DATA 240-21, Fall 2024

Bonus Assignment #1

Release on Oct 3rd , 2024
Due 11:59pm on Oct 22nd, 2024

Notes

This assignment should be submitted in Canvas as a format of ipython notebook (bonusHW1.ipynb).

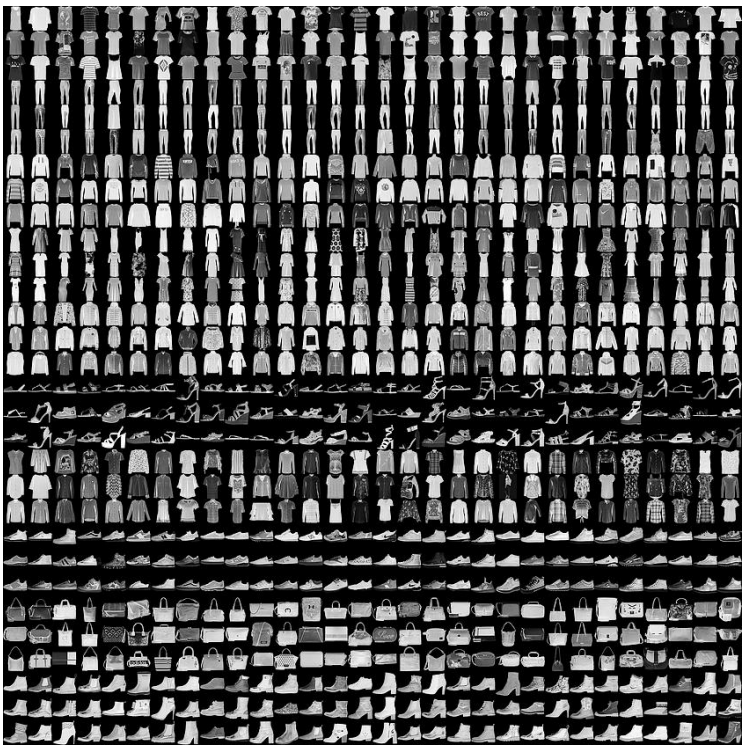
This bonus assignment is optional, not mandatory. You would get additional credit.

No late assignments will be accepted. Do not accept any other format. Minimum penalty is 2pts with acceptable excuse.

You may collaborate on homework but must **independently** write code/solutions. Copying and other forms of cheating will not be tolerated and will result in a **zero score** for the homework (minimal penalty) or a failing grade for the course. Your work will be graded in terms of correctness, completeness, and clarity, not just the answer. Thus, correct answers with no or poorly written supporting steps may receive very little credit.

*NOTE: Please do not use any package/library including scikit-learn library except NumPy, Pandas, Matplotlib, and **SVD** library.*

Please download fashion_mnist1.csv.
The dataset has 10K rows and 785 columns.



1. (2 pts) Dimension reduction using principal component analysis (PCA)

If the data is highly dimensional, you can use PCA to find a reduced-rank approximation of the data that can be visualized easily.

1.1(1pt) Using Singular-Value Decomposition (SVD) method, please decompose 1st and 2nd principal components and project them to the data.
Please draw 2D plot using the 1st and 2nd principal components.
Please legend different colors for the 10 labels in the graph.

1.2(1pt) As we discussed during the class, we can compress images using PCA.

$$A \approx U_k \sum_k V_k^T$$

Please select 10 images. The selected images should have different labels. Please compress the images using k=2, 5, 10. Then, visualize the original images and the compressed images with different k.

2. (5 pts) Dimension reduction using t-SNE

2.1(2pts) Please build t-SNE algorithm from scratch based on the below equations.

$$P_{j|i} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right)}$$

$$q_{j|i} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

$$C = KL(P||Q) = \sum_i \sum_j P_{j|i} log \frac{P_{j|i}}{q_{j|i}}$$

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + \|y_i - y_j\|^2\right)^{-1}$$

$$y^{(t)} = y^{(t-1)} + \lambda \frac{\partial C}{\partial y} + \alpha(t)\left(y^{(t-1)} - y^{(t-2)}\right)$$

Please cite if you are referring to any source for the algorithms.
There are many hyperparameters to optimize, such as initialization (random seed), learning rate λ , momentum $\alpha(t)$, iteration number, and perplexity.

2.2(1.5pts) Using the t-SNE method, please reduce the 784 dimensions to 2 dimensions.
Please try at least 5 different hyperparameters conditions.

For each hyperparameter condition, please calculate its corresponding D and J
Please calculate the sum of the distance D among the 10 centroids. Each centroid corresponds to each label. Since there are 10 centroids, you should calculate distance for 45 pairs.

$$D = \sum_{i,j} \|y_i - y_j\|^2$$

Also, please calculate the objective function J.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^j - c_j\|^2$$

Your goal is to maximize the distance D among the centroids and to minimize the objective function J by optimizing the hyperparameters.

2.3(1.5pts)
Please draw 2D plot for the 5 hyperparameter conditions.
Please legend different colors for the 10 labels in the graph.