Clustering

* **Unsupervised Learning**: It uses an unlabeled training set. Unlike supervised learning, we only have a dataset of features and no corresponding expected results.
* **Use cases for Clustering**: Market segmentation, social network analysis, computer cluster organization, and astronomical data analysis.
* **K-Means Algorithm**:
  1. Initialize two random points as cluster centroids.
  2. Assign each data point to the nearest centroid.
  3. Move centroids to the average of their respective cluster.
  4. Repeat steps 2 and 3 until convergence.
* **Variables in K-means**:
  1. **K**: Number of clusters
  2. **Training Set**: x(1), x(2),…, x(m)
  3. x(i) is an element of R^n, representing the feature vectors.
  4. The algorithm works without using the x0=1 convention.
* **Cluster Assignment**:
  1. Assign data points based on the closest centroid using the equation: c(i) = argmin\_k ||x(i) - μ\_k||^2.
* **Move Centroid**:
  1. Update centroid based on the average of data points in its cluster.
* **Optimization Objective**:
  1. The distortion (cost function) is: J(c, μ) = (1/m) \* sum(||x(i) - μ\_c(i)||^2).
* **Random Initialization**:
  1. Ensure K < m.
  2. Randomly pick K unique training examples for initializing centroids.
* **Choosing the Number of Clusters**:
  1. Elbow method: Identify where the cost function starts to flatten as K increases.
  2. Alternatively, choose K based on performance in a downstream task.
* **Drawbacks of K-Means**: Can give unexpected results in certain situations.

**ML: Dimensionality Reduction**

* **Motivations**:
  1. **Data Compression**: Reduce redundant data, save memory, and speed up algorithms.
  2. **Visualization**: Display data in 2D or 3D for better understanding.
* **Principal Component Analysis (PCA)**:
  1. Most popular dimensionality reduction method.
  2. Reduces the dimensionality by finding a lower-dimensional surface onto which the data projects well.
  3. **PCA vs. Linear Regression**: PCA minimizes orthogonal distance from data points to the projection surface, while linear regression minimizes vertical distances to predict a target.
  4. **PCA Algorithm**:
     1. **Preprocessing**: Feature scaling and mean normalization.
     2. Compute covariance matrix Σ.
     3. Compute eigenvectors of Σ using singular value decomposition (SVD).
     4. Extract the first k columns of the U matrix from SVD for dimensionality reduction.

This recap provides an overview of the topics covered in Week 8 Lecture on clustering and dimensionality reduction. The K-means algorithm is introduced as a method for unsupervised learning, and PCA is explained as a technique for reducing data dimensions. Both concepts are vital in machine learning for data categorization and simplification, respectively.

Top of Form

Bottom of Form