

Clusterization K-Mean Model: Argentine Stock Market M-28

Ezequiel Eliano¹

¹Economista Ministerio de Hacienda, UCA y UBA

July 29, 2018

Abstract

En este paper se revisaran 5 variables de las principales acciones del mercado bursátil Argentino (Merval 28 al cierre de 2017), utilizando el modelo de clusterización K-Mean analizando los diversos tipos de agrupamientos y esquemas de jerarquización de las acciones. Analizando los diversos patrones normalizados de la variación interanual, volumen operado, varianza del retorno y covarianza del Merval y del dolar se pretende comprender los diversos movimientos de las acciones generando un análisis en cuanto a la metodología diferentes, pero con complementariedad a los clásicos estudios de análisis técnico y *fundamentals*. La comprensión del comportamiento busatil de las principales acciones del panel del Merval contribuye a entender los principales movimientos y mecanismos de acción tendientes a generar un mayor volumen y capitalización del mercado local.

In this papaer will be reviewed five variable for the main using the K-Mean clusterization.....

Metodología de Clusterización

K-means (MacQueen, 1967) es una técnica de análisis cluster que trata de establecer una partición en K grupos o clusters sobre un conjunto de N objetos $\{O_1, \dots, O_N\}$ de los que disponemos de una información multivariante P-dimensional. Partiendo de la matriz de datos $X_{N \times P}$, la función que se pretende minimizar en el proceso de clasificación es la suma total de cuadrados de los errores (TESS), cuya expresión viene dada por:

$$TESS_k = \sum E^2_{(k)} = \sum \sum I(O_i \in C_k) d^2_{i(k)} \quad (1)$$

siendo $\sum E^2_{(k)}$ la suma de cuadrados de los errores (ESS) para el cluster C_k , $I[O_i \in C_k] = 1$ si el objeto O_i ha sido asignado a C_k , $I[O_i \in C_k] = 0$ si O_i no ha sido asignado a C_k , y $d_{i(k)}$ la distancia euclídea al cuadrado de cada objeto al centroide de C_k :

En la practica, dada una particion inicial en K clusters, la tecnica se basa en el siguiente algoritmo iterativo:

1. Calculo de las posiciones de los centroides $x(k)$ de los K clusters.
2. Para cada objeto, calculo de su distancia a los K centroides, $d_{i(k)}$.
3. Reasignacion de cada objeto al cluster cuyo centroide es el mas proximo. Es un hecho destacable que la solucion (particion) final depende de la configuracion inicial de los clusters elegida, siendo posible la convergencia a un minimo local de TESS. Una opcion recomendable y que suele ofrecer buenos resultados es la de realizar un analisis cluster jerarquico y elegir como particion inicial la obtenida con un nivel de disimilaridad que aplicado al arbol ultrametrico conduzca al numero de grupos deseado.

<http://biodiver.bio.ub.es/veganaweb/bvegana/SEI02001.pdf>

La informacion provista por `X` paneles de datos o cluster agrupado por el metodo de k -mean, tiene como objetivo dividir los puntos en K grupos para minimizar la suma de cuadrados desde los puntos hasta el centro del cluster asignado. Como minimo todos los centros del cluster estan en la media de sus conjuntos de Voronoi (el conjunto de puntos de datos mas cercano al centro del cluster).

El algoritmo de (Hartigan and Wong, 1979) es usado por default. En este sentido, algunos autores usan k -means para referirse a un algoritmo especifico en lugar del metodo general: mas comunmente el algoritmo dado por, pero aveces es provisto por (MacQueen, 1967), pero tambien por (Lloyd, 1982) y (Forgy, 1965). El algoritmo de Hartigan-Wong generalmente hace un mejor trabajo que los anteriores, pero probando varios experimentos aleatorios comienza (`nstart > 1`) es comunmente recomendado. En casos raros, cuando algunos de los puntos (filas de `x`) son extremadamente cercanos, el algoritmo puede no converger en la etapa “Transferencia rapida”, senalizando una advertencia (y devolviendo `ifault = 4`). El ligero redondeo de los datos puede ser aconsejable en ese caso.

Excepto por el metodo Lloyd-Forgy, siempre se devolveran k clusters si se especifica un numero. Si se suministra una matriz inicial de centros, es posible que ningun punto sea el mas cercano a uno o mas centros, lo que actualmente es un error para el metodo Hartigan-Wong.

The data given by `x` are clustered by the k -means method, which aims to partition the points into k groups such that the sum of squares from points to the assigned cluster centres is minimized. At the minimum, all cluster centres are at the mean of their Voronoi sets (the set of data points which are nearest to the cluster centre).

The algorithm of (Hartigan and Wong, 1979) is used by default. Note that some authors use k -means to refer to a specific algorithm rather than the general method: most commonly the algorithm given by MacQueen (1967) but sometimes that given by (Lloyd, 1982) and (Forgy, 1965). The Hartigan–Wong algorithm generally does a better job than either of those, but trying several random starts (`nstart > 1`) is often recommended. In rare cases, when some of the points (rows of `x`) are extremely close, the algorithm may not converge in the “Quick-Transfer” stage, signalling a warning (and returning `ifault = 4`). Slight rounding of the data may be advisable in that case. For ease of programmatic exploration, $k=1$ is allowed, notably returning the center and `withinss`.

Except for the Lloyd–Forgy method, k clusters will always be returned if a number is specified. If an initial matrix of centres is supplied, it is possible that no point will be closest to one or more centres, which is currently an error for the Hartigan–Wong method

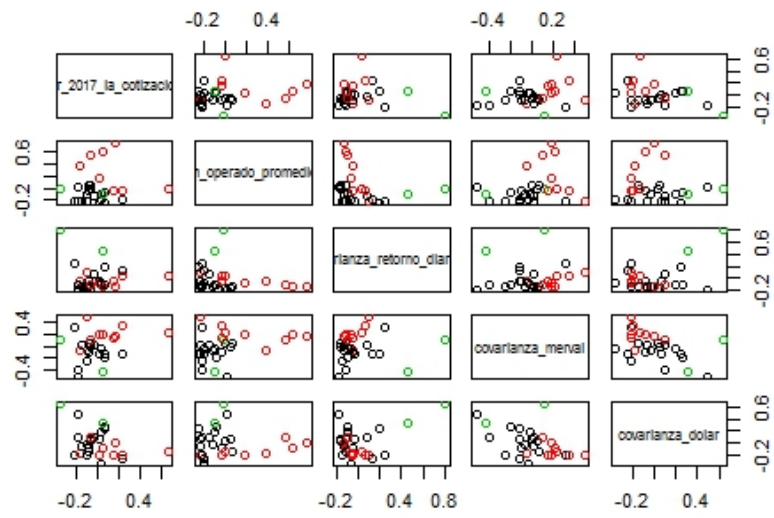


Figure 1: Plot K-mean para 5 variables

References

- E. W. Forgy. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, 21(2):362, jun 1965. doi: 10.2307/2528096. URL <https://doi.org/10.2307%2F2528096>.
- J. A. Hartigan and M. A. Wong. Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, 28(1):100, 1979. doi: 10.2307/2346830. URL <https://doi.org/10.2307%2F2346830>.
- S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, mar 1982. doi: 10.1109/tit.1982.1056489. URL <https://doi.org/10.1109%2Ftit.1982.1056489>.
- J. MacQueen. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. *Berkeley, CA: University of California Press.*, 1, pp. 281–297., 1967.