Discovery Learning 2: The Feature Detective

What Makes Things Similar?

Machine Learning for Smarter Innovation - Pre-Lecture Activity

Learning Objectives

By completing this activity, you will discover:

- How different features lead to different groupings
- Why feature selection is critical for clustering
- What happens when we have too many features (curse of dimensionality)

The Startup Dataset

You're a venture capitalist analyzing 20 startups. Each has multiple features - but which ones matter for grouping?

Startup Name	Industry	Employees	Funding ($M)	Age (yrs)	Growth (%/yr)
TechnoAI	AI/ML	45	12.5	2	150
GreenLeaf	CleanTech	28	5.2	3	85
DataFlow	Analytics	67	22.0	5	110
HealthHub	HealthTech	34	8.7	2	125
EduSmart	EdTech	23	3.5	1	200
FinanceForward	FinTech	89	45.0	6	95
CloudNine	Cloud	56	18.0	4	105
BioGenix	BioTech	42	15.0	3	90
CyberShield	Security	38	9.8	2	140
MarketMind	MarTech	19	2.1	1	175
RoboWorks	Robotics	51	20.0	4	100
SocialSphere	Social Media	72	28.0	5	80
GameOn	Gaming	41	11.0	3	120
FoodTech Plus	FoodTech	26	4.5	2	135
AutoDrive	Autonomous	63	35.0	5	88
VirtualSpace	VR/AR	31	7.2	2	155
LogisticsPro	Logistics	47	13.5	4	98
AgriTech Now	AgTech	22	3.0	1	180
RetailRevolution	E-commerce	58	24.0	6	75
BlockChainBase	Blockchain	36	10.0	3	115

Exercise 1: Single Feature Clustering

Task A: Cluster by Industry Only

Group the startups into 3 clusters based ONLY on their industry type. Write the startup names in each cluster:

Cluster 1: Tech/Software

Cluster 2: Life/Science

Cluster 3: Business/Service

Task B: Cluster by Funding Only

Now group the SAME startups into 3 clusters based ONLY on funding amount:

- Low funding: ¡ $10M
- Medium funding: $10M - $25M
- High funding: ¿ $25M

Low: ¡ $10M	**Medium: $10-25M**	**High: ¿ $25M**

Discovery Question 1

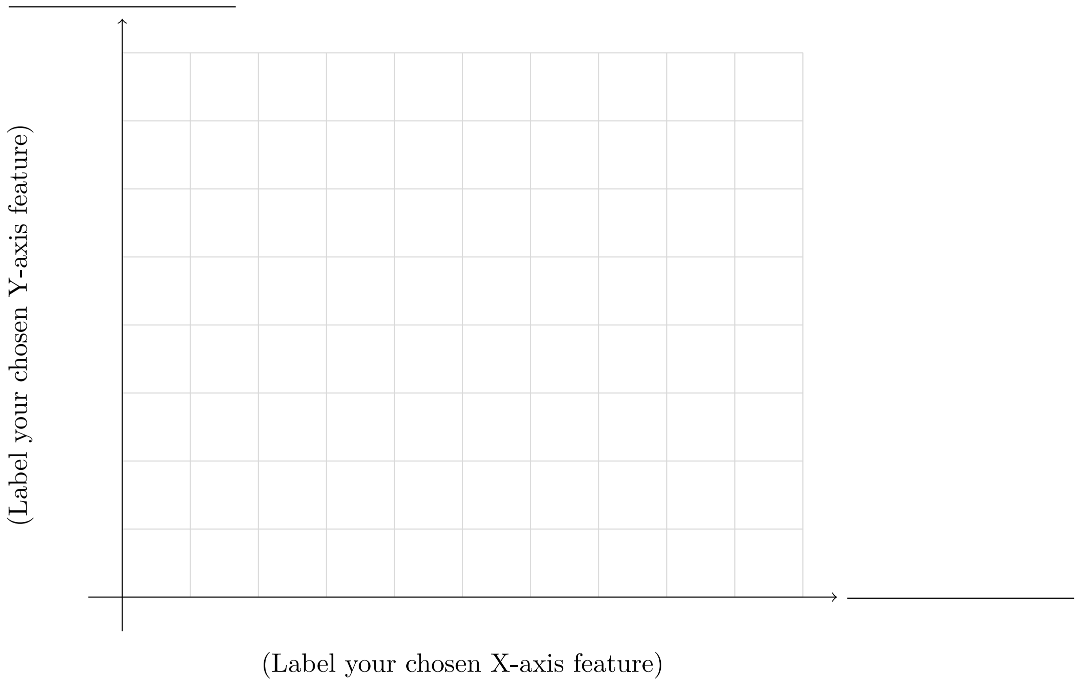
Did the same companies end up together in both clustering approaches?
Give an example of two companies that were together in Task A but separated in Task B:
Company 1: ________________ Company 2: ________________

Why did this happen? ____

Exercise 2: Multi-Feature Clustering

Task C: Your Choice - Combine Features

Create your own 3 clusters using ANY combination of features. Plot on the grid below:



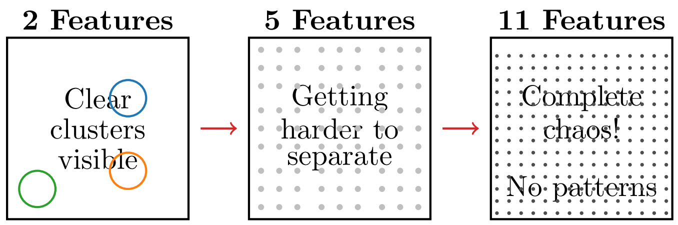
My clustering criteria: ______________________________

Exercise 3: The Curse of Dimensionality

Task D: Feature Overload

Imagine adding 5 more features to each startup:

- Customer satisfaction score (1-10)
- Office locations (1-20)
- Product lines (1-50)
- Patents filed (0-100)
- Social media followers (100-1M)



The Curse of Dimensionality

As we add more features, the "distance" between any two startups becomes less meaningful. With 11 features, every startup seems equally far from every other startup!

Reflection: Feature Importance Ranking

If you could only use 3 features to cluster startups for investment decisions, which would you choose?

Most Important: 1. ____________

Important: 2. ____________

Useful: 3. ____________

Why these three? ____

____

Final Discovery Questions

1. **Pattern Recognition:** Looking at your different clusterings, which feature created the most meaningful groups? Why?

   ____

   ____

2. **Scale Challenge:** With 20 startups and 6 features, you could manage this manually. What about 10,000 startups with 50 features each?

   ____

   ____

3. **Missing Data:** What if some startups didn't report all features (e.g., private funding amounts)? How would this affect clustering?

   ____

   ____

Prepare for Next Class

You've discovered that clustering depends entirely on which features we choose and how we weight them. In our next lecture, you'll learn how machine learning algorithms can:

- Automatically find the most important features
- Handle hundreds of dimensions
- Discover patterns humans would never see

Think about: How would you teach a computer to decide which features matter most?