
Part I: Machine Learning

Feature Transformation

1. **Purpose of Feature Transformation in ML:**
   Feature transformation helps improve model performance by transforming data to reduce noise, simplify structures, and highlight key patterns, often used in dimensionality reduction.
2. **PCA Algorithm and Dimensionality Reduction:**
   PCA (Principal Component Analysis) reduces dimensions by finding directions (principal components) of maximum variance in data. It projects data points along these directions, capturing the most significant variation. For example, it can reduce features in a dataset from many variables to just two or three.
3. **Advantages and Disadvantages of PCA:**
   - **Advantages:** Reduces data dimensionality, decreases computation time, mitigates overfitting.
   - **Disadvantages:** Hard to interpret transformed features; works only with linear relationships.
4. **LDA Algorithm and Application in Classification:**
   LDA (Linear Discriminant Analysis) maximizes class separation by finding a linear combination of features that separates classes. In the Iris dataset, LDA can classify species by separating them based on discriminative features.
5. **PCA vs. LDA – When to Use Which:**
   PCA is unsupervised and used to capture variance, while LDA is supervised, focusing on class separation. Choose PCA for data compression and noise reduction, and LDA for enhancing class separability in classification tasks.

Regression Analysis

1. **What is Regression Analysis?**
   Regression is a statistical method for predicting a continuous outcome based on input variables. It's used in areas like forecasting and risk assessment.
2. **Predicting Uber Ride Prices Using Regression:**
   Use variables like distance, time of day, and demand level to predict prices, applying linear regression to find a line of best fit for price prediction.
3. **Types of Regression Models:**
   - **Linear Regression:** Fits a straight line to minimize errors.
   - **Ridge Regression:** Adds a penalty on coefficients to reduce overfitting.
   - **Lasso Regression:** Similar to ridge but can shrink some coefficients to zero, effectively selecting features.
4. **Regression Model Evaluation Metrics:**
   Common metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared for accuracy and goodness-of-fit.
5. **Handling Outliers in Regression Analysis:**

Outliers can be removed, transformed, or handled by using robust models that minimize their effect.

6. **Univariate and Bivariate Analysis in Diabetes Dataset:**
    - **Univariate Analysis:** Analyzes one variable's distribution (e.g., blood sugar levels).
    - **Bivariate Analysis:** Studies relationships between two variables, like blood sugar and age, using scatter plots or correlation coefficients.

Classification Analysis

1. **What is Classification Analysis?**
   Classification predicts discrete labels (e.g., spam or not spam) and is different from regression, which predicts continuous values.
2. **SVM for Handwritten Digit Classification:**
   Support Vector Machines (SVM) find a hyperplane to separate classes in high-dimensional space. For digit classification, SVM separates images based on pixel values.
3. **K-Nearest Neighbors (KNN) Algorithm:**
   KNN classifies data by finding 'k' nearest points and assigns the most common class among them. Performance metrics include:
    - **Confusion Matrix:** Shows true vs. predicted classes.
    - **Accuracy:** Ratio of correct predictions.
    - **Error Rate:** Ratio of incorrect predictions.
4. **Advantages and Limitations of SVM and KNN:**
    - **SVM Advantages:** Effective in high dimensions, robust with a clear margin.
    - **SVM Limitations:** Not ideal for large datasets.
    - **KNN Advantages:** Simple, no training required.
    - **KNN Limitations:** Slow with large datasets, sensitive to noisy data.

Clustering Analysis

1. **What is Cluster Analysis?**
   Cluster analysis groups data into clusters based on similarity. Applications include market segmentation and image compression.
2. **K-Means and the Iris Dataset:**
   K-Means partitions data into 'k' clusters by minimizing the distance between points and their cluster centroids. For the Iris dataset, it can cluster species based on petal and sepal dimensions.
3. **Optimal Number of Clusters (Elbow Method):**
   The elbow method plots the within-cluster sum of squares (WCSS) for different 'k' values, where the "elbow" point indicates the optimal 'k'.
4. **K-Medoids and Silhouette Method:**
   K-Medoids is similar to K-Means but uses medoids to reduce sensitivity to outliers. The silhouette method calculates a score for each point to assess clustering quality,

indicating the best number of clusters.

