CS 6220 Data Mining — Assignment 3

Exploring Data with PCA

For this assignment, you will be performing dimension reduction of data.

Objectives:

1. Employ Principal Component Analysis (PCA) for data reduction
2. Visualize and interpret results
3. Compare PCA with another data reduction technique of your choice

Submission:
Submit your ipynb on Canvas.

Grading Criteria:
Follow the instructions in the pdf and complete each task. You will be graded on the application of the module's topics, 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 IRIS dataset. The dataset consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other. To know more about the dataset, you can look at (https://archive.ics.uci.edu/ml/datasets/iris).

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 `sscikit-learn`. You can use the following code snippet.

```
data = datasets.load_iris()
X = data.data
y = data.target
```

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

1. Apply PCA projection of the features of IRIS dataset in 2 dimensions
2. Show how much variance ratio is explained by the reduced dimension.

Part 2 - Another Reduction Method [35 Points]:

1. Apply another reduction method of your choice on the features of IRIS 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?