IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – III
Attribute normalization, standardization and dimension reduction of data

Student's Name: Vikas Dangi **Mobile No: 9406661661**

Roll Number: B20238 **Branch: EE**

1 a.

Table 1 Minimum and maximum attribute values before and after normalization

S. No.	Attribute	Before normalization		After normalization	
		Minimum	Maximum	Minimum	Maximum
1	pregs	0.000	13.000	5.0	12.0
2	plas	44.000	199.000	5.0	12.0
3	pres (in mm Hg)	38.000	106.000	5.0	12.0
4	skin (in mm)	0.000	63.000	5.0	12.0
5	test (in mu U/mL)	0.000	318.000	5.0	12.0
6	BMI (in kg/m^2)	18.200	50.000	5.0	12.0
7	pedi	0.078	1.191	5.0	12.0
8	Age (in years)	21.000	66.000	5.0	12.0

Inferences:

1. Outliers are unusual values in your dataset they can distort statistical analyses and violate their assumptions. And after normalization also, the outliers remain outliers. Only the range is changed so we remove them before normalization.
2. By subtracting the minimum value of the feature and then dividing by the range, the min-max technique (also known as normalization) rescales the feature to a hard and fast range of [0,1]. The.min() and.max() functions in Pandas can be used to provide min-max scaling.
3. Data normalization consists of remodeling numeric columns to a standard scale so as all the attributes are in the same range and thus accurate to give results when it comes to parameters which depends on multiple parameters.

b.

Table 2 Mean and standard deviation before and after standardization

S. No.	Attribute	Before standardization		After standardization	
		Mean	Std. Deviation	Mean	Std. Dev
1	pregs	3.786	3.271	0	1.0
2	plas	121.656	30.438	0	1.0
3	pres (in mm Hg)	72.197	11.147	0	1.0
4	skin (in mm)	20.437	15.698	0	1.0
5	test (in mu U/mL)	60.919	77.636	0	1.0
6	BMI (in kg/m^2)	32.199	6.4106	0	1.0
7	pedi	0.427	0.245	0	1.0
8	Age (in years)	32.760	11.055	0	1.0

Inferences:

1. We rescale an original variable to have a mean of zero and standard deviation of one using Z score standardization. The data is largely scattered over a range before it but we standardize it with mean=0 and variance=1.
2. When comparing measurements with various units, standardizing the features around the center 0 with a standard deviation of 1 is critical. Variables assessed on different scales do not contribute equally to the analysis and may result in a bias.

2. a

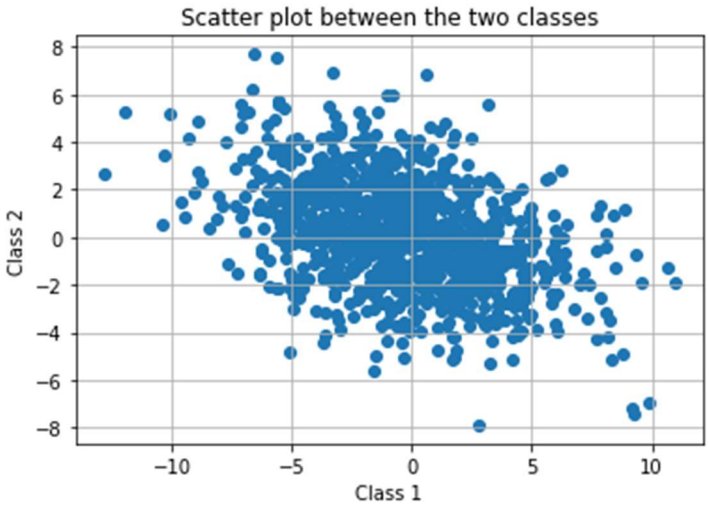

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. Attribute 1 is negatively correlated to attribute 2 based as the data seems to have a negative slope and one decreases with the increase in other.
2. The density of the graph mostly lies around the mean and in the range of [-5,5] for class 1 and [-4,4] for class2.

b.

