DATA SCIENCE AND MACHINE LEARNING – COMPLETE NOTES

UNIT 1: INTRODUCTION TO DATA SCIENCE

What is Data Science?

Data Science is an interdisciplinary field that uses statistics, mathematics, programming, and domain knowledge to extract meaningful insights from structured and unstructured data. It involves data collection, cleaning, analysis, visualization, and model building to support decision-making.

Difference Between AI, Machine Learning, and Data Science

- **Artificial Intelligence (AI):** Broad field focused on creating systems that mimic human intelligence (reasoning, learning, decision-making).
- **Machine Learning (ML):** Subset of AI that enables systems to learn patterns from data without explicit programming.
- **Data Science:** Focuses on extracting insights from data using ML, statistics, and visualization.

Basic Introduction to Python

Python is a high-level, interpreted programming language widely used in data science due to its simplicity and extensive libraries.

Popular libraries: - NumPy – numerical computing - Pandas – data manipulation - Matplotlib, Seaborn – visualization - Scikit-learn – machine learning

Google Colab

Google Colab is a cloud-based Jupyter notebook environment. Features: - Free GPU/TPU - No installation required - Easy sharing

Popular Dataset Repositories

- Kaggle
- UCI Machine Learning Repository
- Google Dataset Search
- OpenML

UNIT 2: DATA PREPROCESSING AND EXPLORATION

Data Preprocessing

Data preprocessing converts raw data into a usable format. Steps include: - Handling missing values - Encoding categorical data - Scaling and normalization

Data Scaling

- **Min-Max Scaling:** Scales data between 0 and 1
- **Standardization:** Mean = 0, Std = 1

Similarity and Dissimilarity Measures

- Euclidean Distance
- Manhattan Distance
- Cosine Similarity

Sampling

Sampling selects a subset of data from a population. Types: - Random sampling - Stratified sampling

Quantization

Reduces the number of distinct values in data.

Filtering

Removes noise or irrelevant data.

Data Transformation and Merging

- Log transformation
- One-hot encoding
- Merge, join, concatenate datasets

Data Visualization

Common plots: - Histogram - Box plot - Scatter plot

Principal Component Analysis (PCA)

PCA reduces dimensionality while preserving variance.

Correlation

Measures relationship between variables. - Pearson - Spearman

Chi-Square Test

Used to test association between categorical variables.

UNIT 3: REGRESSION ANALYSIS

Regression Analysis

Regression models relationships between dependent and independent variables.

Linear Regression

Models linear relationship: Y = b0 + b1X

Generalized Linear Models (GLM)

Extends linear regression for non-normal data. Examples: - Logistic regression - Poisson regression

Regularized Regression

Controls overfitting. - Ridge Regression (L2) - Lasso Regression (L1)

Cross Validation

Technique to evaluate model performance. Common type: k-fold cross validation.

Training and Testing Dataset

- Training set: used to train model
- Testing set: used to evaluate model

Nonlinear Regression

Models nonlinear relationships.

Ridge Regression

Adds penalty term to reduce large coefficients.

Latent Variables

Hidden variables not directly observed.

Structural Equation Modelling (SEM)

Combines factor analysis and regression to model complex relationships.

UNIT 4: TIME SERIES AND FORECASTING

Forecasting

Predicting future values based on historical data.

Time Series Data

Data collected over time intervals. Components: - Trend - Seasonality - Cyclic - Noise

Stationarity

Statistical properties remain constant over time.

Seasonality

Repeating patterns at fixed intervals.

Autoregressive Models (AR)

Current value depends on past values.

Moving Average (MA)

Uses past errors for prediction.

ARIMA Model

Combination of AR, I, MA.

Recurrent Models

Neural networks for sequential data. Example: RNN, LSTM.

UNIT 5: CLASSIFICATION

Classification

Supervised learning where output is categorical.

Linear Discriminant Analysis (LDA)

Finds linear combinations that best separate classes.

Support Vector Machine (SVM)

Finds optimal hyperplane separating classes.

Decision Trees

Tree-based model using feature splits.

Key concepts: - Entropy - Information Gain - Gini Index

Advantages: - Easy to interpret - No scaling needed

UNIT 6: CLUSTERING

Clustering

Unsupervised learning to group similar data points.

Types of Clustering

- Partitioning
- Hierarchical
- Density-based
- Grid-based
- Model-based

K-Means Clustering

Partitions data into K clusters by minimizing variance.

Hierarchical Clustering

Creates dendrogram using bottom-up or top-down approach.

DBSCAN

Density-based clustering handling noise.

Grid-Based Clustering

Uses grid structures for fast clustering.

Model-Based Clustering

Assumes data generated from probabilistic models. Example: Gaussian Mixture Model.

Clustering Evaluation

- Silhouette Score
- Davies-Bouldin Index

FINAL NOTES

- Data preprocessing is critical for model performance
- Model selection depends on data type and problem
- Visualization helps in understanding patterns
- Evaluation metrics are essential for validation

END OF NOTES