Regime Detection via Unsupervised Learning from Order Book and Volume Data

Praneel B. Satare

1 Introduction

This report presents a regime analysis conducted on financial time-series data using unsupervised learning techniques. The objective was to segment the market behavior into distinct regimes, thereby enhancing understanding of market conditions and supporting strategic decision-making in trading environments.

2 Data Description

The data consisted of two sources spanning four full trading days:

- **Order Book Data:** Included top 20 levels of bid-ask quotes, capturing detailed information about market depth and liquidity.
- **Trade Data:** Comprised executed trades with details like trade price, volume, timestamp, and aggressor side.

From these, several hand-crafted features were extracted. Additionally, a new feature named `trade_direction` was introduced:

- 1: Buy dominant
- -1: Sell dominant
- 0: No clear direction

These features were standardized and used as inputs to the clustering models.

3 Methodology

3.1 Feature Engineering

Derived features capture market microstructure characteristics such as volatility, spread, imbalance, and trade intensity.

3.2 Dimensionality Reduction

Principal Component Analysis (PCA) was used to reduce dimensionality and denoise the feature space before clustering. This preserved most of the variance while reducing computation.

3.3 Clustering Algorithms

Two clustering techniques were employed:

- **K-Means Clustering**
- **Gaussian Mixture Models (GMM)**

3.4 Evaluation Metrics

Performance was assessed using:

- **Silhouette Score:** Measures cohesion and separation of clusters (higher is better).
- **Davies-Bouldin Index:** Measures cluster similarity (lower is better).

4 Model Evaluation and Selection

Table 1: Clustering Performance Metrics

Algorithm	Silhouette Score	Davies-Bouldin Index
K-Means	0.4599	0.8914
GMM	0.0714	3.2176

K-Means outperformed GMM across both evaluation metrics and was selected for detailed regime analysis.

5 Cluster Distribution

Table 2: K-Means Cluster Sizes

Cluster ID	Count
0	11,338
1	12,631
2	20,319

6 Regime Interpretation

The following metrics were calculated per cluster:

Table 3: Cluster-wise Market Metrics

Cluster	Volatility	Spread	Liquidity	Buy Ratio
0	0.0702	0.2971	-0.0175	0.2934
1	0.0899	0.7758	-0.0066	0.3758
2	0.0573	-0.6481	0.0139	0.2756

- **Cluster 0: Mean-Reverting & Moderately Volatile**
  Moderate volatility and spread suggest range-bound movement, with minor liquidity stress. Suitable for mean-reversion strategies.
- **Cluster 1: Trending, Volatile & Illiquid**
  Highest volatility and spread, low liquidity, and strong buy ratio suggest aggressive directional moves, ideal for trend-following strategies.
- **Cluster 2: Stable, Liquid & Narrow Spread**
  Low volatility and spread, high liquidity indicate a calm, efficient market. May represent consolidation or low-activity phases.

7 Visualizations

The following plots were created for analysis and validation:

- Cluster assignments from K-Means and GMM.
- Time-series evolution of regimes alongside price and volatility.
- 2D t-SNE projection of clusters to visualize separation.

8 Conclusion

The analysis successfully identified three meaningful market regimes using K-Means clustering. These regimes reflect differing volatility, liquidity, and trade dynamics, which can be leveraged for strategy selection and risk control.