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. The spread of data based is more towards the eigen vector having eigen value 14 compared to eigen value 4.
2. The density decreases as we move away from the eigen vector intersection axes.

c.

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. One eigen value has magnitude of 14 and other has 4. The eigen values show us the variance of the data in the direction of the corresponding eigen vector. The eigen vector with value 14 is more spread as seen clearly from the graph.
2. The variance of the data along the Eigen axes is almost equal to the eigen values, as the eigen values represent the variance of the newly projected data and our newly projected data takes projection depending on the spread too the values are very close to one another. We can see that the spread of projected data is equal in the direction of the eigen vector.

d. Reconstruction error = 0.000

Inferences:

1. The magnitude of reconstruction error tells the quality of reconstruction the more components we take for PCA the less is the error. The error is almost zero if we reconstructed it using all the components.

3 a.

Table 3 Variance and Eigenvalues of the projected data along the two directions

Direction	Variance	Eigenvalue
1	1.98986869	1.99246305
2	1.85100888	1.85342219

Inferences:

1. The Eigen values are almost equal to the Variance of the projected data.
2. All the data is being projected on the eigen vector and the magnitude of the spread of that projected data comes out to be equal to the eigen value.
3. The Eigen values scale the vectors in order to reconstruct the original data based on the characteristic vectors.

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. It doesn't look like they are correlated.
2. The components after PCA are such that they are projected towards eigen vectors so they are uncorrelated as eigen vectors are perpendicular.

b.

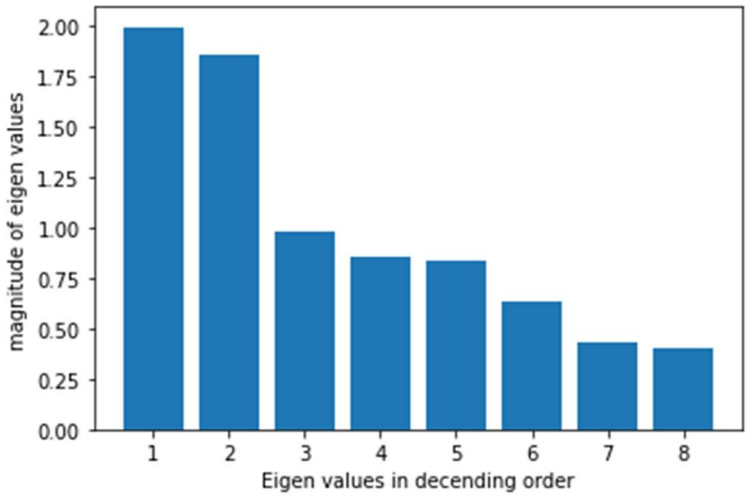

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. The subsequent Eigenvalues decrease rapidly.
2. After the Eigenvalue 1.853 the rate of decrease changes substantially.

c.

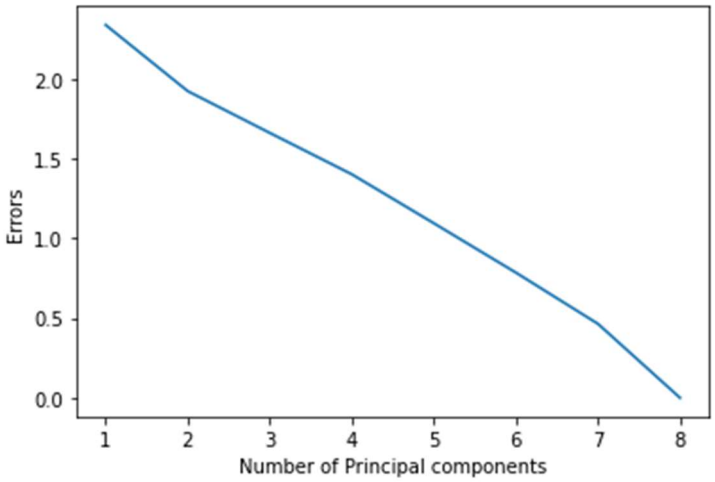

Figure 7 Line plot to demonstrate reconstruction error vs. components

Inferences:

1. The magnitude of reconstruction error tells the quality of reconstruction the more components we take for PCA the less is the error. The error is almost zero if we reconstructed it using all the components.
2. We can see from the graph that the error decreases as we take a greater number of components.

