CS 663 - Machine Learning Spring, 2020

Lab 03 - Principal Component Analysis

As seen in during class, PCA is a technique for dimensionality reduction and explanation of variance. In this lab, your task is to recreate the visual work in class and show how you are able to apply the technique to a previously-unseen dataset.

Data

There are three datasets for this lab:

- The Wisconsin Diagnostic Breast Cancer (WDBC) dataset
- The Gisette dataset
- The Dorothea dataset

Process

Your implementation must:

- Use the starter code and fill in your name and USF email in the readme
- Read the data from and apply PCA to two of the datasets: the WDBC dataset and EITHER the Gisette or Dorothea dataset
- Optionally apply a normalisation technique to the data
- Show how instances and classes are related in the datasets
- Show the percentage of variance explained, starting from the first principal component
- Use a Jupyter notebook or Python script to complete your implementation.

Grading

Grades for this assignment will be determined by the grader as follows:

- 100% = Code functions, is well-documented and clearly shows the relationship between fares and survival.
- 75% = Code functions but is not well-documented or does not clearly show the relationship between instances and classes, and percentage of variance.
- 50% = Code functions but is not well-documented -AND- does not clearly show the relationship between instances and classes, and percentage of variance.
- 0% = No submission / code does not function.

Submission

Submit your code to Github, and submit the link to your repo on Canvas. Late submissions will be penalized or not accepted.
The deadline is 11:59 PM on February 10, 2020.