Regime Detection via Unsupervised Learning

Objective

The aim of this project is to identify and segment distinct market regimes using unsupervised learning techniques. The segmentation is based on real-time order book and trade volume features. The regimes aim to capture variations across:

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

Data Sources

- **depth20.csv:** Contains top 20 levels of bid/ask price and quantity from the order book.
- **aggTrade.csv:** Aggregated buy/sell trade volumes over time.

Feature Engineering

Key market behavior features were manually engineered:

Liquidity & Depth

- Bid-Ask Spread = ask_1_price - bid_1_price
- Order Book Imbalance = (bid_qty_1 - ask_qty_1) / (bid_qty_1 + ask_qty_1)
- Mid Price = (ask_1_price + bid_1_price) / 2

Volatility & Price Action

- Log Mid Price Return = $\log(\frac{mid_t}{mid_{t-1}})$
- Rolling Volatility = Standard deviation of log returns over 10 timestamps

Volume-Based

- Volume Imbalance = Buy volume - Sell volume

Preprocessing

- Data merged on timestamp key
- NaNs removed from rolling operations
- Features normalized using Z-score normalization
- PCA applied to reduce dimensions to 2D for clustering and visualization

Clustering Techniques Used

Three clustering methods were applied:

1. **KMeans (n=4)**: Hard clustering based on Euclidean distance.
2. **Gaussian Mixture Models (GMM)**: Soft clustering using probabilistic assignments.
3. **HDBSCAN**: Density-based clustering that detects noise and variable density clusters.

Evaluation Metrics

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

Silhouette Score: Measures cohesion vs separation (higher is better).
Davies-Bouldin Index: Measures intra-cluster compactness vs inter-cluster distance (lower is better).

Insights and Regime Interpretation

Cluster	Characteristics	Possible Regime
0	Low spread, low volatility	Stable & Liquid
1	High spread, high volatility	Illiquid & Volatile
2	Low spread, high buy volume imbalance	Trending & Liquid (Bullish)
3	Moderate volatility, sell-side pressure	Mean-Reverting

Conclusion

This project demonstrates how unsupervised learning techniques can effectively detect market regimes using real-time data. KMeans gave compact and interpretable clusters, while HDBSCAN offered flexibility by identifying noise and irregular behavior. These regime insights are useful for risk management, trading strategies, and anomaly detection in financial markets.