Table 4 Covariance matrix for dimensionally reduced data (l=2)

	x1	x2
x1	1.992463e+00	-1.197362e-15
x2	-1.197362e-15	1.853422e+00

Table 5 Covariance matrix for dimensionally reduced data (l=3)

	x1	x2	x3
x1	1.992463e+00	-4.238244e-16	1.528547e-16
x2	-4.238244e-16	1.853422e+00	1.25e629e-16
x3	1.528547e-16	1.250629e-16	9.818791e-01

Table 6 Covariance matrix for dimensionally reduced data (l=4)

	x1	x2	x3	x4
x1	1.992463e+00	8.823885e-16	-8.105931e-17	1.111671e-16
x2	8.823885e-16	1.853422e+00	5.824690e-16	1.296949e-16
x3	-8.105931e-17	5.824690e-16	9.818791e -01	-1.343269e-16
x4	1.111671e-16	1.296949e-16	-1. 343269e-16	8.583073e-01

Table 7 Covariance matrix for dimensionally reduced data (l=5)

	x1	x2	x3	x4	x5
x1	1.992463e+00	6.762662e-16	4.770919e-16	7.874333e-17	1.065351e-16
x2	6.762662e-16	1.853422e+00	1.043349e-15	-1.447488e -16	-3.126573e-17
x3	4.770919e-16	1.043349e-15	9.818791e-01	-1.424328e-16	-4.134025e-16
x4	-7.874333e-17	-1.447488e-16	-1.424328e-16	8.583073e-01	4.713020e-16
x5	1.06535le-16	-3.126573e-17	-4.134025e-16	4.713020e-16	8.387496e-01

Table 8 Covariance matrix for dimensionally reduced data (l=6)

	x1	x2	x3	x4	x5	x6
x1	1.992463e+00	3.485550e-15	-6.253147e -17	-5.048837e-16	7.642735e-17	1.482227e-16
x2	3.48555ee-15	1.853422e+00	-3.693988e-16	1.968583e -17	5.442554e-17	6.803192e-18
x3	-6.253147e -17	-3.693988e-16	9.818791e-01	4.666700e-16	-6.994260e-16	-4.863558e-16
x4	-5.048837e-16	1.968583e-17	4.666700e -16	8.583073e-01	-1.082721e-16	4.215084e-16
x5	7.642735e-17	5.442554e-17	-6.994260e - 16	-1.082721e-16	8.387496e - 01	3.462390e-16
x6	-1.482227e-16	6. 803192e-18	4.863558e-16	4.215084e-16	3.462390e-16	6.364084e-01

Table 9 Covariance matrix for dimensionally reduced data (l=7)

	x1	x2	x3	x4	x5	x6	x7
x1	1.992463e+00	4.516161e-16	4.46984e-16	2.40862e-16	4.72460e-16	2.45494e-16	-2.4549e-16
x2	4.51616le-16	1.853422e+00	4.20350e-16	-3.70557e-17	-3.45081e-16	2.19005e-16	-1.9251e-16
x3	4.469842e-16	4.203504e-16	9.81879e-01	1.78331e-16	-5.32675e-16	2.17702e-16	1.9686e-17
x4	2.408619e-16	-3.705568e-17	1.78330e-16	8.58307e-01	-2.15965e-16	4.56248e-16	-2.4317e-17
x5	4.724600e-16	-3.450811e-16	-5.32675e-16	-2.15965e-16	8.38749e-01	-2.5359e-16	-2.6084e-16
x6	-2.454939e-16	-2.190049e-16	2.17702e-16	4.56248e-16	-2.53599e-16	6.36408e-01	-2.1654e-16
x7	-2.454939e-16	1.925159e-16	1.96858e-17	-2.43178e-17	-2.60837e-16	-2.1654e-16	4.34143e-01

