FORMULA BOOK

(A Comprehensive guide for Statistical Mathematics in ML & DL)

1. Descriptive Statistics: Measures and Formulas

Descriptive statistics are essential tools for summarizing and understanding data. They provide insights into central tendencies, variability, and the shape of data distributions. Here's an in-depth explanation of key descriptive statistics measures along with their formulas and the meanings of each term:

1.1. Range:

Explanation: Range measures the difference between the maximum and minimum values in a dataset.

Formula: $(\text{Range} = \text{Max} - \text{Min})$

1.2. Mean $(\bar{x})$:

Explanation: The mean (average) is the sum of all data values divided by the total number of values. It represents the typical value of the dataset.

Formula: $(\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n})$

- (x_i): Each individual data value
- (n): Total number of data values

1.3. Median:

Explanation: The median is the middle value in a sorted dataset. It's less affected by extreme values compared to the mean.

Formula $(\boldsymbol{Odd(n)})$: $\text{Median} = \text{Value of } \frac{n+1}{2}\text{-th observation}$

Formula $(\boldsymbol{Even(n)})$: $\text{Median} = \frac{\text{Average of } \frac{n}{2}\text{-th and } \frac{n}{2}+1\text{-th observations}}{2}$

- (n): Total number of data values

1.4. Mode:

Explanation: The mode is the value that appears most frequently in the dataset.

Formula: Mode is the value with the highest frequency.

1.5. Interquartile Range (IQR):

Explanation: IQR measures the spread of the middle 50% of data, providing a robust measure of variability.

Formula: $(\text{IQR} = Q3 - Q1)$

Where:

- $(Q3)$: Third quartile (75th percentile)
- $(Q1)$: First quartile (25th percentile)

1.6. Variance (σ^2):

Explanation: Variance measures how much data values differ from the mean. It quantifies the spread of data.

Population Formula: $(\text{Population Variance} = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N})$

Sample Formula: $(\text{Sample Variance} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1})$

- (x_i): Each individual data value
- (μ): Population mean
- $(\bar{x})$: Sample mean
- (N): Total number of data values in population
- (n): Total number of data values in sample

1.7. Standard Deviation (σ):

Explanation: Standard deviation is the square root of the variance. It indicates the average distance of data points from the mean.

Population Formula: $(\text{Population SD} = \sqrt{\text{Population Variance}})$

Sample Formula: $(\text{Sample SD} = \sqrt{\text{Sample Variance}})$

- $\text{Variance}(\boldsymbol{\sigma^2})$: Variance of the dataset

1.8. Skewness:

Explanation: Skewness measures the asymmetry of the data distribution.

Formula:

Pearson's First Coefficient of Skewness: $(\text{Skewness} = \frac{3\times(\text{Mean}-\text{Median})}{\text{Standard Deviation}})$

- (Mean): Mean of the dataset
- (Median): Median of the dataset
- $(\text{Standard Deviation})$: Standard deviation of the dataset

Fisher-Pearson Coefficient of Skewness: $\left(\text{Skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}^3}\right)$

- (x_i): Each individual data value
- $(\bar{x})$: Mean of the dataset
- (n): Total number of data values

Types of Skewness:

- **Positive Skewness:** When the tail on the right side of the distribution is longer or fatter, indicating more high values. (Skewness > 0)
- **Negative Skewness:** When the tail on the left side of the distribution is longer or fatter, indicating more low values. (Skewness < 0)

1.9. Kurtosis:

Explanation: Kurtosis measures the shape of the data distribution. High kurtosis indicates more extreme values.

Formula: $\left(\text{Kurtosis} = \frac{\sum_{i=1}^{n}(x_i-\text{Mean})^4}{n\times\text{Standard Deviation}^4}\right)$

- (x_i): Each individual data values
- (Mean): Mean of the dataset
- (n): Total number of data values
- $(\text{Standard Deviation})$: Standard deviation of the dataset

Types of Kurtosis:

- **Leptokurtic:** Distribution with higher peak and heavier tails than a normal distribution. (Kurtosis > 3)
- **Mesokurtic:** Distribution with similar shape as a normal distribution. (Kurtosis = 3)
- **Platykurtic:** Distribution with lower peak and lighter tails than a normal distribution. (Kurtosis < 3)

1.10. Coefficient of Variation (CV):

Explanation: The coefficient of variation (CV) measures the relative variability of data by comparing the standard deviation to the mean. It helps assess the variation in relation to the mean.

