Assignment 4

April 22, 2021

Question 1: 15 points

- Load the MNIST dataset and split it into a training set and a test set: take the first 60,000 instances for training, and the remaining 10,000 for testing.
- Train a Random Forest classifier on the dataset and time how long it takes, then evaluate the resulting model on the test set.
- Next, use PCA to reduce the dataset's dimensionality, with an explained variance ratio of 95%.
- Train a new Random Forest classifier on the reduced dataset and see how long it takes.
- Was training much faster?
- Next, evaluate the classifier on the test set. How does it compare to the previous classifier?

Question 2: 15 points

- Use t-SNE to reduce the MNIST dataset down to two dimensions and plot the result using Matplotlib. You can use a scatterplot using 10 different colors to represent each image's target class. Alternatively, you can replace each dot in the scatterplot with the corresponding instance's class (a digit from 0 to 9), or even plot scaled-down versions of the digit images themselves (if you plot all digits, the visualization will be too cluttered, so you should either draw a random sample or plot an instance only if no other instance has already been plotted at a close distance). You should get a nice visualization with well-separated clusters of digits.
- Try using other dimensionality reduction algorithms such as PCA, LLE, or MDS and compare the resulting visualizations.

Question 3: 15 points

- The classic Olivetti faces dataset contains 400 grayscale 64×64pixel images of faces. Each image is flattened to a 1D vector of size 4,096. 40 different people were photographed (10 times each), and the usual task is to train a model that can predict which person is represented in each picture.
- Load the dataset using the `sklearn.datasets.fetch_olivetti_faces()` function, then split it into a training set, a validation set, and a test set (note that the dataset is already scaled between 0 and 1). Since the dataset is quite small, you probably want to use stratified sampling to ensure that there are the same number of images per person in each set.
- Next, cluster the images using K-Means, and ensure that you have a good number of clusters using silhouette score.
- Visualize the clusters: do you see similar faces in each cluster?

Question 4: 15 points

- Continuing with the Olivetti faces dataset, train a classifier to predict which person is represented in each picture, and evaluate it on the validation set.
- Next, use K-Means as a dimensionality reduction tool, and train a classifier on the reduced set.
- Search for the number of clusters that allows the classifier to get the best performance: what performance can you reach?
- What if you append the features from the reduced set to the original features (again, searching for the best number of clusters)?

Requirements

These are the general rules for submitting the homework not only for this assignment.

1. Submit jupyter notebook(s) in ipynb and html formats: html format allows the grader to easily read it in canvas while ipynb allows to execute it if something is not clear from html.
2. The notebook should be well formatted:
    - Use markdown to break it into sections, bullet points, etc.
        - See `l1_n3.ipynb` for example how to use it.
        - If you double click on markdown cell, you can see the code.

- To show the formatted text, execute the corresponding markdown cell.
- Start the notebook with a markdown title cell of the form: '# Assignment X, Jane Doe, date'.
- When answering a particular question from the assignment, start the answer with markdown cell
  '## Question Y'
  that indicates what question you are answering.

3. Name your notebook (and the corresponding html file) according to the following template: `Assignment_X_Jane_Doe.ipynb`

4. If you are submitting several notebooks, append to the name `'_partZ'` and explain in canvas what each notebook is about.

5. There should not be any failed cells in the notebook. Every cell should work and have the results of the execution.

6. The notebook should contain only what is needed to answer a particular question and nothing else. More is as bad as less.

7. All the plots, if needed for the assignment, should be of production quality with readable labels, titles, etc. Also, try to find the most approriate type of plot to clearly demonstrate your statement.

8. Due date: 04/30/2021