Allison Molitor

M.L. Learning Journal

Start Date: 01/05/2022

Most Recent Edit: 08/24/2023

Machine Learning Pattern Recognition Models

Cluster Modeling

Clustering is an unsupervised machine-learning algorithm that aims to uncover patterns and group objects on unlabeled data. These grouped objects are formed into clusters in which objects within the same group exhibit higher similarity to one another compared to those in different groups. These clusters can be formed based on shared features, attributes, or characteristics. Clustering allows for the organization of data and the discovery of relationships that the naked eye may be unable to delineate.

- K Means Clustering
  - Mathematical Representation
    - $SS_k = \sum_{i \in \text{Cluster } k} (x_i - \bar{x}_k)^2 + (y_i - \bar{y}_k)^2$
      - SSk = The measure of how close the data points within cluster 'k' are to the centroid of that cluster.
      - Σ i ∈ Cluster k = Sum of squared distances from the cluster
      - (xi - x(bar)k)^2 = The coordinates of the ith data point minus the centroid of cluster ‘k’
      - (yi - y(bar)k)^2 = The coordinates of the ith data point minus the centroid of cluster ‘k’
    - Complexity is 0(n * K * I * d)
      - n = number of points
      - K = number of clusters
      - I = number of iterations
      - d = number of attributes
  - Real World Example and Mathematical Computation

Customer	Annual Income (in $1000s)	Annual Spending (in $ 1000s)

1	20	10
2	40	20
3	35	15
4	70	25
5	55	10

K = 2
Initial centroid 1: (30, 15)
Initial centroid 2: (40, 25)
Distance to centroid 1(customer 1): $\sqrt{(20 - 30)^2 + (10 - 15)^2} = 11.2$
Distance to centroid 2(customer 1): $\sqrt{(20 - 40)^2 + (10 - 25)^2} = 25$
Distance to centroid 1(customer 2): $\sqrt{(40 - 30)^2 + (20 - 15)^2} = 11.2$
Distance to centroid 2(customer 2): $\sqrt{(40 - 40)^2 + (20 - 25)^2} = 5$
Distance to centroid 1(customer 3): $\sqrt{(35 - 30)^2 + (15 - 15)^2} = 5$
Distance to centroid 2(customer 3): $\sqrt{(35 - 40)^2 + (15 - 25)^2} = 11.2$
Distance to centroid 1(customer 4): $\sqrt{(70 - 30)^2 + (25 - 15)^2} = 41.2$
Distance to centroid 2(customer 4): $\sqrt{(70 - 40)^2 + (25 - 25)^2} = 30$
Distance to centroid 1(customer 5): $\sqrt{(55 - 30)^2 + (10 - 15)^2} = 25.5$
Distance to centroid 2(customer 5): $\sqrt{(55 - 40)^2 + (10 - 25)^2} = 21.2$

Results:

- Cluster 1 has customers 1 and 3, while Cluster 2 has customers 2, 4, and 5
- Cluster 1 consists of customers with lower income and moderate spending, and Cluster 2 consists of customers with higher income and higher spending
    - Primary Advantages
        - Efficient
            - Scales well
            - Efficiency on large datasets
            - Quick converging rate
        - Simplicity
            - Straightforward
            - Easy to interpret and visualize
        - Works well with spherical clusters
    - Primary Disadvantages
        - Amount of clusters
            - Need to specify the number of clusters beforehand

        - Choosing the appropriate k is important
    - Non-Globular
    - Outliers
        - Can significantly affect centroids
    - Assumes Equal Variance
        - Which may not always hold for all datasets

- Hierarchical Clustering
    - Real World Example and Mathematical Computation

Person	Height (in)	Weight (lbs)
A	72	180
B	69	173
C	58	115
D	67	140

Distance between A and B: $\sqrt{(72 - 69)^2 + (180 - 173)^2} = 7.6$
Distance between A and C: $\sqrt{(72 - 58)^2 + (180 - 115)^2} = 66.5$
Distance between A and D: $\sqrt{(72 - 67)^2 + (180 - 140)^2} = 40.3$
Distance between B and C: $\sqrt{(69 - 58)^2 + (173 - 115)^2} = 59$
Distance between B and D: $\sqrt{(69 - 67)^2 + (173 - 140)^2} = 33.1$
Distance between C and D: $\sqrt{(58 - 67)^2 + (115 - 140)^2} = 26.6$
Closest data points: A and B

	A, B	C	D
A, B	0.00		
C	66.5	0.00	
D	40.3	26.6	0.00

	A, B, D	C
A, B, D	0.00	

C	66.5	0.00

Final Merged Cluster (using single linkage): {A, B, D} , {C}

- Primary Advantages
  - Flexible Shape
    - Does not assume any shape or distribution
  - Amount of clusters
    - No need to specify the number of clusters
  - Can handle outliers well
- Primary Disadvantages
  - Can have high computational complexity
  - Sensitive to noise
  - Difficulty with large datasets
    - Limited scalability
- Common Uses
  - User behavior analysis
    - Grouping users based on interactions
  - Network analysis
    - Analyze patterns and group similar network flows
  - Software testing
    - Identify redundant or similar test cases
  - Data exploration
    - Explore relationships between datasets
    - Identify hidden patterns

Principle Component Analysis

Principle Component Analysis (PCA) is a statistical method that summarizes large datasets by condensing them into smaller sets of summary indices. These smaller sets can be more easily visualized and analyzed. It simplifies data by finding new variables, or principal components, that capture the most important information from the original data. PCA is useful for minimizing residual variance in the least squares and maximizing the variance of the projection coordinates. It is further useful in identifying correlations and patterns between data points.

- Real-World Example with Steps
  - Portfolio Optimization
    - Collect data

            - Data is on the daily returns of various stocks over a certain period of time
        - Preprocessing
            - Calculate the daily returns of each stock. Each daily return is a data point
        - Covariance Matrix
            - Captures how the returns of each stock move. Diagonal elements portray the variances of individual returns, and off-diagonal portray covariances between pairs of returns.
        - Eigenvalues and Eigenvectors
            - Find the directions with the highest variances (eigenvectors) and the amount of variance explained by each eigenvector (eigenvalues)
        - Principal Component Selection
            - Choose which eigenvectors capture a significant portion of the total variance
- Primary Advantages
    - Dimensionality Reduction
        - Reduces size of data while retaining the most important information
    - Feature Selection
        - Removes noise
        - Removes least relevant features
    - Visualization
        - Projects data to make it easier to identify patterns, clusters, or outliers
- Primary Disadvantages
    - Loss of Information
        - There can be some degree of information loss
    - Sensitive to Scaling
    - Computational complexity
        - Memory-intensive
- Common Uses
    - Performance analysis
    - Dimensionality Reduction
        - Reduce features in large datasets
        - Cluster analysis
    - Fraud detection
    - Face recognition/Image processing