DATA SCIENCE

1. Introduction to Data Science

(a) What is Data Science?

(b) Fundamental fields of study related Data Science

(c) Data Science and related Terminologies

(i) Big Data

(ii) Data Mining

(iii) Artificial Intelligence

(iv) Machine Learning

(v) Deep Learning

(d) Applications of Data Science

(e) Data Science Process Model

II. Data Collection and Cleaning

2.1 Data Sources

Different types of data sources: structured, unstructured, and semi-structured

Web scraping, API integration, and database connections

2.2 Data Cleaning and Preprocessing

Handling missing data

Outlier detection and treatment

Data normalization and standardization

2.3 Exploratory Data Analysis (EDA)

Descriptive statistics

Data visualization using Matplotlib and Seaborn

Identifying patterns and trends

2. Statistics

(I) Introduction to Statistics

- ✓ Data
- ✓ Meanings of variables
- ✓ Central Tendency
- ✓ Measure of Location and Dispersion
- ✓ Skewness and Kurtosis

(II) Data and Sampling distributions

- ✓ Random Sampling and Sample Bias
- ✓ Bias
- ✓ Random Selection
- ✓ Size versus Quality: When Does Size Matter?
- ✓ Sample Mean versus Population Mean
- ✓ Sampling Distribution of a Statistic
- ✓ Central Limit Theorem
- ✓ Standard Error
- ✓ Normal Distribution
- ✓ Standard Normal and QQ-Plots
- ✓ Long-Tailed Distributions
- ✓ Further Reading
- ✓ Student's t-Distribution
- ✓ Further Reading
- ✓ Binomial Distribution
- ✓ Further Reading
- ✓ Poisson and Related Distributions
- ✓ Poisson Distributions
- ✓ Exponential Distribution
- ✓ Estimating the Failure Rate

(III) Statistical Experiments and Significance Testing

- ✓ A/B Testing
- ✓ Hypothesis Tests
- ✓ Permutation Test
- ✓ Permutation Tests: The Bottom Line for Data Science
- ✓ t-Tests
- ✓ Multiple Testing
- ✓ Two-Way ANOVA
- ✓ Chi-Square Test

(IV) Regression and Prediction

- ✓ Simple Linear Regression
- ✓ The Regression Equation
- ✓ Fitted Values and Residuals
- ✓ Least Squares
- ✓ Prediction versus Explanation (Profiling)
- ✓ Multiple Linear Regression
- ✓ Example: King County Housing Data
- ✓ Assessing the Model
- ✓ Cross-Validation
- ✓ Model Selection and Stepwise Regression
- ✓ Weighted Regression
- ✓ Prediction Using Regression
- ✓ Interpreting the Regression Equation
- ✓ Correlated Predictors
- ✓ Testing the Assumptions: Regression Diagnostics
- ✓ Outliers

(V) Classification

- ✓ Naive Bayes
- ✓ Why Exact Bayesian Classification Is Impractical
- ✓ The Naive Solution
- ✓ Numeric Predictor Variables
- ✓ Further Reading
- ✓ Discriminant Analysis
- ✓ Covariance Matrix
- ✓ Fisher's Linear Discriminant
- ✓ A Simple Example
- ✓ Further Reading
- ✓ Logistic Regression
- ✓ Logistic Response Function and Logit
- ✓ Logistic Regression and the GLM
- ✓ Generalized Linear Models
- ✓ Predicted Values from Logistic Regression
- ✓ Interpreting the Coefficients and Odds Ratios
- ✓ Linear and Logistic Regression: Similarities and Differences
- ✓ Assessing the Model
- ✓ Further Reading
- ✓ Evaluating Classification Models
- ✓ Confusion Matrix

- ✓ The Rare Class Problem
- ✓ Precision, Recall, and Specificity
- ✓ ROC Curve
- ✓ AUC
- ✓ Lift
- ✓ Further Reading
- ✓ Strategies for Imbalanced Data
- ✓ Undersampling
- ✓ Oversampling and Up/Down Weighting
- ✓ Data Generation
- ✓ Cost-Based Classification
- ✓ Exploring the Predictions

