Assignment 1: Dimensionality Reduction using PCA

Important Information

School of Engineering Technology and Applied Science

Information and Communication Engineering Technology

Unsupervised and Reinforcement Learning (COMP257)

- **Dimensionality Reduction using PCA (10%)**
- **Due Date: Friday of Week 3 by 11:59 pm EST (late penalty at 10 points per day)**
- **Upload your assignment here: Assignment 1: Dimensionality Reduction using PCA (/d2l/common/dialogs/quickLink/quickLink.d2l?ou=694716&type=dropbox&rcode=CENCOL-3439540)**

Instructions

- You are free to choose any toolkits to solve the problems at hand (e.g., TensorFlow, Sci-Learn, etc.)
- All written reports and codes are to be maintained on a repository of your choice such as Github. The course instructor will discuss and exchange with you information to get access to your code.
- The video presentation will be required as part of the submission that documents the steps taken to obtain the results.
- **IMPORTANT NOTES:**
  - 1 point will be deducted for each incident that does not conform to the requirements (e.g., code not properly formatted, comments not relevant to support documentation of code, missing code documentation, etc.).
  - All points will be deducted for submission of nonsensical code (i.e., code that doesn't contribute to the relevancy of the task at hand). This is question-specific.

Questions

- **Read the questions below carefully.**

Question 1

[50 points]

1. Retrieve and load the mnist_784 (https://www.openml.org/d/554) dataset of 70,000 instances. [5 points]
2. Display each digit. [5 points]
3. Use PCA to retrieve the 1^{th} and 2^{nd} principal component and output their *explained variance ratio*. [5 points]
4. Plot the projections of the 1^{th} and 2^{nd} principal component onto a 1D hyperplane. [5 points]
5. Use Incremental PCA to reduce the dimensionality of the MNIST dataset down to 154 dimensions. [10 points]
6. Display the original and compressed digits from (5). [5 points]
7. Create a video discussing the code and result for each question. Discuss challenges you confronted and solutions to overcoming them, if applicable [15 points]

Question 2

[50 points]

1. Generate Swiss roll dataset. [5 points]
2. Plot the resulting generated Swiss roll dataset. [2 points]
3. Use Kernel PCA (kPCA) with linear kernel (2 points), a RBF kernel (2 points), and a sigmoid kernel (2 points). [6 points]
4. Plot the kPCA results of applying the linear kernel (2 points), a RBF kernel (2 points), and a sigmoid kernel (2 points) from (3). Explain and compare the results [6 points]
5. Using kPCA and a kernel of your choice, apply Logistic Regression for classification. Use *GridSearchCV* to find the best kernel and *gamma* value for kPCA in order to get the best classification accuracy at the end of the pipeline. Print out best parameters found by *GridSearchCV*. [14 points]
6. Plot the results from using *GridSearchCV* in (5). [2 points]
7. Create a video discussing the code and result for each question. Discuss challenges you confronted and solutions to overcoming them, if applicable [15 points]