HOMEWORK 4:
UNSUPERVISED LEARNING

CPEN 355 @ UBC
TAs: Chun-Yin Huang (Primary), Wenlong Deng

March 25th, 2024

Instructions

- **Homework Submission:** Submit your code and report to Canvas. You will use Co-lab to implement the coding tasks. Please check Piazza for updates about the homework.
  - Upload a zip file containing two files: Your report in .PDF format and your notebook in .Ipynb format.
  - To ensure the reproducibility of your results: (1) set a seed for numpy and python random modules on top of your Colab notebook (2) restart and run all the cells of your notebook once before submission.
- **Collaboration policy**: The purpose of student collaboration is to facilitate learning, not to circumvent it. Studying the material in groups is strongly encouraged. It is also allowed to seek help from other students in understanding the material needed to solve a particular homework problem, provided no written notes (including code) are shared, or are taken at that time, and provided learning is facilitated, not circumvented. The actual solution must be done by each student alone. The presence or absence of any form of help or collaboration, whether given or received, must be explicitly stated and disclosed in full by all involved.

1 Coding practice [10 pts]

There are two problems in this coding practice. In the first problem, you will study how different numbers of principal components represent the images visually. For the second problem, you will use logistic regression to predict the class label of images using the principal components representation of the images and examine how the classification error changes with the number of principal components used. We will explore the MNIST handwriting digits loaded from tensorflow. More details and hints are available in 'HW4.ipynb.'

1.1 PCA for dimension reduction (3pt)

(a) For $k = 0, 10, 20, 30, 40, 49$, use k-th principal components ONLY, namely, $\hat{x}_i = \bar{X} + (\phi_k^\top x_i)\phi_k$, where $\hat{x}_i$ is the reconstructed image, $\bar{X}$ is the averaged image, and ϕ_k is the k-th principle axis for MNIST 0's to approximately reconstruct the image selected above. Note that we index from 0, namely 0-th principal component is the first one. Display the reconstruction for each value of k. To display the set of images compactly, you may want to use the 'plot_images' function defined in 'HW3.ipynb.'

1.2 PCA for classification (7 Pts)

(a) Load in the MNIST data with the labels as y and the images as x by running the next cell in the notebook. Create a subset of the data by keeping only the images that have the label of either 4 or 9. Use Principal Components Analysis (PCA) to project the data onto the first two principal components, and create a plot of the projected data color-coded by the label. Does the plot make sense? Explain in a couple of sentences. You don't need to do image reconstruction for this question.

(b) Why not use more principal components? For $k = 2, 3, 4, ..., 15$, use PCA to project the data onto k principal components. For each k, you will end up with k dimensional representation for each data point. Then use the k dimensional representations and logistic regression to build a model to classify images as 4 or 9, and calculate the accuracy of the model. Create a plot of accuracy as a function of k, the number of principal components used. Does the plot make sense? Explain in a few sentences. You don't need to do reconstruction for this question.
Note: you need to report the accuracy on the *test set*, not the training set.

2 Multiple Choice Questions [2 pts]

(a) Given a set of seven one-dimensional data points $\{1, 2, 3, 4, 5, 6, 7\}$, suppose we run the K-Means algorithm to cluster them into two groups. The initial cluster centers are set at 1.8 and 2.8. Determine the final cluster centers after the K-Means algorithm converges. *Please include your intermediate derivations leading to the final answer.*

(A) 1.5 and 5.0

(B) 2.0 and 5.5

(C) 2.5 and 6.0

(D) 3.0 and 5.0

(b) Determine which of the following statements about PCA are *true*.

(A) PCA is a deterministic algorithm, and it will always produce the same results for the same input data.

(B) The first principal component captures the least variance in the data.

(C) PCA is a linear technique for dimensionality reduction.

(D) PCA is only applicable to categorical data.

Note

1. Remember to submit your assignment by 11:59pm of due date. Late submission will affect your scores.
2. If you submit multiple times, ONLY the content and time-stamp of the latest one would be considered.
3. We strictly follow the rules of UBC Academic Misconduct.