GSCM 410/510 – Summer 2021

Short-Answer Solutions Week 7

Directions

It is OK to work together on the Homework. Of course, make sure that everyone working together on an assignment is learning the material. Each person still turns in their own homework even if exactly the same as that of other group members. Learning this material well enough to do the homework problems is how you succeed in this course.

Short-Answer Problems

These concepts can appear on the short-answer part of the tests. As part of this homework, answer the following questions, usually just several sentences that include the definition.

1. *K-means cluster analysis minimizes the cluster inertia for each cluster. What is cluster inertia?*
   A key assessment of cluster analysis is how far a point lies from the center of the cluster to which it is assigned. The sum of the squared distances on each data value from the cluster center (square of Euclidean distance) for a point is an index of how far the point is from the center. Cluster inertia is total within-cluster variation, the sum of the individual within-cluster variations for all samples (objects) assigned to a cluster.

2. *A cluster should demonstrate cohesion and separation.*
   a. What are these concepts?
   Cohesion is a measure of how close a point is to its own cluster. Separation is a measure of how far a point is to the nearest cluster.
   b. What fit index simultaneously assesses these two concepts? How is it interpreted?
   The silhouette index assesses both cohesion and separation. Values close to 1 indicate cohesion and separation. Values of 0 indicate the point is on a cluster boundary. A negative value indicates the point likely does not belong in the cluster.

3. *What is the Pythagorean theorem so important to cluster analysis?*
   The standard measurement of distance between points, the basis for a cluster analysis where the coordinates of each point are the data values for a single sample, is Euclidean distance. That distance is computed with the application of the Pythagorean theorem to all *p* dimensions (number of features).

4. *What is a cluster centroid and how is it computed?*
   A cluster centroid is the center of the cluster. Each coordinate is the mean of the corresponding coordinate for each of the samples assigned to the cluster.

5. *Why is it good practice to investigate multiple initial configurations of centroids when pursuing a K-means cluster analysis?*
   There is no analytic solution for the optimum cluster solution even given the number of

clusters. Instead iteration is used. And iteration begins with an initial, trial solution, usually by more or less randomly seeding the centroids of the initial clusters. Unfortunately, the final solution can depend on the initial configuration because the iterative process cannot guarantee the best optimal solution, only an optimal solution relative to the starting points. Exploring solutions from different starting points can avoid accepting a solution less desirable than another solution obtained from a different starting configuration.

6. *a. What is the distinction between supervised learning and unsupervised learning?*
Supervised learning uses features X to forecast target y. There is a correct answer from which to evaluate the success of the analysis. Unsupervised learning uncovers structure without a target by which to aim for.
b. How can unsupervised learning be a preliminary step to supervised learning?
The structure uncovered buy a cluster analysis, for example, can then be used, perhaps with some modification, to define a target for future supervised learning.

7. *Express in words the calculation of Euclidean distance between two points calculated over p features. [Write the formula if you wish, but do describe verbally.]*
The calculation of Euclidean distance is the application of the Pythagorean theorem to the coordinates (data values) of the points. Take the difference of the data values for each feature across the two points, square the difference, sum the squared differences, and then take the square root.

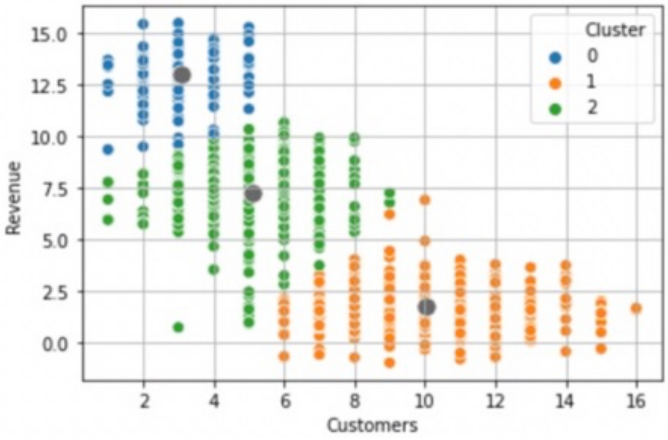
8. *Why is it important to standardize (or otherwise normalize) the data before pursuing a K-Means cluster analysis? Explain specifically of the importance in terms of K-Means.*
For example, length measured in millimeters yields much larger numbers than length measured in meters. Corresponding distances between points across a variety of variables would be much different depending on the unit, which different cluster analysis results obtained. Standardized ensures that the metrics of the variables are about the same. In particular, the same distribution of standard scores are obtained regardless if the length of something is measured in millimeters, meters, inches, feet, or whatever.

9. *Briefly describe the statistical procedure to assess the best number of clusters for a given data set.*
Following a theme of machine learning, find the best by empirically testing alternatives, and then select the best. The concept is hyper-parameter tuning. Systematically evaluate a variety of number of cluster solutions.

10. *What is the distinction between a parameter of a model, and a hyper-parameter? Give an example of each.*
A parameter is a characteristic of a specific model estimated for that model, such as specifying Euclidean distance for a cluster analysis, or a slope coefficient in a regression model. A hyper-parameter is a characteristic of a model that is changed over a range of analyses for the corresponding models, such as the number of clusters for a k-means cluster analysis.

11. *As a supplier you supply almost 1200 products to retail locations. You wish to develop an inventory strategy for your warehouse based on type of product. You analyze two features for*

each product, revenue and the number of customers during the last year who purchased the product. The resulting cluster analysis solution follows.



