67978: A Needle in a Data Haystack
Introduction to Data Science
Homework #2: Similar items, Clustering, Community Detection
Due: **18 Dec, 11:59pm, on Moodle**

Read the instructions **carefully** (that's a good idea in general).

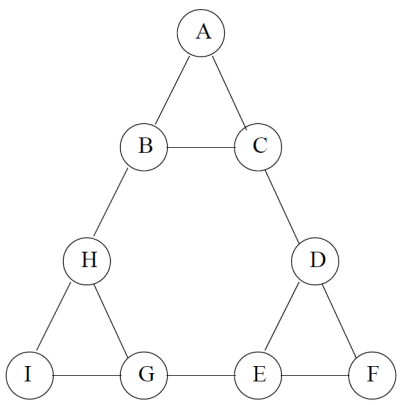
- There will be two submission links, for theoretical and practical parts.
- Each person submits their own theoretical part. The theoretical part should be a single file in pdf format only (no docx or jpg) named **ex2t_ID.pdf** (ID is your ID).
- If you are submitting handwritten answers, make sure they are crystal clear.
- The practical part must be submitted in **groups of three**. In the submission link, enter your partners (using the **add partner** button). In addition, all three people should write the name and id of their partners in the theoretical part pdf.
- In the practical part, submit a single ZIP file named **ex2p.zip**. The zip should contain a folder named **code** and a folder called **output**. Make sure to submit only the relevant files.
- Points may be reduced for submissions that fail to comply.
- The **meta** question should be answered through a questionnaire. The link will be posted on the moodle.
- Make sure you follow the News forum and HW forum for any updates.

Problem 1 (Finding Similar Items).

(a) Prove or disprove: if the Jaccard similarity of two columns is 0, then minhashing always gives a correct estimate of the Jaccard similarity.

(b) One might expect that we could estimate the Jaccard similarity of columns without using all possible permutations of rows. For example, we could only allow cyclic permutations; start at a randomly chosen row r, which becomes the first in the order, followed by rows $r+1$, $r+2$, and so on, down to the last row, and then continuing with the first row, second row, and so on, down to row $r-1$. There are only n such permutations if there are n rows. However, these permutations are not sufficient to estimate the Jaccard similarity correctly. Give an example of a two-column matrix where averaging over all the cyclic permutations does not give the Jaccard similarity. Compute Jaccard and average similarity.

Problem 2 (Communities).



(a) Use the Girvan-Newman approach to find the number of shortest paths from each of the following nodes that pass through each of the edges. (1) Node H (2) Node I.

(b) Using symmetry, these are all the calculations you need to compute the betweenness of each edge. Do the calculation.

(c) Using your results for (b), pick a threshold and remove the edges with higher betweenness. What is the threshold? What are the communities?

(d) Compute the Laplacian matrix for this graph, find the second-smallest eigenvalue and its eigenvector. What communities does it suggest?

Problem 3 (k-means on the real line). In this question we assume single dimension points. We apply k-means with the objective $\sum_{j=1..n} \min_{i=1..k} ||x_j - c_i||^2$, measuring the sum of squared distances from each point x_j to its nearest center.

(a) Consider the case where k = 3 and we have 4 data points $x_1 = 1$, $x_2 = 3$, $x_3 = 6$, $x_4 = 7$. What is the optimal clustering for this data? What is the corresponding value of the objective?

(b) Show that there exists a suboptimal cluster assignment for the data in part (a) that the algorithm will not be able to improve (show the assignment, why it is suboptimal and explain why it will not be improved).

(c) Assume we sort our data points such that $x_1 \leq x_2 \leq ...x_n$. Prove that in an optimal cluster assignment each cluster corresponds to some interval of points.

(d) Design an $O(kn^2)$ algorithm for k-means on the real line.

Problem 4 (Clustering Part I (Coding question)).

Synthetic data is information that's artificially manufactured. Synthetic data is created algorithmically, and it is used as a stand-in for test datasets, to validate models and, increasingly, to train machine learning models. It is very useful for testing clustering methods.

Your goal is to generate synthetic data. For each task, generate 300 random points (in $\mathbb{R}^2$. That is, 2D) and plot them. Repeat this **TWICE** (so two plots, 300 points each for each task).

(a) Uniform distribution, $x \in [-10, 1], y \in [17, 35]$

(b) Gaussian with center at [5,1] and std=3.

(c) Three Gaussians with centers at $[i, -i]$ and std=$0.5 \times i$ $(i = 1, 2, 5)$.

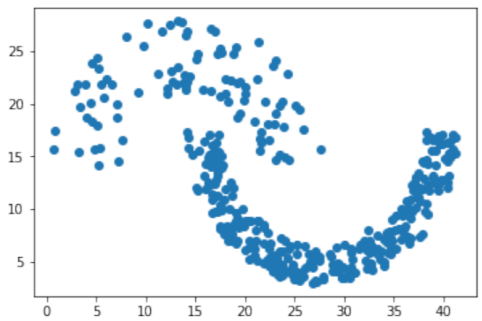
(d) A circle inside a ring.

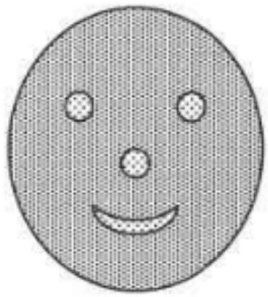
(e) The first letter of your first names (so the data should look like a noisy version of "NDH", for your own letters; if they are the same letter, use last names. :)).

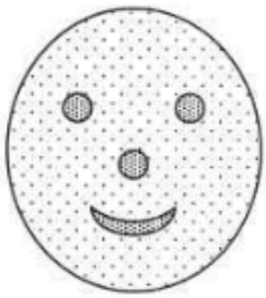
(f) Two moons, one sparser than the other (see example in the Figure, left)

(g) Face, with nose, mouth and eyes sparser than the face (fig, middle)

(h) Face, with nose, mouth and eyes denser than the face (fig, right)







Problem 5 (Clustering Part II (Coding Question)).

For the datasets you generated above (one for each (a)-(h), not the pair), run the following algorithms and plot the result (use your original plot, with colors of points indicating the cluster id). You can use existing implementations of the algorithms.

(a) K-means. Plot for $k = 2, 3, 4$.

(b) Perform a hierarchical clustering. Assume clusters are represented by their centroid (average). At each step merge the clusters with the closest centroids. Plot for $k = 2, 3, 4$.

(c) Perform a hierarchical clustering. Assume clusters are represented by their centroid (average). At each step merge the two clusters whose resulting cluster has the smallest diameter. Plot for $k = 2, 3, 4$.

(d) DBSCAN: set parameters so you get two different numbers of clusters (and plot both). Report parameters.

Problem 6 (Meta). How long (in hours) did this assignment take? Please answer using the link, not in the pdf.