UNSUPERVISED LEARNING

Cluster Analysis & Anomaly Detection

Class Notes

Data Analytics, Predictive Analytics & Data Mining

Gannon University | Dahlkemper School of Business

Spring 2026

Prepared by: Dr. Benyawarath “Yaa” Nithithanatchinnapat

What You Will Learn

This module covers unsupervised learning techniques that help you discover hidden patterns in data without a target variable. You will learn to segment data into meaningful groups using cluster analysis and identify unusual observations using anomaly detection. These techniques are essential for customer segmentation, fraud detection, quality control, and exploratory data analysis.

Contents

1. The Big Picture: What is Unsupervised Learning?

So far in this course, we have focused on supervised learning, where we predict a known target variable (like whether a customer will churn, or how much revenue a product will generate). Unsupervised learning flips the script. There is no target. No labels. No right answer to check against.

Instead, the algorithm explores the data on its own and looks for structure: groups of similar observations, unusual patterns, or hidden relationships. Think of it as letting the data tell its own story.

Supervised vs. Unsupervised: The Key Difference

Supervised learning asks: "Based on what I know, what will happen next?" Unsupervised learning asks: "What patterns exist in this data that I haven't seen yet?" One predicts. The other discovers.

Why Does This Matter for Business?

Unsupervised learning is everywhere in business, even if people do not always call it that. Here are a few common applications:

- Customer segmentation: Grouping customers by purchasing behavior so marketing can target each group differently
- Fraud detection: Identifying transactions that look nothing like normal activity
- Quality control: Flagging manufactured products that deviate from specifications
- Healthcare: Finding subgroups of patients who respond differently to treatments
- Network security: Detecting unusual patterns in server logs that could signal a cyberattack

Pro Tip: When a stakeholder says "We want to understand our customers better" or "Help us find what doesn't belong," they are usually describing an unsupervised learning problem.

The Two Techniques We Will Cover

This module focuses on two of the most widely used unsupervised techniques:

Technique	Core Question	Business Example
Cluster Analysis	"Who are they?" (What natural groups exist?)	Segmenting 50,000 retail customers into 4-6 actionable groups
Anomaly Detection	"What doesn't belong?" (What looks unusual?)	Flagging 23 suspicious credit card transactions out of 1 million

2. Cluster Analysis: Finding Natural Groups

Cluster analysis is about discovering natural groupings in your data. The goal is to put similar observations together and keep dissimilar ones apart. No one tells the algorithm what the groups are; it figures them out from the patterns in the data.

2.1 Core Concepts

What Makes a Good Cluster?

Two properties define a good set of clusters:

- **High intra-cluster similarity:** Members within the same cluster should be very similar to each other.
- **High inter-cluster dissimilarity:** Members in different clusters should be clearly different from each other.

Think about customer segments at a grocery store. A useful segmentation might separate "budget-conscious families" from "health-focused singles" from "premium foodies." Each group is internally consistent, and the groups are meaningfully distinct.

Distance: How We Measure Similarity

Clustering algorithms need a way to measure how "close" or "far apart" two data points are. The most common distance measures include:

Distance Metric	How It Works	Best Used When
Euclidean	Straight-line distance between two points	Variables are on similar scales and continuous
Manhattan	Sum of absolute differences (city-block distance)	High-dimensional data or when outliers are a concern
Cosine	Angle between two vectors (ignores magnitude)	Text data or when direction matters more than size

Why Standardization Matters

Imagine clustering customers using income (ranging from $20,000 to $200,000) and age (ranging from 18 to 80). Without standardization, income would dominate the distance calculation simply because its numbers are bigger. Always standardize your variables before clustering so each feature contributes equally.

2.2 K-Means Clustering

K-Means is the most widely used clustering algorithm. It is fast, intuitive, and works well for many business problems.

How K-Means Works

The algorithm follows a simple iterative process:

1. Choose K (the number of clusters you want).
2. Randomly place K "centroids" (cluster centers) in the data space.
3. Assign every data point to the nearest centroid.
4. Recalculate each centroid as the average of all points assigned to it.
5. Repeat steps 3-4 until the assignments stop changing (convergence).

