due in 1wk 6d

Submission Phase

1. Do assignment ☐ (/algorithmicthink-001/human_grading/view/courses/972071/assessments/33/submissions)

Evaluation Phase

2. Evaluate peers 🔒 (/algorithmicthink-001/human_grading/view/courses/972071/assessments/33/peerGradingSets)

Results Phase

3. See results 🔒 (/algorithmicthink-001/human_grading/view/courses/972071/assessments/33/results/mine)

☐ In accordance with the Honor Code, I certify that my answers here are my own work, and that I have appropriately acknowledged all external sources (if any) that were used in this work.

Save draft | Submit for grading

Overview

In Project 3, you implemented two methods for clustering sets of data. In this Application, we will analyze the performance of these two methods on various subsets of our county-level cancer risk data set. In particular, we will compare these two clustering methods in three areas:

- **Efficiency** - Which method computes clusterings more efficiently?
- **Automation** - Which method requires less human supervision to generate reasonable clusterings?
- **Quality** - Which method generates clusterings with less error?

Important: Please use Cousera's "Attach a file" button to attach your plots/images for this Application as required. In particular, please do not host your solution plots/images on 3rd party sites. This practice exposes your peers to extra security risks and has the potential for abuse since the contents of a link to an external site may be modified after the hard deadline. Failure to follow this policy may lead to your plots/images being counted as "not submitted".

Efficiency

The next four questions will consider the efficiency of hierarchical clustering and k-means clustering. Note that successfully computing the 3108 county images for Questions 2 and 3 in desktop Python may require some fine tuning of your code for one or both methods.

Question 1 (2 pts)

Write a function `gen_random_clusters(num_clusters)` that creates a list of clusters where each cluster in this list corresponds to one randomly generated point in the square with corners $(\pm 1, \pm 1)$. Use this function and your favorite Python timing code to compute the running times of the functions `slow_closest_pairs` and `fast_closest_pair` for lists of clusters of size 2 to 200.

Once you have computed the running times for both functions, plot the result as two curves combined in a single plot. (Use a line plot for each curve.) The horizontal axis for your plot should be the the number of initial clusters while the vertical axis should be the running time of the function in seconds. Please include a legend in your plot that distinguishes the two curves.

Once you are satisfied with your plot, upload your plot in the box below using the "Attach a file" button (the button is disabled under the 'html' edit mode; you must be under the 'Rich' edit mode for the button to be enabled). Your plot will be assessed based on the answers to the following questions:

- Does the plot follow the formatting guidelines (https://class.coursera.org/algorithmicthink-001/wiki/ides?page=plotting) for plots? Does the plot include a legend?
- Do the two curves in the plot have the correct shapes?

B | I | Link | <code> | Math | Edit: Rich | Preview

Attach a file *(supports: txt, png, jpg, gif, pdf)*

Question 2 (1 pt)

Use `alg_project3_viz` to create an image of the 15 clusters generated by applying hierarchical clustering to the 3108 county cancer risk data set. Once you are satisfied with your image, upload your image in the box below using the "Attach a file" button (the button is disabled under the 'html' edit mode; you must be under the 'Rich' edit mode for the button to be enabled).

Your submitted image will be assessed based on whether it matches our solution image. You do not need to include axes, axis labels, or a title for this image.

B | I | Link | <code> | Math | Edit: Rich | Preview

Attach a file *(supports: txt, png, jpg, gif, pdf)*

Question 3 (1 pt)

Use `alg_project3_viz` to create an image of the 15 clusters generated by applying 5 iterations of k-means clustering to the 3108 county cancer risk data set. As in Project 3, the initial clusters should correspond to the 15 counties with the largest populations. Once you are satisfied with your image, upload your image in the box below using the "Attach a file" button (the button is disabled under the 'html' edit mode; you must be under the 'Rich' edit mode for the button to be enabled).

Your submitted image will be assessed based on whether it matches our solution image. You do not need to include axes, axis labels, or a title for this image.

