Regime Detection via Unsupervised Learning

Objective

This project focuses on identifying distinct market regimes using unsupervised learning. The segmentation is based on high-frequency order book and volume data. The aim is to detect behavioral patterns such as:

- **Trend:** Trending vs Mean-Reverting
- **Volatility:** Volatile vs Stable
- **Liquidity:** Liquid vs Illiquid

Data and Tools

- Dataset: Order book and trade volume data per timestamp.
- Environment: Google Colab
- Libraries: `pandas`, `numpy`, `scikit-learn`, `matplotlib`, `seaborn`, `umap-learn`

Feature Engineering

The following features were engineered to capture market behavior:

Liquidity & Depth

- **Bid-Ask Spread:** $\text{spread} = \text{ask\_1\_price} - \text{bid\_1\_price}$
- **Order Book Imbalance:** $\text{imbalance} = \frac{\text{bid\_qty\_1} - \text{ask\_qty\_1}}{\text{bid\_qty\_1} + \text{ask\_qty\_1}}$
- **Mid Price:** $\text{mid\_price} = \frac{\text{ask\_1\_price} + \text{bid\_1\_price}}{2}$

Volatility & Price Action

- **Log Return:** $\log\_return = \log\left(\frac{\text{mid}_t}{\text{mid}_{t-1}}\right)$
- **Rolling Volatility:** Standard deviation of log returns over a 10-point window

Volume Features

- **Volume Imbalance:** Buy volume - Sell volume

Preprocessing

- Features were normalized using Z-score normalization
- Dimensionality reduction was performed using:
  - PCA (Principal Component Analysis)
  - UMAP (Uniform Manifold Approximation and Projection)

Clustering Methods

The following clustering algorithms were applied:

1. **KMeans (k=4)** - Hard clustering using Euclidean distance
2. **Gaussian Mixture Model (GMM)** - Soft clustering using probability
3. **HDBSCAN** - Density-based clustering with noise detection

Evaluation Metrics

Silhouette Score: Measures how well samples fit in their own cluster vs others (higher is better).
Davies-Bouldin Index: Measures intra-cluster compactness vs inter-cluster distance (lower is better).

Algorithm	Silhouette Score	Davies-Bouldin Index
KMeans	0.2636	**1.20**
GMM	**0.2789**	2.15

Visualizations

- PCA and UMAP scatter plots showing clusters from KMeans and GMM
- Time series of cluster labels for regime transition visualization
- Cluster-wise mean feature plots for interpretation

Insights and Regime Interpretation

Cluster	Characteristics	Interpretation
0	Low spread, low volatility	Stable & Liquid
1	High volatility, imbalance fluctuations	Volatile & Illiquid
2	Sharp upward price movements	Bullish Trend & Liquid
3	Negative log returns, sell volume pressure	Mean-Reverting or Bearish

Regime Change Insights

- Time-based plots reveal the persistence or transition of regimes.
- HDBSCAN identified rare or transitional points as noise, increasing robustness.
- Regime switching patterns suggest potential predictive signals.

Conclusion

This project successfully segments market behavior into meaningful regimes using unsupervised learning. KMeans provided clear cluster boundaries, GMM allowed for probabilistic flexibility, and HDBSCAN added robustness through noise detection. These techniques enable enhanced understanding of market dynamics and can support future forecasting or risk management strategies.