Formula: $\left(\text{CV} = \frac{\text{Standard Deviation}}{\text{Mean}}\right)$

- $(\text{Standard Deviation})$: $(\sigma)(Sigma)$
- (Mean): $(\mu)(Mu)$

1.11. Correlation (r):

Explanation: Correlation measures the strength and direction of a linear relationship between two variables. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation).

Formula: $\text{Correlation} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)\times\sigma_x\times\sigma_y}$

- (x_i): Each individual data value of variable (x)
- $(\bar{x})$: Mean of variable (x)
- (y_i): Each individual data value of variable (y)
- $(\bar{y})$: Mean of variable (y)
- (n): Total number of data values
- (σ_x): Standard Deviation of variable (x)
- (σ_y): Standard Deviation of variable (y)

1.12. Covariance $(\text{Cov}(x, y))$:

Explanation: Covariance measures the extent to which two variables change together. A positive covariance indicates they tend to increase together, while a negative covariance suggests one decreases as the other increases.

Population Formula: $\left(\text{Population Covariance} = \frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{N}\right)$

Sample Formula: $\left(\text{Sample Covariance} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}\right)$

- (x_i): Each individual data value of variable (x)
- (y_i): Each individual data value of variable (y)
- (μ_x): Population mean of variable (x)
- (μ_y): Population mean of variable (y)
- $(\bar{x})$: Sample mean of variable (x)
- $(\bar{y})$: Sample mean of variable (y)
- (N): Total number of data values in population
- (n): Total number of data values in sample

2. Inferential Statistics: Principles and Concepts

Inferential statistics involves drawing conclusions about a population based on sample data. These techniques help us make predictions, test hypotheses, and make informed decisions about larger populations. Here's an exploration of the key principles and concepts of inferential statistics:

2.1. Population and Sample:

Explanation: A population includes all individuals or items of interest, while a sample is a subset of the population. Inferential statistics aims to make inferences about the population using information from a sample.

- Population: The entire group under consideration.
- Sample: A representative subset of the population.

2.2. Sampling Methods:

Explanation: Different sampling methods are used to select samples from populations. Common methods include random sampling, stratified sampling, and cluster sampling.

- Random Sampling: Every individual in the population has an equal chance of being selected.
- Stratified Sampling: Dividing the population into subgroups (strata) and then selecting samples from each stratum.
- Cluster Sampling: Dividing the population into clusters and then randomly selecting clusters for sampling.

2.3. Central Limit Theorem:

Explanation: The Central Limit Theorem states that regardless of the population distribution, the distribution of the sample means will tend to be normal for a sufficiently large sample size.

Implication: This theorem is the basis for many inferential techniques as it allows us to make assumptions about sample means.

Formula (Central Limit Theorem): If (n) is sufficiently large, the distribution of sample means $(\bar{x})$ is approximately normal with mean (μ) and standard deviation $(\frac{\sigma}{\sqrt{n}})$.

- (n): Sample size
- $(\bar{x})$: Sample mean
- (μ): Population mean
- (σ): Population standard deviation

2.4. Hypothesis Testing:

Explanation: Hypothesis testing involves making a statistical decision about a population parameter based on sample data. It helps us determine if an observed effect is statistically significant.

- Null Hypothesis (H_0): The default assumption that there is no significant difference or effect.
- Alternative Hypothesis (H_1) or (H_a): The hypothesis that contradicts the null hypothesis.

Type I and Type II Errors:

- Type I Error (False Positive): Incorrectly rejecting a true null hypothesis.
- Type II Error (False Negative): Failing to reject a false null hypothesis.

2.5. P-Value:

Explanation: The p-value is a measure of the evidence against the null hypothesis. It helps us decide whether to reject the null hypothesis.

Interpretation: A lower p-value indicates stronger evidence against the null hypothesis.

- $p<\alpha$: Reject the null hypothesis (Evidence supports the alternative hypothesis).
- $p\geq\alpha$: Fail to reject the null hypothesis (Insufficient evidence to support the alternative hypothesis).

$\alpha=0.05$: This is a widely used significance level, corresponding to a 5% chance of making a Type I error.

2.6. Confidence Intervals:

Explanation: A confidence interval is a range of values around a sample statistic that is likely to contain the population parameter. It provides a measure of the precision of an estimate.

Interpretation: A 95% confidence interval, for example, means that if we repeatedly sample and calculate confidence intervals, we expect about 95% of those intervals to contain the true population parameter.

