To,

IITD-AIA Foundation on Smart Manufacturing

Subject: Weekly Progress Report for Week-7

Respected sir,

Following is the required progress report to the best of my knowledge considering relevant topics to be covered:

- Clustering
- Text Analysis
- Dimensionality Reduction
- Hierarchical Clustring
- Elbow Technique

Day 18:

Clustering is a data analysis technique that groups similar data points together based on their characteristics.

It helps identify patterns and relationships in the data without prior knowledge of the groups.

Different clustering algorithms.

1. K-means: Divides data into k clusters by minimizing distances to cluster centroids.
2. Hierarchical: Builds a hierarchy of clusters using a bottom-up or top-down approach.
3. DBSCAN: Forms clusters based on data density and identifies regions with different densities.
4. GMM (Gaussian Mixture Models): Assumes data comes from a mixture of Gaussian distributions.
5. Mean Shift: Shifts towards dense regions to find clusters.
6. Spectral: Transforms data into a lower-dimensional space using eigenvalues and eigenvectors**.**

Day 19:

It helps extract valuable information from text and is used in various fields for automation,

decision-making, and understanding textual data.

Steps:

1. Preprocessing: Clean and format the text by removing irrelevant characters and normalizing words.
2. Tokenization: Divide the text into smaller units such as words, phrases, or sentences.
3. Part-of-speech tagging: Label each word with its grammatical category.
4. Named entity recognition: Identify and classify named entities in the text.
5. Sentiment analysis: Determine the sentiment expressed in the text (positive, negative, or neutral).
6. Topic modeling: Discover the main topics or themes discussed in the text.
7. Text classification: Categorize text documents into predefined classes or categories.
8. Text summarization: Condense the text into a shorter version while preserving key information.
9. Word embeddings: Represent words as dense vectors to capture meaning and context.

Day 20 :

The goal of dimensionality reduction is to simplify the data, making it easier to analyze, visualize,

and process, while also addressing potential issues like the curse of dimensionality.

Main approaches for dimensionality reduction :

1. Feature Selection: Choosing important features, discarding the rest. The selected features should capture the most relevant information and patterns in the data.
2. Feature Extraction: Creating new, compressed features from the original data.These new features are a compressed representation of the data and aim to retain essential information

https://www.kaggle.com/competitions/open-problems-multimodal/discussion/348233

Day 21:

Hierarchical clustering is a data analysis technique used to group similar data points into clusters based on their similarities.

Two main types of hierarchical clustering:

1. Agglomerative Hierarchical Clustering: Bottom-up approach, starts with each data point as its own cluster, and merges the closest clusters iteratively.
2. Divisive Hierarchical Clustering: Top-down approach, starts with all data points in one cluster, and recursively divides it into smaller clusters.

Hierarchical Clustering algorithm.

1. Start with each data point as its own cluster.
2. Calculate the similarity between all pairs of clusters.
3. Identify the two closest clusters based on the similarity measure.
4. Merge the two closest clusters into a single cluster.
5. Update the similarity matrix with the new merged cluster.
6. Repeat steps 2-5 until all data points belong to a single cluster or a stopping criterion is met.
7. Represent the clustering process as a dendrogram to visualize the hierarchy of clusters.
8. Decide the number of clusters by cutting the dendrogram at a certain height.

Day 22:

The Elbow Technique finds the best number of clusters for data by identifying the point on a plot where adding more clusters doesn't significantly improve results. It helps strike the right balance between accuracy and simplicity .

STEPS:

1. Choose a range of possible cluster counts (K values).
2. Run K-means clustering for each K value.
3. Calculate the sum of squared distances (inertia) for each clustering result.
4. Plot the K values against their corresponding inertias.
5. Look for the "elbow point" on the plot, where the inertia starts to level off.
6. The K value at the elbow point is the optimal number of clusters for the dataset.

Day 23:

Elbow Technique implementation by example:

```
import numpy as np
```

```
import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

data = np.array(data)

K_values = range(1, 11)

inertia_values = []

for K in K_values:

    kmeans = KMeans(n_clusters=K)

    kmeans.fit(data)

    inertia_values.append(kmeans.inertia_)

plt.plot(K_values, inertia_values, marker='o')

plt.xlabel('Number of Clusters (K)')

plt.ylabel('Sum of Squared Distances (Inertia)')

plt.title('Elbow Curve for K-means Clustering')

plt.show()
```

Reference:

https://www.youtube.com/watch?v=CLKW6uWJtTc&pp=ygUVZWxib3cgdGVjaG5pcXVlIGluIG1s