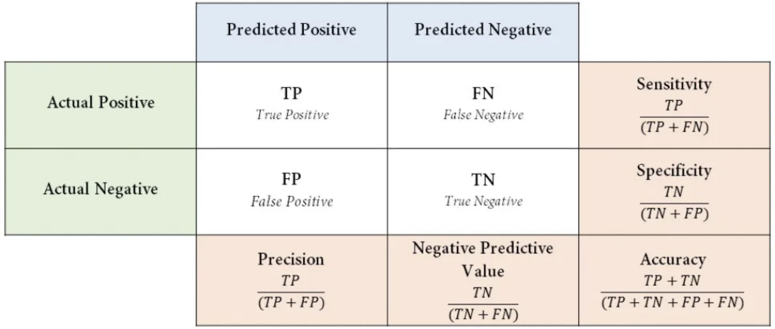
Recall is calculated as TP / (TP + FN). It measures the proportion of actual positive instances that were correctly predicted by the model.

Specificity (True Negative Rate):

Specificity is calculated as TN / (TN + FP). It measures the proportion of actual negative instances that were correctly predicted by the model.

F1-Score:

The F1-Score is the harmonic mean of precision and recall. It balances the trade-off between false positives and false negatives.



	Predicted Positive	Predicted Negative	
Actual Positive	TP *True Positive*	FN *False Negative*	Sensitivity $\frac{TP}{(TP+FN)}$
Actual Negative	FP *False Positive*	TN *True Negative*	Specificity $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

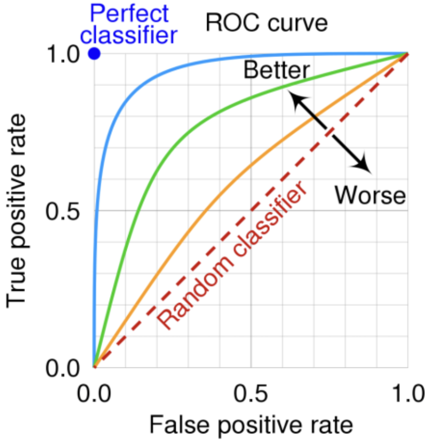
Demo : Metrics Derived from Confusion Matrix

ROC Curve

The Receiver Operating Characteristic (ROC) curve is a **graphical representation** commonly used to assess the performance of binary classification models.

The ROC curve is created by plotting the **TPR (sensitivity or recall)** on the y-axis and the **FPR (1-specificity)** on the x-axis. Here's how it works:

- The TPR measures the model's ability to correctly identify positive cases.
- The FPR measures the model's tendency to produce false alarms for negative cases.



The ROC curve is generated by **varying the decision threshold** of the model, which determines the point at which it classifies an instance as positive or negative.

By calculating the **TPR and FPR at each point**, you create the ROC curve.

A model with **perfect classification** would have a ROC curve that hugs the top-left corner (TPR = 1, FPR = 0), while a **random classifier** model would have a curve that approximates a diagonal line (45-degree angle).

AUC ROC Score

AUC-ROC (Area Under the ROC Curve):
AUC-ROC is a single scalar value that quantifies the performance of a classification model. It represents the area under the ROC curve and ranges from 0 to 1.

- An AUC-ROC of 0.5 indicates that the model's performance is no better than random chance.

- **Higher AUC-ROC values** indicate better model performance. A model with an AUC-ROC closer to 1 is better at separating positive and negative instances.

- It is especially useful when dealing with **imbalanced datasets**, where one class significantly outweighs the other.

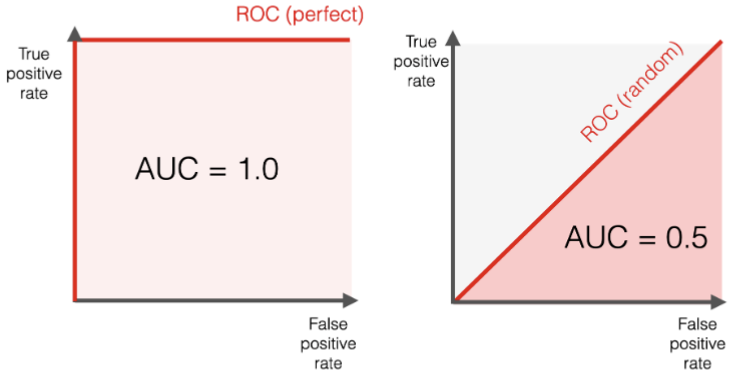
Interpretation of model's performance based on AUC-ROC:

- **0.5 < AUC-ROC < 0.7**: The model's performance is generally poor.