Pro Tip: Think of centroids as the "prototype" or "most typical member" of each cluster. They represent the average values for all the attributes in that group.

K-Means Assumptions

K-Means works best when:

- Clusters are roughly spherical (round-ish groups, not elongated or irregular shapes)
- Clusters are roughly similar in size
- All variables are continuous and standardized
- You have a reasonable idea of how many clusters to expect

How to Choose K: The Elbow Method

One of the biggest challenges with K-Means is deciding how many clusters to create. The Elbow Method helps:

6. Run K-Means for K = 1, 2, 3, ... up to some reasonable number.
7. For each K, record the total within-cluster sum of squares (WCSS), which measures how tightly packed the clusters are.
8. Plot K on the x-axis and WCSS on the y-axis.
9. Look for the "elbow," the point where adding more clusters stops significantly reducing WCSS.

Business Sense Trumps Math

The elbow method gives you a statistical suggestion, but the final decision should make business sense. Having 47 customer segments is mathematically possible but operationally useless. Aim for 3-8 groups that your stakeholders can actually act on.

2.3 Hierarchical Clustering

Hierarchical clustering builds a tree of clusters (called a dendrogram) rather than requiring you to specify K upfront.

Two Approaches

- **Agglomerative (bottom-up):** Start with every observation as its own cluster. Merge the two closest clusters. Repeat until everything is in one big cluster.
- **Divisive (top-down):** Start with all observations in one cluster. Split the least cohesive cluster. Repeat until every observation is on its own.

Agglomerative is far more common in practice. The dendrogram it produces is a powerful visual tool: you can "cut" it at different heights to get different numbers of clusters.

Advantages Over K-Means

- No need to pre-specify K; you can decide after examining the dendrogram
- Can handle non-spherical cluster shapes
- Produces a visual hierarchy that shows relationships between clusters

Disadvantages

- Computationally expensive for large datasets (not ideal for 100,000+ observations)
- Once two clusters merge, the decision cannot be undone

2.4 Evaluating Cluster Quality

Since there is no target variable, evaluating clusters is trickier than evaluating a classification or regression model. Here are the main approaches:

Method	What It Measures	Interpretation
Silhouette Score	How similar each point is to its own cluster vs. the nearest other cluster	Ranges from -1 to +1. Higher is better. Above 0.5 is generally good.
Elbow Method (WCSS)	Total within-cluster variation	Look for the elbow; lower WCSS means tighter clusters.
Business Validation	Whether the clusters make practical sense	Can stakeholders name and act on each cluster?

Pro Tip: The best evaluation is often the simplest: show your clusters to a business stakeholder and ask, "Do these groups make sense? Can you do something different for each one?" If the answer is yes, your clustering is working.

2.5 Cluster Analysis in SAS Viya

In SAS Visual Analytics, cluster analysis is available as an interactive object:

10. Add a "Cluster" object to your Visual Analytics report canvas.
11. Assign the measure variables you want to use for clustering (for example, age, income, purchase frequency).
12. SAS will generate clusters and display a parallel coordinates plot showing how each cluster differs across variables.
13. You can derive a Cluster ID variable to use in future analysis or modeling.

In SAS Model Studio, clustering can be included as a pipeline node for more advanced workflows, including automatic data preprocessing and comparison with other techniques.

3. Anomaly Detection: Finding What Doesn't Belong

Anomaly detection (also called outlier detection) is about identifying data points that are significantly different from the majority. These unusual observations can signal important events: fraud, equipment failure, data entry errors, or emerging trends.

3.1 Why Anomaly Detection Is Different

Anomaly detection sits at an interesting intersection of supervised and unsupervised learning. It is technically unsupervised because in most real-world cases, you do not have labeled examples of "normal" vs. "anomalous" data. But it behaves a bit like classification because the output is often a prediction: "Is this observation typical (1) or anomalous (0)?"