(VI) Statistical Machine Learning

- ✓ K-Nearest Neighbors
- ✓ A Small Example: Predicting Loan Default
- ✓ Distance Metrics
- ✓ One Hot Encoder
- ✓ Standardization (Normalization, Z-Scores)
- ✓ Choosing K
- ✓ KNN as a Feature Engine
- ✓ Tree Models
- ✓ A Simple Example
- ✓ The Recursive Partitioning Algorithm
- ✓ Measuring Homogeneity or Impurity
- ✓ Stopping the Tree from Growing
- ✓ Predicting a Continuous Value
- ✓ How Trees Are Used
- ✓ Further Reading
- ✓ Bagging and the Random Forest
- ✓ Bagging
- ✓ Random Forest
- ✓ Variable Importance
- ✓ Hyperparameters
- ✓ Boosting
- ✓ The Boosting Algorithm
- ✓ Regularization: Avoiding Overfitting
- ✓ Hyperparameters and Cross-Validation

(VII) Unsupervised Learning

- ✓ Principal Components Analysis
- ✓ A Simple Example

- ✓ Computing the Principal Components
- ✓ Interpreting Principal Components
- ✓ K-Means Clustering
- ✓ A Simple Example
- ✓ K-Means Algorithm
- ✓ Interpreting the Clusters
- ✓ Selecting the Number of Clusters
- ✓ Hierarchical Clustering
- ✓ Model-Based Clustering
- ✓ Multivariate Normal Distribution
- ✓ Scaling and Categorical Variables
- ✓ Scaling the Variables
- ✓ Dominant Variables
- ✓ Categorical Data and Gower's Distance
- ✓ Problems with Clustering Mixed Data

3. PYTHON

1: Introduction to Python

- Introduction to Python as a programming language
- Installing Python and essential libraries (e.g., NumPy, Pandas, Matplotlib)
- Basic data types, variables, and operators

2: Control Structures

- Conditional statements (if-else)
- Loops (for, while)
- Functions and modules
- Handling exceptions

3: Data Structures in Python

- Lists, tuples, and dictionaries
- Working with arrays and matrices (NumPy)
- Data manipulation with Pandas

4: Data Cleaning and Preprocessing

- Handling missing data
- Data normalization and standardization
- Data aggregation and merging

5: Exploratory Data Analysis (EDA)

- Descriptive statistics
- Data visualization with Matplotlib and Seaborn
- Univariate and bivariate analysis

6: Data Visualization

- Advanced data visualization techniques
- Plot customization and aesthetics
- Interactive visualization with Plotly

7: Numpy

8:Pandas

4. Machine Learning

Module 1: Introduction to Machine Learning

1.1 Understanding Machine Learning

Definition and Types: Supervised, Unsupervised, Reinforcement Learning

Applications and use cases

1.2 Basics of Statistics for Machine Learning

Descriptive and Inferential Statistics

Probability distributions and central limit theorem

1.3 Python and Machine Learning

Setting up Python environment

Introduction to key libraries: NumPy, Pandas, Matplotlib, Scikit-learn

Module 2: Supervised Learning

2.1 Regression

Linear Regression

Polynomial Regression

Evaluation metrics: Mean Squared Error, R-squared

2.2 Classification

Logistic Regression

Decision Trees and Random Forests

Support Vector Machines

Performance metrics: Accuracy, Precision, Recall, F1 Score

2.3 Model Evaluation and Validation

Cross-validation techniques

Overfitting and underfitting

Hyperparameter tuning

Module 3: Unsupervised Learning

3.1 Clustering

K-Means Clustering

Hierarchical Clustering

DBSCAN

3.2 Dimensionality Reduction

Principal Component Analysis (PCA)

t-Distributed Stochastic Neighbor Embedding (t-SNE)

Module 4: Introduction to Deep Learning

4.1 Neural Networks

Perceptrons and Multi-layer Perceptrons (MLP)

