CSC 240/440 - 2017 Spring: Homework 4

Hand in the hardcopy before the class on Apr. 18

Requirement

Due to the request from some students, the homework is posted online right now, but would be updated probably every week until it is formally released. (A) or (G) indicates questions for all or just graduate students. Undergraduate students are not required to do (G) questions, but they can get bonus points from that. Please hand in the hardcopy of your homework before the class.

The homework must be completed individually. However, you are encouraged to discuss the general algorithms and ideas with classmates in order to help you answer the questions. If you work with one or more other people on the general discussion of the assignment questions, please record their names over every question they participated.

However, the following behaviors will receive heavy penalties (lose all points and apply the honest policy explained in syllabus)

- explicitly tell somebody else the answers;
- explicitly copy answers or code fragments from anyone or anywhere;
- allow your answers to be copied;
- get code from Web.

Please also indicate how many late days you want to apply to your submission (Check the late policy in the Syllabus). All late submission without indicating late days or running out the late days cannot be accepted. For medical reasons, if the homework in time cannot be submitted on time, you have to submit the certificate with your homework.

1 (A) Hierarchical clustering (6 points)

This is a programming question. The input data is a set of 2-D points. You are asked to implement the hierarchical clustering algorithm to divide the input data into two groups. The reasonable clustering result is that the points in the center is a group while the points surrounding them is the other group. You should properly choose your cluster distance.
Please report the following results:

- the distance measure you used in the algorithm;
- your clustering result shown in a figure (you should use different signs for data points in different clusters; Do not use Photoshop or similar softwares);
- your answer to the following question: can you get the same result using K-means? Explain your conclusion.

You should upload your code to Blackboard and report your results as hard copy (hand in it together with your answers for other questions).

2 (A) PCA (5 points)

This is a programming question about PCA. Consider the handwritten digit dataset (the same as you used in your last homework).
For this question, you only use the gray-scale images of hand-drawn digits with “0, 6, and 8”. Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255, inclusive. Each data set contains 785 columns. The first column (Column 0) is the label (You should not use it in this question! You need to pretend that you do not know the label): the digit drawn by the user. The rest of the columns contain the pixel-values of the associated image in the following way:

\[
\begin{bmatrix}
\text{Col.1} & \text{Col.2} & \text{Col.3} & \ldots & \text{Col.28} \\
\text{Col.29} & \text{Col.30} & \text{Col.31} & \ldots & \text{Col.56} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\text{Col.757} & \text{Col.758} & \text{Col.759} & \ldots & \text{Col.784}
\end{bmatrix}
\]

Let \( N \) be the total number of images with 0, 6, or 8. Let \( x_n \in \mathbb{R}^{784} \) denote a vectorized image and \( \{x_n\}_{n=1}^N \) be the set of all vectorized images. Apply PCA to compute the low dimension representation \( y_n \)'s for all \( x_n \)'s and draw all points \( y_n \)'s in a figure. The dimension of \( y_n \) is chosen to 2. (Note that so far you should pretend not to know the true labels for images when you implement K-means. But when you draw the figure, you can indicate points by its true label to see if your dimension reduction makes sense or not.)

You should upload your code to Blackboard and report your results as hard copy (hand it together with your answers for other questions).

3 (A) K-means (5 points)

This is a followed question. Given the low dimension representation learned from the former question, you are asked to apply the K-means algorithm to cluster all low dimensional representation points \( y_n \). Choose the number of clusters to be 3.

Please report the following results:

- the distortion function value 
  \[
  S = \sum_{n \in \text{cluster 1}} \|y_n - c_1\|^2 + \sum_{n \in \text{cluster 2}} \|y_n - c_2\|^2 + \sum_{n \in \text{cluster 3}} \|y_n - c_3\|^2
  \]
  where \( c_1, c_2 \) and \( c_3 \) are the centers for three clusters you obtain;

- your clustering result shown in a figure (you should use different marks to denote data points in different clusters; Do not use Photoshop and similar softwares).

- the comparison to the figure you print out in the previous question and write down your comments.

You should upload your code to Blackboard and report your results as hard copy (hand in it together with your answers for other questions).