- **0.7 ≤ AUC-ROC < 0.8**: The model's performance is fair, but there is room for improvement.

- **0.8 ≤ AUC-ROC < 0.9**: Good, model can effectively distinguish between positive and negative instances.

- **AUC-ROC ≥ 0.9**: Excellent, and the model demonstrates a high degree of accuracy in classifying instances.



Demo : ROC Curve and AUC-ROC Score

Performance Metrics for Classification Models

Performance Metrics

Performance metrics for classification models are essential tools for assessing how well a machine learning model classifies data into distinct categories or classes.

Key Classification Performance Metrics

- Accuracy
- Precision
- Recall (Sensitivity or True Positive Rate)
- F1-Score
- Specificity (True Negative Rate)
- Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC-ROC)

Metrics Selection

When selecting performance metrics for a classification problem, consider the specific objectives and constraints of your task, as well as the characteristics of your dataset.

Pop Quiz

Q. If a model has a higher Adjusted R-squared compared to another model, it means:

a. The first model is better regardless of the number of predictors

b. The first model is better at fitting the data but might be overfitting

c. The second model is better because it has more predictors

d. The first model is better because it explains a larger proportion of the variance

Pop Quiz

Q. If a model has a higher Adjusted R-squared compared to another model, it means:

a. The first model is better regardless of the number of predictors

b. The first model is better at fitting the data but might be overfitting

c. The second model is better because it has more predictors

d. The first model is better because it explains a larger proportion of the variance

Q&A
?
?
?
?
?
?
?

Logistic Regression

Introduction to Logistic Regression

Logistic regression is a statistical method used for binary classification, which means it predicts one of two possible outcomes, such as yes/no, 1/0, or true/false.

Logistic regression is primarily used for binary classification problems. It answers questions like:

- Will a customer buy a product (yes/no)?
- Will an email be classified as spam (yes/no)?

Probability Estimation :
Logistic regression estimates the probability that the dependent variable (outcome) belongs to a particular category.

Use Cases :

Credit scoring, Spam detection, Disease diagnosis, Customer churn prediction

Odds and Odds Ratio

Odds and Odds ratio are fundamental concepts to interpret the results of logistic regression models and understand the relationship between predictor variables and the probability of the event of interest occurring.

Odds

In logistic regression, odds represent the likelihood of an event occurring relative to the likelihood of it not occurring.
Odds (O) = P(Event occurs) / P(Event does not occur)
Odds can range from 0 to positive infinity.

Odds Ratio

The odds ratio (OR) is used to compare the odds of an event occurring in one group to the odds in another group.
Odds Ratio (OR) = (Odds in Group A) / (Odds in Group B)
For a one-unit change in a predictor variable, the odds ratio tells how much the odds of the event occurring increase or decrease.

Interpretation

- An odds ratio of less than 1 suggests that an increase in predictor variables is associated with lower odds of events occurring and vice versa.
- The magnitude of the odds ratio indicates the strength of the association, while its direction (greater or less than 1) indicates the direction of the effect.

Demo : Odds and Odds Ratio

Sigmoid Function

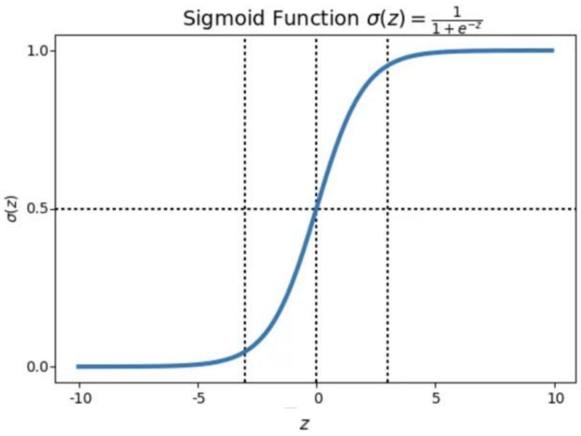
The sigmoid function, also known as the logistic function, is a mathematical function that maps a linear combination of input features to a probability score between 0 and 1.