Table 10 Covariance matrix for dimensionally reduced data (l=8)

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1.99246e+00	4.51616e-16	4.46984e-16	2.40861e-16	4.72460-16	2.45493e-16	-2.45493e-16	-1.01903e-16
x2	4.51616e-16	1.85342e+00	4.20350e-16	-3.70557e-17	-3.45081e-16	-2.19005e-16	-1.92516e-16	-1.15799e-18
x3	4.46984e-16	4.20350e-16	9.81879e-01	1.78330e-16	-5.32675e-16	2.17702e-16	1.96858e-17	-1.79488e-17
x4	2.40862e-16	-3.70556e-17	1.78330e-16	8.58307e-01	-2.15965e-16	4.56248e-16	-2.43177e-17	-7.41113e-17
x5	4.72460e-16	-3.45081e-16	-5.32675e-16	-2.15965e-16	8.38749e-01	-2.53599e-16	-2.60837e-16	-4.34246e-16
x6	-2.4549e-16	-2.19005e-16	2.177e2e-16	4.56248e-16	-2.53599e-16	6.36408e-01	-2.16544e-16	2.95287e-16
x7	-2.4549e-16	-1.92516e-16	1.96858e-17	2.43178e-17	-2.608373e-16	-2.16544e-16	4.34143e-01	-8.82967e-18
x8	-1.8190e-16	-1.15799e-18	-1.79488e-17	7.41113e-17	-4.34246e-16	2.95287e-16	-8.82967e-18	4.04627e-01

Inferences:

1. The off-diagonal elements are almost **zero** the reason for the observed trend is that they are uncorrelated and independent.
2. The diagonal elements show us the variance of that element and the off diagonals show us the correlation of two variables. As the components are independent after PCA analysis the correlation tends to zero.
3. The diagonal values are getting reduced significantly.
4. With the number of components, the value is decreasing which shows us that the variance decreases with the components having low values of eigen values as they have less data spread and hence less variance,
5. From the magnitude of diagonal elements first captures data variations the best due to large eigen values.

6. From the value of diagonal elements, the number of components that shall give the optimum reconstruction along with dimensionality reduction is two or if we want three these are providing us the maximum variance and hence most of the information is in here.
7. The magnitude of the 1st diagonal element (topmost left corner) in each of the obtained covariance matrices is the same as it denotes the variance of the data spread along the first eigen vector it is justified to remain independent from the other components.
8. The magnitude of the 2nd diagonal element (topmost left corner) in each of the obtained covariance matrices is the same as it denotes the variance of the data spread along the second eigen vector it is justified to remain independent from the other components.
9. The 3rd, 4th, 5th, 6th, and 7th diagonal elements across covariance matrices are all same.

d.

Table 11 Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	1.000000	0.117692	0.208953	-0.096720	-0.108474	0.028339	0.004519	0.560768
plas	0.117692	1.000000	0.204539	0.060034	0.179566	0.228245	0.081613	0.274264
pres (in mm Hg)	0.208953	0.204539	1.000000	0.025645	-0.050956	0.271560	0.022495	0.326372
skin (in mm)	-0.09672	0.060034	0.025645	1.000000	0.472714	0.373726	0.152763	-0.101397
test (in mu U/mL)	-0.10847	0.179566	-0.050956	0.472714	1.000000	0.171503	0.198580	-0.073726
BMI (in kg/m^2)	0.028339	0.228245	0.271560	0.373726	0.171503	1.000000	0.123776	0.077668
pedi	0.004519	0.081613	0.022495	0.152763	0.198580	0.123776	1.000000	0.036109
Age (in years)	0.560768	0.274264	0.326372	-0.101397	-0.073726	0.077668	0.036109	1.000000

Inferences:

1. The off-diagonal values are non-zero unlike the covariance matrix obtained after PCA l=8 reduction.
2. The values are all equal in this case unlike in the past case where the variance was continuously getting distributed among all the components.
3. IN PCA the variance gets distributed in such a way that we can have components which we can discard and pick up the most significant ones for data reduction.

Thank You