Note: There are multiple defensible answers to the b) especially.

a. *Interpret each of the three clusters. [one or sentences each]*
Blue: Products have high revenue and small number of customers.
Green: Products have relatively high revenue, and more customers.
Green: Each product has more customers for which to be concerned, but lower generated revenue.
Orange: Each product has many customers, but generates low revenue.

b. *What inventory strategy do you recommend for each group of products? [a few sentences each]*
Blue: Highest level of priority. Always maintain adequate inventory so never run out.
Green: High level of priority. Maintain adequate inventory.
Black: Lower priority. Try to order enough inventory, revenue is low.
Orange: Lowest priority. The product ships to many customers, each sale with low revenue. Again, try to order enough inventory, but the consequence of needing to re-order

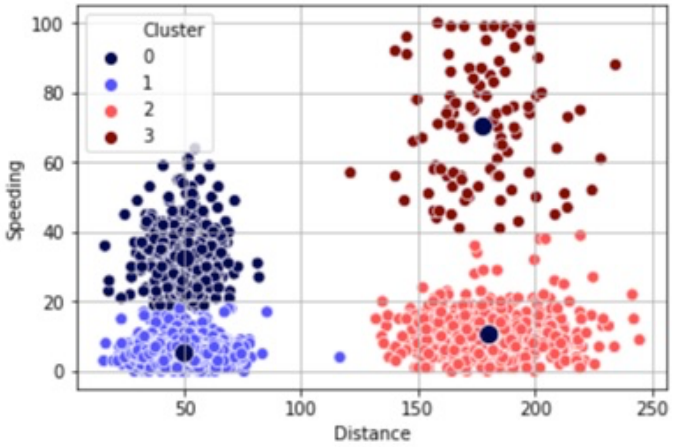
12. *For the above 4-cluster solution, the fit analysis across number of clusters yielded the following:*

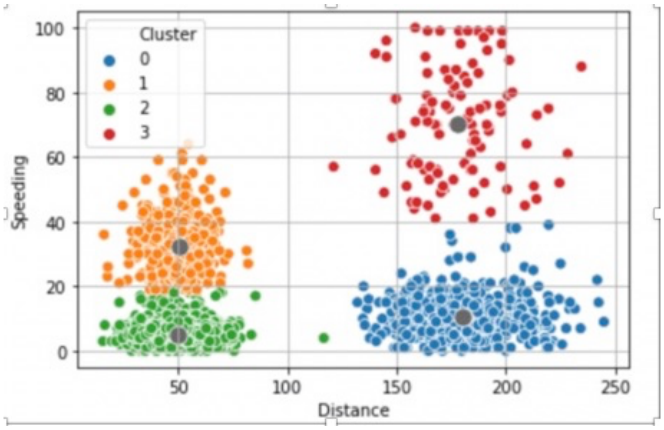
```
nc                2       3       4       5       6       7       8       9      10
----------------------------------------------------------------------------------------
Silhouette    0.623   0.587   0.509   0.459   0.441   0.426   0.407   0.417   0.406
Inertia       668.6   419.6   269.1   218.4   174.6   154.3   137.4   122.0   112.0
```

Justify the fours-cluster solution from these fit indices. Specify how you balance Silhouette and Inertia.
Inertia drops much for each increase in the number of clusters. Bu the drop amount itself decreases moving from 3 to 4 clusters. The silhouette index is almost 0.6 for 3 clusters, then drops almost down to 0.5 for 4 clusters.

13. *The prevalence of customer loyalty cards for grocery stores provides much data for analysis. To do develop a targeted advertising campaign, a typical large-scale grocery chain did a cluster analysis of its customers according to their purchases at a given location. What are some likely clusters do you think such an analysis might uncover?*
    Some possible answers include "junk-food junkie", "health-food foodie", "baby food and baby supplies", and "beer stop time".

14. *You are tasked with analyzing drivers who transport materials to various points in your supply chain. Because drivers are driving company trucks, you can collect real-time information regarding their driving habits for each trip. One analysis is of distance driven in miles and percentage of the time the driver was driving over 5mph of the speed limit. The resulting 4-cluster solution of several hundred trips by multiple drivers is shown below.*





   a. *Interpret each of the four clusters. [one or two sentences each]*
      Orange: Speed in low mileage, likely urban, deliveries.
      Green: Mostly drive the speed limit for urban deliveries.
      Blue: Most drive the speed for long-distance deliveries.
      Red: Speed for long-distance deliveries.

Green and Blue clusters of delivery trips do not have much speeding, with the clusters differentiated by the distance driven. The Orange cluster consists of trips in the city, yet still speeding. The Red cluster consists of longer distance trips, with regular speeding, some at 100% of the time.

b. *[Not asked, but of interest.] What additional variable would be useful in this analysis to make the results actionable?*

Would be good to know the driver for each trip. That would allow the analysis to identify which drivers are speeding.

c. For the above 4-cluster solution, the fit analysis across number of clusters yielded the following:

nc	2	3	4	5	6	7	8	9	10
Silhouette	0.708	0.764	0.784	0.734	0.734	0.664	0.359	0.36	0.36
Inertia	3911.93	1756.55	739.153	619.404	502.037	437.872	374.474	337.009	305.45

Justify the four-cluster solution from these fit indices. Specify how you balance Silhouette and Inertia.

Huge drop-offs of inertia from 2 to 3 to 4 clusters, then not so much from 4 to 5 clusters. For silhouette, the largest value is recorded for four clusters. Clearly a four-cluster solution.