Sigmoid Function Formula:

$\sigma(z)$ is the estimated probability lying between 0 to 1.

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

z is the linear combination of predictor variables and their associated coefficients in logistic regression.



Properties of the Sigmoid Function:

S-Shaped Curve: The curve starts at 0 as *z* approaches negative infinity and approaches 1 as *z* approaches positive infinity.

Range: The output of the sigmoid function always falls in the range [0, 1], making it suitable for modeling probabilities.

Midpoint: The midpoint of the curve, where S(*z*)=0.5, corresponds to *z*=0. If *z* is greater than 0, the probability of an event is greater than 0.5, and vice versa.

Role of Sigmoid Function in Logistic Regression

The sigmoid function is used to transform the linear combination of predictor variables into a probability score.

The linear combination *z* is calculated as follows where: $z = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_n \cdot x_n$

z is log-odds or logit, representing a linear combination of coefficients (β) and predictor variables (x).

- B_0 is the intercept term. B_1,β_2,...,β_n are coefficients associated with each predictor variable.
- x_1,x_2,...,x_n are the values of the predictor variables.

The log-odds (*z*) is then passed through the sigmoid function to obtain the estimated probability of the event occurring:

Where: $P(Y = 1) = S(z) = \frac{1}{1+e^{-z}}$

P(*Y*=1) is the probability that the event (e.g., "yes" in a binary classification problem) occurs.

Decision Boundary:

The decision boundary in logistic regression is where z=0, which corresponds to (z)=0.5. This boundary separates the feature space into two regions, one where the estimated probability is greater than 0.5 (positive class) and the other where it is less than 0.5 (negative class).

Demo : Sigmoid Function

Logit Function

The logit function, also known as log-odds, represents the natural logarithm of the odds of an event occurring.

Significance : It allows the model a linear relationship between predictor variables and log-odds of the outcome variable.

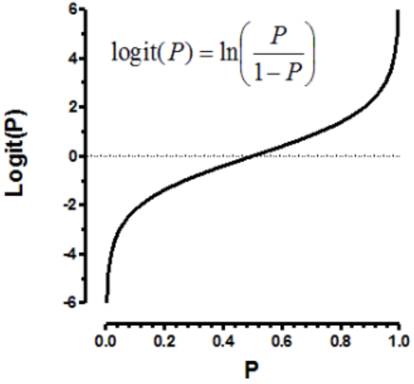
Usage : The logit function is used to transform a linear combination of predictor variables into a linear model.

Beta coefficients (β) associated with predictor variables in linear models indicate the direction and strength of their influence on the log-odds.

Logit Function Formula:

$$\mathrm{Logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

- Logit(p) is the logit function, representing the natural logarithm of the odds.
- p is the probability of the event occurring, and ln denotes the natural logarithm.



🗎 Demo : Logit Function

Properties and Interpretation of Logit Function

Properties of Logit Function

- **Range:** The logit function can take any real value from negative infinity to positive infinity.
- **Sensitivity to Probabilities:** Small changes in **p** near 0 or 1 lead to large changes in the logit value.
- **Symmetry:** The logit function is symmetric around **p = 0.5**.

Interpretation

- The coefficients (*β*) associated with predictor variables reflect the change in log-odds for a one-unit change in predictor variable while holding other variables constant.
- Exponentiating these coefficients (i.e., $e^{\beta^{i}}$) gives the **odds ratio**, which quantifies the change in odds of the event for a one-unit change in the predictor.

Pop Quiz

Q. What is the range of the sigmoid function S(z), where z represents the linear combination of predictor variables in logistic regression?

a. [0, 1]

b. (-∞, ∞)

c. [0, 0.5]

d. [-1, 1]

Pop Quiz

Q. What is the range of the sigmoid function S(z), where z represents the linear combination of predictor variables in logistic regression?

a. [0, 1]

b. (-∞, ∞)

c. [0, 0.5]

d. [-1, 1]

Maximum Likelihood Estimates

Maximum Likelihood Estimation (MLE) is used to estimate parameters of model by finding values that maximize likelihood of observing given data.