B I Link <code> Math Edit: Rich Preview

Attach a file *(supports: txt, png, jpg, gif, pdf)*

Question 4 (1 pt)

Which clustering method is faster when the number of output clusters is either a small fixed number or a small fraction of the number of input clusters? Provide a short explanation in terms of the asymptotic running times of both methods. You should assume that `hierarchical_clustering` uses `fast_closest_pair` and that k-means clustering always uses a small fixed number of iterations.

B I Link <code> Math Edit: Rich Preview

Automation

In the next five questions, we will compare the level of human supervision required for each method.

Question 5 (1 pt)

Use `alg_project3_viz` to create an image of the 9 clusters generated by applying hierarchical clustering to the 111 county cancer risk data set. Once you are satisfied with your image, upload your image in the box below using the "Attach a file" button (the button is disabled under the 'html' edit mode; you must be under the 'Rich' edit mode for the button to be enabled).

Your submitted image will be assessed based on whether it matches our solution image. You do not need to include axes, axes labels, or a title for this image.

B I Link <code> Math Edit: Rich Preview

Attach a file *(supports: txt, png, jpg, gif, pdf)*

Question 6 (1 pt)

Use `alg_project3_viz` to create an image of the 9 clusters generated by applying 5 iterations of k-means clustering to the 111 county cancer risk data set. As in Project 3, the initial clusters should correspond to the 9 counties with the largest populations. Once you are satisfied with your image, upload your image in the box below using the "Attach a file" button (the button is disabled under the 'html' edit mode; you must be under the 'Rich' edit mode for the button to be enabled).

Your submitted image will be assessed based on whether it matches our solution image. You do not need to include axes, axes labels, or a title for this image.

B | I | Link | <code> | Math | Edit: Rich | Preview

Attach a file *(supports: txt, png, jpg, gif, pdf)*

Question 7 (1 pt)

The clusterings that you computed in Questions 5 and 6 illustrate that not all clusterings are equal. In particular, some clusterings are better than others. One way to make this concept more precise is to formulate a mathematical measure of the error associated with a cluster. Given a cluster C, its *error* is the sum of the squares of the distances from each county in the cluster to the cluster's center, weighted by each county's population. If p_i is the position of the county and w_i is its population, the cluster's error is:

$$error(C) = \sum_{p_i \in C} w_i (d_{p_i c})^2$$

where c is the center of the cluster C. The `Cluster` class includes a method `cluster_error(data_table)` that takes a `Cluster` object and the original data table associated with the counties in the cluster and computes the error associated with a given cluster.

Given a list of clusters L, the *distortion* of the clustering is the sum of the errors associated with its clusters.

$$distortion(L) = \sum_{C \in L} error(C).$$

Write a function `compute_distortion(cluster_list)` that takes a list of clusters and uses `cluster_error` to compute its distortion. Now, use `compute_distortion` to compute the distortions of the two clusterings in questions 5 and 6. Enter the values for the distortions (with at least four significant digits) for these two clusterings in the box below. Clearly indicate the clusterings to which each value corresponds.

As a check on the correctness of your code, the distortions associated with the 16 output clusters produced by hierarchical clustering and k-means clustering on the 290 county data set are approximately 2.575×10^{11} and 2.323×10^{11}, respectively.

B I Link <code> Math Edit: Rich Preview

Question 8 (1 pt)

Examine the clusterings generated in Questions 5 and 6. In particular, focus your attention on the number and shape of the clusters located on the west coast of the USA.

Describe the difference between the shapes of the clusters produced by these two methods on the west coast of the USA. What caused one method to produce a clustering with a much higher distortion? To help you answer this question, you should consider how k-means clustering generates its initial clustering in this case.

In explaining your answer, you may need to review the geography of the west coast of the USA (http://en.wikipedia.org/wiki/West_Coast_of_the_United_States).

B I Link <code> Math Edit: Rich Preview

Question 9 (1 pt)

Based on your answer to Question 8, which method (hierarchical clustering or k-means clustering) requires less human supervision to produce clusterings with relatively low distortion?

B I Link <code> Math Edit: Rich Preview

Quality

In the last two questions, you will analyze the quality of the clusterings produced by each method as measured by their distortion.