Formula (Confidence Interval for Mean): $\bar{x} \pm Z \times \frac{s}{\sqrt{n}}$

- $(\bar{x})$: Sample mean
- (Z): Z-score based on desired confidence level
- (s): Sample standard deviation
- (n): Sample size

2.7. Z-Test and T-Test:

Explanation: Z-test and t-test are used to compare a sample mean to a known population mean (Z-test) or to compare two sample means (T-test).

- Z-Test: Used when the population standard deviation is known.
- T-Test: Used when the population standard deviation is unknown and must be estimated from the sample.

Formula (T-Test): $(t = \frac{\bar{x}-\mu}{s/\sqrt{n}})$

- $(\bar{x})$: Sample mean
- (μ): Population mean
- (s): Sample standard deviation
- (n): Sample size

2.8. Degrees of Freedom:

Explanation: Degrees of freedom represent the number of values in the final calculation of a statistic that are free to vary.

Formula (Degrees of Freedom for T-Test): $n - 1$

- (n): Sample size

3. Sum of Squares and Adjusted R-squared: Evaluating Variability and Model Fit

Sum of Squares is a pivotal concept in statistics that assists in comprehending data distribution, and it plays a key role in assessing the goodness of fit of regression models. This section delves into various types of sum of squares, introduces R-squared, and further discusses the significance of adjusted R-squared in enhancing model evaluation.

3.1. Total Sum of Squares (SST):

Explanation: Total Sum of Squares measures the total variability in the dependent variable.

Formula: $(\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2)$

- (y_i): Each individual observed dependent variable value
- $(\bar{y})$: Mean of the dependent variable

3.2. Regression Sum of Squares (SSR):

Explanation: Regression Sum of Squares represents the variability explained by the regression model.

Formula: $(\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2)$

- $(\hat{y}_i)$: Predicted value of the dependent variable from the regression model

- $(\bar{y})$: Mean of the dependent variable

3.3. Error Sum of Squares (SSE):

Explanation: Error Sum of Squares measures the unexplained variability in the dependent variable.

Formula: $(\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2)$

- (y_i): Each individual observed dependent variable value
- $(\hat{y}_i)$: Predicted value of the dependent variable from the regression model

3.4. Coefficient of Determination (R-squared):

Explanation: R-squared assesses the proportion of variability in the dependent variable explained by the regression model.

Formula: $(R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}})$

Interpretation:

- $(R^2 = 0)$: The regression model explains none of the variability.
- $(R^2 = 1)$: The regression model explains all the variability.

3.5. Adjusted R-squared:

Explanation: Adjusted R-squared accounts for the number of predictors in the model, preventing overfitting by penalizing the addition of irrelevant predictors.

Formula: $(\text{Adjusted } R^2 = 1 - \frac{\frac{\text{SSE}}{n-k-1}}{\frac{\text{SST}}{n-1}})$

- (n): Total number of data points
- (k): Number of predictors in the model

3.6. F-statistic:

Explanation: The F-statistic tests the overall significance of the model. A higher F-statistic suggests the model's coefficients are jointly significant.

Formula (F-statistic): $(F = \frac{\text{MSR}}{\text{MSE}})$

- (MSR): Mean Square Regression (SSR divided by degrees of freedom)
- (MSE): Mean Square Error (SSE divided by degrees of freedom)

04. Regression Analysis: Exploring Relationships and Interpreting Results

Regression analysis is a powerful statistical technique that enables us to model relationships between variables. It plays a crucial role in understanding how changes in one variable can influence another. This section provides an in-depth exploration of both simple linear regression and multiple linear regression, covering interpretation of coefficients and results.

4.1. Simple Linear Regression:

Explanation: Simple linear regression models the relationship between a dependent variable and a single independent variable. It assumes a linear relationship between the variables.

Formula: $y = \beta_0 + \beta_1 x + \varepsilon$

- (y): Dependent variable
- (x): Independent variable
- (β_0): Intercept term, the expected value of (y) when ($x = 0$)
- (β_1): Slope coefficient, the change in (y) for a one-unit change in (x)
- (ε): Error term

4.2. Multiple Linear Regression:

Explanation: Multiple linear regression models the relationship between a dependent variable and multiple independent variables. It extends simple linear regression to consider multiple predictors.