Significance :The **likelihood function** assumes that outcome variable (Y) follows a Bernoulli distribution, which is a probability distribution for binary outcomes (0 or 1).

Formula : The likelihood function measures how likely it is to observe the actual outcomes given the predicted probabilities.
The likelihood function for logistic regression is given by: $L(\beta) = \prod_{i=1}^{n} \left(p_i^{y_i} \cdot (1 - p_i)^{1-y_i}\right)$

- L(β) is the likelihood function. β represents the vector of coefficients (parameters) we want to estimate.
- n is the number of observations (data points). p_i is the predicted probability of the event occurring for the ith observation.
- y_i is the actual outcome for the ith observation (0 or 1).

Log-Likelihood Function : The log-likelihood function is often used to simplify calculations and work with sums instead of products. It's the natural logarithm of the likelihood function:

$$\ln L(\beta) = \sum_{i=1}^{n} (y_i \cdot \ln(p_i) + (1 - y_i) \cdot \ln(1 - p_i))$$

Maximizing the log-likelihood is equivalent to maximizing the likelihood function.

Interpretation of Maximum Likelihood Estimates

Finding Maximum Likelihood Estimation (MLE)

- To find MLE estimates for β, optimization techniques such as gradient descent or numerical optimization algorithms are used.
- These algorithms iteratively adjust parameter values until they converge to the values that maximize the log-likelihood.

Interpretation of MLE Coefficients

The coefficients (β) represent the model's best fit to the observed data. These coefficients tell us:

- The direction (positive or negative) of the relationship between each predictor variable and log-odds of the event occurring.
- The magnitude of the impact of each predictor variable on the log-odds.

For example :

- If β_1 is positive, as the predictor variable increases, log-odds of the event occurring increase and vice versa.

Demo : Maximum Likelihood Estimates

Interpretation of Beta Coefficients

Direction and Magnitude of Coefficient

The sign of a beta coefficient indicates direction of the relationship between the predictor variable and log-odds of an event occurring:

- A positive coefficient implies a positive effect on the probability of the event, and a negative coefficient implies a negative effect.

The magnitude of a beta coefficient is directly proportional to the strength of the relationship between the predictor variable and the log odds.

Confidence Intervals

- **Confidence intervals** provide a range of values within which we are reasonably confident that the true coefficient or odds ratio lies.

- A wide confidence interval indicates uncertainty about an estimate, while a narrow interval suggests greater confidence.

How to Build a Logistic Regression Model?

Building a logistic regression model involves several steps, from data preparation to model evaluation.

Data Collection and Pre-processing

Exploratory Data Analysis

Feature Selection and Engineering

Model Training (Parameter Estimation)

Model Evaluation

Model Fine Tuning (To avoid overfitting)

Model Interpretation

Pop Quiz

Q. In MLE, what does the likelihood function measure?

a. The probability of observing the data given the parameters

b. The probability of finding the maximum likelihood estimate

c. The distribution of the data

a. The difference between observed and predicted values

Pop Quiz

Q. In MLE, what does the likelihood function measure?

- **a. The probability of observing the data given the parameters**
- b. The probability of finding the maximum likelihood estimate
- c. The distribution of the data
- d. The difference between observed and predicted values

Q&A
?
?
?
?
?
?
?

Activity 1

Pre-requisites:

Familiarity with confusion matrix and calculations of classification evaluation metrics (e.g., accuracy, precision, recall, F1-score).

Scenario:

You are a data scientist working for a healthcare analytics company, and you have developed a machine learning model to predict whether a patient has a specific medical condition based on various health attributes. In this real-world scenario, you will evaluate the model's performance using a confusion matrix and compute different classification metrics.

Data:

You have a dataset of 200 patient records with the following information:
Actual Medical Condition :
Positive Cases (Patients with the medical condition): 60,
Negative Cases (Patients without the medical condition): 140
Model Predictions :
True Positives (TP): 45
True Negatives (TN): 120
False Positives (FP): 10
False Negatives (FN): 25

Activity 1

Expected Outcome:

Calculate and interpret various classification metrics, including accuracy, precision, recall, and F1-score, using the provided confusion matrix. These metrics will help you assess the model's performance in identifying patients with the medical condition and making informed decisions about its suitability for clinical use.

