Perform K-means Clustering Manually, With K = 2, On A Small Example With = 6 Observations And P = 2 Features:

Perform K-means Clustering Manually, With K = 2, On A Small Example With = 6 Observations And P = 2 Features:

Clustering is a fundamental technique in unsupervised machine learning used to group similar data points based on their features. Among various clustering algorithms, K-means stands out for its simplicity and efficiency, especially when dealing with large datasets. However, understanding how K-means works under the hood can deepen your grasp of the algorithm's mechanics and help you apply it more effectively. In this article, we'll walk through a detailed, step-by-step manual implementation of K-means clustering with K=2, on a small dataset comprising 6 observations with 2 features each. This hands-on approach will demystify the process, illustrating each stage from initial centroid selection to convergence.

Understanding the Basics of K-means Clustering

Before diving into the manual computation, it’s essential to grasp the core concepts of K-means clustering:

What is K-means?

- An iterative partitioning algorithm that aims to divide data into K clusters.
- Each cluster is represented by its centroid (mean of points in that cluster).
- The goal is to minimize the within-cluster sum of squares (WCSS), also known as inertia.

Key Components

- Number of clusters (K): Predefined; here, K=2.
- Centroids: The center points of each cluster.
- Assignments: Which data point belongs to which cluster based on proximity to centroids.
- Distance metric: Typically Euclidean distance.

High-level Steps of K-means

1. Initialize centroids.
2. Assign data points to the nearest centroid.
3. Update centroids based on current cluster memberships.
4. Repeat steps 2 and 3 until convergence (no change in assignments or centroids).

Preparing the Dataset

Let's consider a small dataset with 6 observations and 2 features for simplicity:

Observation	Feature 1 (X1)	Feature 2 (X2)
1	1.0	2.0
2	1.5	1.8
3	5.0	8.0
4	8.0	8.0
5	1.2	0.6
6	9.0	11.0

This dataset appears to have two natural groupings: observations 1, 2, and 5 seem close, and observations 3, 4, 6 seem to form another cluster.

Step-by-Step Manual Implementation of K-means (K=2)

Let's proceed through the clustering process manually.

Step 1: Initialization of Centroids

The initial step involves selecting two initial centroids. Strategies can vary:

- Random selection of data points.
- Selecting points that are far apart for better initial separation.

For simplicity, we'll choose:

- Centroid 1 (C1): Observation 1 — (1.0, 2.0)
- Centroid 2 (C2): Observation 4 — (8.0, 8.0)

This choice provides a good initial spread across the data space.

Step 2: Assign Data Points to Nearest Centroid

Calculate the Euclidean distance of each observation to each centroid:

Distance formula:

$$
d = \sqrt{(X_1 - C_{X1})^2 + (X_2 - C_{X2})^2}
$$

Calculations:

Obs	(X1, X2)	Distance to C1	Distance to C2	Assigned Cluster
1	(1.0, 2.0)	0.0 (self)	$\sqrt{(1-8)^2 + (2-8)^2} = \sqrt{49 + 36} = \sqrt{85} \approx 9.22$	Cluster 1
2	(1.5, 1.8)	$\sqrt{(1.5-1)^2 + (1.8-2)^2} = \sqrt{0.25 + 0.04} = \sqrt{0.29} \approx 0.54$	$\sqrt{(1.5-8)^2 + (1.8-8)^2} = \sqrt{42.25 + 38.44} = \sqrt{80.69} \approx 8.99$	Cluster 1
3	(5.0, 8.0)	$\sqrt{(5-1)^2 + (8-2)^2} = \sqrt{16 + 36} = \sqrt{52} \approx 7.21$	$\sqrt{(5-8)^2 + (8-8)^2} = \sqrt{9 + 0} = 3.0$	Cluster 2
4	(8.0, 8.0)	$\sqrt{(8-1)^2 + (8-2)^2} = \sqrt{49 + 36} = \sqrt{85} \approx 9.22$	0.0 (self)	Cluster 2
5	(1.2, 0.6)	$\sqrt{(1.2-1)^2 + (0.6-2)^2} = \sqrt{0.04 + 1.96} = \sqrt{2.0} \approx 1.41$	$\sqrt{(1.2-8)^2 + (0.6-8)^2} = \sqrt{44.89 + 54.76} = \sqrt{99.65} \approx 9.98$	Cluster 1
6	(9.0, 11.0)	$\sqrt{(9-1)^2 + (11-2)^2} = \sqrt{64 + 81} = \sqrt{145} \approx 12.04$	$\sqrt{(9-8)^2 + (11-8)^2} = \sqrt{1 + 9} = \sqrt{10} \approx 3.16$	Cluster 2

Cluster Assignments after first iteration:

- Cluster 1: Observations 1, 2, 5
- Cluster 2: Observations 3, 4, 6

Step 3: Recompute Centroids

Calculate the mean (average) of features within each cluster:

Cluster 1 (Observations 1, 2, 5):

