

Unsupervised Learning - Basic Handout

Machine Learning for Smarter Innovation

1 Unsupervised Learning - Basic Handout

Target Audience: Beginners with no ML background **Duration:** 25 minutes reading **Level:** Basic (no math required)

1.1 What Is Unsupervised Learning?

Think of it like organizing a messy closet. Nobody tells you which items go together - you discover patterns yourself based on colors, seasons, or occasions.

Supervised vs Unsupervised: - **Supervised:** Teacher gives you answers to learn from - **Unsupervised:** No answers - find patterns on your own

1.2 Real-World Examples

1.2.1 Customer Segmentation

- Group customers by behavior (not demographics)
- Discover: “Weekend browsers” vs “Quick buyers” vs “Sale hunters”

1.2.2 Anomaly Detection

- Find unusual transactions (fraud detection)
- Spot equipment failures before they happen

1.2.3 Document Organization

- Group news articles by topic automatically
- Organize customer feedback into themes

1.2.4 Recommendation Systems

- “Customers like you also bought. . .”
 - Spotify’s Discover Weekly playlists
-

1.3 Three Main Types

1.3.1 1. Clustering

Goal: Group similar items together - K-means: You choose number of groups - DBSCAN: Algorithm finds natural groupings - Hierarchical: Creates tree of relationships

Use when: You want to find natural segments

1.3.2 2. Dimensionality Reduction

Goal: Simplify complex data while keeping patterns - PCA: Find most important features - t-SNE: Create visual maps of high-dimensional data

Use when: Too many features to analyze

1.3.3 3. Association Rules

Goal: Find items that appear together - Market basket: “People who buy X also buy Y” - Example: Diapers and beer on Friday evenings

Use when: Finding hidden relationships

1.4 When to Use Unsupervised Learning

1.4.1 Good Fit:

- Exploring new datasets
- No labeled training data available
- Looking for hidden patterns
- Reducing data complexity
- Generating features for other models

1.4.2 Poor Fit:

- Need specific predictions
 - Have clear right/wrong answers
 - Very small datasets (under 100 items)
 - Need highly interpretable results
-

1.5 Quick Start Checklist

1.5.1 Before You Begin:

- ☐ Define your exploration goal
- ☐ Ensure data is cleaned and scaled
- ☐ Remove or handle missing values
- ☐ Identify which features to include

1.5.2 Your First Project:

- ☐ Start with K-means clustering (K=3)

- ☐ Visualize results with scatter plots
- ☐ Try different K values (2-10)
- ☐ Name your discovered groups

1.5.3 Validate Results:

- ☐ Do groups make business sense?
 - ☐ Are groups distinct and meaningful?
 - ☐ Can you take action on findings?
-

1.6 Common Pitfalls

1. **Forgetting to scale data:** Features with larger values dominate
 2. **Choosing K randomly:** Use elbow method or silhouette score
 3. **Ignoring outliers:** They can distort cluster centers
 4. **Over-interpreting:** Not all patterns are meaningful
 5. **No domain validation:** Always check with experts
-

1.7 Tools for Beginners

1.7.1 No-Code Options:

- **Orange3:** Visual drag-and-drop ML
- **KNIME:** Workflow-based analytics
- **Tableau:** Built-in clustering

1.7.2 Python Libraries:

- **scikit-learn:** Standard ML library
 - **pandas:** Data manipulation
 - **matplotlib/seaborn:** Visualization
-

1.8 Key Terms

Term	Simple Definition
Clustering	Grouping similar items
K-means	Choose K groups, find centers
DBSCAN	Density-based grouping
PCA	Reduce dimensions, keep variance
Silhouette	Quality score (-1 to 1, higher better)
Elbow method	Graph to choose optimal K

1.9 Next Steps

1. **Try:** Cluster a simple dataset (iris, wine)
2. **Explore:** Visualize your clusters
3. **Compare:** Try $K=2, 3, 4, 5$ and compare
4. **Read:** Intermediate handout for implementation details

Unsupervised learning is about discovery. Let the data tell its story, then use your expertise to interpret what it means.