Steps:

1. Calculate Accuracy: Measure the proportion of correct predictions.
2. Calculate Precision: Assess the model's ability to avoid false positives.
3. Calculate Recall: Evaluate the model's ability to capture true positives.
4. Calculate F1-Score: Find the harmonic mean of precision and recall.
5. Interpret Results: Analyze metrics for model performance.
6. Make Informed Decisions: Decide model suitability for real-world use in healthcare.

Activity 2

Pre-requisites:

Familiarity with logistic regression and understanding of interpreting coefficient values in linear models.

Scenario:

You are a data scientist working for a financial institution, and you have developed a logistic regression model to predict whether a credit card application will be approved or denied based on a single independent variable, the applicant's credit score. In this real-world scenario, you will interpret the provided beta coefficient for the credit score variable to understand its impact on credit card approval decisions.

Data:

You have a dataset of 200 credit card applications with the following information:

- Credit Card Approval:
    - Approved Applications: 120
    - Denied Applications: 80
- Independent Variable:
    - Credit Scores (ranging from 300 to 850)
- Model Coefficient (Beta):
    - Beta Coefficient for Credit Score (Given): β_1 = 0.02, which means that for each one-unit increase in credit score, the log odds of credit card approval increase by 0.02.

Activity 2

Expected Outcome:

Interpret the given beta coefficient (β_1) for the credit score variable in logistic regression and understand how it influences the odds of credit card approval.

Steps:

1. Interpret Beta Coefficient: Understand its meaning in terms of log-odds change.
2. Calculate Odds Ratio: Use the beta coefficient to compute the odds ratio. (OR = $e^{\beta^{i}}$)
3. Interpret Odds Ratio: Understand how a one-unit change impacts the odds.
4. Conclusion: Discuss practical implications for credit card approval decisions.

Summary

✓ Classification is a branch of supervised machine learning that helps in categorizing data into predefined classes or categories based on certain features.

✓ Metrics like accuracy, precision, recall, F1-score, and ROC AUC help assess the model's effectiveness in making correct predictions.

✓ Logistic regression is a statistical method used for binary and multi-class classification.
It involves modeling the probability of an event occurring using the sigmoid function.

✓ Beta coefficients in logistic regression represent the change in the log-odds of the dependent variable for a one-unit change in the predictor variable.

Next Session:
Logistic Regression – Case Study

THANK YOU!

Please complete your assessments and review the self-learning content for this session on the **PRISM** portal.

knowledgehut
upGrad

knowledgehut
upGrad

Case Study on Logistic Regression

Pre-requisites

Hope you have gone through the self-learning content for this session on the PRISM portal.

By the End of this Session, You Will:

- Learn the significance of logistic regression in making predictions about real world problems
- Explore real world data using pre-processing steps before creating a logistic regression model
- Conduct extensive exploratory data analysis before creating the model and writing its conclusions
- Find best performing model by comparing model evaluation and performance metrics

Recap

Poll Time

Q. What is the primary purpose of logistic regression?

a. To predict continuous numerical values

b. To classify data into two or more discrete categories

c. To calculate probabilities of events

d. To perform feature selection

Poll Time

Q. What is the primary purpose of logistic regression?

a. To predict continuous numerical values

b. To classify data into two or more discrete categories

c. To calculate probabilities of events

d. To perform feature selection

Case Study on Logistic Regression

Case Study – Problem Statement

Problem Statement

- One of the prestigious banks helps customers with their financial needs. It maintains a database of the customer and their details.
- Recently, the bank has observed that there have been a lot of customers ending their financial relationship with the bank. This phenomenon in any business is called customer churn.
- This is leading to a lot of loss for the bank due to the end of the existing financial business of customers with the bank.
- The bank needs your help to reduce the churn rate of the customers by seeing the financial habits of the customers.
- They want you to create a classification model using Logistic Regression to predict if the customer will churn or not.
- The dataset comprises several columns, including credit score, salary, age, gender, tenure, etc., and information on 1000+ customers.

