Data Mining Machine Learning

Md. Abdullah-Al-Kafi
Lecturer
Daffodil International University

Dataset

We have a small dataset of 5 food items with two features: **calories** (in kcal) and **protein content** (in grams).

Food Item	Calories (kcal)	Protein (g)
Apple	52	0.3
Chicken	239	27
Banana	89	1.1
Beef	250	26
Carrot	41	0.9

We will cluster this dataset into $K = 2$ clusters.

Step 1: Initial Centroids

We start by choosing two initial centroids:

$$C_1 = (52, 0.3) \quad \text{(Apple)}, \quad C_2 = (250, 26) \quad \text{(Beef)}$$

Step 2: Assign Each Point to the Nearest Centroid

We use the Euclidean distance formula:

$$\text{distance}(X, C) = \sqrt{(X_{\text{calories}} - C_{\text{calories}})^2 + (X_{\text{protein}} - C_{\text{protein}})^2}$$

Calculating distances between each point and the centroids:

- **Apple** (52, 0.3)

  $$\text{Distance to } C_1 = 0, \quad \text{Distance to } C_2 = \sqrt{(52-250)^2 + (0.3-26)^2} \approx 198.05$$

  Assigned to C_1.

- **Chicken** (239, 27)

  Distance to $C_1 = \sqrt{(239-52)^2 + (27-0.3)^2} \approx 188.52$, Distance to $C_2 = \sqrt{(239-250)^2 + (27-26)^2} \approx 11.05$

  Assigned to C_2.

- **Banana** (89, 1.1)

  Distance to $C_1 = \sqrt{(89-52)^2 + (1.1-0.3)^2} \approx 37.02$, Distance to $C_2 = \sqrt{(89-250)^2 + (1.1-26)^2} \approx 161.92$

  Assigned to C_1.

- **Beef** (250, 26)

$$\text{Distance to } C_1 = 198.05, \quad \text{Distance to } C_2 = 0$$

  Assigned to C_2.

- **Carrot** (41, 0.9)

  Distance to $C_1 = \sqrt{(41-52)^2 + (0.9-0.3)^2} \approx 11.05$, Distance to $C_2 = \sqrt{(41-250)^2 + (0.9-26)^2} \approx 210.00$

  Assigned to C_1.

After this round of assignments, we have the following clusters:

- Cluster 1: Apple, Banana, Carrot
- Cluster 2: Chicken, Beef

Step 3: Update Centroids

Calculating the new centroids by averaging the values in each cluster:

- **New Centroid for Cluster 1** (Apple, Banana, Carrot):

$$\text{Calories} = \frac{52+89+41}{3} = 60.67, \quad \text{Protein} = \frac{0.3+1.1+0.9}{3} = 0.77$$

- **New Centroid for Cluster 2** (Chicken, Beef):

$$\text{Calories} = \frac{239+250}{2} = 244.5, \quad \text{Protein} = \frac{27+26}{2} = 26.5$$

Step 4: Reassign Points

Calculating distances with the updated centroids:

- **Apple** (52, 0.3)

$$\text{Distance to new } C_1 = 8.67, \quad \text{Distance to new } C_2 = 193.33$$

  Assigned to C_1.

- **Chicken** (239, 27)

$$\text{Distance to new } C_1 = 180.22, \quad \text{Distance to new } C_2 = 5.57$$

  Assigned to C_2.

- **Banana** (89, 1.1)

$$\text{Distance to new } C_1 = 28.42, \quad \text{Distance to new } C_2 = 157.29$$

  Assigned to C_1.

- **Beef** (250, 26)

$$\text{Distance to new } C_1 = 190.48, \quad \text{Distance to new } C_2 = 5.57$$

  Assigned to C_2.

- **Carrot** (41, 0.9)

$$\text{Distance to new } C_1 = 19.64, \quad \text{Distance to new } C_2 = 204.08$$

  Assigned to C_1.

Convergence

The assignments did not change, so the algorithm has converged.