- $X1_{C1}$: $\frac{1.0 + 1.5 + 1.2}{3} = \frac{3.7}{3} \approx 1.23$
- $X2_{C1}$: $\frac{2.0 + 1.8 + 0.6}{3} = \frac{4.4}{3} \approx 1.47$

Cluster 2 (Observations 3, 4, 6):

- $X1_{C2}$: $\frac{5.0 + 8.0 + 9.0}{3} = \frac{22.0}{3} \approx 7.33$
- $X2_{C2}$: $\frac{8.0 + 8.0 + 11.0}{3} = \frac{27.0}{3} = 9.0$

New centroids:

- C1: (1.23, 1.47)
- C2: (7.33, 9.0)

Step 4: Reassign Data Points Based on New Centroids

Calculate distances again:

Obs	(X1, X2)	Distance to C1	Distance to C2	Assignment

| 1 | (1.0, 2.0) | \(\sqrt{(1-1.23)^2 + (2-1.47)^2} \approx 0.56\) | \(\sqrt{(1-7.33)^2 + (2-9)^2} \approx 7.11\) | Cluster 1 |
| 2 | (1.5, 1.8) | \(\sqrt{(1.5-1.23)^2 + (1.8-1.47)^2} \approx 0.39\) | \(\sqrt{(1.5-7.33)^2 + (1.8-9)^2} \approx 7.62\) | Cluster 1 |
| 3 | (5.0

Frequently Asked Questions

What is the purpose of performing K-means clustering manually on a small dataset?

Performing K-means manually helps in understanding the step-by-step process of clustering, including initialization, assignment, and updating centroids, which enhances conceptual clarity especially with small datasets.

How do you choose initial centroids for K=2 on a small dataset?

Initial centroids can be chosen randomly from the data points or by selecting two points that are well-separated, ensuring a good starting point for the algorithm.

In a small dataset with 6 observations and 2 features, how do you assign observations to clusters during each iteration?

You calculate the distance of each observation to each centroid (usually Euclidean distance) and assign each point to the cluster of the nearest centroid.

How do you update the centroids after assigning observations to clusters?

For each cluster, compute the mean of all observations assigned to that cluster across each feature, and set this mean as the new centroid.

What is the stopping criterion for manual K-means clustering in this example?

The process stops when the centroids no longer change significantly between iterations or after a fixed number of iterations, indicating convergence.

Can you explain how to visualize the clustering process for a small 2-feature dataset?

Yes, by plotting the observations on a 2D scatter plot, marking the centroids, and showing the assignment of points to clusters in each iteration, you can visualize how clusters evolve.

What are potential challenges when performing K-means manually on small datasets?

Challenges include choosing good initial centroids, handling ties in assignments, and ensuring convergence, especially when data points are close or overlapping.

Why is K=2 a common choice for small examples, and how does it simplify manual calculations?

K=2 simplifies the process by reducing complexity, making it easier to track changes in centroids and cluster assignments, which aids in understanding the algorithm's mechanics.

How can understanding manual K-means clustering help in applying it to larger, real-world datasets?

Understanding the manual process builds intuition about cluster formation, helps in selecting initial centroids, and improves interpretation of the algorithm's results on complex data.

Additional Resources

Perform K-means Clustering Manually, With K = 2, On A Small Example With 6 Observations And P = 2 Features

---

Introduction

Clustering algorithms serve as foundational tools in unsupervised machine learning, enabling researchers and data analysts to discover patterns and groupings within datasets without predefined labels. Among these, K-means clustering stands out for its simplicity, efficiency, and widespread applicability. While many practitioners rely on computer algorithms to perform K-means, understanding how to execute the process manually fosters deeper comprehension of its mechanics, assumptions, and limitations.

This article takes a detailed, step-by-step approach to perform K-means clustering manually with K=2 on a small dataset consisting of 6 observations with 2 features each. Such an exercise exemplifies the iterative nature of clustering, the importance of initial centroid selection, and the convergence criteria that guide the algorithm to a stable solution.

Theoretical Foundations of K-means Clustering

Before diving into the manual execution, it is essential to clarify the core concepts underpinning K-means clustering:

- Objective: Partition the dataset into K clusters such that the within-cluster sum of squares (WCSS) — the sum of squared distances between data points and their respective cluster centroids — is minimized.

- Key Components:
- Centroids: The mean position of all points within a cluster.
- Assignment step: Assign each data point to the nearest centroid.
- Update step: Recalculate centroids based on current cluster memberships.
- Convergence: The process repeats until assignments no longer change or a maximum number of iterations is reached.

The Data Set and Initial Setup

Consider the following 6 observations with 2 features (e.g., height and weight, or two measurements of some characteristic):

Observation	Feature 1 (X1)	Feature 2 (X2)
1	1.0	2.0
2	1.5	1.8
3	5.0	8.0
4	8.0	8.0
5	1.0	0.6
6	9.0	11.0

Our goal is to split these observations into K=2 clusters via manual execution of the K-means algorithm.