Areas to Focus

Exploratory data Analysis

Outlier treatment, missing value treatment, feature scaling, etc

Mathematics of Logistic Regression

Geometric intuition, sigmoid function etc

Model Evaluation

Confusion matrix, accuracy, ROC-AUC score, F1 score etc

Poll Time

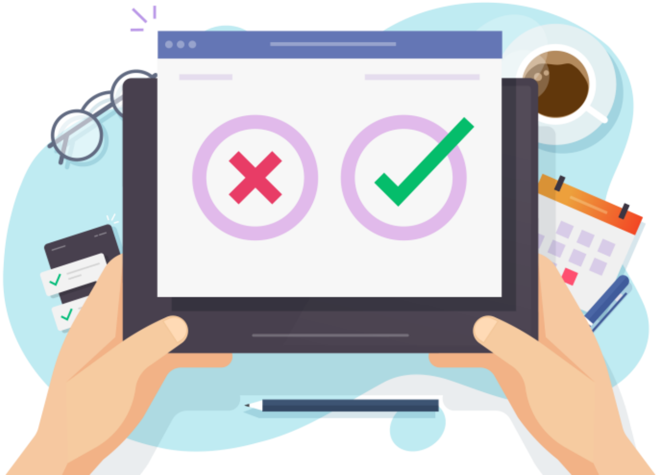
Q. What does the Receiver Operating Characteristic (ROC) curve visualize?

a. The relationship between precision and recall

b. The trade-off between a true positive rate and a false positive rate

c. The distribution of feature importance in the model

d. The accuracy of the model's predictions



Poll Time

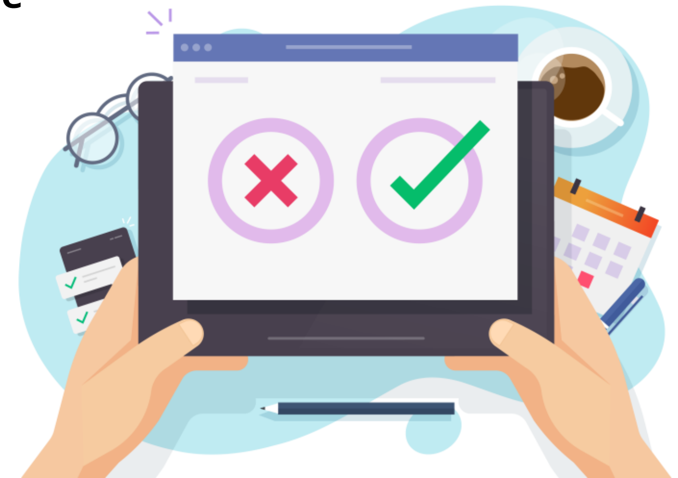
Q. What does the Receiver Operating Characteristic (ROC) curve visualize?

a. The relationship between precision and recall

🟩 **b. The trade-off between a true positive rate and a false positive rate**

c. The distribution of feature importance in the model

d. The accuracy of the model's predictions



Q&A

Hands-on: Case Study Questions

Poll Time

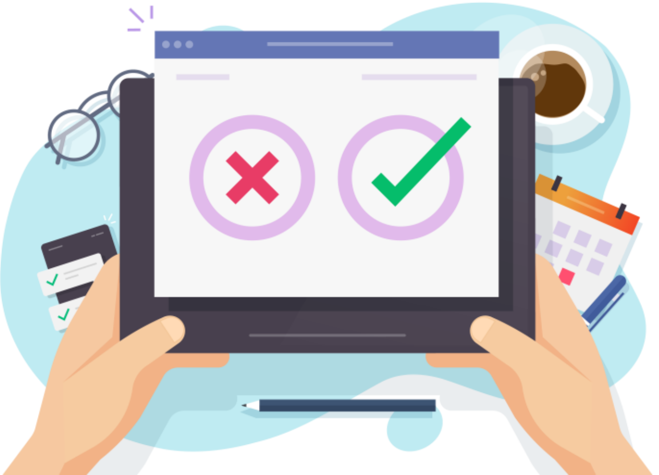
Q. In a binary classification problem, what does True Positive (TP) represent?

a. Predicted positive cases that are actually positive

b. Predicted positive cases that are actually negative

c. Predicted negative cases that are actually positive

d. Predicted negative cases that are actually negative



Poll Time

Q. In a binary classification problem, what does True Positive (TP) represent?

- **a. Predicted positive cases that are actually positive**
- b. Predicted positive cases that are actually negative
- c. Predicted negative cases that are actually positive
- d. Predicted negative cases that are actually negative

Q&A

Activity 1

Pre-requisites:

Familiarity with Python pandas library and logistic regression concepts.