Ensemble Learning

1. **What is Ensemble Learning?**
   Ensemble learning combines multiple models to improve accuracy, reduce overfitting, and boost robustness.
2. **Random Forest Classifier for Car Safety Prediction:**
   Random Forest uses multiple decision trees, each trained on different data samples, and averages results. It's effective for car safety because it can handle varied and large feature sets.
3. **Voting Mechanisms in Ensemble Learning:**
   - **Hard Voting:** Majority voting.
   - **Soft Voting:** Averages probability scores from each model.
4. **Comparison of AdaBoost, Gradient Boosting, and XGBoost:**
   - **AdaBoost:** Adds weak learners sequentially, focusing on misclassified samples.
   - **Gradient Boosting:** Minimizes loss by adding models that correct previous errors.
   - **XGBoost:** Faster, with additional tuning options and regularization for improved performance.
5. **Evaluating Ensemble Models:**
   Use metrics like accuracy, precision, recall, F1-score, and AUC-ROC for performance assessment.

Reinforcement Learning

1. **What is Reinforcement Learning?**
   RL involves an agent learning by trial and error within an environment, different from supervised (labeled data) and unsupervised (no labels) learning.
2. **Maze Environment in RL:**
   A maze environment challenges an agent to find an optimal path to a goal, using rewards to reinforce correct paths.
3. **Taxi Problem in RL:**
   The agent learns how to pick up and drop off passengers in an optimal route by maximizing cumulative rewards.
4. **Building Tic-Tac-Toe with RL:**
   Use Q-learning or similar RL approaches to teach an agent strategies for optimal moves, by rewarding wins and penalizing losses.

Part II: Data Modeling and Visualization

Interacting with Web APIs

1. **What are Web APIs?**
   APIs allow interaction with external services. They're used to fetch data (like weather) for analysis.
2. **Process with OpenWeatherMap API:**
   Request data through HTTP requests, parse JSON responses, and visualize or analyze the data.
3. **Analyzing Weather Data from OpenWeatherMap:**
   Clean and process API data, then use visualization to show trends or correlations.

Data Cleaning and Preparation

1. **Importance of Data Cleaning:**
   Ensures data quality by removing errors, inconsistencies, and improving model accuracy.
2. **Strategies for Handling Missing Values:**
   Options include deletion, imputation, or replacing with statistical values like mean or median.
3. **Handling Outliers:**
   Use techniques like IQR-based removal or transformation to reduce the impact of extreme values.
4. **Feature Engineering and Customer Churn:**
   Create new features like "number of months active" or "number of customer support calls" to predict churn probability.

Data Wrangling

1. **What is Data Wrangling?**
   It prepares data for analysis, including cleaning, transforming, and merging. Unlike cleaning, wrangling includes structuring the data.
2. **Data Wrangling on Real Estate Dataset:**
   Includes handling missing values, standardizing formats, and encoding features for easy analysis.
3. **Encoding Categorical Variables:**
   Convert categorical data into numerical using techniques like one-hot encoding or label encoding.
4. **Identifying and Handling Outliers in Real Estate Data:**
   Use visualizations like box plots or Z-scores to identify outliers and handle them based on analysis goals.

Data Visualization using matplotlib

1. **Types of Visualizations:**
   Line plots for trends, bar charts for comparisons, and scatter plots for relationships. Choose based on the data type and analysis goals.
2. **Visualizing AQI Trends:**
   Line plots can show AQI changes over time, bar charts for city comparisons, and scatter plots for correlation.
3. **Customization with matplotlib:**
   You can change colors, labels, titles, and add legends to improve clarity.
4. **Box Plots and Violin Plots:**
   Box plots show data spread and outliers; violin plots add density information for a fuller distribution view.

Data Aggregation

1. **What is Data Aggregation?**
   Aggregation combines data, like summing sales by region, to provide meaningful insights.
2. **Analyzing Sales by Region:**
   Aggregate sales data by regions, then use mean or sum to analyze performance.
3. **Aggregation Functions and Applications:**
   Functions like sum, mean, median are used for summarizing large datasets.
4. **Visualizing Aggregated Data:**
   Use bar charts, pie charts, or heatmaps in matplotlib to show aggregated results.

Part III: Mini Project

1. **Objectives and Scope of Mini-Project:**
   Explain your goals, whether they are predictive modeling, data analysis, or a specific research question.
2. **Problem Statement and Relevance:**
   Discuss why the problem matters (e.g., customer churn, sales forecasting) and what benefits it brings.
3. **Methodology and Tools:**
   Outline data sources, model choice, and tools used (e.g., Python libraries like sklearn, pandas).
4. **Key Findings and Insights:**
   Summarize insights or patterns discovered, model accuracy, and future recommendations.