Formula: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$

- (y): Dependent variable
- ($x_1, x_2, \ldots, x_n$): Independent variables
- ($\beta_0, \beta_1, \beta_2, \ldots, \beta_n$): Coefficients of the independent variables
- (ε): Error term

4.3. Ordinary Least Squares (OLS):

Explanation: OLS is a method to estimate the coefficients that minimize the sum of squared differences between observed and predicted values.

Formula: $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_n} = \operatorname{argmin} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)\right)^2$

- ($\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_n}$): Estimated coefficients
- (n): Number of observations

4.4. Standard Error:

Explanation: The standard error quantifies the accuracy of coefficient estimates in the regression model.

Formula: Standard Error $(SE) = (\frac{\text{Residual standard deviation}}{\sqrt{n}})$

Where:

- Residual standard deviation: Measures the spread of residuals (observed - predicted values)
- (n): Number of observations

4.5. T-Statistic:

Explanation: The t-statistic assesses the significance of individual coefficients in the regression model.

Formula: $(t = \frac{\text{Coefficient Estimate}}{\text{Standard Error}})$

- Coefficient Estimate: The estimated coefficient for a variable.
- Standard Error: The standard deviation of the coefficient estimate.

5. Loss Function: Measuring Discrepancy in Regression

In the context of regression analysis, a loss function is a crucial concept that quantifies the discrepancy between predicted values and actual target values. It guides the model towards finding the best-fitting parameters by minimizing this discrepancy.

5.1. Mean Squared Error (MSE):

Explanation: MSE is one of the most common loss functions in regression. It calculates the average squared difference between predicted and actual values.

Formula: $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- (n): Total number of data points
- (y_i): Actual target value of the (i)-th data point
- $(\hat{y}_i)$: Predicted value by the model for the (i)-th data point

5.2. Root Mean Squared Error (RMSE):

Explanation: RMSE is the square root of MSE and provides a measure of the average prediction error in the same units as the target variable.

Formula: $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

5.3. Mean Absolute Error (MAE):

Explanation: MAE calculates the average absolute difference between predicted and actual values.

Formula: $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$

5.4. Huber Loss:

Explanation: Huber loss combines the characteristics of both MSE and MAE, providing a balance between robustness to outliers and smoothness.

Formula: $Hube\,\mathrm{r}(r) = \begin{cases} \frac{1}{2}(r)^2 & \text{if } |r| \le \delta \\ \delta\left(|r| - \frac{1}{2}\delta\right) & \text{otherwise} \end{cases}$

- $(r = y_i - \hat{y}_i)$
- (δ): A threshold parameter

06. Feature Selection Techniques: Enhancing Model Performance and Interpretability

Feature selection is a crucial step in machine learning and statistical analysis, aiming to choose a relevant subset of features from a larger set. This process helps improve model performance, reduce computational complexity, and enhance the interpretability of results.

6.1. FILTER METHODS:

6.1.1. Chi-Square Test:

Explanation: The chi-square test assesses the independence between categorical variables in a contingency table.

Formula: $\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

- (O_{ij}): Observed frequency in cell (i, j)
- (E_{ij}): Expected frequency in cell (i, j)

6.2. WRAPPER METHODS:

6.2.1. Backward Elimination:

Explanation: Backward elimination starts with all features and iteratively removes the least significant one based on p-values from the model until only significant features remain.

Process: Fit the model, find the feature with the highest p-value, remove it, refit the model, and repeat until all features are significant.

6.2.2. Forward Selection:

Explanation: Forward selection starts with an empty model and iteratively adds the most significant feature based on p-values until no more significant features can be added.

Process: Fit the model with one feature, add the feature with the lowest p-value, refit the model, and repeat until no more features are significant.

6.2.3. Recursive Feature Elimination (RFE):

Explanation: RFE recursively fits the model and removes the least significant feature based on coefficients or feature importance until the desired number of features is reached.

Process: Fit the model, identify the least important feature, remove it, refit the model, and repeat until the desired feature count is achieved.

6.3. EMBEDDED METHODS:

6.3.1. L1 Regularization (Lasso):

Explanation: Lasso adds a penalty term based on the absolute values of coefficients (w) to the cost function, encouraging sparsity in feature selection.

Formula: $J(w) = \text{Loss}(w) + \alpha||w||$

- $(\text{Loss}(w))$: Loss function
- (α): Regularization strength
- ($|w|$): L1 norm (sum of absolute values) of coefficients (w)

6.3.2. L2 Regularization (Ridge):

Explanation: Ridge adds a penalty term based on the squared values of coefficients (w) to the cost function, promoting small but non-zero coefficients.

