Summer 2018 EE690 Machine Learning Dr. L. Jololian

Assignment #4 (Midterm)

Principle Component Analysis

**Due Date: Thursday, July 26**

**Problem 1**

For this problem, you are asked to use the Python algorithms developed in class for the Principle Component Analysis (PCA) implementation to investigate dimensionality reduction of the “breast cancer” dataset provided by Scikit-Learn.

1. You should implement PCA in two different ways:
2. using the eigenvector decomposition of the covariance matrix
3. using Singular Value Decomposition (SVD) method

in each case show the Eigen values and Eigen vectors.

1. Provide a plot that shows the level of data variance retained as a function of dimensionality reduction.
2. Using the first two Eigen vectors, show a two-dimensional plot of the dataset with the reduced number of features. The points on the plot should be color-coded based on their values.
3. How many Eigen vectors should be used to retain 95% of the data variance.

**Problem 2**

For this problem, you are to implement in Python the Naïve Bayes algorithm (based on the class notes) to classify news headlines.

There are approximately 125k news headlines from the year 2013 to 2018 obtained from HuffPost. Each record consists of six attributes:

authors: authors who published the article

category: the category of the news article

date: date article was published

headline: the headline of the news article

article\_link: link to the original news article. Useful for collecting supplementary data

short\_description: a short description of the news article

The objective is to identify the type of news based on headlines and short descriptions.

Compare you results with an alternate implementation, using the Scikit-Learn implementation of the Naïve Bayes.

To read the data file, you can use the code:

import pandas as pd

df = pd.read\_json('News\_Category\_Dataset.json', lines=True)

The following is a sample of the data ,

| **authors** | **category** | **date** | **headline** | **link** | **short\_description** |
| --- | --- | --- | --- | --- | --- |

0 Melissa CRIME 2018-05-26 There Were 2 <https://www.huff>... She left her husband. He

Jeltsen Mass Shootings killed their children...

In Texas Last Week...

**Problem 3**

Implement in Python the Naïve Bayes algorithm for the digits dataset in the Scikit-Learn library. Compare your results against the results obtained by using the Scikit-Learn implementation of the Naïve Bayes algorithm.