AI 221: Machine Exercise 5

Instructions:

- Read and answer each problem using computer code. This MEX should be done *individually*.
- Each item should be answered as either a Jupyter Notebook or a MATLAB Live Script, exported as a single PDF file for the entire MEX. Make sure to HIGHLIGHT your final answers.
- When done, submit the PDF file through UVLE.

8x8 Handwritten Digits Visualization: Continued

This is a continuation of the previous MEX.

Load the `load_digits` data from sklearn.datasets, then do the following:

a. **[50 pts]** Normalize the X data using Standard Scaler. Then, project all the X data into 2 dimensions using 6 dimensionality reduction techniques:
    1. Local Linear Embedding (n_neighbors = 200, random_state = 0)
    2. t-SNE (perplexity = 50, random_state = 0)
    3. Isomap (n_neighbors = 200)
    4. Laplacian Eigenmap (n_neighbors = 200)
    5. Kernel PCA (kernel = 'rbf', gamma = 0.01)
    6. PCA

    The points should then be colored according to the digit labels, y.

    Which of the methods produced clear clusters of data points?

b. In this item, we will perform classification with and without dimensionality reduction. First, split the data into 70% training and 30% testing, *stratified* according to class label. Compare the results between the following methods:

    1. **[25 pts]** Make a pipeline using StandardScaler, Kernel PCA (kernel = 'sigmoid', n_components = 40), and SVC (default hyper-parameters). Fit the pipeline on the training set, then report the accuracy and F1-score on the test set.

    2. **[25 pts]** Make a pipeline using StandardScaler and SVC (default hyper-parameters) alone. Fit the pipeline on the training set, then report the accuracy and F1-score on the test set.

        Which of the two methods had a better test performance for classification? Why?

END OF EXERCISE