Formula: $R(w) = \text{Loss}(w) + \alpha||w||^2$

- $(\text{Loss}(w))$: Loss function

- (α): Regularization strength
- ($|w|^2$): L2 norm (Euclidean norm) squared of coefficients (w)

6.3.3. Elastic Net:

Explanation: Elastic Net combines L1 and L2 regularization, providing a balance between feature selection (Lasso) and coefficient stability (Ridge).

Formula: $R(\omega) = \text{Loss}(\omega) + \alpha_1||\omega||^2 + \alpha_2||\omega||$

- ($\text{Loss}(\omega)$): Loss function
- (α_1): L2 regularization strength
- (α_2): L1 regularization strength
- ($|\omega|^2$): L2 norm (Euclidean norm) squared of coefficients (ω)
- ($|\omega|$): L1 norm (sum of absolute values) of coefficients (ω)

7. Gradient Descent:

Explanation: Gradient descent is an optimization algorithm used for minimizing the cost function in various machine learning models.

Formula: $\theta_j = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j}$

- (θ_j): Model parameter to update
- (α): Learning rate
- ($J(\theta)$): Cost function

7.1. Stochastic Gradient Descent (SGD):

Explanation: SGD is a variant of gradient descent that updates model parameters using only one training example at a time, making it faster for large datasets.

Formula: $\theta_j = \theta_j - \alpha \frac{\partial J(\theta, x_i, y_i)}{\partial \theta_j}$

- (x_i): Training example
- (y_i): True label of (x_i)

7.2. Batch Gradient Descent:

Explanation: Batch Gradient Descent is an optimization algorithm used to minimize the cost function by updating model parameters using the gradient of the entire training dataset.

Formula: $\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \frac{\partial J(\theta, x_i, y_i)}{\partial \theta_j}$

- (θ_j): Model parameter to update
- (α): Learning rate
- (m): Total number of training examples
- $\frac{1}{m}\sum_{i=1}^{m}\frac{\partial J(\theta,x_i,y_i)}{\partial \theta_j}$: Average gradient of the cost function over the entire dataset

8. Feature Scaling Techniques:

Feature scaling is a preprocessing step in machine learning that ensures all input features or variables are on a similar scale. Scaling features can improve the performance and convergence of various machine learning algorithms by reducing the impact of differing magnitudes among features.

8.1. Normalization:

Explanation: Normalization, also known as Min-Max scaling, transforms features to a common range between 0 and 1.

Formula: $(x_{\text{normalized}} = \frac{x-\min(x)}{\max(x)-\min(x)})$

- (x): Original feature value
- $(\min(x))$: Minimum value of the feature
- $(\max(x))$: Maximum value of the feature

8.2. Standardization:

Explanation: Standardization, also known as Z-score normalization, scales features to have a mean of 0 and a standard deviation of 1.

Formula: $(x_{\text{standardized}} = \frac{x-\text{mean}(x)}{\text{std}(x)})$

- (x): Original feature value
- $(\text{mean}(x))$: Mean of the feature
- $(\text{std}(x))$: Standard deviation of the feature

9. Dimensionality Reduction Techniques:

Dimensionality reduction techniques are used to reduce the number of features or variables in a dataset while maintaining as much of the original information as possible. These techniques are particularly useful when dealing with high-dimensional data to improve computation efficiency, visualization, and model performance.

9.1. Principal Component Analysis (PCA):

Explanation: PCA is a widely used dimensionality reduction technique that transforms data into a new coordinate system where the new dimensions, called principal components, capture the maximum variance present in the original data.

Process:

- Calculate the covariance matrix:

$$C = \frac{1}{n-1}\sum_{i=1}^{n} \begin{bmatrix} (x_i - \bar{x})(x_i - \bar{x}) & (x_i - \bar{x})(y_i - \bar{y}) \\ (y_i - \bar{y})(x_i - \bar{x}) & (y_i - \bar{y})(y_i - \bar{y}) \end{bmatrix}$$

- Characteristic Equation: $[\det(C - \lambda I) = 0]$
- Compute eigenvectors and eigenvalues: $(Cv = \lambda v)$
- Sort eigenvalues in descending order to determine significant principal components.
- Project data onto new coordinate system

$$[x_i' = v_1 \cdot (x_i - \bar{x}) + v_2 \cdot (y_i - \bar{y})]$$
$$[y_i' = v_1 \cdot (x_i - \bar{x}) + v_2 \cdot (y_i - \bar{y})]$$

Application: Reducing the number of features while preserving as much variance as possible.

