CS 6220 Data Mining — Assignment 3

Exploring Data with PCA

For this assignment, you will be using a PCA to reduce the dimension of data.

Objectives:

1. Employ data reduction techniques such as principal component analysis
2. Visualize and interpret results
3. Compare PCA with another data reduction technique

Submission:
Through the assignment submission portal on Canvas, submit your ipynb with a pdf of your assignment solution; no need to zip the files.

Grading Criteria:
Follow the instructions in the pdf, and complete each task. You will be graded on the application of the module's topics, the completeness of your answers to the questions in the assignment notebook, and the clarity of your writing and code.

Dataset:
For this assignment, you will use the Fashion MNIST dataset. Fashion-MNIST is a dataset of Zalando's article images—consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes. To know more about the dataset, you can look here (https://github.com/zalandoresearch/fashion-mnist).

What You Need to Do

You will be using `scikit-learn` to apply PCA on the dataset. Also for the simplicity, you can download the dataset from `scikit-learn`. You can use the following code snippet.

```
data, labels = fetch_openml('Fashion-MNIST', version=1, return_X_y=True)
```

Part 1 - PCA [40 Points]:
You will need to use PCA, which is implemented in `scikit-learn`. See this link for documentation here (http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html).

1. Apply PCA projection of the features of the Fashion MNIST dataset in 2 dimensions.
2. Show scatter plot of the reduced dimension. Use separate color for each class of the data.
3. Show how much variance ratio is explained by the reduced dimension.

Part 2 - Another Reduction Method [35 Points]:

1. Apply another reduction method on the features of the Fashion MNIST dataset in 2 dimensions.
2. Show scatter plot of the reduced dimension. Use separate color for each class of the data.
3. Show how much variance ratio is explained by the reduced dimension.

Part 3 Conceptual Question [25 Points]:
Answer the following question in the same ipython notebook.

1. Compare the variance ratio explained by the 2-dimensions of the methods you have used. Which is better?
2. Compare the scatter plot of the two methods after reduction. Which is a better method for separating the different classes of data?
3. What is the primary difference between the two methods? Which method works better in this case and why?