The One-Class Problem

In many fraud detection scenarios, you have millions of legitimate transactions but only a handful of confirmed fraud cases. This is called a one-class problem: you have plenty of examples of what "normal" looks like, but very few (or zero) examples of what "abnormal" looks like. Anomaly detection algorithms learn the boundaries of "normal" and flag anything outside those boundaries.

3.2 Types of Anomalies

Not all anomalies are the same. Understanding the type helps you choose the right approach:

Type	Description	Example
Point Anomaly	A single data point that is far from the rest of the data	A credit card charge of $15,000 when the average is $50
Contextual Anomaly	A point that is anomalous in a specific context but normal otherwise	A temperature of 90°F is normal in July but anomalous in January
Collective Anomaly	A group of data points that are anomalous together but individually might look normal	A series of small transactions from different locations within 10 minutes

3.3 Common Approaches

Statistical Methods

The simplest approach uses basic statistics. If a value falls more than 2-3 standard deviations from the mean, it may be anomalous. Z-scores and the Interquartile Range (IQR) method are classic examples. These are easy to implement but only work well for simple, normally distributed data.

Distance-Based Methods

These methods use the same distance concepts from clustering. Points that are far from their neighbors are flagged as anomalies. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular algorithm that simultaneously clusters data and identifies outliers as noise points.

One-Class SVM

Support Vector Machine (SVM) models can be trained on only "normal" data. The algorithm learns a boundary around the normal observations. New data points that fall outside this boundary are flagged as anomalous. This is one of the most powerful approaches when labeled anomaly data is scarce.

In SAS and Oracle ML, One-Class SVM produces:

- **Prediction = 1:** The case is typical (normal).
- **Prediction = 0:** The case is anomalous.
- A probability score indicating confidence in the classification.

Clustering-Based Anomaly Detection

You can also use clustering itself to detect anomalies. After building a cluster model, points that do not fit well into any cluster (low cluster membership probability) may be anomalous. This is a practical two-for-one approach: you get both customer segments and anomaly flags from the same analysis.

Pro Tip: In practice, the best anomaly detection systems combine multiple methods. A bank might use statistical rules for obvious fraud (charges over $10,000), clustering for pattern-based detection, and One-Class SVM for subtle anomalies.

3.4 Business Applications of Anomaly Detection

Industry	Application	What Gets Flagged
Financial Services	Fraud detection	Unusual transaction patterns, suspicious account activity
Healthcare	Clinical monitoring	Abnormal patient vital signs, unexpected lab results
Manufacturing	Quality control	Products outside specification tolerances, equipment sensor anomalies
Cybersecurity	Intrusion detection	Unusual network traffic, unauthorized access patterns
Retail	Inventory management	Unexpected demand spikes, shrinkage patterns
Insurance	Claims fraud	Claims with unusual characteristics compared to typical claims

3.5 Setting the Outlier Rate

Most anomaly detection algorithms require you to set an expected outlier rate. This is your best estimate of what percentage of your data is likely to be anomalous. For example:

- Credit card fraud: roughly 0.1% to 1% of transactions
- Manufacturing defects: typically 1% to 5%
- Insurance fraud: roughly 5% to 10% of claims

Getting this rate right matters. Set it too low and you miss real anomalies. Set it too high and you generate too many false alarms, which erodes trust in the system.

The False Alarm Trade-Off

In anomaly detection, you are constantly balancing two risks. Missing a real anomaly (a false negative) can be costly: undetected fraud, missed equipment failure. But flagging too many normal observations as anomalies (false positives) wastes investigator time and can cause "alert fatigue" where people start ignoring the alerts altogether.

