Aprendizagem 2021

Lab 10: Dimensionality Reduction

Practical exercises

1. Given the following training data

	y_1	y_2
$\mathbf{x}_1$	0	0
$\mathbf{x}_2$	4	0
$\mathbf{x}_3$	2	1
$\mathbf{x}_4$	6	3

a) Compute the K-L transformation

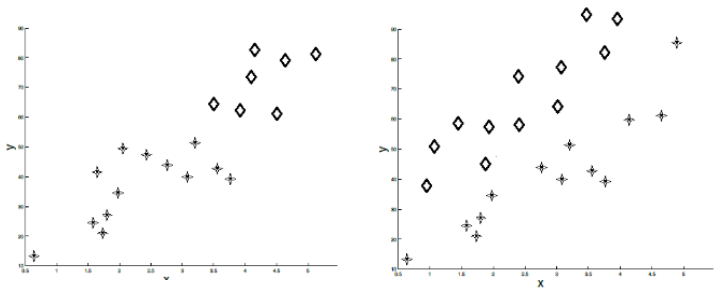
b) What is the rotation applied to go from the original to the eigenvector coordinate system?

c) Which eigenvector is most significant?

d) Can we apply the Kaiser criterion?

e) Map the points onto the most significant dimension

2. Given the following datasets where observations are in $\mathbb{R}^2$ and belong to one of two classes:



Which principal components can accurately discriminate the class per dataset?

3. The following top-7 eigenvalues explain 90% of the variation of dataset X:

λ_1=20, λ_2=10, λ_3=5, λ_4=4, λ_5=3, λ_6=2, λ_7=1

What is the most accurate information regarding X:

i. X has less than 7 attributes
ii. X has 7 attributes
iii. X has more than 7 attributes
iv. X has more than 11 attributes

4. Given a set of data points in $\mathbb{R}^3$, the following covariance matrix was obtained:

$$\begin{bmatrix} 91.43 & 171.92 & 297.99 \\ & 373.92 & 545.21 \\ & & 1297.26 \end{bmatrix}$$

as well as the following eigenvectors retrieved:

$$\boldsymbol{u}_1 = \begin{pmatrix} 0.2179 \\ 0.4145 \\ 0.8836 \end{pmatrix}, \boldsymbol{u}_2 = \begin{pmatrix} -0.2466 \\ -0.8525 \\ 0.4608 \end{pmatrix}, \boldsymbol{u}_3 = \begin{pmatrix} 0.9443 \\ -0.3183 \\ -0.0836 \end{pmatrix}$$

Please select the more complete answer:

i. eigenvalue λ1 is approximately 1626
ii. eigenvalue λ2 is approximately 129
iii. eigenvalues λ1 and λ2 explain >99% of the variation in data
iv. all of the above

5. Given the following dataset:

	y_1	y_2
x_1	1	-1
x_2	0	1
x_3	-1	0

and the corresponding eigenvectors and eigenvalues:

λ_1=3/2 and λ_2=1/2

$$\boldsymbol{u}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \boldsymbol{u}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

a) Transform the input data using PCA
b) [*optional*] Assess the data recovery error when considering the most informative component only

Programming quest

6. Considering the *housing* dataset available at https://web.ist.utl.pt/~rmch/dscience/data/housing.arff

a. How many principal components are necessary to explain 90% of data's variability?
b. Compare the accuracy of one of the covered classifiers in the original *versus* reduced data space. Why we are unable to observe considerable improvements?

Resource: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html