THE CENTER FOR ARTIFICIAL INTELLIGENCE
IN BUSINESS ANALYTICS & FINANCIAL TECHNOLOGY

Recursive K-Means Clustering for Time Series

Ryan Holmes (ryan.holmes@ask2.ai) – Advisors: Federico Klinkert, Ali Hirsa

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Abstract

Clustering is an unsupervised learning technique used to group data points based on some measure of similarity. Algorithms such as K-Means and Hierarchical clustering are designed to operate on one dataset. New challenges arise when extending clustering models to incorporate a time series. Two models, Closest Previous Means and Epsilon method, are proposed to alleviate some of issues found in recursive K-Means clustering methods.

Standard K-Means Clustering

Algorithm 1 K-Means Algorithm (1)

1: Choose a distance metric
2: Choose a dataset X
3: Choose k initial centroids $\mathcal{C} = \{c_1, ..., c_k\}$ randomly from X ▷ Unless given initial centroids by Algorithm 2, 3
4: **for** $i \in \{1, ..., k\}$ **do**
5: Set cluster C_i to be the set of all points in X closer to c_i that other centroids c_j $\forall j \neq i$ by distance metric
6: Set c_i to be the center of mass for all points in C_i where $c_i = \frac{1}{\|C_i\|} \sum_{x \in C_i} x$
7: **end for**
8: Repeat steps 3-6 until $\mathcal{C}$ no longer changes

Clusters computed by standard recursive K-Means clustering techniques are highly dependent on the random initialization step at each time increment, resulting in unstable clustering behaviors and difficulty in labeling clusters over different time increments.

Model Advantages comparing to Standard K-Means

- If the data does not change over time, then the clusters themselves will not change
- Both methods are deterministic after the first time increment, significantly reducing the effects of random noise on the clustering results
- No chance of mislabeling clusters as the labels are directly pulled from the previous time period
- Reduces the variation in cluster movement over time resulting in more stable and reliable results

Challenges

K-Means serves as the basis of Closest Previous Means and Epsilon methods. When making modifications to K-Means, there are two important rules to follow:

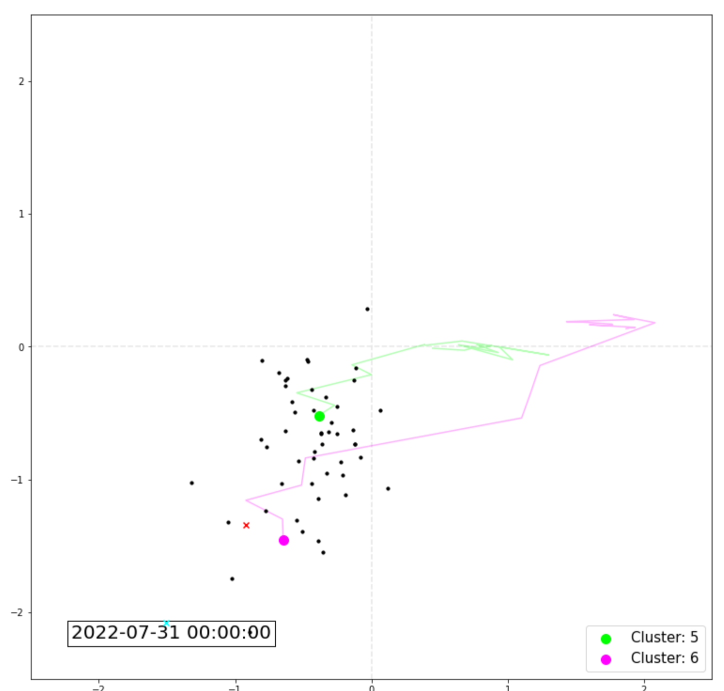
- Each centroid must have at least one data point closest to it throughout each step of iteration
- The centroids (means) can never share the same location in space of another centroid