10. Bias-Variance:

Bias-Variance trade-off is a fundamental concept in machine learning that aims to strike a balance between the model's ability to capture complex patterns (low bias) and its susceptibility to noise (low variance).

10.1. Overfitting:

Explanation: Overfitting occurs when a model captures noise and random fluctuations in the training data, resulting in poor performance on new, unseen data.

Characteristics: The model fits the training data extremely well but performs poorly on validation or test data due to its sensitivity to noise.

Solutions: Regularization techniques, reducing model complexity, increasing training data, and feature selection can mitigate overfitting.

10.2. Underfitting:

Explanation: Underfitting happens when a model is too simple to capture the underlying patterns in the data, leading to poor performance on both training and new data.

Characteristics: The model lacks the complexity to accurately represent the relationships between variables, resulting in high bias.

Solutions: Increasing model complexity, adding relevant features, and using more sophisticated algorithms can help address underfitting.

10.3. Bias:

Explanation: Bias refers to the error introduced by approximating a real-world problem, which may be complex, by a simplified model. High bias can lead to underfitting, where the model fails to capture underlying patterns in the data.

Implication: Models with high bias tend to oversimplify the problem and generalize poorly to new data.

10.4. Variance:

Explanation: Variance measures the model's sensitivity to small fluctuations in the training dataset. High variance can lead to overfitting, where the model captures noise in the training data.

Implication: Models with high variance fit the training data very closely but may fail to generalize to new, unseen data.

10.5. Bias-Variance Trade-off:

Explanation: The goal is to find a balance between bias and variance that results in the best predictive accuracy. Reducing bias may increase variance, and vice versa. The trade-off aims to minimize the model's overall error.

10.6. Cross-Validation:

Explanation: Cross-validation is a technique to assess a model's performance on unseen data by partitioning the dataset into training and validation sets multiple times.

Benefits: Cross-validation helps detect overfitting and provides a more accurate estimate of a model's generalization ability.

11. Non-Linear Regression:

Explanation: Nonlinear regression is a statistical method used to model relationships between variables when the relationship cannot be adequately described by a linear equation. In linear regression, the goal is to find a linear equation that best fits the data, whereas in nonlinear regression, the goal is to find a nonlinear function that provides the best fit.

11.1. Polynomial Regression:

Explanation: Polynomial regression is a type of nonlinear regression technique used to model relationships between variables using polynomial functions. In polynomial regression, the relationship between the independent variable (x) and the dependent variable(y) is approximated using a polynomial equation of a specified degree.

Formula: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \epsilon$

- (y) is the dependent variable you're trying to predict.
- (x) is the independent variable.
- $(\beta_0, \beta_1, \ldots, \beta_n)$ are the coefficients of the polynomial terms.
- (n) is the degree of the polynomial, indicating the highest power of (x) in the equation.
- (ϵ) represents the error term.

11.2. Support Vector Regressor:

Explanation: Support Vector Regression (SVR) is a supervised machine learning technique used for regression tasks. Its objective is to determine a regression function that effectively fits the data while considering a specified margin of error.

11.2.1 Hyper Plane:

Explanation: In SVR, the hyperplane represents the regression function that predicts target values based on input features. Unlike its role in classification, where hyperplanes separate classes, SVR's hyperplane closely aligns with data points within a defined margin.

Hyperplane: y = wx + b

Decision Boundary: wx + b = +a, wx + b= -a

subject to: -a < y - y_actual < a

- y: Predicted target value.
- x: Input features.
- w: Weight vector defining the hyperplane.
- b: Bias term or intercept.
- a: Allowable range around the hyperplane within which prediction errors are acceptable.

11.3. KNN Regressor:

Explanation: KNN Regressor is a supervised machine learning algorithm used for regression tasks. It works by finding the 'k' nearest data points (neighbours) to a

given input data point and then predicts the target value based on the average or weighted average of the target values of these nearest neighbours. The value of 'k' is a hyperparameter that needs to be defined beforehand.

11.4. Decision Tree Regressor:

Explanation: Decision Tree Regressor is another supervised learning algorithm used for regression tasks. It builds a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents a predicted output. The tree is constructed by recursively partitioning the data into subsets based on feature thresholds, aiming to minimize the variance of the target variable within each subset.

