Module 6: Machine Learning – Unsupervised Learning

Data Science with Python – BSc Course

Why Unsupervised Learning?

Renaissance Technologies, the most successful hedge fund in history, clusters market regimes to adjust trading strategies. Robo-advisors like Betterment and Wealthfront segment millions of clients by risk profile — no human labels required. In supervised learning, you need labeled data: fraud/not fraud, default/no default. But much of finance has no labels. What defines a market regime? How many types of investors exist? The algorithm must find structure on its own. Unsupervised learning discovers patterns, segments customers, detects anomalies, and reduces complexity — all without being told what to look for.

Finding structure in unlabeled data — the hidden patterns that drive markets

Why This Matters

- **Market regime detection:** Bull, bear, sideways markets — no ground truth labels exist
- **Customer segmentation:** Robo-advisors group investors by behavior for personalized advice
- **Anomaly detection:** Unusual trading patterns, outliers, potential fraud
- **Dimensionality reduction:** Hundreds of features reduced to manageable few for modeling

When labels don't exist, unsupervised learning reveals the hidden structure

What You'll Learn

By the end of this module, you will be able to:

- Cluster data using K-Means and hierarchical methods
- Reduce dimensionality with PCA for visualization and modeling
- Build end-to-end ML pipelines from raw data to deployment
- Choose appropriate clustering metrics and validate results
- Apply unsupervised techniques to real financial problems

From high-dimensional data to actionable insights without labeled examples

Lesson Roadmap

Lesson	Topic	Narrative Arc	Focus
L29	K-Means Clustering	"Finding the Hidden Groups"	Sorting without labels
L30	Hierarchical Clustering	"Building a Family Tree"	Dendrograms, linkage
L31	PCA	"Seeing the Big Picture"	Compression, loadings
L32	ML Pipeline	"The Assembly Line"	Preprocessing to deploy

Module Story: This module removes the safety net: no more labels. L29 discovers hidden groups in raw data. L30 reveals the full hierarchy of relationships. L31 compresses high-dimensional chaos into a few meaningful directions. L32 assembles all the pieces into a leak-free, production-ready pipeline.

Four lessons, one journey — from unlabeled data to production ML

Key Concepts

- **K-Means Clustering** – Partition data into K groups by minimizing within-cluster variance
- **Hierarchical Clustering & Dendrograms** – Build cluster trees without specifying K upfront
- **PCA Dimensionality Reduction** – Extract principal components explaining most variance
- **Complete ML Pipeline** – Data preprocessing, feature engineering, model training, deployment
- **Cluster Validation** – Silhouette scores, elbow method, business validation

These tools power customer segmentation, regime detection, and portfolio optimization

Finance Application

Scenario: Client Segmentation Engine
Using the techniques from this module, you will:

- Cluster bank customers by transaction behavior (spending patterns, savings rate, volatility)
- Visualize segments with PCA to identify distinct client groups
- Build a complete pipeline from raw transaction data to actionable client groups
- Enable targeted product offers based on segment characteristics

This is the exact approach that robo-advisors and private banks use to personalize services at scale.

Segment customers without manual labeling — let the data reveal the groups

Who Uses This?

- **Hedge Funds** – Renaissance Technologies, Man Group use regime clustering for strategy switching
- **Robo-Advisors** – Betterment, Wealthfront segment clients for personalized portfolios
- **Risk Management** – PCA for yield curve modeling — first 3 components explain 95% of variance
- **Compliance** – Anomaly detection for insider trading, unusual transaction patterns

Unsupervised learning reveals structure that even experts struggle to define manually

What's Next: Module 7 – Deep Learning

Traditional ML has limits. Linear models assume straight-line relationships. Decision trees struggle with high-dimensional interactions. K-Means requires you to specify the number of clusters.
Deep learning pushes beyond these constraints — learning complex, non-linear patterns that simpler models cannot capture.
Module 7 introduces neural networks: perceptrons, multi-layer networks, backpropagation, and regularization. From simple building blocks to networks that process alternative data at scale.

When traditional ML plateaus, deep learning breaks through the ceiling

Let's Begin!

First up: L29 – K-Means Clustering

Finding groups in data without being told what to look for.