Scenario:

You work in the HR department of a large company, and you're tasked with addressing the issue of employee attrition. Your goal is to build a logistic regression model to predict whether an employee is likely to leave the company based on a single independent variable: their satisfaction level. By identifying employees at risk of attrition, the company can implement retention strategies to reduce turnover.

Data:

```
import pandas as pd
df = pd.DataFrame({'EmployeeID': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
'SatisfactionLevel': [0.75, 0.62, 0.45, 0.82, 0.37, 0.90, 0.60, 0.78, 0.55, 0.42],
 'Attrition': [0, 1, 0, 0, 1, 0, 1, 0, 1, 0]})
```

Activity 1

Expected Outcome:

Built a logistic regression model using the dataset. Evaluated the model's performance using key metrics such as accuracy, precision, recall, and ROC-AUC.
Gained insights into employees at risk of attrition based on their satisfaction levels.

Steps:

1. Load and explore the dataset
2. Split the dataset into the feature (SatisfactionLevel) and target (Attrition) variables.
3. Build a logistic regression model using `LogisticRegression’ class from scikit-learn.
4. Calculate key evaluation metrics: accuracy, precision, recall, ROC-AUC.
5. Interpret the metrics to understand the model's performance in identifying employees at risk of attrition based on their satisfaction levels.

Activity 2

Pre-requisites :

Familiarity with the python pandas library and logistic regression evaluation metrics.

Scenario :

You are working for a marketing department at a subscription-based online streaming service. Your challenge is to build a logistic regression model to predict whether a website visitor will subscribe to a premium membership based on their browsing behavior. By identifying potential subscribers, the company can tailor marketing efforts to increase conversion rates.

Data:

```
import pandas as pd
df = pd.DataFrame({'VisitorID': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], 'TimeSpent(minutes)': [12, 25, 8, 43, 56, 32, 18, 27, 5,
38],
'PagesViewed': [2, 4, 1, 6, 8, 3, 2, 5, 1, 7], ‘NumberVideoPreviewsWatched': [0, 1, 0, 2, 3, 1, 0, 2, 0, 4], 'Subscription':
[0, 1, 0, 1, 1, 0, 0, 1, 0, 1]}).
```

Activity 2

Expected Outcome:

- Built a logistic regression model using visitor behavior metrics as independent variables.
- Evaluated the model's performance using key metrics such as accuracy, precision, recall, and ROC-AUC.
- Gained insights into potential subscribers based on browsing behavior.

Steps:

1. Load and explore the dataset
2. Use visitor behavior metrics (time spent, pages viewed, video previews) as independent variables and Subscription as target variable (where 1 means subscribed, 0 means not subscribed)
3. Build a logistic regression model using `LogisticRegression’ class from scikit-learn.
4. Calculate key evaluation metrics : accuracy, precision, recall, ROC-AUC.
5. Interpret the metrics to understand the model's performance in identifying potential subscribers based on browsing behavior.

Summary

- ✓ Logistic regression is used for binary classification problems, where the goal is to predict one of two possible outcomes, such as yes/no or true/false.
- ✓ Logistic regression helps in identifying key variables impacting the variable we are trying to predict while making accurate classification.
- ✓ The coefficients of logistic regression indicate the strength and direction of the relationship between input variables and the probability of the outcome. Positive coefficients increase the odds of the event, while negative coefficients decrease them.
- ✓ Evaluation metrics provide insights into how well the logistic regression model is performing and help you make informed decisions about its effectiveness.

Session Feedback

Next Session:
Introduction to K Nearest Neighbors (KNN) Algorithm

THANK YOU!

Please complete your assessments and review the self-learning content for this session on the **PRISM** portal.

knowledgehut
upGrad