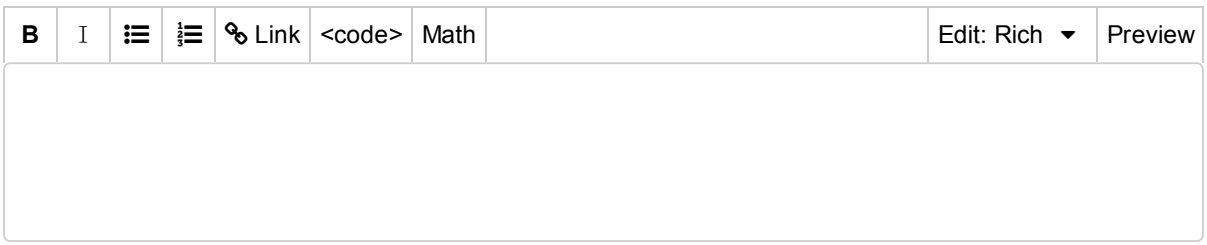
Question 10 (4 pts)

Compute the distortion of the list of clusters produced by hierarchical clustering and k-means clustering (using 5 iterations) on the 111, 290, and 896 county data sets, respectively, where the number of output clusters ranges from 6 to 20 (inclusive). **Important note:** To compute the distortion for all 15 output clusterings produced by `hierarchical_clustering`, you should consider integrating the computation of the distortion for each intermediate clustering directly into `hierarchical_cluster`. Otherwise, you will introduce an unnecessary factor of 15 into the computation of the distortions for the hierarchical clusters.

Once you have computed these distortions for both clustering methods, create three separate plots (one for each data set) that compare the distortion of the clusterings produced by both methods. Each plot should include two curves drawn as line plots. The horizontal axis for each plot should indicate the number of output clusters while the vertical axis should indicate the distortion associated with each output clustering. For each plot, include a title that indicates the data set used in creating the plots and a legend that distinguishes the two curves.

Once you are satisfied with your plots, upload them in the box below using the "Attach a file" button (the button is disabled under the 'html' edit mode; you must be under the 'Rich' edit mode for the button to be enabled). Your plots will be assessed based on the answers to the following questions:

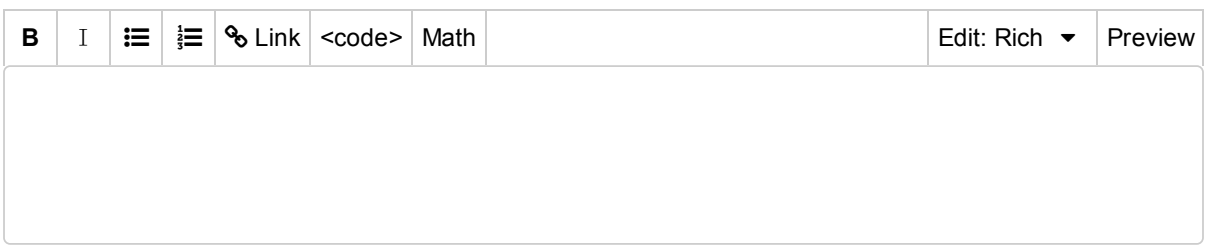
- Do the plots follow the formatting guidelines (https://class.coursera.org/algorithmicthink-001/wiki/ides?page=plotting) for plots? Does the title of each plot indicate which data was used to create the plot? Do the plots include a legend?
- Do the two curves in each plot have the correct shapes?



Attach a file *(supports: txt, png, jpg, gif, pdf)*

Question 11 (1 pt)

For each data set (111, 290, and 896 counties), does one clustering method consistently produce lower distortion clusterings when the number of output clusters is in the range 6 to 20? Is so, indicate on which data set(s) one method is superior to the other.



Overall evaluation (1 pt Extra Credit)

Which clustering method would you prefer when analyzing these data sets? Provide a summary of each method's strengths and weaknesses on these data sets in the three areas considered in this application.

Your summary should be at least a paragraph in length (4 sentences minimum).

B | I | Link | <code> | Math | Edit: Rich | Preview

In accordance with the Honor Code, I certify that my answers here are my own work, and that I have appropriately acknowledged all external sources (if any) that were used in this work.

Save draft | Submit for grading