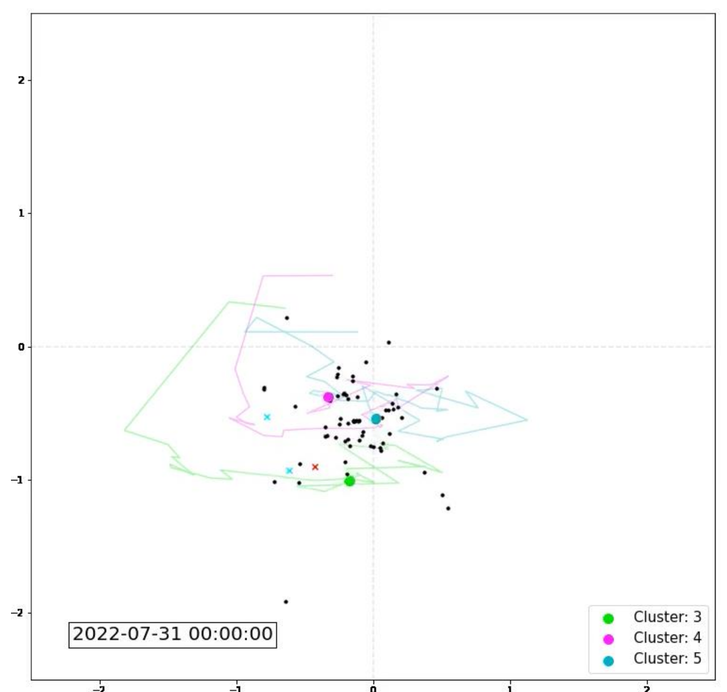
Closest Previous Means (CPM)

Algorithm 2 Closest Previous Means

Choose a fixed number of clusters k
Set $t = 1$
Set time T as the maximum date index in time series
Choose a time series data set X indexed by $t, ..., T$
Compute K-Means on X_t with random initial centroids ▷ See Algorithm 1
Set $\{\mu_1, ..., \mu_k\}_t = \{c_1, ..., c_k\}$ computed by K-Means algorithm
Sort $\{\mu_1, ..., \mu_k\}_t$ by number of data points in each cluster by ascending order ▷ Ascending for stability purposes
for $t = 2, ..., T$ **do**
  for $\tilde{\mu_1} \in \{\tilde{\mu_1}, ..., \tilde{\mu_k}\}_{t-1}$ **do** ▷ $\tilde{\mu}$ indicates sorted μ
    Compute data point $x \in X_t$ closest to $\tilde{\mu_i}$ where x has not already been chosen by a different $\tilde{\mu}$
    Set centroid $c_i = x$
  end for
  Compute K-Means on X_t with initial centroids $\{c_1, ..., c_k\}_t$ ▷ See Algorithm 1
  Set $\{\mu_1, ..., \mu_k\}_t = \{c_1, ..., c_k\}$ computed by K-Means algorithm
  Sort $\{\mu_1, ..., \mu_k\}_t$ by number of data points in each cluster by ascending order
end for

The main idea of Closest Previous Means is to replace the random initialization step of K-Means by selecting the data point closest to each of the means from the previous period. By doing this we eliminate the need for cluster labeling since we are directly using the labels from the previous period in a deterministic way. The best way to reduce variance and promote model stability is to give the centroids with the fewest data points in the previous period the first priority in choosing its closest data point.



