Unsupervised Learning

Clustering

Anomaly Detection

Dimensionality Reduction

Julia Lenc

Content

General

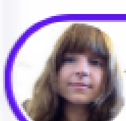
- What is Unsupervised Learning
- When to use
- Methods overview

Review of methods and their business applications

- Clustering (K-Means, Hierarchical, DBSCAN, GMM, Mean Shift, SOM)
- Dimensionality reduction (PCA, SOM, Autoencoders)
- Anomaly detection (Autoencoders)

Guidance

- Choosing the method based on business priorities and available data
- 6 critical things to remember
- Links to free, beginner-friendly tutorials



Julia Lenc

What is Unsupervised Learning?

Definition

Unsupervised learning is a group of methods that find hidden patterns or structures in data without pre-defined labels or outcomes. It helps to discover natural groupings, reduce data complexity and identify opportunities or risks.

Areas

💡 Pre-development: segmentation, portfolio strategy

🎁 Development: feature optimization, resource allocation

🛒 Go-to-Market: personalized marketing, distribution channels, pricing strategies by customer segment

💜 Post launch: continutions improvement

Julia Lenc

When to use?

🔍 Clustering = targeting, personalization

Finding natural groupings in data where items within the same group are similar to each other and different from items in other groups.

📊 Dimensionality reduction = feature shock prevention

Simplifying complex data by reducing the number of variables while preserving important information.

⚠️ Anomaly detection = data quality checks

Identifying data points that don't conform to expected patterns or behaviors.

Julia Lenc

Methods

🔍 Clustering

- K-means clustering
- Hierarchical clustering
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Gaussian Mixture Models (GMM)
- Mean Shift
- Self-organizing maps (SOM)

📊 Dimensionality reduction

- Principal Component Analysis (PCA)
- SOM
- Autoencoders

⚠️ Anomaly detection

- DBSCAN
- GMM
- Autoencoders

Julia Lenc

K-Means Clustering

Purpose

Groups similar data points together based on their features. Predefined number of clusters!

Business applications

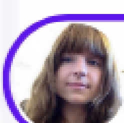
- Customer and product segmentation
- Recommenders

Example

Group customers based on purchase frequency and spending -> segments “Budget Shoppers”, “Frequent Buyers”, “Big Spenders” -> personalized marketing

How it works

- Clusters are based on a distance to a central point (centroid). Each point belongs to nearest centroid.
- Sensitive to outliers and initial clusters placement.



Julia Lenc

Hierarchical Clustering

Purpose

Groups similar data points together based on their features. No need to predefine the number of clusters.

Business applications

- Functional segmentation (“Jobs to be done”)
- Latent customer needs

Example

Cluster customers based on “Jobs to be done” -> identify segments like “Performance”, “Outdoor”, “Casual” -> tailor product offers and marketing.

How it works

- Clusters form step-by-step in a tree structure (dendrogram). Agglomerative clustering - up from individual points. Divisive clustering (splitting down).

Julia Lenc

DBSCAN

Purpose

Clusters data points based on density. Detects outliers. No need to predefine the number of clusters.

Business applications

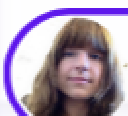
- Geospatial clustering (identifying hubs, hotspots)
- Fraud/anomaly detection

Example

Identify hiking trail hotspots from GPS data -> segment as “Busy Trails”, “Remote Areas” and “Noise (Outliers)” → improved trail recommendations and visitor guidance.

How it works

- Groups data points based on dense regions separated by sparse ones.
- Requires two parameters: the neighborhood radius and minimum points for forming a dense cluster.



Julia Lenc

Mean Shift

Purpose
Clusters data points by iteratively shifting towards areas of higher density (modes). Detects outliers. No need to predefine the number of clusters.

Business applications
- Optimizing store locations
- Recommenders based on “Jobs to be done”

Example (Starbucks case)
Find optimal store locations by grouping customer density in urban areas -> “High-Density Regions” and “Emerging Markets” -> guide coffee shops placement.

How it works
- Computes the "center of mass" for each cluster, moving data points iteratively toward the nearest high-density region. Utilizes a kernel function to estimate the density around each point. Clustering stops when the cluster centers stabilize.

Julia Lenc

GMM

Purpose

Models the data as a mixture of several Gaussian distributions, allowing for soft clustering where data points can belong to multiple clusters with different probabilities.

Business applications

Blended targeting via combining primary and seconday segment (hybrid content, lifecycle or seasonal targeting)

Example (online retailers)

Assign to customers probabilities to fit each segment (e.g., 70% "Frequent Buyers" and 30% "Window Shoppers") to define primary and secondary target → optimize seasonal campaigns.

How it works

- Assumes data is generated from a combination of multiple Gaussian distributions. Estimates the parameters of each Gaussian using the Expectation-Maximization (EM) algorithm. Assigns probabilities of cluster membership to each data point based on its distance from each Gaussian's center.

Julia Lenc

SOM

Purpose

Visualizes high-dimensional data in a simplified 2D map while preserving topological relationships, enabling insight discovery and dimensionality reduction.

