Discovery Learning: When K-Means Fails

Hands-on Exploration of Clustering Limitations

Machine Learning for Smarter Innovation - Week 1

Learning Objectives

By completing this discovery exercise, you will:

- Identify scenarios where K-means clustering fails
- Understand why these failures occur
- Discover which alternative algorithms to use
- Develop intuition for algorithm selection

Part 1: Observation Exercise

Look at the chart showing 6 different data patterns. For each pattern, answer:

Pattern Analysis

1. **Crescent Shapes (Technology Evolution Chains)**
   - What shape do you see? _______________
   - Why might circles fail here? _________________________
   - Real-world example: ______________________________
2. **Nested Circles (Core vs Peripheral Innovation)**
   - Describe the structure: _________________________
   - What's the K-means assumption violated? ____________________
   - Business analogy: ______________________________
3. **Chain Patterns (Innovation Pipelines)**
   - What's the data distribution? _________________________
   - K-means draws what shape? _________________________
   - Industry example: ______________________________

Part 2: Prediction Challenge

Before looking at the solutions, predict how K-means would cluster these patterns:

Your Predictions

Pattern 4: Different Densities
Draw where you think K-means would split:

Pattern 5: With Outliers
Circle the outliers K-means would misassign:

Think About

- Does K-means handle density differences?
- Can K-means ignore outliers?
- What assumptions does K-means make?
- When should you NOT use K-means?

Part 3: Algorithm Matching

Match each problematic pattern with its best alternative algorithm:

Pattern Type	Match	Algorithm Options
Non-spherical shapes	___	A. Hierarchical Clustering
Different densities	___	B. DBSCAN
Connected components	___	C. Gaussian Mixture Model
With outliers	___	D. DBSCAN
Nested structures	___	E. Spectral Clustering
Elongated clusters	___	F. GMM with full covariance

Part 4: Real-World Application

Consider your own innovation data or business problem:

Your Scenario

1. Describe your data: ______________________________
2. Expected cluster shapes: ______________________________
3. Potential outliers?: ______________________________
4. Density variations?: ______________________________
5. Your algorithm choice: ______________________________
6. Why this choice?: ______________________________

Key Takeaways

- K-means assumes **spherical** clusters of **similar size**
- K-means is sensitive to **outliers** and **initialization**
- DBSCAN finds **arbitrary shapes** and handles **noise**
- Hierarchical clustering shows **relationships** at multiple scales
- GMM allows **overlapping** clusters with **soft assignments**

Challenge Question

If you had customer behavior data with clear weekday vs. weekend patterns, seasonal variations, and some unusual one-time events, which clustering approach would you use and why?