11.5. Random Forest Regressor:

Explanation: Random Forest Regressor is an ensemble learning algorithm that combines multiple Decision Trees to improve predictive accuracy and control overfitting. It builds a collection of decision trees by training on different subsets of the data and features. The final prediction is obtained by averaging the predictions of individual trees (for regression tasks).

12. Classification:

12.1. Classification Table:

Explanation: A classification table, also known as a confusion matrix or error matrix, is a fundamental tool for assessing the performance of a classification model. It visually summarizes the comparison between predicted and actual class labels for a dataset.

12.1.1. Confusion Matrix:

Explanation: A confusion matrix is a tabular representation that provides a detailed view of the performance of a classification model. It breaks down the predictions into four categories: true positives, true negatives, false positives, and false negatives.

		PREDICTED CLASS	
		Positive	Negative
ACTUAL CLASS		TP	FN
		FP	TN

- True Positive (TP): Instances correctly predicted as positive.
- True Negative (TN): Instances correctly predicted as negative.
- False Positive (FP): Instances incorrectly predicted as positive (Type I error).
- False Negative (FN): Instances incorrectly predicted as negative (Type II error).

12.1.2. Precision:

Explanation: Precision is a performance metric that quantifies the accuracy of positive predictions made by a model. It indicates how well the model avoids false positive errors.

Formula: $\text{Precision} = \frac{TP}{TP+FP}$

12.1.3. Recall (Sensitivity):

Explanation: Recall, also known as sensitivity or true positive rate, assesses the model's ability to correctly identify positive instances from the total actual positive instances.

Formula: $\text{Recall} = \frac{TP}{TP+FN}$

12.1.4. Specificity:

Explanation: Specificity measures the model's ability to correctly identify negative instances out of the total actual negative instances.

Formula: $\text{Specificity} = \frac{TN}{TN+FP}$

12.1.5. F1-Score:

Explanation: The F1-score is a balanced metric that combines both precision and recall into a single value. It is especially useful when the class distribution is imbalanced.

Formula: $\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$

12.2. Logistic Regression:

Explanation: Logistic Regression is a widely used statistical method for binary classification tasks, where the goal is to predict a categorical outcome that has two classes (e.g., yes/no, 0/1, true/false). Despite its name, logistic regression is actually a classification algorithm, not a regression algorithm.

12.2.1. Sigmoid Function:

Explanation: The Sigmoid function is a commonly used activation function in machine learning and neural networks. It's valued between 0 and 1 and is often employed to introduce non-linearity to models. The Sigmoid function's characteristic S-shaped curve allows it to map input values to a probability-like output.

Formula: $\sigma(x) = \frac{1}{1+e^{-x}}$

- σ(x): The output value of the Sigmoid function for input x.
- e: The base of the natural logarithm (approximately 2.71828).
- x: The input value to the Sigmoid function.

12.3. Support Vector Classifier:

Explanation: It's a type of Support Vector Machine (SVM) that focuses on classifying data points into different categories or classes based on their features. The goal of SVC is to find a hyperplane that best separates the data into distinct classes while maximizing the margin between the classes.

12.3.1. Optimization Function:

Explanation: The optimization function of a Support Vector Classifier (SVC) aims to find the parameters of the decision boundary that best separates the data while minimizing the classification error and controlling the complexity of the model.

Minimize: $\frac{1}{2} \cdot |\mathbf{w}|^2 + C \cdot \sum_{i=1}^{N} \xi_i$

Subject to $(y_i \cdot (\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1 - \xi_i)$

$(\xi_i \geq 0)$

for all $(i = 1, \ldots, N)$, where (N) is the number of data points.

- $(\mathbf{w})$ is the weight vector.
- $(|\mathbf{w}|^2)$ is the squared L2 norm of the weight vector, controlling the complexity of the model.
- (C) is the regularization parameter that balances the trade-off between maximizing the margin and minimizing the classification error.
- (y_i) is the label of the (i)-th data point.
- $(\mathbf{x_i})$ is the feature vector of the (i)-th data point.
- (b) is the bias term.
- (ξ_i) is the slack variable associated with the (i)-th data point, allowing for some misclassification.

12.3.2. Kernel Function:

Explanation: A kernel function is used to implicitly map the input data points into a higher-dimensional feature space.