4. Connecting the Dots: The Business Framework

4.1 Supervised vs. Unsupervised: A Decision Framework

How do you know when to use unsupervised learning? Ask yourself these questions:

Question	If Yes...	If No...
Do I have a specific target I want to predict?	Use supervised learning (regression, classification)	Consider unsupervised learning
Do I want to discover groups I don't know about yet?	Cluster analysis	Maybe you need something else
Do I need to find unusual observations?	Anomaly detection	Standard analysis may suffice
Do I have labeled training data?	Supervised is likely best	Unsupervised may be your only option

4.2 The PAIR Framework for Unsupervised Learning

Even though unsupervised learning has no target variable, you still need to connect your analysis to business decisions. Use the PAIR framework:

- **P - Pattern:** What pattern did the algorithm discover? (e.g., 5 distinct customer segments)
- **A - Action:** What will the business do differently? (e.g., create targeted marketing campaigns for each segment)
- **I - Impact:** What is the expected business value? (e.g., 15% increase in email click-through rate)
- **R - Risk:** What could go wrong? (e.g., segments may shift over time and need refreshing)

Pro Tip: When presenting cluster analysis or anomaly detection results to stakeholders, always lead with the action and impact. They care less about the algorithm and more about what to do next.

4.3 Common Pitfalls to Avoid

- **Skipping standardization:** This is the number one mistake in clustering. Variables with larger ranges will dominate the distance calculations.
- **Choosing K without justification:** "I picked 5 because it seemed reasonable" is not a valid approach. Use the elbow method, silhouette scores, and business logic together.
- **Forgetting to profile clusters:** Clusters are only useful if you can describe who is in each group and why they are different.

- **Treating anomaly scores as certainties:** A flagged anomaly needs investigation, not automatic action. It is a lead, not a verdict.
- **Ignoring data quality:** Missing values and outliers affect unsupervised methods even more than supervised ones. Clean your data thoroughly.

5. Key Terms & Definitions

Term	Definition
Unsupervised Learning	Machine learning techniques that discover patterns in data without a predefined target variable.
Cluster Analysis	The process of grouping observations into subsets (clusters) where members within a cluster are more similar to each other than to those in other clusters.
K-Means	A distance-based clustering algorithm that partitions data into K clusters by minimizing the within-cluster sum of squares.
Centroid	The center point of a cluster, calculated as the mean of all observations assigned to that cluster.
Hierarchical Clustering	A clustering method that builds a tree-like structure (dendrogram) showing nested groupings from individual observations up to one large cluster.
Dendrogram	A tree diagram that visualizes the arrangement of clusters produced by hierarchical clustering.
Elbow Method	A technique for choosing the optimal number of clusters by plotting within-cluster sum of squares against the number of clusters and looking for a bend.
Silhouette Score	A measure of how similar an object is to its own cluster compared to other clusters, ranging from -1 to +1.
Anomaly Detection	The process of identifying data points that differ significantly from the majority of the data.
One-Class SVM	A Support Vector Machine trained on only "normal" data that identifies observations outside the learned boundary as anomalies.
Outlier Rate	The expected percentage of anomalous observations in the data, used to calibrate anomaly detection algorithms.
Standardization	Transforming variables to have a mean of 0 and standard deviation of 1 so all features contribute equally to distance calculations.
WCSS	Within-Cluster Sum of Squares; the total variance within each cluster. Lower WCSS means tighter, more cohesive clusters.
PAIR Framework	A business analytics framework: Pattern, Action, Impact, Risk. Used to connect analytical findings to business decisions.

6. Self-Check Questions

Test your understanding of the key concepts from this module. Try answering these before reviewing the material above.

14. What is the fundamental difference between supervised and unsupervised learning?

15. Why is standardization critical before running K-Means clustering?

16. Describe the K-Means algorithm in your own words. What are the key steps?

17. What is the Elbow Method, and what does it help you decide?

18. A marketing director asks you to segment their 100,000 customers. The elbow method suggests K=12, but the director says "We can only manage 4 campaigns." What do you do?

19. Name two advantages of hierarchical clustering over K-Means.

20. Explain the difference between a point anomaly and a contextual anomaly. Give an example of each.

21. Your anomaly detection model flags 500 out of 100,000 credit card transactions. After investigation, only 50 are actual fraud. Is this a good model? Why or why not?

22. How can clustering be used as a tool for anomaly detection?

23. Using the PAIR framework, describe how a retailer could use cluster analysis to improve holiday marketing.

End of Class Notes