Business applications

- Clustering: churn reduction via hyper-personalized offers
- Dimensionality reduction: simplified customer data for better analysis and strategy planning.

Example (telecom)

Creating micro segments for hyper-personalization, (“Mobile Gaming Enthusiasts”, “Work-from-Anywhere Professionals”) instead of simple segmentation (Low / Medium / Heavy users).

How it works

- Transforms multidimensional customer data (e.g., app usage, call patterns, demographics) into a manageable visual output. Uses unsupervised learning to cluster customers while preserving their relationships. Finds hidden correlations and reduces dimensionality without losing context.

Julia Lenc

PCA

Purpose

Reduces high-dimensional data into a smaller set of components that retain the most important information, enabling easier analysis, vizualization and preventing feature shock.

Business applications

- Streamlined product design (most impactful features).
- Simplifying marketing campaigns to avoid feature shock.

Example (software)

Identify key features for product design and core areas for positioning, e.g., ease of use and speed → develop software and its campaign around these features, e.g.” Faster, easier, smarter”.

How it works

- Finds the directions (principal components) in which the data varies the most. Projects the data onto these components to explain most of the variability with fewer dimensions. Enables companies to focus on what matters most by trimming irrelevant or redundant information.

Julia Lenc

Autoencoders

Purpose

Leverages neural networks to learn a compressed representation of input data for dimensionality reduction and identifies patterns that deviate significantly from expected for anomaly detection.

Business applications

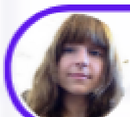
- Dimensionality reduction: extracting key features from financial data to simplify risk modeling .
- Anomaly detection: identifying fraudulent transactions.

Example (banking)

Flag anomalies in high-dimensional transaction datasets (e.g., repeated small withdrawals, unusual spending patterns) to enable early detection of *potential* fraud.

How it works

- Trains a neural network to encode input data into a simplified representation (bottleneck) and then reconstructs it from this reduced representation. The bottleneck captures only essential features of the data. Large reconstruction errors signal deviations, as anomalies fail to fit the learned patterns.



Julia Lenc

Choosing Clustering Method

Visualize the data, run Exploatory Data Analysis (EDA)

- Identify patterns. Assess cluster shapes, densities, and outliers.

Align with business goals

Fast prototyping:

Defined number of clusters → **K-Means**

Uncertain number of clusters or need hierarchy → **Hierarchical clustering**

Geographical insights or rare behaviors → **DBSCAN, Mean Shift**

Personalized recommendations for users with multiple preferences → **GMM**

Hyperpersonalization, clustering+dimensionality reduction->**SOM**

Align with data

Separated, spherical groups of similar size known k → **K-Means**

K is not known, need visual hierarchy → **Hierarchical clustering**

Irregular cluster shapes, outliers → **DBSCAN, Mean Shift**

Overlapping clusters with soft boundaries → **GMM**

Complex features topology → **SOM**

Test and validate different algorithms!

Julia Lenc

Choosing Dimensionality Reduction Method

Visualize the data, run Exploatory Data Analysis (EDA)

- Identify redundant, high-dimensional features and patterns.

Align with business goals

Fast feature selection → **PCA**

Data denoising, learning task specific patterns → **Autoencoders**

Clustering enhancement, feature ergonomics → **SOM**

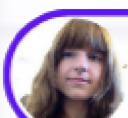
Align with data

Linear relationships → **PCA**

Non-linear data. You need topologically interpretable results or insightful visualizations. Moderate-sized datasets → **SOM**

Large, noisy or highly complex dataset. You need tailored, task-specific embeddings → **Autoencoders**

Test and validate different algorithms!



Julia Lenc

6 Things to Remember

1. Data preprocessing

Unsupervised methods are sensitive to data quality and scaling

2. Hyperparameter tuning

The simpler is the method, the more it relies on hyperparameters:

- Use elbow method to choose the number of clusters (K-Means)
- Use explained variance to choose the number of features (PCA)

3. Model training and execution

Consider model scalability and computational needs.

4. There is no ground truth in unsupervised learning

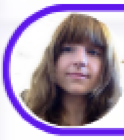
Use domain expertise. Use Silhouette Score and Davies-Bouldin Index for more objective evaluation. Visualize when possible.

5. Reproducibility

Document all model steps.

6. Business context and “favourite methods”

Always put results in business context. Understand most actionable methods in your sector, e.g., K-Means in FMCG, Autoencoders in Banking / Ensurance.



Julia Lenc

Useful Links

1. Core methods (link)

   Free tutorial on K-Means, Hierarcical Clustering and PCA.

2. Unsupervised learning with Python (link)

   Free tutorial, very beginner-friendly.

3. Ultimate Machine Learning Bootcamp (link)

   Free tutorial, very beginner-friendly. Machine Learning. Python.

4. YouTube learning channels:

- Edureka
- Derek Banas
- Stat Quest
- Joma Tech
- Analytics Vidya

Did you find it useful?

Save

Share

Follow

Analytics, Market Research, Machine Learning and AI

Julia Lenc