Clustering

Prof. Hariprasad Kodamana

December 4, 2025

Unsupervised Learning

1. We now consider unsupervised learning, where we are given only output data, without any input data.
2. The goal is to discover "interesting structure" in the data; this is sometimes called knowledge discovery.
3. Unlike supervised learning, we are not given the desired output for each input; instead, we formalize our task as one of density estimation and build models of the form $p(x_i|\theta)$.
4. There are two differences from the supervised case: (a) we have written $p(x_i|\theta)$ instead of $p(y_i|x_i, \theta)$, that is, supervised learning is conditional density estimation, whereas unsupervised learning is unconditional density estimation.
5. x_i is a vector of features, so we need to create multivariate probability models. By contrast, in supervised learning, x_i is typically a single variable that we aim to predict.

Unsupervised Learning

1. Unsupervised learning is arguably more typical of human and animal learning.
2. It is also more widely applicable than supervised learning, since it does not require a human expert to manually label the data, and labeled data is expensive to acquire.
3. Some canonical examples of supervised clustering: clustering data into groups, discovering latent factors or dimension reduction, discovering graph structure.

Clustering

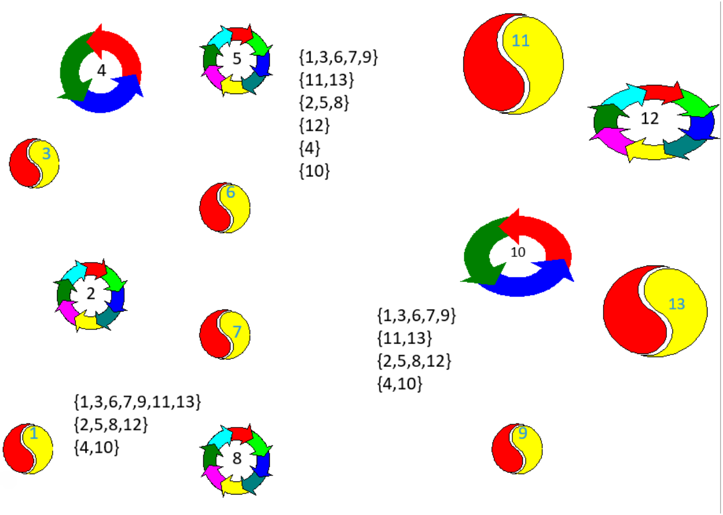
Why Clustering?

1. To discover groupings in a data set.

Examples:

1. Biology - classification of plants and animals given their features.
2. City-planning - Identifying groups of houses according to their house type, value, and geographical location.
3. Earthquake studies - Clustering observed earthquake epicenters to identify dangerous zones.
4. Insurance - Identifying groups of motor insurance policy holders with a high average claim cost, identifying fraud.
5. Libraries - Book ordering.
6. Marketing - Finding groups of customers with similar behavior, given a large database of customer data containing their properties and past buying records.
7. WWW - Document classification, clustering weblog data to discover groups of similar access patterns.

Cluster These Objects



{1,3,6,7,9}
{11,13}
{2,5,8}
{12}
{4}
{10}

{1,3,6,7,9}
{11,13}
{2,5,8,12}
{4,10}

{1,3,6,7,9,11,13}
{2,5,8,12}
{4,10}

K-Means Clustering

1. We begin by considering the problem of identifying groups, or clusters, of data points in a multidimensional space.
2. Suppose we have a data set $\{x_1, \cdots, x_N\}$ consisting of N observations of a random D-dimensional Euclidean variable x.
3. Our goal is to partition the data set into some number K of clusters, where we shall suppose for the moment that the value of K is given.
4. Intuitively, we might think of a cluster as comprising a group of data points whose inter-point distances are small compared with the distances to points outside of the cluster.
5. Let μ_k, where $k = 1, \cdots, K$, be a prototype (representing the centers of the clusters) associated with the kth cluster.
6. For each data point x_n, we introduce a corresponding set of binary indicator variables $r_{nk} \in \{0, 1\}$, where $k = 1, \cdots, K$ describing which of the K clusters the data point x_n is assigned to.
7. If data point x_n is assigned to cluster k then $r_{nk} = 1$, and $r_{nj} = 0$ for $j \neq k$. This is known as the 1-of-K coding scheme.

