Weekly Report: Nusrat Jahan Nuha-Alpha AI

Week-05 (May 19 - May 23)

1. Course Documentation

Unsupervised Learning, Recommenders, Reinforcement Learning

Key Takeaways:

- **Clustering:** Clustering is an unsupervised learning technique used to group similar data points based on feature similarity. It helps discover hidden patterns and natural groupings in data without predefined labels.
- **K-Means Intuition**: K-Means aims to partition data into K clusters where each point belongs to the cluster with the nearest mean. The goal is to minimize intra-cluster variance while maximizing inter-cluster separation.
- **K-Means Algorithm:** The algorithm iteratively assigns points to the nearest cluster centroid, then updates the centroids based on current assignments. This process continues until convergence or minimal centroid movement.
- **Optimization Objective in K-Means:** K-Means minimizes the within-cluster sum of squares (WCSS)—the squared distance between each point and its assigned cluster centroid. This ensures tight, compact clusters.
- **Initializing K-Means:** K-Means initialization is critical; poor starting centroids can lead to suboptimal clustering. The **K-Means**++ initialization method improves performance by spreading out initial centroids.

- **Choosing the Number of Clusters:** The **Elbow Method** and **Silhouette Score** are common techniques to determine the optimal number of clusters. They help balance model complexity with clustering performance.
- **Finding Unusual Events (Anomalies)**: Anomalies are data points that significantly differ from the norm. Clustering and statistical models can help detect such outliers, which may indicate rare or suspicious events.
- **Gaussian Distribution:** The Gaussian (normal) distribution is a bell-shaped curve used to model continuous data. It plays a key role in probabilistic methods like anomaly detection and PCA.
- **Anomaly Detection Algorithm**: This algorithm models normal data distribution and flags outliers with low probability under the learned distribution. It often assumes Gaussian distribution for features.
- **Collaborative Filtering vs Content-Based Filtering:** Collaborative filtering recommends items based on user-user or item-item similarities using historical interactions. Content-based filtering uses item attributes and user preferences to generate recommendations.
- **PCA Algorithm:** Principal Component Analysis (PCA) reduces dimensionality by projecting data onto directions (principal components) that maximize variance. It helps in data compression and noise reduction.
- **Deep Learning for Content-Based Filtering:** Deep learning enhances content-based filtering by learning complex user-item feature representations using neural networks. It improves recommendations in high-dimensional and sparse datasets.

2. ISPToolbox

Research and Documentation Progress

Overview: ISPToolbox provides a comprehensive suite of tools and services aimed at simplifying the process of planning, deploying, and operating community-based ISPs. It offers functionalities such as market evaluation, network cost comparison, line-of-sight analysis, and digital surface model exports. The platform is hosted on AWS, utilizing services like EC2, S3, and PostgreSQL, with Redis for key-value storage. The backend is built using Django, complemented by Celery for asynchronous tasks, Django REST Framework library for API generation, and Channels for WebSocket support.

Core Functionalities Developed:

- Market Evaluator: Likely to evaluate demand/supply, cost, and feasibility of ISP ventures.
- Network Cost Comparison: Compares infrastructure and operational costs for different network setups.
- Line-of-Sight (LoS) Check: Analyzes topography to validate wireless transmission feasibility.
- DSM Export: Digital Surface Model exports for geographical and signal planning.

Codebase Analysis

Networking Tool

LOS Check

The purpose of the ISPToolbox LOS Checking tool is to determine if there's a clear path between the access point and the potential customer's hub using LiDAR in 3D modeling. It helps to visualize what the obstacles are in the way. So that the position of the towers can be adjusted accordingly.

The `DSMEngine` class is responsible for generating Digital Surface Models (DSMs) — raster images where each pixel represents the maximum height (Z value) in a given area — from LiDAR point cloud data stored in EPT format.

These DSMs are used for:

Line-of-sight (LoS) analysis

Wireless planning (viewshed tools)

Terrain visualization in 3D tools

3. **Project on Test Driven Development (TDD)**
    - Implemented TDD principles in the Tesla-S stock prediction project to ensure model correctness and robustness.
    - Wrote comprehensive unit tests for baseline ML models along with LSTM, GRU and custom feature engineering functions.

- Utilized Python's built-in `unittest` framework for structuring test cases.
- Used assert methods (e.g., assertEqual, assertTrue, assertAlmostEqual) to verify model outputs, shape consistency, data transformations, and metric thresholds.