Pairs trading via unsupervised learning

ADAMBI IDRISSOU Soulemane
&
ADEDE Ezechiel

University Mohammed VI Polytechnic

October 27, 2023

Plan

Introduction

General idea

The article discusses the use of unsupervised learning to classify trading stocks in pairs. Three classification algorithms were used: k-means, DBSCAN and agglomerative classification.

Introduction

Motivation

Several models had been used and these were essentially based on supervised learning methods. However, the portfolios constructed by these models are mostly made up of stocks that are not strongly related. This is because these models only use past data to make predictions. It is therefore these inadequacies that this article corrects in which the authors used unsupervised learning methods for the construction of pairs trading stock portfolios. This article differs from others because it is based not only on past trading data but also on the characteristics of the concerned company.

The algorithms

k-means clustering

It is a method of vector quantization that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. k-means clustering minimizes within-cluster variances (squared Euclidean distances).

The algorithms

Density-based spatial clustering of applications with noise (DBSCAN)

DBSCAN is a data clustering algorithm .It is a density-based clustering non-parametric algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away).

The algorithms

Agglomerative clustering

It is a bottom-up approach. Initially, each data point is a cluster in its own right, other pairs of clusters are joined over time to ultimately obtain the desired number of clusters.

Methodology

The feature set includes 48 performance factors and 78 firm characteristics generated monthly for the sample period from December 1979 to November 2020. The out-of- sample period is from January 1980 to December 2020 The data used are those of the American stock market available from the Center for Research in Security Prices (CRSP). These data were pre-processed using the PCA method.

Clustering methods

K-means

By setting the number of clusters to k and denoting by n the total number of points to be grouped, the maximization problem is defined as:

$$W = \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} ||x^i - \mu_k||^2$$

where x^i refers to *ith* data point, μ_k the centroid of cluster k , $w_{ik} = 1$ if x^i belongs to cluster k , otherwise $w_{ik} = 0$, and N is the total number of data points.

Clustering methods

DBSCAN

Two parameters are taken into account for this algorithm. This is the minimum number of points per cluster and the maximum distance between two points in the same cluster.

lustering Methods

Agglomerative clustering

The parameters taken into account are the number of clusters that we want to form and the maximum distance between two points to be considered as being in the same cluster.

Results obtained

Comparing the performance of these three clustering methods for creating longshort portfolios, it is clear that agglomerative clustering stands out significantly from the others. It has an average annual return of 24.8%, an annual Sharpe ratio of 2.69 and a maximum loss limited to 12.3% over a period of just two months. In contrast, K-means clustering and DBSCAN obtain annual Sharpe ratios of 2.34 and 2.04 respectively, indicating a strong potential for unsupervised learning in pairwise trading. They significantly outperform the S&P500 index as well as the conventional short-term reversal strategy.