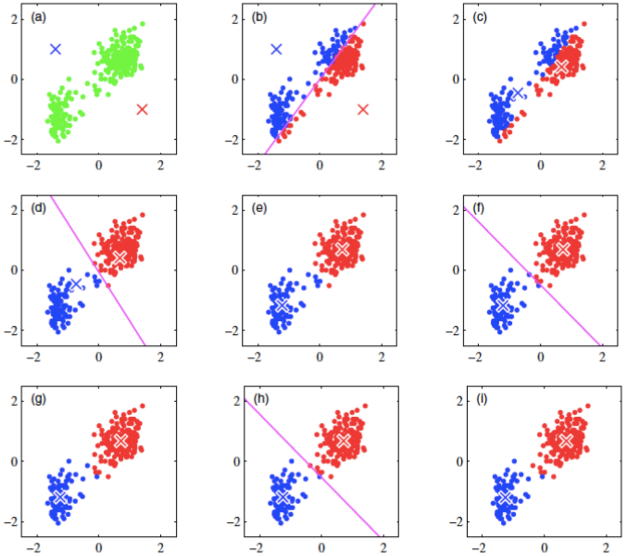
K-Means Clustering

1. Minimize distortion measure, given by:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2$$

   which represents the sum of the squares of the distances of each data point to its assigned vector μ_k.
2. Our goal is to find values for the r_{nk} and the μ_k to minimize J.
3. We can perform the optimization through an iterative procedure, in which each iteration involves two successive steps corresponding to optimizations with respect to r_{nk} and μ_k.
4. First, we choose some initial values for the μ_k. Then we minimize J with respect to the r_{nk}, keeping the μ_k fixed.
5. In the second phase, we minimize J with respect to the μ_k, keeping r_{nk} fixed.
6. This two-stage optimization is then repeated until convergence.

K-Means Clustering



K-Means Clustering

1. Consider first the determination of the r_{nk}.
2. The terms involving different n are independent and so we can optimize for each n separately by choosing r_{nk} to be 1 for whichever value of k gives the minimum value $\|x_n - \mu_k\|^2$.

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|x_n - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

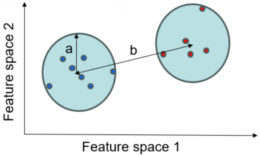
3. Now consider the optimization of the μ_k with the r_{nk} held fixed.
4. The objective function J is a quadratic function of μ_k, and it can be minimized by setting its derivative with respect to μ_k to zero, giving:

$$2\sum_{n=1}^{N} r_{nk}\|x_n - \mu_k\| = 0 \Rightarrow \mu_k = \frac{\sum_{n=1}^{N} r_{nk}x_n}{\sum_{n=1}^{N} r_{nk}}$$

5. Set μ_k equal to the mean of all the data points x_n assigned to cluster k.

Silhouette Analysis on K-Means Clustering

1. The separation distance between the resulting clusters can be analyzed using Silhouette analysis.
2. The silhouette plot represents the relative closeness of each point in one cluster to points in the neighboring clusters. Therefore, it can lead to a number of visually distinct clusters. The Silhouette coefficient varies in a range of $\{-1, 1\}$.
3. Silhouette coefficients close to $+1$ indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters, and negative values indicate that those samples might have been assigned to the wrong cluster.

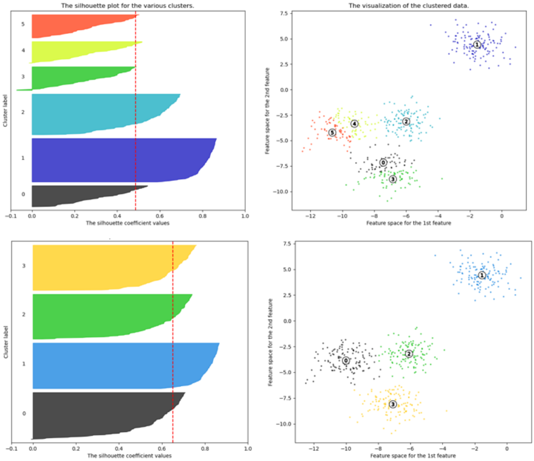


$$\text{Silhouette Score} = \frac{b - a}{\max(a, b)}$$

a: Intra-cluster distance
b: Inter-cluster distance

Silhouette Analysis on K-Means Clustering

Silhouette analysis for KMeans clustering on sample data with n_clusters = 6



Shuffled Dataset

Point	Coordinates	Original Label	True Cluster
A	(8, 7)	P4	Cluster 2
B	(3, 4)	P2	Cluster 1
C	(8, 8)	P6	Cluster 2
D	(2, 3)	P1	Cluster 1
E	(9, 8)	P5	Cluster 2
F	(2, 4)	P3	Cluster 1

True Cluster Structure

- Cluster 1 (High): A, C, E
- Cluster 2 (Low): B, D, F
- Balanced clusters: 3 points each

Iteration 1: Initialization

Initial Centroids

$$\mu_1^{(0)} = A = (8, 7) \quad \mu_2^{(0)} = B = (3, 4)$$

Point	To $\mu_1 = (8, 7)$	To $\mu_2 = (3, 4)$
A(8,7)	0	34
B(3,4)	34	0
C(8,8)	1	41
D(2,3)	52	2
E(9,8)	2	52
F(2,4)	45	1

Cluster Assignment

- Cluster 1: A, C, E
- Cluster 2: B, D, F

$$Distortion\ J_1 = 0 + 0 + 1 + 2 + 2 + 1 = 6$$

Iteration 1: Update Centroids

New Centroids

$$\mu_1^{(1)} = \frac{(8,7) + (8,8) + (9,8)}{3} = (8.3333, 7.6667)$$

$$\mu_2^{(1)} = \frac{(3,4) + (2,3) + (2,4)}{3} = (2.3333, 3.6667)$$

Visual Representation

Cluster 1 : A(8,7), C(8,8), E(9,8)
Centroid: (8.33, 7.67)
Cluster 2 : B(3,4), D(2,3), F(2,4)

