model_study:101_wk_6_Kmeans_clustering

Seung Hyun Sung

11/18/2021

DS4B 101-R: R FOR BUSINESS ANALYSIS —-

K-Means Clustering & Dimensionality Reduction

K-Means Clustering Concept

- Step1: Select the number of clusters you want to identify in your data. This is the "K" in"K-means clustering"
  - Elbow point: select the optimal number of "K" group
- Step2: Randomly select 3 distinct data points
  - The initial clusters
- Step3: Measure the distance between the 1st point and the three initial clusters
- Step4: Assign the first point to the nearest cluster
- Step5: Calculate the mean of each cluster
- Step6: Repeat (measure distances between new data to continuosuly adjusting new means of initial clusters)

We can assess the quality of the clustering by adding up the **variation within each cluster**.

Since k-means clustering can't "see" the best clustering, its only option is to keep track of these clusters, and their variance, and do the whole thing over again with **different starting points.**

- How? - reclusters based on the new means. It repeats until the clusters no longer change.

At this point, K-means clustering knows that the 2nd clustering is the best clustering so far. But it does not know if it is the best overall, so it will do a few more clusters (it does as many times as you tell it to do) and then come back and return that one if it is still the best.

K optimal: elbow point

- Plot the reduction in variance per value for K
  - x = Number of clusters(K), y = Reduction is Variation/ or Variation
- if the ideal K =3. Huge reduction in variation with K = 3 will be seen, but after that, the variation does not go down as quickly.
- This is called an "elbow plot" and you can pick optimal "K", by finding the "elbow" in the plot

Hierachical Clustering

- Hierarchical clustering often associated with heatmaps!! very important
  - why? it organises heat map based on their similarities, hence the correlation visualise much more effectively.
  - heat maps: the columns represent different samples, the rows represent measurments from different genes.
  - Hierarchical clustering orders the row and/or the columns based on simailarity.
  - This makes it easy to see correlations in the data
  - Hierarchical clustering is usually accompanied by a "dendrogram"
  - It indicates both the similarity and the order that the clusters were formed.

Similarity - How do we define...

- the method for determining simialrity is arbitrarily chosen. However, the Euclidian distance between genes is used a lot. Most cases, Euclidian distance is default.
- Choice of distance matrix is arbitrary... There is no scientific reason to choose one and not the other.
- Pick the one that gives you more insight your data.

Ways to compare clusters

- The average of each cluster (called **centroid**)
- The closet point in each cluster (called **single-linkage**)
- The furthest point in each cluster (called **complete-linkage**)
- If use R, default setting complete-linkage is the default setting for the hclust() function