Homework 8

Nov. 16, 2023

Task:

Experiment with methods of dimensionality reduction and visualization.

Goal:

- Find packages of PCA and UMAP and apply them on the dataset of handwritten digits. Explore the usage of PCA for dimensionality reduction and 2D visualization. Explore the usage of UMAP for 2D visualization of the original data and the low-dimensional data after PCA.

Data:

Please check the “data2forEx8+” folder for the following files of 60,000 handwritten digits and their labels:

train-images-idx3-ubyte.gz
train-labels-idx1-ubyte.gz

You can use the package *idx2numpy* in Python to load the idx files.

Requirements:

1) Apply PCA on the image data. Draw the plot of the represented variances of the principal components to decide the number of principal components that can represent >80% of the variance in the data. Draw the 2D plots of the first PC1 vs. PC2 and PC1 vs. PC3. Use colors to label the points of different digits on the plots to visually study the distribution of the 10 digits on the 2D plots. Discuss your observations (e.g., which classes can be more easily discriminated on which PCs).
2) Apply PCA on the images of only one digit and draw the plot of PC1 vs. PC2. Choose some representative data points on the PC plot and show their original images. Discuss your observations (e.g., whether there are some relations in the style of the writing among images shown on different locations on the plot).
3) Apply UMAP on the original image data and the first few PCs you selected. Use colors to label the points of different digits. Adjust some hyper-parameters (especially “n_neighbors”, “min_dist” and “spread”) in UMAP to produce multiple visualization results, to explore their influence on the visualization results. Discuss your observations.

Experiment Report:

- Write an experiment report to describe and analyze the experiment observations. The report should also include the short essay on parameter choices.
- Provide detailed supplementary materials that should include at least the following:
  - A readme file containing information on all supplementary files, programming environment and parameters used in the experiments (if any)
  - Source codes (should let TAs to be able to run the code and reproduce your experiments)
  - Experiment result files

Due date: Nov. 29 (Wed.) 23:59 Beijing time