K-means Clustering

Contents

Introduction

K-Means Clustering is a type of unsupervised machine learning that groups data on the basis of similarities. Recall that in supervised machine learning we provide the algorithm with features or variables that we would like it to associate with labels or the outcome in which we would like it to predict or classify. In unsupervised machine learning, we only provide the model with features and then it "learns" the associations on its own.

K-Means is one technique for finding subgroups within datasets. One difference in K-Means versus that of other clustering methods is that in K-Means, we have a predetermined amount of clusters and some other techniques do not require that we predefine the number of clusters.

The algorithm begins by randomly assigning each data point to a specific cluster with no one data point being in any two clusters. It then calculates the centroid, or mean of these points.

The object of the algorithm is to reduce the total within-cluster variation. In other words, we want to place each point into a specific cluster, measure the distances from the centroid of that cluster and then take the squared sum of these to get the total within-cluster variation. Our goal is to reduce this value.

The process of assigning data points and calculating the squared distances is continued until there are no more changes in the components of the clusters, or in other words, we have optimally reduced the in-cluster variation.

Working

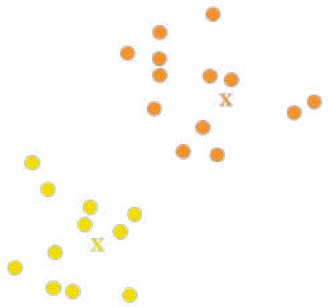
Let us take a look at how the K-means algorithm works. These are the steps which the algorithm goes through:

- First of all, the algorithm is fed with K random data points.

- Then, we select some k points for clustering. The distance between selected k points and other points is calculated, which tells how far the data points lie in the graph.

- Now, the clusters are made out of the points closer to each other.

- Then for each cluster, a centroid is calculated again and we repeat the clustering process with the centroid points.



- This repetition of clustering will continue until the centroid becomes relatively stable.

Application

K-means algorithm is widely used, and in the trading world, if you want to know the importance of k-means, you don't have to look further than the implementation of statistical arbitrage.

Statistical Arbitrage is one of the most recognizable quantitative trading strategies. Though several variations exist, the basic premise is that despite two securities being random walks, their relationship may not be random, thus yielding a trading opportunity. A key concern of implementing any version of statistical arbitrage is the process of pairs trading.

Pairs trading is a market-neutral strategy, where a pair or a basket of stocks are selected and a long position is taken on one leg and a short position on the another.
The strategy builds on mean reversion theory, which states that the price ratios of correlated/cointegrated stocks revert back to their long term mean after unexplained considerable deviations in the prices. In a pair, where the prices of its constituents generally move together and when one stock outperforms the other, the outperforming stock is sold (short position) and the underperforming stock is bought (long position) with an expectation that they will revert back to their mean.

To better understand the strength of using a technique like K-Means for Statistical Arbitrage, we'll do a walk-through of trading a Statistical Arbitrage strategy if there was no K-Means.

First, let's identify the key components of any Statistical Arbitrage trading strategy.

1. We must identify assets that have a tradable relationship
2. We must calculate the Z-Score of the spread of these assets, as well as the hedge ratio for position sizing
3. We generate buy and sell decisions when the Z-Score exceeds some upper or lower bound

To begin, we need some pairs to trade. But we can't trade Statistical Arbitrage without knowing whether or not the pairs we select are cointegrated. Cointegration simply means that the statistical properties

between our two assets are stable. Even if the two assets move randomly, we can count on the relations in between them to be constant, or at least most of the time.

Traditionally, when solving the problem of pairs trading, in a world with no K-Means, we must find pairs by brute force, or trial and error. This was usually done by grouping stocks together that were merely in the same sector or industry. The idea was that if these stocks were of companies in similar industries, thus having similarities in their operations, their stocks should move similarly as well. But, as we shall see, this is not necessarily the case. Despite two stocks being related on a fundamental level, this doesn't necessarily mean that they will provide a tradable relationship.

Advantages & Disadvantages

K-means method of machine learning is neither all complicated nor entirely sorted. There are a few advantages and disadvantages associated with this algorithm.

Advantages

- It is easy to implement k-means and find out unknown groups from complex datasets
- Easily adjusts to changes by adjusting the cluster segment
- Tackles problems associated with huge datasets
- The cluster descriptions are easy to interpret and it ensures clustering accuracy
- Ideal for exploring raw or unknown data for predictions

Disadvantages

- K-means works best with clusters having a spherical shape, whereas, clusters with a complicated geographical shape lead the K-means to perform poorly
- While forming clusters, K-means excludes those data points which are far from a cluster that it actually belongs to
- It leaves the data points in small clusters away from the centroid while trying to focus on large clusters. This way, it may not be able to figure out the clusters as they should be.
- The output is always in a uniform size even though the input data is provided with different sizes.

After some quizzes to brush up your knowledge, you will learn about data quality ahead in the next section on data quality & feature engineering.