CS 412 Assignment 1

Solution

Guangyu Zhou

Netid: gzhou6

1. Basic statistical description about mid-term scores are presented below:

(a) & (b)

For max, min, quartile Q1, Q3 and median, first sort the midterm score array and then pick the rank index of each relatively. In ascending order, min is Score[1], max is Score[n], Q1 is Score[n/4], Q3 is Score[3n/4] and median is Score[n/2]. Here are the values:

min = 37 **max = 100**

Q1 = 68 **Median = 77** **Q3 = 87**

(c) For mean score, we just sum up all and divide by the total number.

mean = 76.6810

(d) For modes, we traverse to count the occurrence of each score value and pick those with the most occurrence.

There are 2 Modes: 83 and 72

(e) For Empirical Variance, by definition is just $E[(x-E(x))^2]$

Empirical Variance= 173.5308

2.

(a) Jaccard coefficient = K&J/(K&J+K&!J+J&!K)

= 79/(79+22+58)=0. 4969

(b) Here are the Minkowski distance with different h:

1) H=1 dist1= 5700

2) H=2 dist2 = 9497

3) H=infinity dist_Infinity = 1

(c) Cosine similarity:

$Cos(d_1,d_2)$ =(d1d2)/(|d1||d2|)= 0.8449

3.

(a)

As z-score normalization is $Z=(X-\text{mean}(x))/\text{std}_{empirical}$ in this case:

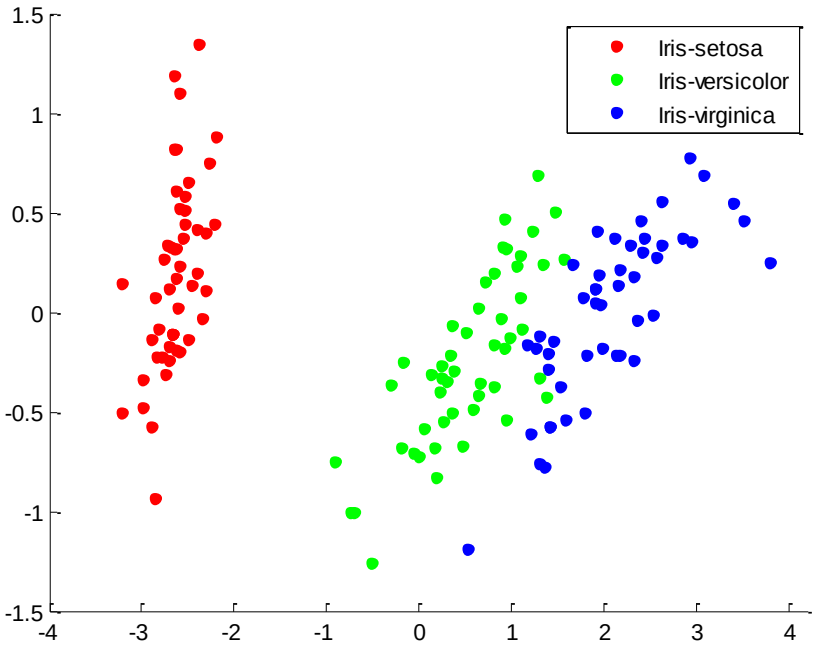
Before Normalization	After Normalization
mean = 76.6810	meanN = -7.1054e-018
expVar = 173.5308	expVarN = 1.0000

(b) Corresponding score of 90 is

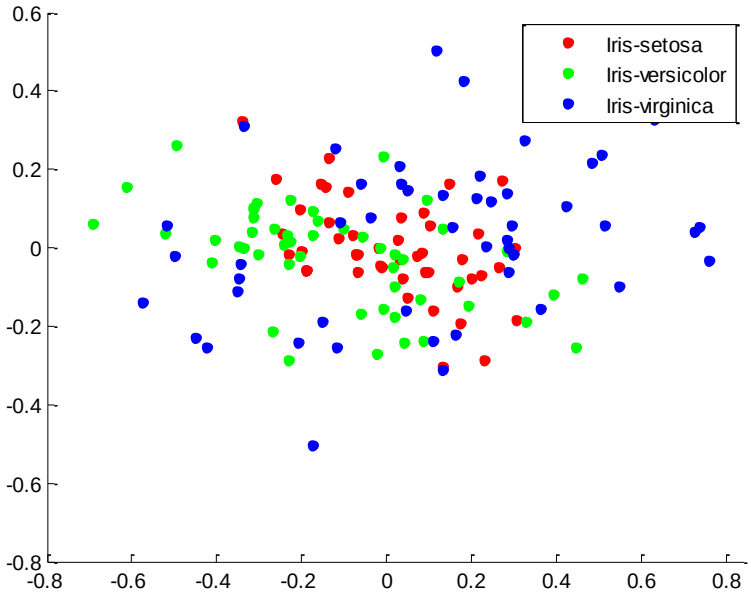
Z(90) = (90-76.6810)/ 13.17=1.011

4.

(a) Plot under the projection based on the first and second principal component.



(b) Plot under the projection based on the third and fourth principal component.



In comparison of this two graphs, following conclusion can be draw:

1) Under the projection of the first two principle components, three groups of data are clustering under each class and they are lining on a one direction, which can be seen as correlation.
2) Under the other two dimensional spaces, for which I chose the 3rd and 4th components, there is no more clustering among groups and no more correlations in each group.

5.

(a)

First we construct X_{2*3} based on our 3 data points:

$$X=\begin{pmatrix}-1 & 0 & 1\\ 1 & 0 & -1\end{pmatrix}$$

Then compute $C_x = \frac{1}{N}(XX^T) = \begin{pmatrix}2/3 & -2/3\\ -2/3 & 2/3\end{pmatrix}$, where N = 3

Now we need to find the eigenvectors of C_x, we construct:

$$\det(C_x - \lambda X)=0$$

We get $\lambda=4/3$ or $\lambda=0$

Case $\lambda=4/3$: $e_1=\begin{pmatrix}1\\ -1\end{pmatrix}$ change to unit vector is $e_1=\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{-1}{\sqrt{2}}\end{pmatrix}$

Case $\lambda=0$: $e_2=\begin{pmatrix}1\\ 1\end{pmatrix}$ change to unit vector is $e_2=\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}$

Based on the eigenvalue, we have a sorted $P = (e_1, e_2)$

So the first principal component is $e_1=\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{-1}{\sqrt{2}}\end{pmatrix}$

(b)

To get the 1-d subspace based on e_1, we first let $P_1= e_1^T=\left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right)$

Then the projection $Y_1 = P_1*X = (-\sqrt{2}, 0, \sqrt{2})$

And the variance is Var = $\frac{2+2}{3} = \frac{4}{3}$

(c) $X_{Reconstruct} = e_1* Y_1=\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{-1}{\sqrt{2}}\end{pmatrix}* (-\sqrt{2}, 0, \sqrt{2})= \begin{pmatrix}-1 & 0 & 1\\ 1 & 0 & -1\end{pmatrix} = X$

So L2 error is 0.