Module 12 - Bonus Challenges

§M12: Dimensionality Reduction

- Below are open-ended bonus challenges; solving them is not required but can help you better understand ML/AI in the context of engineering, and how to use them in practical cases.
- Bonus points earned in all homework assignments will be averaged (6 bonus points for each assignment) and then directly added to your final score to calculate your final letter grade.

Challenge 1.1. In many machine learning tasks, dimensionality reduction is often combined with other techniques as a form of feature engineering. By reducing the number of dimensions in high-dimensional data, we can simplify models, reduce computational costs, and improve performance. In this homework, you will explore how dimensionality reduction can be used to improve clustering on high-dimensional image data.

You will apply PCA to the Fashion MNIST dataset, reducing the dimensionality of the data, and then perform K-means clustering to group the images into 10 clusters. Herein, we use only a subset (1,000 samples) of the Fashion MNIST dataset[1]. This subset, stored in the file `m12.npz`, contains 1,000 grayscale images of size 28x28 pixels, representing 10 categories of clothing such as T-shirts, shoes, and pants. The images are stored in the variable `X`, while `y` contains clothing classes for reference, but you can ignore these labels for this task.

Your tasks include: (*6pts*)

1. **Dataset loading and standardization**
   - Load the data, and standardize the data using `StandardScaler` from `sklearn` to normalize the pixel values of the images.
2. **PCA for dimensionality reduction**
   - Apply PCA to reduce the dimensionality of the dataset.
   - Choose the appropriate number of principal components based on the explained variance, retaining around 95% of the variance in the data.

3. **K-means Clustering**

   - After dimensionality reduction, perform K-means clustering in the reduced-dimensional space with 10 clusters, corresponding to the 10 clothing categories in the dataset.

4. **Visualization**

   - For each cluster, display 5 images:
     (a) One image reconstructed from the cluster center in the reduced PCA space.
     (b) The remaining four images will be random samples from the same cluster.
     (c) Remember to rescale your images using the StandardScaler.
   - Your images will look like Figure 1.

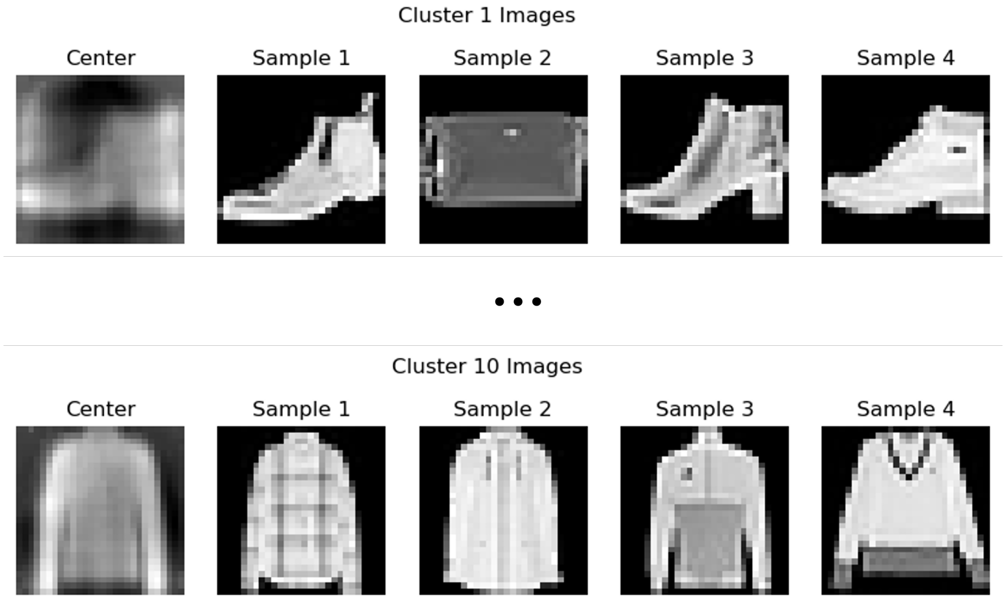

Figure 1: Example of visualization

5. Submit your Jupyter Notebook file with appropriate comments.

References

[1] Han Xiao, Kashif Rasul, and Roland Vollgraf. “Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms”. In: *CoRR* abs/1708.07747 (2017). arXiv: 1708.07747. URL: http://arxiv.org/abs/1708.07747.