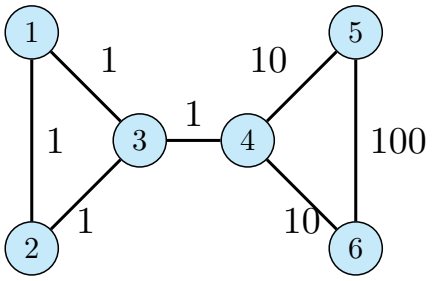
Homework 4 for CSI 431

Due: Mon, Dec 3 at 23:59:59

All homeworks are individual assignments. This means: write your own solutions and do not copy code/solutions from peers or online. Should academic dishonesty be detected, the proper reporting protocols will be invoked (see Syllabus for details).

Instructions: Submit two files. One should be a write-up of all solutions and observations, as *Solution.pdf*. The second should be an archive *Code.zip* containing code and any relevant results files.

1. **[50 pts.] Spectral Clustering:** Consider the following similarity graph with similarity values annotated on edges.



   (a) **[5 pts.] Cuts:** Consider the following cuts:

$$\text{cut 1: } S_1 = \{1,2,3\}, T_1 = \{4,5,6\}$$

$$\text{cut 2: } S_2 = \{1\}, T_2 = \{2,3,4,5,6\}$$

$$\text{cut 3: } S_3 = \{1,2,3,4\}, T_3 = \{5,6\}$$

   Compute:
   (1) the cut weight $W(S_i, T_i)$;
   (2) the ratio cut: $\frac{W(S_i,T_i)}{|S_i|} + \frac{W(S_i,T_i)}{|T_i|}$; and
   (3) the normalized cut $\frac{W(S_i,T_i)}{vol(S_i)} + \frac{W(S_i,T_i)}{vol(T_i)}$
   for each of the cuts ($i = 1 \ldots 3$). How do the three cuts rank (from smallest to highest) according to each of the cut criteria?

   (b) **[5 pts.] Laplacians:** Compute the adjacency matrix A, degree matrix D, Laplacian matrix $L = D - A$ and Symmetric Laplacian matrix $L_s = D^{-1/2} L D^{-1/2}$ of the graph above (do this by hand and show them in your Solutions file).

   (c) **[20 pts.] Eigen vectors (requires coding):** Encode L_s and L (you obtained above) into a python program and perform eigen decomposition to obtain the eigenvectors corresponding to the **smallest non-zero eigenvalue** of both L and L_s. (Hint: Use *numpy.linalg.eig*(X) to compute the eigen decomposition of L and L_s.) Let u be the aforementioned eigenvector for L and u_s for L_S. Report the values corresponding to each nodes in u and u_s. Plot these values, where on the x axis you have the node id (1 to 6) and on the y axis you have the corresponding value of u and u_s.

   Hint: Note that since numpy uses numerical methods to find eigenvalue/eigenvector pairs, numerical errors can lead to the first eigenvalue being non-zero but rather some small number like 10^{-15}*. This is still actually a zero eigenvalue.*

   (d) **[10 pts.] Hierarchical clusters from eigenvectors:** Here we will do agglomerative single-link clustering (on paper, NOT in a program) of the nodes represented by their u values and by their u_s values computed in the previous part. Note that we will get two dendrograms, one for u and one for u_s and in each the nodes will be one-dimensional points. Draw the two dendrograms.

(e) **[10 pts.] Hierarchical clusters from eigenvectors:** Cut the dendrograms above to obtain two clusters. Show the resulting partitions in your solution file. What are their cut weights, normalized cuts and ratio cuts? Discuss why we obtain these partitions by relating the Laplacians to the specific cut measures.

2. **[20 pts.]** There is a reason why we typically use the mean of data points as cluster representative when we do k-means with L2 distance measure: the mean minimizes the sum of squared distances to all points. In this question we will convince ourselves theoretically and by an example that these are good choices.

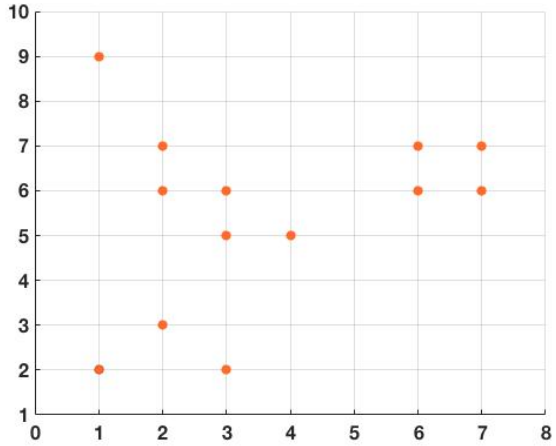
   (a) **[10 pts.] Example :** Consider the following set of numbers $\{0, 1, 2, 2, 10\}$. Compute the mean μ and median m. Compute the sum of squared distances to the mean $\sum_i (x_i - \mu)^2$ and the sum of squared distances to the median $\sum_i (x_i - m)^2$. Which one is bigger?

   (b) **[10 pts.]** Show that the sum of squared distances from the mean is smaller than that to the median in general, i.e.

   $$\sum_i (x_i - \mu)^2 \leq \sum_i (x_i - m)^2.$$

   Hint: you need to mathematically prove it, not use specific numbers as in the previous question.

3. **[30 pts.] Density-based Clustering:** Considering the data points in the figure below, answer the following questions.



   (a) **[5 pts.]** L_∞ **:** Using $\epsilon = 1$, $minpts = 3$, and L_∞ distance, find all core, border and noise points. Hint: The distance between points according to L_∞ is defined as the maximum absolute difference in any of the dimensions, i.e. $d_\infty(x, y) = max_{i=1...d}|x_i - y_i|$

   (b) **[8 pts.]** L_∞ **Clusters :** Find all the clusters found by DBSCAN using the setting in part (a) and show them on the figure.

   (c) **[5 pts.]** L_2 **:** Using $\epsilon = 1$, $minpts = 3$, and L_2 distance, find all core, border and noise points. Compare your results with part (a). What differences do you see? Why?

   (d) **[12 pts.]** L_2 **Clusters :** Find all the clusters found by DBSCAN using the setting in part (c) and show them on the figure. Compare your results with part (b). What differences do you see? Why?