Activation functions

Backpropagation algorithm

4.2 Convolutional Neural Networks (CNNs)

Image classification and feature extraction

Transfer learning

4.3 Recurrent Neural Networks (RNNs)

Sequence data and applications

Long Short-Term Memory (LSTM) networks

Module 5: Model Deployment and Scaling

5.1 Model Deployment

Containerization with Docker

Deployment on cloud platforms: AWS, Azure, Google Cloud

5.2 Scaling Machine Learning Projects

Collaborative tools and version control for team projects

Best practices for scalability and efficiency

Module 6: Advanced Topics in Machine Learning

6.1 Ensemble Learning

Bagging and Boosting

Stacking models

6.2 Hyperparameter Optimization

Grid Search and Random Search

Bayesian Optimization

6.3 Explainable AI

Interpreting machine learning models

Addressing bias and fairness

Module 7: Project

7.1 Project Definition

Identifying a real-world problem or dataset

Defining project objectives and scope

7.2 Implementation

Applying learned techniques and methodologies

Iterative development and improvement

7.3 Presentation and Documentation

Communicating results effectively

Creating a comprehensive project report

Natural Language processing

Introduction

Recognizing NLP Applications

Tokenizing text

Module 1: Introduction to Natural Language Processing

1.1 What is NLP?

Definition and scope

Applications of NLP in real-world scenarios

1.2 Key Components of NLP

Tokenization

Part-of-speech tagging

Named Entity Recognition (NER)

1.3 Challenges in NLP

Ambiguity and context

Handling different languages and dialects

Module 2: Text Preprocessing

- 2.1 Cleaning and Normalization
  - Removing noise and unnecessary characters
  - Lowercasing and stemming
- 2.2 Stopwords and Punctuation
  - Identifying and removing common stopwords
  - Handling punctuation in text data
- 2.3 Feature Engineering for NLP
  - Bag-of-Words (BoW) representation
  - TF-IDF (Term Frequency-Inverse Document Frequency)

Module 3: Text Representation with Word Embeddings

- 3.1 Word Embeddings Overview
  - Introduction to Word2Vec, GloVe, and FastText
  - Word Embeddings vs. Traditional Representations
- 3.2 Training Word Embeddings
  - Word2Vec and GloVe training processes
  - Hyperparameter tuning for word embeddings

Module 4: Named Entity Recognition (NER) and Part-of-Speech (POS) Tagging

- 4.1 Named Entity Recognition
  - Types of entities: Person, Organization, Location, etc.
  - NER algorithms and techniques
- 4.2 Part-of-Speech Tagging
  - Understanding POS tags
  - POS tagging techniques and applications

Module 5: Syntax and Parsing

- 5.1 Syntax Trees
  - Representing sentence structure
  - Dependency Parsing vs. Constituency Parsing

5.2 Parsing Techniques

Overview of parsing algorithms

Implementing parsers in NLP applications

Module 6: Sentiment Analysis

6.1 Understanding Sentiment

Sentiment scales and classifications

Challenges in sentiment analysis

6.2 Sentiment Analysis Techniques

Rule-based approaches

Machine learning-based approaches

Module 7: Advanced NLP Techniques

7.1 Sequence-to-Sequence Models

Introduction to encoder-decoder architectures

Applications in machine translation and summarization

7.2 Attention Mechanism

Addressing long-range dependencies in NLP

Transformer models and BERT

Module 8: NLP Applications and Case Studies

8.1 Text Summarization

Extractive vs. abstractive summarization

Case studies on news and document summarization

8.2 Machine Translation

Overview of machine translation systems

Case studies on language translation applications

Module 9: Ethics in NLP

9.1 Bias and Fairness

Addressing biases in NLP models

Ensuring fairness in language processing applications

9.2 Privacy Concerns

Handling sensitive information in NLP tasks

GDPR and other privacy regulations

Module 10: Project

10.1 Project Definition

Identifying a real-world NLP problem

Defining project objectives and scope

10.2 Implementation

Applying learned techniques to solve the NLP problem

Iterative development and improvement

10.3 Presentation and Documentation

Communicating project results effectively

Creating a comprehensive project report showcasing NLP implementation