CS 434: Assignment 4

Due May 28th 11:59PM, 2019

General instructions.

1. You can work in team of up to 3 people. Each team will only need to submit one copy of the source code and report.
2. You need to submit your source code (self contained, well documented and with clear instruction for how to run) and a report via TEACH. In your submission, please clearly indicate your team members' information.
3. Your code will be tested on flip. Please make sure that your program can run on flip without any issue.
4. Be sure to answer all the questions in your report. Your report should be typed, submitted in the pdf format. You will be graded based on both your code as well as the report. In particular, the clarity and quality of the report will be worth 10 % of the pts. So please write your report in clear and concise manner. Clearly label your figures, legends, and tables.

1 Data description

In this implementation assignment you will explore clustering. The provided data set contains handwritten digits. Data file has 6000 rows, where every row is a 784 dimensional vector that represents a particular digit.

2 Non-hierarchical clustering - K-Means algorithm

1. (25 pts) Implement the K-means algorithm. Run your K-means algorithm with $k = 2$. To verify that your algorithm actually converges, please plot the objective of the K-means algorithm (i.e., the SSE) as a function of the iterations. From one run to another run, this curve may look different. Just present the results of a typical run.
2. (25 pts) Now apply your K-means implementation to this data with different values of k (consider values $2, 3, \cdots, 10$). For each value of k, please run your algorithm 10 times, each time with a different random initialization, record the lowest SSE value achieved in these 10 repetitions for each

value of k. Plot the recorded SSE values against the changing k value. What do you think would be a proper k value based on this curve? Please provide justification for your choice.

Write your code so that it can be run with the following command:

python kmeans.py k

Your code should run the K-means algorithm with the specified value of k for a finite number of iterations. For each iteration, your code should print out the SSE at the current iteration. Notice that this format will be used to test your program. You should include the plots specified in q2.1 and q2.2 in your report.

3 Principal Component Analysis

1. (15 pts) Implement Principal Component Analysis for dimension reduction. Specifically, your program needs to compute the mean and covariance matrix of the data, and compute the top ten eigen-vectors with ten largest eigen-values of the Covariance matrix (you can use existing functions in numpy to compute the eigen-values and eigen-vectors). Report the eigen-values in decreasing order.

   Write your code so that it can be run with the following command:

   python pca_1.py

   The output should include:

   - Top-ten eigen-values in decreasing order.

2. (15 pts) Plot the mean image, and each of the top ten eigen-vectors. To make the image for eigen-vectors viewable, you should re-scale each eigen-vector by its maximum value. Inspect the resulting images, what do you think they each capture?

   Write your code so that it can be run with the following command:

   python pca_2.py

   The output should include:

   - For each of the top ten eigen-vectors, plot the mean image and the eigen-vector. Include these images in your report.

3. (10 pts) Use the top 10 eigen-vectors to project each image to 10 dimensions. Identify for each dimension the image that has the largest value in that dimension and plot it. Compare the image with its corresponding eigen-vector image, what do you observe? What do you think the reduced 10-dimensional representation is capturing in this case?

   Write your code so that it can be run with the following command:

   python pca_3.py

   The output should include:

- For each dimension, plot the image that has the largest value in that dimension. Include these images in your report.