Centroid: (2.33, 3.67)

Iteration 2: Recalculation

Distances from New Centroids

$$\begin{aligned}
\text{From } \mu_1^{(1)} = (8.3333, 7.6667) : \\
& d^2(A) = 0.5556,\ d^2(B) = 41.8889,\ d^2(C) = 0.2222, \\
& d^2(D) = 61.8889,\ d^2(E) = 0.5556,\ d^2(F) = 53.5556
\end{aligned}$$

$$\begin{aligned}
\text{From } \mu_2^{(1)} = (2.3333, 3.6667) : \\
& d^2(A) = 43.2222,\ d^2(B) = 0.5556,\ d^2(C) = 50.8889, \\
& d^2(D) = 0.5556,\ d^2(E) = 63.2222,\ d^2(F) = 0.2222
\end{aligned}$$

Iteration 2: Recalculation

Assignments Unchanged

- Cluster 1: A, C, E
- Cluster 2: B, D, F

Distortion J_2

$$J_2 = 0.5556 + 0.5556 + 0.2222 + 0.5556 + 0.5556 + 0.2222 = 2.6667$$

Convergence and Final Results

Convergence Check

- Initial distortion: $J_1 = 6.0000$
- Final distortion: $J_2 = 2.6667$
- Reduction: 55.56%
- Cluster assignments unchanged
- **Algorithm converged in 2 iterations**

Convergence and Final Results

Final Clustering

Cluster	Points	Centroid
Cluster 1	A(8,7), C(8,8), E(9,8)	(8.3333, 7.6667)
Cluster 2	B(3,4), D(2,3), F(2,4)	(2.3333, 3.6667)

Performance

- Perfect recovery of true cluster structure
- Balanced clusters (3 points each)
- Low final distortion

Silhouette Analysis: Example Calculation

For Point A(8,7) in Cluster 1

$$a(A) = \text{average distance to other points in Cluster 1}$$

$$= \frac{d(A,C) + d(A,E)}{2} = \frac{1+1}{2} = 1$$

$$b(A) = \text{average distance to Cluster 2}$$

$$= \frac{d(A,B) + d(A,D) + d(A,F)}{3}$$

$$= \frac{5.8310 + 7.2111 + 6.7082}{3} \approx 6.5834$$

$$s(A) = \frac{b(A) - a(A)}{\max\{a(A), b(A)\}} = \frac{6.5834 - 1}{6.5834} \approx 0.8482$$

Silhouette Analysis: Example Calculation

Interpretation

- $a(i)$: Cohesion (smaller is better)
- $b(i)$: Separation (larger is better)
- $s(i)$: Overall quality (-1 to 1)

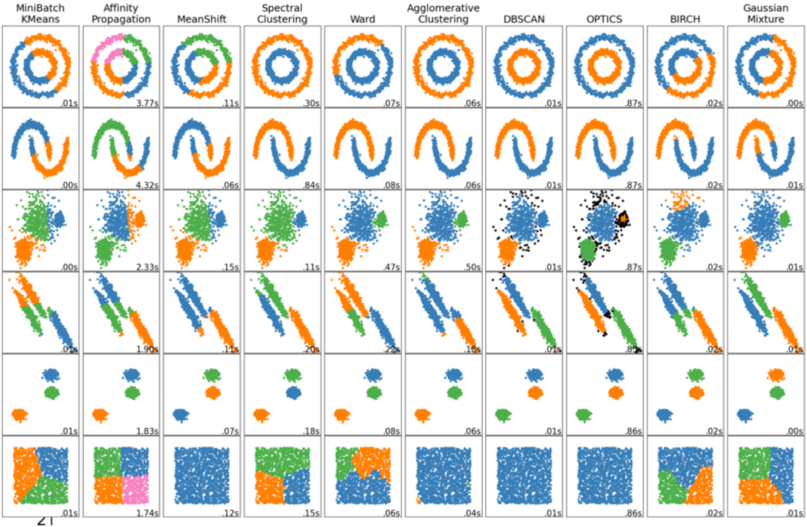
Complete Silhouette Scores

Point	Cluster	$a(i)$	$b(i)$	$s(i)$
A	1	1.0000	6.5834	0.8482
B	2	1.2071	6.4817	0.8138
C	1	1.0000	7.1415	0.8600
D	2	1.2071	7.8745	0.8468
E	1	1.0000	7.9586	0.8743
F	2	1.0000	7.3272	0.8635
Average Silhouette Score				0.8511

Interpretation of Scores

- All scores > 0.8: Excellent clustering
- Average $0.8511 > 0.7$: Strong cluster structure
- No negative scores: No misclassified points
- Consistent high scores: Uniform cluster quality

Different Clustering Models



21 https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

Thank You