Step 1: Initialization of Centroids

The initial choice of centroids can significantly influence the final clustering. For simplicity, select two observations at random:

- Centroid 1 (C1): Observation 1 -> (1.0, 2.0)
- Centroid 2 (C2): Observation 4 -> (8.0, 8.0)

Alternatively, other initializations are possible, but for this example, these choices allow us to observe the clustering process clearly.

Step 2: Assignment of Observations to Clusters

Calculate the Euclidean distance from each observation to each centroid:

$$
d = \sqrt{(X_1 - C_{X})^2 + (X_2 - C_{Y})^2}
$$

Observation 1: (1.0, 2.0)

- Distance to C1: $\sqrt{(1.0-1.0)^2 + (2.0-2.0)^2} = 0$

- Distance to C2: $\sqrt{(1.0-8.0)^2 + (2.0-8.0)^2} = \sqrt{7^2 + 6^2} = \sqrt{49 + 36} = \sqrt{85} \approx 9.22$

Observation 2: (1.5, 1.8)

- To C1: $\sqrt{(1.5-1.0)^2 + (1.8-2.0)^2} = \sqrt{0.5^2 + (-0.2)^2} = \sqrt{0.25 + 0.04} = \sqrt{0.29} \approx 0.54$

- To C2: $\sqrt{(1.5-8)^2 + (1.8-8)^2} = \sqrt{(-6.5)^2 + (-6.2)^2} = \sqrt{42.25 + 38.44} = \sqrt{80.69} \approx 8.99$

Observation 3: (5.0, 8.0)

- To C1: $\sqrt{(5.0-1.0)^2 + (8.0-2.0)^2} = \sqrt{4^2 + 6^2} = \sqrt{16 + 36} = \sqrt{52} \approx 7.21$

- To C2: $\sqrt{(5.0-8)^2 + (8.0-8)^2} = \sqrt{(-3)^2 + 0} = 3$

Observation 4: (8.0, 8.0)

- To C1: $\sqrt{(8-1)^2 + (8-2)^2} = \sqrt{7^2 + 6^2} = \sqrt{49 + 36} = \sqrt{85} \approx 9.22$

- To C2: $\sqrt{(8-8)^2 + (8-8)^2} = 0$

Observation 5: (1.0, 0.6)

- To C1: $\sqrt{(1-1)^2 + (0.6-2)^2} = \sqrt{0 + (-1.4)^2} = 1.4$

- To C2: $\sqrt{(1-8)^2 + (0.6-8)^2} = \sqrt{(-7)^2 + (-7.4)^2} = \sqrt{49 + 54.76} \approx \sqrt{103.76} \approx 10.19$

Observation 6: (9.0, 11.0)

- To C1: $\sqrt{(9-1)^2 + (11-2)^2} = \sqrt{8^2 + 9^2} = \sqrt{64 + 81} = \sqrt{145} \approx 12.04$

- To C2: $\sqrt{(9-8)^2 + (11-8)^2} = \sqrt{1^2 + 3^2} = \sqrt{1 + 9} = \sqrt{10} \approx 3.16$

Cluster Assignments After First Iteration:

Observation	Assigned to Cluster 1 or 2?
1	Cluster 1 (distance 0)
2	Cluster 1 (0.54 vs 8.99)
3	Cluster 2 (7.21 vs 3)
4	Cluster 2 (9.22 vs 0)

5	Cluster 1 (1.4 vs 10.19)
6	Cluster 2 (12.04 vs 3.16)

Updated Clusters:

- Cluster 1: Observations 1, 2, 5

- Cluster 2: Observations 3, 4, 6

Step 3: Recalculating Centroids

Calculate the mean of features for each cluster:

Cluster 1 (Obs 1, 2, 5):

- $\text{X1} = \frac{1.0 + 1.5 + 1.0}{3} = \frac{3.5}{3} \approx 1.17$

- $\text{X2} = \frac{2.0 + 1.8 + 0.6}{3} = \frac{4.4}{3} \approx 1.47$

Cluster 2 (Obs 3, 4, 6):

- $\text{X1} = \frac{5.0 + 8.0 + 9.0}{3} = \frac{22.0}{3} \approx 7.33$

- $\text{X2} = \frac{8.0 + 8.0 + 11.0}{3} = \frac{27.0}{3} = 9.0$

New Centroids:

- C1: (1.17, 1.47)

- C2: (7.33, 9.0)

Step 4: Reassignment Based on Updated Centroids

Calculate distances again:

Observation 1: (1.0, 2.0)

- To C1: \(\sqrt{(1-1.17)^2 + (2-1.47)^

Perform K Means Clustering Manually With K 2 On A Small Example With 6 Observations And P 2 Features

Perform K Means Clustering Manually With K 2 On A Small Example With 6 Observations And P 2 Features

Related Articles

- Japan Can Produce 80 Tablets Or 200 Cellphones Per Day. South Korea Can Produce 90 Tablets And 270 Cellphones
- T/F. Eventually, The Scrum Master Becomes Unnecessary When The Development Team Matures Through The Project.
- How Much MORE (in Grams) Of Solute A Would Remain In The Water Layer If A Solution Of 3.0 G Of A Is Dissolved

Back to Home