Machine Learning Quick Notes

PART 1: Supervised Learning Concepts

1 Test Data

Definition:

Test data is a subset of the dataset used to evaluate the performance of a trained machine learning model.

It checks how well the model generalizes to unseen data (data not used during training).

Example: If you have 1000 samples, you might use 80% for training and 20% for testing.

Purpose: To estimate the accuracy and reliability of the model.

2 Training Data

Definition:

Training data is the portion of the dataset used by the algorithm to learn patterns, relationships, and parameters.

The model uses this data to fit itself.

Example: For a linear regression, training data is used to find the best-fitting line between input (X) and output (Y).

3 RMSE (Root Mean Squared Error)

Definition:

RMSE measures the average magnitude of prediction errors — how far predictions are from actual values.

Formula:

$RMSE = \sqrt{((1/n) * \Sigma (y_i - \hat{y}_i)^2)}$

where

y_i = actual value

$\hat{y}_i$ = predicted value

n = number of samples

Lower RMSE = better model performance.

4 R^2 Score (Coefficient of Determination)

Definition:

R^2 score shows how much variance in the dependent variable (Y) is explained by the model.

Formula:

$R^2 = 1 - (\Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2)$

where

y_i = actual value

$\hat{y}_i$ = predicted value

$\bar{y}$ = mean of actual values

5 Linear Regression

Definition:

Linear Regression is a supervised learning algorithm used for predicting continuous values by finding a linear relationship between input (X) and output (Y).

Equation:

$Y = mX + c$

Goal: Minimize the error between predicted and actual values using methods like Least Squares.

6 Random Forest Regression

Definition:

Random Forest is an ensemble algorithm that builds multiple Decision Trees and averages their results to make predictions.

Steps:

1. Create many random subsets of the training data.

2. Train a Decision Tree on each subset.

3. Combine (average) their predictions for final output.

Advantages:

- Reduces overfitting

- More accurate than a single Decision Tree

- Works well with nonlinear data

7 Confusion Matrix

Definition:

A 2×2 table used to evaluate classification models.

Structure:

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Formulas:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

8 KNN (K-Nearest Neighbors) Classification

Definition:

KNN is a non-parametric classification algorithm that assigns a class to a data point based on the majority class among its k nearest neighbors.

Steps:

1. Choose the number of neighbors (k)

2. Calculate the distance (usually Euclidean) between the new point and all training points

3. Select the k nearest neighbors

4. Perform majority voting

5. Assign the most common class

Formula:

Euclidean Distance = $\sqrt{((x_1 - c_1)^2 + (x_2 - c_2)^2 + ... + (x_n - c_n)^2)}$

9️⃣ SVM (Support Vector Machine) Classification

Definition:

SVM is a supervised learning algorithm that finds the best decision boundary (hyperplane) to separate classes.

Steps:

1. Plot data points in n-dimensional space.

2. Find the hyperplane that maximizes the margin between classes.

3. These nearest points are called Support Vectors.

4. Use kernel functions (RBF, Polynomial) for non-linear separation.

5. Classify new points based on which side of the hyperplane they fall on.

PART 2: Evaluation Metrics and Clustering

1 Accuracy

Definition:

Accuracy measures the overall correctness of a classification model.

Formula:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

2 Error Rate

Definition:

Error Rate is the percentage of incorrect predictions made by the model.

Formula:

Error Rate = (FP + FN) / (TP + TN + FP + FN)

or

Error Rate = 1 - Accuracy

3 Precision

Definition:

Precision tells how many of the predicted positive values are actually correct.

Formula:

Precision = TP / (TP + FP)

4 Recall (Sensitivity)

Definition:

Recall measures how many actual positives the model correctly identified.

Formula:

Recall = TP / (TP + FN)

5 Normalization

Definition:

Normalization is a data preprocessing technique used to scale numeric features so that they have a uniform range.

Formula (Min-Max Normalization):

$X' = (X - Xmin) / (Xmax - Xmin)$

6 K-Means Clustering

Definition:

K-Means is an unsupervised algorithm used to group data into K clusters.

Steps:

1. Choose number of clusters K

2. Initialize K random centroids

3. Assign points to nearest centroid

4. Recalculate centroids

5. Repeat until centroids stabilize

Formula (Euclidean Distance):

$d = \sqrt{((x_1 - c_1)^2 + (x_2 - c_2)^2 + ... + (x_n - c_n)^2)}$

7 Elbow Method

Definition:

Used to find the optimal number of clusters (K) for K-Means.

Steps:

1. Run K-Means for different K values

2. Compute WCSS for each K

3. Plot K vs WCSS

4. The 'elbow point' where WCSS drops slowly indicates the best K

8 WCSS (Within-Cluster Sum of Squares)

Definition:

WCSS measures how close data points are to the centroid of their cluster.

Formula:

$WCSS = \Sigma_k \Sigma_{i \in C_k} (x_i - \mu_k)^2$

Machine Learning Quick Notes

PART 1: Supervised Learning Concepts

RMSE (Root Mean Squared Error)

Definition:

RMSE measures the average magnitude of prediction errors — how far predictions are from actual values.

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

R^2 Score (Coefficient of Determination)

Definition:

R^2 score shows how much variance in the dependent variable (Y) is explained by the model.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Linear Regression

Equation representing a linear relationship between input and output.

$$Y = mX + c$$

Confusion Matrix

Formulas for classification metrics based on TP, TN, FP, FN values.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

KNN (K-Nearest Neighbors)

KNN classifies a data point based on majority class among its nearest neighbors.

$$d = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \ldots + (x_n - c_n)^2}$$

PART 2: Evaluation Metrics and Clustering

Accuracy

Overall correctness of model predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Error Rate

Percentage of incorrect predictions.

$$Error\ Rate = \frac{FP + FN}{TP + TN + FP + FN} = 1 - Accuracy$$

Precision

How many of the predicted positives are actually correct.

$$Precision = \frac{TP}{TP + FP}$$

Recall

How many actual positives are correctly identified.

$$Recall = \frac{TP}{TP + FN}$$

Normalization

Scaling features to a common range (usually 0–1).

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

K-Means Clustering

Unsupervised algorithm that groups data into K clusters by minimizing distance to centroids.

$$d = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \ldots + (x_n - c_n)^2}$$

WCSS (Within-Cluster Sum of Squares)

Measures compactness of clusters (lower = better).

$$WCSS = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \mu_k)^2$$

R^2, RMSE, Accuracy Combined Summary

Summary of key regression and classification metrics.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

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

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$