Final Clusters

- Cluster 1 (Low-calorie, low-protein foods): Apple, Banana, Carrot
- Cluster 2 (High-calorie, high-protein foods): Chicken, Beef

The final centroids are:

- Centroid 1: (Calories = 60.67, Protein = 0.77)
- Centroid 2: (Calories = 244.5, Protein = 26.5)

Theory

Md. Abdullah-Al-Kafi
Lecturer
Daffodil International University

1. Fundamentals of K-Means

- **Q1:** Explain the objective of the K-Means clustering algorithm. How does K-Means define the "best" clustering?
- **Q2:** Describe the iterative process of the K-Means algorithm. What are the main steps, and what role does each step play in refining the clusters?
- **Q3:** Why is Euclidean distance commonly used in K-Means clustering? What are the implications of using Euclidean distance in terms of the cluster shapes it creates?
- **Q4:** Discuss the initialization of centroids in K-Means. How does centroid initialization affect the outcome of the clustering process?
- **Q5:** Define what is meant by "convergence" in K-Means. How can we determine that K-Means has converged?

2. Theoretical Properties and Assumptions

- **Q6:** What are the primary assumptions made by K-Means clustering regarding the data? How do these assumptions impact the types of datasets for which K-Means is most suitable?
- **Q7:** K-Means aims to minimize the within-cluster sum of squares (WCSS). Write down the mathematical formula for WCSS and explain each term.
- **Q8:** Is K-Means guaranteed to find the global minimum of the objective function? Why or why not?
- **Q9:** Explain why K-Means clustering is sensitive to outliers. How do outliers affect the centroids and the resulting clusters?
- **Q10:** Discuss the computational complexity of K-Means. How does the complexity scale with respect to the number of data points, features, and clusters?

3. Practical Considerations and Challenges

- **Q11:** What is the "K-Means++" initialization method, and how does it improve upon random initialization?
- **Q12:** Describe how the choice of K (number of clusters) affects the clustering results in K-Means. What are some methods used to select an appropriate K value?
- **Q13:** How does K-Means clustering handle high-dimensional data? Discuss any potential challenges or limitations when applying K-Means in high-dimensional spaces.
- **Q14:** Explain what is meant by "cluster inertia" or "within-cluster sum of squares." How can this metric be used to assess the quality of the clusters?
- **Q15:** Describe the limitations of K-Means for clustering non-spherical clusters. What kinds of data distributions can lead to suboptimal clustering when using K-Means?

4. Extensions and Variants of K-Means

- **Q16:** What is the K-Medoids algorithm, and how does it differ from K-Means? In what scenarios might K-Medoids be preferred over K-Means?
- **Q17:** Explain how fuzzy C-Means extends the K-Means algorithm. What are the main differences between hard clustering in K-Means and soft clustering in fuzzy C-Means?
- **Q18:** Describe the Mini-Batch K-Means algorithm. Why is it considered suitable for large datasets, and how does it compare to standard K-Means in terms of accuracy and efficiency?
- **Q19:** What is "bisecting K-Means"? How does this hierarchical approach address certain limitations of standard K-Means?
- **Q20:** Compare K-Means clustering with other clustering algorithms, such as DBSCAN and hierarchical clustering. What are the strengths and weaknesses of K-Means relative to these methods?

5. Applications and Real-World Scenarios

- **Q21:** Discuss a practical application of K-Means clustering in customer segmentation. How might a business benefit from segmenting customers into clusters?
- **Q22:** Explain how K-Means clustering could be applied to image compression. Describe the role of cluster centroids in reducing the image's color palette.
- **Q23:** Describe the use of K-Means in anomaly detection. How can K-Means clusters be leveraged to identify outliers or anomalies in data?
- **Q24:** How can K-Means clustering be used in recommendation systems? Describe a use case where K-Means can help group items or users based on similarity.
- **Q25:** In the context of bioinformatics, how can K-Means clustering assist in gene expression analysis? Describe the potential benefits of clustering genes with similar expression patterns.

6. Interpretation and Evaluation of Clusters

- **Q26:** Define "cluster interpretability" in the context of K-Means clustering. What challenges might arise when trying to interpret the resulting clusters?
- **Q27:** What is the "elbow method," and how is it used to select the optimal number of clusters in K-Means? What are its limitations?
- **Q28:** Explain the "Silhouette Score" and how it can be used to evaluate the quality of K-Means clusters. How does the score provide insight into cluster cohesion and separation?
- **Q29:** Describe how cross-validation can be applied to assess the performance of K-Means clustering. What metrics can be used, and what challenges exist in using cross-validation for unsupervised learning?
- **Q30:** In the context of K-Means clustering, what is a "boundary point"? How can boundary points affect the stability of clusters, and how might they be addressed?