COM S 474/574 Spring 2020
Introduction to Machine Learning

Homework 6

1 Directions:

- **Due: Friday April 24, 2020 at 10pm.** Late submissions will be accepted for 24 hours after that time, with a 15% penalty. [note: we will keep office hours the same days/times]
- Upload the homework to Canvas as a pdf file. Answers to problem 1 can be handwritten, but writing must be neat and the scan should be high-quality image. Other responses should be typed or computer generated.
- Any non-administrative questions must be asked in office hours or (if a brief response is sufficient) Piazza.

2 Problems

Problem 1. [25 points] Ch 10 # 2 "Suppose that we have four…"

Show your calculations. Recall that "single linkage" means we define the distance of two clusters C_1 and C_2 as the minimum distance of any pair of elements

$$dist(C_1, C_2) \ := \ \min_{a \in C_1, b \in C_2} dist(a, b)$$

and that "complete linkage" means we define the distance of two clusters C_1 and C_2 as the maximum distance of any pair of elements

$$dist(C_1, C_2) \ := \ \max_{a \in C_1, b \in C_2} dist(a, b).$$

Problem 2. [20 points] On canvas there is Python code `HW6-PCA-template`, saved as a Jupyter notebook and a webpage. It applies PCA to a digit dataset. Each image is 8 by 8 pixels and gray scale, thus can be stored as a 64 dimensional vector.

A. Make a plot of the cumulative explained variance. You should get a curve that starts close to 0 and monotonically increases to 1 (reaching 1 at least when all dimensions are used).

B. Report the cumulative explained variance for the best two components.

C. Make images of the first five samples using just PCA with `n_components=2`, and display them side by side next to the original images. (eg we find the best two-dimensional projection for the data set, then project back to visualize how much/what information was preserved.)

D. In 2-3 sentences, comment on how visually similar (or not) the images made after PCA transformation are to the originals.

E. Repeat the previous 3 steps with 32 components (half that of the original data).

Problem 3. [20 points] We will apply clustering to the HW 3 data. Just use the `HW3train.csv`. Example Python code is posted as `HW6-cluster-template`.

A. Make a scatter-plot of the data, coloring each data point black.

B. For `n_clusters=range(1,16)`, apply K-means clustering. Make a scatter plot for each. You only need to include 3 of them in your homework submission. Select the three pictures whose clusters you think look the best.

C. For bottom-up hierarchical clustering (aka 'Agglomerative Clustering'), make a dendogram using 'single' linkage.

D. Manually select and report a distance threshold for single linkage. Look for regions in the dendogram where there are few mergers (eg a big vertical gap in distance threshold between mergers). Use that to make a scatter plot of the data clustered based on that threshold.

E. Repeat the previous two steps, using 'average' linkage.

F. Repeat the previous two steps, using 'complete' linkage.

G. Using 1-3 sentences for each clustering method (k-means, single linkage aggl., average linkage aggl., and complete linkage aggl.), comment on the clusters found and how they compare to the other methods.
In 1-3 sentences, which clustering method do you think resulted in the best clusters and why?

If you need more than 15 colors (eg you pick thresholds resulting in more than 15 clusters), you can pick some at `https://htmlcolorcodes.com/` and add their hex values to the `colors` list in the code.