Clustering Oil Commodities Futures

Evan Zhang

Purdue University

ECE 50836

Professor Jing Gao

December 1st, 2024

Introduction

Oil commodities and futures contracts play a critical role in global markets, serving as a key indicator of economic activity and influencing energy prices, transportation costs, and production expenses across industries. Futures contracts are a tool for entities to use to strategize for ways to work with the market by being able to hold a sale to a price ahead of time regardless of market activity. The task at hand involves being able to find trends in the market that may not be obvious due to the high complexity and reasons the market will fluctatuate. Oil prices can be based on various features such as location, contract type, product, process, and time/date of the contract. By analyzing historical data, like finding price trends, regional influences, and contract details, we can develop a model that can potentially accurately understand the market. This prediction will support strategic decision-making in trading, procurement, and risk management, helping to identify profitable opportunities and minimize exposure to price fluctuations.

Formulation

Since the market data for oil commodities is unlabeled, we do not have predefined class labels such as "buy," "sell," or "hold" for the data. Therefore, we cannot directly apply supervised learning techniques. Instead, to gain insights into the patterns within the data, we will go to unsupervised learning methods Spectral clustering, DBSCAN clustering, and Gaussian Mixture Model clustering. The key with all these methods are done in mind that the data has an unknown topology and these methods are able to handle potentially nonlinear trends albeit with controlled dimensionality. Clustering will help to uncover natural groupings in the data, such as commodities with similar price movements or characteristics like contracts or processes. By analyzing these clusters, we can infer potential patterns or trends in market behavior and better understand the underlying structure of the data.

Dataset

The data was obtained from an API maintained by the US Energy Information Administration, which receives a certain query request for a lot of different types of energy commodities and applications, such as household electricity usage, import data, and (for this project) New York Mercantile Exchange data on futures contracts for oil. The NYMEX data obtained is very clean in terms of no missing values, and contains the following features: period, duoarea, area-name, product, product-name, process, process-name,series, series-description,value, and units. To preprocess, every price was standardized to per gallon, the units feature was removed, and the period was split into year and month features to better capture trends.Since there is a lot of categorical data that tells us about the contracts, process names, etc. the data under categorical columns were mapped to a numerical space so it was useable for some clustering algorithms.

Algorithms

To preprocess, the code first retrieves the database from the USEIA API with an API key, and converts to a data frame. The data processes out the period into month and year, and standardizes prices to per gallon. After that, the dataframe is then modified to change all categorical data into a ordinal encoding using SKlearn's OrdinalEncoder. There are two CSVs to draw from, encoded and non-encoded.

From here, this is where a few unsupervised learning clustering algorithms are used. 1. Agglomerative Hierachical clustering, 2. DBSCAN, 3. Gaussian Mixture Model. The main method of evaluation will be silhouette for Spectral and DBSCAN, and loglikelihood for GMM.

Experimental Results

Spectral Clustering

Implemeneted using SKlearn's implementation of Spectral clustering. Uses a target of 9 clusters, which was chosen due to performance decrease before and after. The clustering method returned a best silhouette score of 0.2534 which indicates above average clustering.

DBSCAN

Implemeneted using SKlearn's implementation of DBSCAN. The epsilon (max distance to be neighbor) value is set to 0.5 after trialing values between 0.1 to 5. The min_samples is set to 3 as there is a drop off in performance between 3 to 4. Essentially, the data takes the mixed data of encoded categorial (ordinal), prices, and time to get a good understanding of the clusters using k-nearest neighbors. The silhouette evaluation received a solid 0.91627 from a scale -1 to 1 with 1 indicating perfect clustering with 1149 unique labels out of 5000 data entries.

Gaussian Mixture Model

Implemeneted using SKlearn's implementation of Gaussian Mixture Model. The n_components used for this model was 10 as the performance peaked at this value. The loglikelihood score returned the highest 172347.83637192386 at 10 components. However, the silhouette score only returns at 0.20212

Discussion

From what the results from each clustering method tells us, DBSCAN performs the best in terms of cluster quality. With the parameters of 0.5 max distance and 3 minimum to be a cluster, the output is a high quality clustering of 0.92 silhouette scoring and 1149 of 5000 unique clusters, we can deduce that for the data is has a lot of variation due even in small groups of samples which can be from how much variation there are in categories along with the pricing.

The quality of clusters is also not just due to overfitting because it doesn't capture each data point near a 1:1 ratio so it is accurately reporting the just how many classifications we can label the data. Because of how much better this algorithm has in quality compared to spectral and GMM clustering, that the market data in general is well-separated and dense but that also the data cannot be clustered linearly or in a certain shape because of the uniqueness of the data.