12.4. KNN Classifier:

Explanation: The k-Nearest Neighbours (KNN) algorithm is a simple and intuitive classification algorithm used in machine learning for both binary and multiclass classification tasks. It's a type of instance-based learning where the algorithm classifies a new data point based on the class of its nearest neighbors in the feature space.

12.4.1. Euclidean Distance:

Explanation: The Euclidean distance is a measure of the straight-line distance between two points in a two-dimensional space.

Formula: The Euclidean distance between two points (x_1, y_1) and (x_2, y_2) is given by: $d = \sqrt{((x_2 - x_1)^2 + (y_2 - y_1)^2)}$

12.5. Decision Tree Classifier:

Explanation: A Decision Tree Classifier is a popular machine learning algorithm used for classification tasks. It's a tree-like structure where each internal node represents a decision based on the value of a specific feature, each branch represents an outcome of that decision, and each leaf node represents a class label.

12.6. Random Forest Classifier:

Explanation: A Random Forest Classifier is an ensemble learning algorithm that combines multiple Decision Tree classifiers to improve predictive accuracy and control overfitting. It works by creating a collection (or forest) of decision trees and making predictions by aggregating the predictions of individual trees.

12.7. Naive Bayes:

Explanation: Naive Bayes is a probabilistic classification algorithm that is based on the principles of probability and Bayes' theorem. It's particularly well-suited for text classification tasks and situations where the assumption of feature independence holds reasonably well.

12.7.1. Conditional Probability:

Explanation: Conditional probability is a fundamental concept in probability theory. It represents the probability of an event occurring given that another event has already occurred. Mathematically, the conditional probability of event A

occurring given that event B has occurred is denoted as $P(A|B)$ and is calculated as:

Formula: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

- $P(A \cap B)$ is the probability that both events A and B occur.
- $P(B)$ is the probability that event B occurs.

12.7.2. Bayes Theorem:

Explanation: Bayes' theorem is a mathematical formula that relates conditional probabilities. It is named after the Reverend Thomas Bayes. The formula can be stated as:

Formula: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

- $P(A|B)$ is the posterior probability of event A occurring given event B.
- $P(B|A)$ is the conditional probability of event B occurring given event A.
- $P(A)$ is the prior probability of event A occurring.
- $P(B)$ is the probability of event B occurring.

12.7.3. Gaussian Naive Bayes:

Explanation: Gaussian Naive Bayes is used for continuous numerical features. The likelihood of each feature for each class is assumed to follow a Gaussian distribution.

Given a class (C_k) and a feature (X_j) with a Gaussian distribution, the likelihood $P(X_j|C_k)$ can be calculated using the Gaussian probability density function:

$$P(X_j|C_k) = \frac{1}{\sqrt{2\pi\sigma_{jk}^2}} \exp\left(-\frac{(x_j - \mu_{jk})^2}{2\sigma_{jk}^2}\right)$$

- (x_j) is the value of the feature (X_j) for a specific data point.
- (μ_{jk}) is the mean of feature (X_j) for class (C_k).
- (σ_{jk}) is the standard deviation of feature (X_j) for class (C_k).

12.7.4. Multinomial Naive Bayes:

Explanation: Multinomial Naive Bayes is used for discrete features, typically in text classification where features represent word counts or frequencies.

Given a class (C_k) and a feature (X_j) representing the count of a word (w_j), the likelihood $P(X_j|C_k)$ follows a multinomial distribution:

$$P(X_j|C_k) = \frac{(n_{jk} + \alpha_j)}{(n_k + \alpha_j V)}$$

- (n_{jk}) is the count of word (w_j) in documents of class (C_k).
- (n_k) is the total count of all words in documents of class (C_k).
- (α_j) is a smoothing parameter (Laplace smoothing) to handle unseen words.
- (V) is the vocabulary size (total number of unique words).

12.7.5. Bernoulli Naive Bayes:

Explanation: Bernoulli Naive Bayes is used for binary features, often in text classification where features represent word presence or absence.

Given a class (C_k) and a binary feature (X_j) indicating the presence (1) or absence (0) of a word (w_j), the likelihood $(P(X_j|C_k))$ follows a Bernoulli distribution:

$$P(X_j|C_k) = P(X_j = 1|C_k)^{x_j} \cdot \left(1 - P(X_j = 1|C_k)\right)^{(1-x_j)}$$

- (x_j) is the binary value of feature (X_j) (0 or 1).
- $P(X_j = 1|C_k)$ is the probability of word (w_j) occurring in documents of class (C_k).