IC 272: Lab3: Attribute normalization, standardization and dimension reduction of data

Deadline for submission: 19 September 2021, 10:00 PM

You are given the **Pima Indians Diabetes Database** as a csv file (`pima-indians-diabetes.csv`). This data-set is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females with at least 21 years old of Pima Indian heritage. It contains following 9 attributes.

- **pregs**: Number of times pregnant
- **plas**: Plasma glucose concentration 2 hours in an oral glucose tolerance test
- **pres**: Diastolic blood pressure (mm Hg)
- **skin**: Triceps skin fold thickness (mm)
- **test**: 2-Hour serum insulin (mu U/mL)
- **BMI**: Body mass index (weight in kg/(height in m)^2)
- **pedi**: Diabetes pedigree function
- **Age**: Age (years)
- **class**: Class variable (0 or 1)

Write a python program (with pandas) to read the given data and do the following:

1. **Consider only first eight attributes (i.e., excluding class)** for the analysis. Replace the outliers (if at all any) in any attribute with the median of the respective attributes and do the following on outlier corrected data:
    a. Do the **Min-Max normalization** of the outlier corrected data to scale the attribute values in the range 5 to 12. Find the minimum and maximum values before and after performing the Min-Max normalization of the attributes.
    b. Find the mean and standard deviation of the attributes of the outlier corrected data. **Standardize** each selected attribute using the relation $\hat{x}_n = (x_n - \mu)/\sigma$ where μ is mean and σ is standard deviation of that attribute. Compare the mean and standard deviations before and after the standardization.

2. Generate 2-dimensional synthetic data of 1000 samples and let it be denoted as data matrix **D** of size 2x1000. Each sample is independently and identically distributed with bi-variate Gaussian distribution with user entered mean values, $\boldsymbol{\mu} = [0, 0]^T$ and covariance matrix, $\boldsymbol{\Sigma} = \begin{bmatrix} 13 & -3 \\ -3 & 5 \end{bmatrix}$. Perform the followings:
    a. Draw a scatter plot of the data samples.
    b. Compute the eigenvalues and eigenvectors of the covariance matrix and plot the Eigen directions (with arrows/lines) onto the scatter plot of data (Refer Figure 1. Direction may change).
    c. Project the data on to the first and second Eigen direction individually and draw both the scatter plots superimposed on Eigen vectors (Refer Figures 2 and 3. Directions may change).

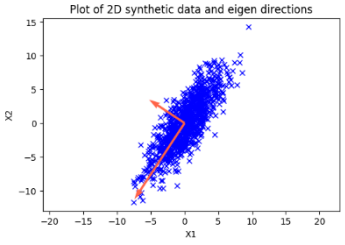

Figure 1

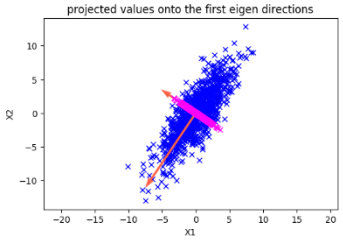

Figure 2

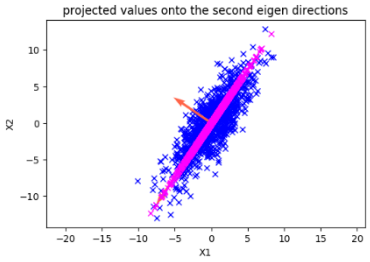

Figure 3

d. Reconstruct the data samples using both eigenvectors, say it $\widehat{\mathbf{D}}$. Estimate the reconstruction error between $\widehat{\mathbf{D}}$ and **D** using mean square error.

3. Data frame used for PCA include only first 8 attributes (**class** attribute is excluded). Perform principal component analysis (PCA) on outlier corrected standardized data (Data frame obtained after Question 1b.) and do the followings:
   a. Reduce the multidimensional (d = 8) data into lower dimensions (l = 2). Print the variance of the projected data along the two directions and compare with the eigenvalues of the two directions of projection. Also show the scatter plot of reduced dimensional data.
   b. Plot all the eigenvalues in the descending order.
   c. Plot the reconstruction errors in terms of RMSE considering the different values of l (=1, 2, ..., 8). The x-axis is the l and y-axis is reconstruction error in RMSE. Print the covariance matrix of each of the l-dimensional representations (l = 2, 3, ..., 8) and comment on the observations.
   d. Give the covariance matrix for the original data (8-dimensional). Compare the covariance matrix for the original data (8-dimensional) with that of the covariance matrix for 8-dimensional representation obtained using PCA with l = 8.

You can use

- `matplotlib` library for plotting.
- `np.random.multivariate_normal` for generating the multivariate normal distribution.
- `numpy.linalg.eig` function to compute the eigenvectors.
- matplot `quiver` function for plotting arrows in Eigen directions.
- `decomposition` from `sklearn` for PCA
- `pca.inverse_transform` for back projection

Write a report that should include the at least the following:

- Answers to these questions (including figures/plots). Note: Clearly label the x and y axis of figures/plots
- Significance of performing normalization and standardization.
- Significance of dimension reduction.
- The inference from the direction of the principal components from the plot.
- Observations from the reconstructed data (on synthetic data).

- Draw the inferences regarding the variance and the reconstruction errors for considering different principal components.
- Any other observations and inferences.

Instructions:

- **For question no. 2 and 3, you use Eigen decomposition of the covariance matrix and PCA respectively**
- **Your python program(s) should be well commented. Comment section at the beginning of the program(s) should include your name, registration number and mobile number.**
- **The python program(s) should be in the file extension .py**
- **Report should be strictly in PDF form. Write the report in word or latex form and then convert to PDF form.**
- **First page of your report must include your name, registration number and mobile number. Use the template of the report given in the assignment.**
- **Upload your program(s) and report in a single zip file. Give the name as <roll_*number*>_Assignment3.zip. Example: *b20001*_Assignment3.zip**
- **Upload the zip file in the link corresponding to your group only.**

In case the program found to be copied from others, both the person who copied and who help for copying will get zero as a penalty.