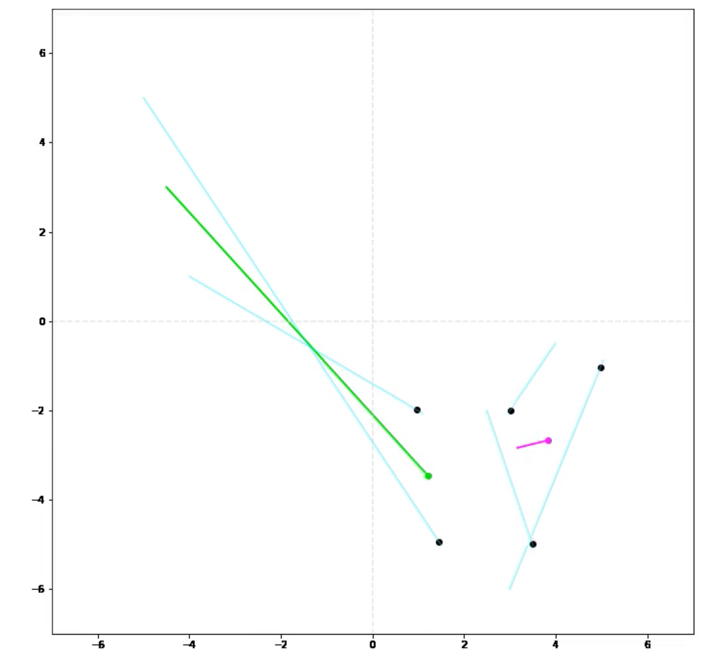


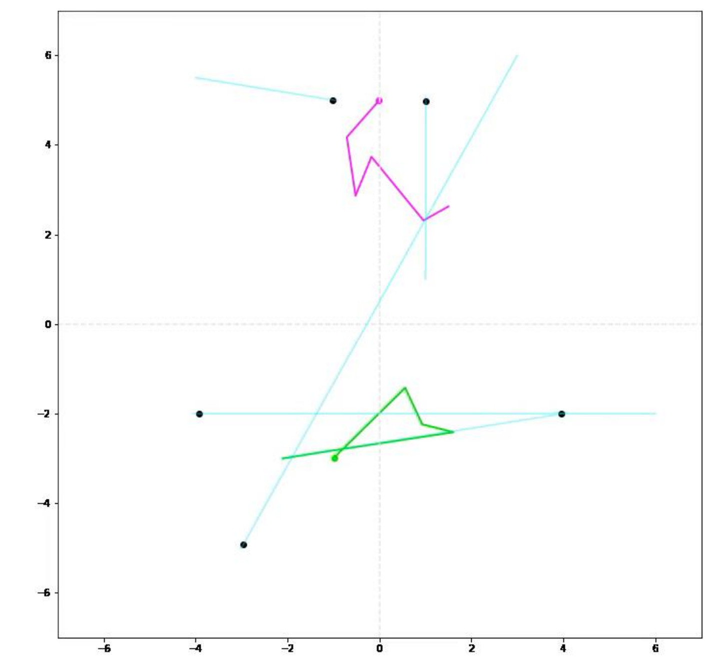
Epsilon Method

Algorithm 3 Epsilon Method

Choose range of partition values $\epsilon_{min}, ..., \epsilon_{max} \in \mathbb{Z}^+$
Choose a fixed number of clusters k
Set time $t = 1$
Set time T as the maximum date index in time series
Choose a time series data set X indexed by $t, ..., T$
Compute K-Means on X_t with random initial centroids ▷ See Algorithm 1
for $t = 2, ..., T$ **do**
  Set endreached = **False**
  for $\epsilon = \epsilon_{min}, ..., \epsilon_{max}$ **do** ▷ Interpolate and run K-Means ϵ times
    if endreached = **False** **then** ▷ Boolean when **True** will end the iteration at time t
      for $p = 1, ..., \epsilon$ **do**
        $\hat{X} = X_{t-1} + \frac{p}{\epsilon}(X_t - X_{t-1})$ ▷ Interpolation Step
        Compute K-Means on $\hat{X}$ with initial centroids as means from $p-1$ step
        if any clusters have 0 data points or any centroids overlap **then**
          Break ▷ Then ϵ partitions fails $\rightarrow$ try $\epsilon + 1$ partitions
        end if
        if $p = \epsilon$ **then**
          Set endreached = **True** ▷ Model successfully clustered X_t with ϵ partitions
        end if
      end for
    end if
  end for
end for

The Epsilon method is designed to linearly interpolate between the data in the previous and current time increment. We then partition this interpolation an epsilon number of times, running K-Means at each partition step. In comparison to Closest Previous Means, the random initialization step is replaced with the actual location of the centroid of the previous step instead of choosing the data point closest in the current period closest to it.





Advantages of CPM over Epsilon Method

- Much faster to run since K-Means is only required to run once at each time increment
- Designed with a focus to model a purely spatial relationship between clusters
- Model designed to work on datasets with missing data, while the Epsilon Method runs into issues if data cannot be linearly interpolated

Advantages of Epsilon Method over CPM

- Interpolating will capture a unique relationship between the funds over time; Closest Previous Means will just see data points as data points and cluster a spatial relationship
- Generally has smaller cluster variance since outlier clusters do not have the chance of a having their closest data points taken by another cluster

Future Work

- Find a way to fix missing value problem in Epsilon Method that will work for all time series data sets
- Implement ways to speed up runtimes of algorithms, especially the Epsilon Method since it is much more computationally intensive

References

1) Arthur, David, and Sergei Vassilvitskii. K-Means++: The Advantages of Careful-Seeding-Stanford-University. https://theory.stanford.edu/~sergei/papers/kMeansPP-soda.pdf.