FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Departamento de Informática

Data Analysis and Mining
2018/2019

Test 1

Duration: 2 hours

April, 23 2019

Important notice: 1) The test consists of **five groups of questions**. 2) You must write your name and student number in each **of the answer sheets**. 3) You can only leave the room **30 minutes after the beginning** of the exam. 4) You must return the answer sheets when you finish the test.

Good work!

Question 1 – Linear Regression [1.0 + 1.0 + 2.0 points]

The economic structure of Major League Baseball (MLB) allows some teams to make substantially more money than others, which allows them to spend much more on player salaries. These teams might therefore be expected to have better players and win more games. Over the course of four years each of the 30 MLB teams were measured each year and the following data were collected for these 120 observations:

wins: number of games the team won for a specific year
payroll: opening day payroll, in millions of dollars, for the team for a specific year
AL: a binary variable for whether the team is in the American League (AL)
(14 of the 30 teams are in the AL; the rest are in the National League (NL))
year: the year in which the measurement was taken

A linear regression model was run to predict the number of *wins* a team had from the ln(*payroll*), and the results in R language are shown below, along with some summary statistics:

```
> summary(model1<-lm(wins~log(payroll),data=mlb))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    44.076     11.654   3.782 0.000246 ***
log(payroll)    7.963      2.505   3.179 0.001887 **
---
Residual standard error: 10.5649 on 118 degrees of freedom
Multiple R-squared:  0.07889,   Adjusted R-squared:  0.07109
F-statistic: 10.11 on 1 and 118 DF,  p-value: 0.001887
> mean(log(mlb$payroll))
[1] 4.6371
> sd(log(mlb$payroll))
[1] 0.3866663
```

1.1 Interpret the slope coefficient in this model.

Solution: A 1-unit change in ln (payroll) is associated with an estimated increase in of 7.963 wins on average. Specifically, a doubling of payroll is associated with an estimated increase in of ln(2)*7.963= 5.52 wins, on average.

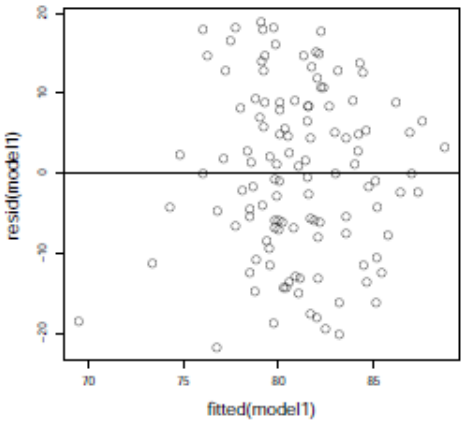
1.2 What is the estimated correlation between variables '*wins*' and ln(*payroll*)?

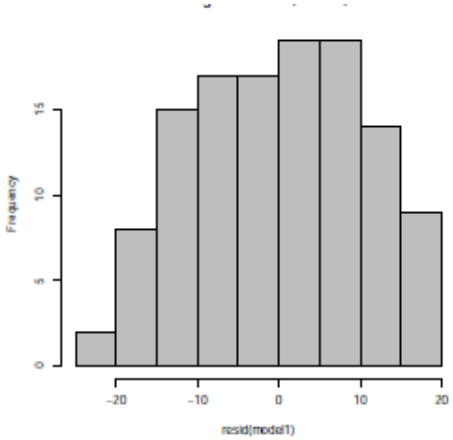
For simple regression we have

$$r = \pm\sqrt{R^2} = \sqrt{0.07889} = 0.2809$$

The sign is positive since the the sign of the slope estimate is positive.

1.3 The figure bellow presents the residual graphs for this model. Discuss, and justify, if the assumptions of 'independence', 'linearity', 'normality' and 'constant variance' verify or not for this regression model.





Solution

Independence: cannot be determined from these graphs, but this assumption is likely violated since there multiple measurements from each team across years.

Linearity: this looks correct, as there is no obvious evidence of curvature in the residuals vs fitted scatterplot to the left.

Normality: the residuals look fairly symmetric at the bery least: just missing any right tail. Inferences will be robust to this since the sample size is n= 120.

Constant Variance: t looks fine as the spread in the vertical direction is pretty consistent.

Question 2 – PCA [1.0 + 1.0 + 1.0 points]

2.1. Which of the following sentences are true:

(i) *Principal component analysis (PCA) can be used with variables of any type: quantitative, qualitative, or a mixture of both types*.

(ii) *The variables subjected to PCA must all have the same units*.

(iii) When the variables have different units they must be subject to normalization before applying PCA.

(iv) *The proportion of explained variance accumulated increases as more principal components are extracted from data*.

(v) *The exploration of all the principal components allows to have a better understanding of the data structure*.

F, T, T, T, F

2.2 Consider the following four points of bidimensional real space: $\boldsymbol{p}_1$= (1,2), $\boldsymbol{p}_2$=(2,1), $\mathbf{p}_3$=(3,4) and $\mathbf{p}_4$=(4,3), which are represented in a 4 × 2 matrix, *P*. The two main eigen vectors of the matrix P^TP are shown as column vectors of matrix *E*

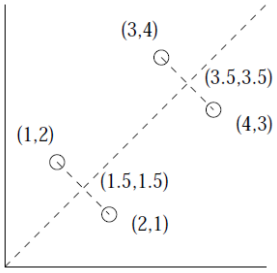
$$E = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

2.2.1 Calculate the coordinates of the four points projected on the axes of the first eigen vector.

$$ME = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 3/\sqrt{2} & 1/\sqrt{2} \\ 3/\sqrt{2} & -1/\sqrt{2} \\ 7/\sqrt{2} & 1/\sqrt{2} \\ 7/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

2.2.2 Refer to the (*eventual*) cluster structure present in data enhanced by PCA transformation for data points obtained in 2.2.1, and the advantage (*or not*) of using PCA.

>>> **The points coincide in pairs. Therefore, the clustering structure is revealed by points** (p_1, p_2) e (p_3, p_4)



2.3. There was conducted a study about the quality of six brands of beer. For that, it was conducted a survey involving 200 beer consumers, asking to classify (with a score 0 to 100), on the importance of the following seven features to decide to buy a certain brand: *low COST*, *high SIZE of the bottle* (volume), *high percentage of ALCOHOL*, the *REPUTATION of the brand*, the *COLOR of the beer*, *nice AROMA*, and *good TASTE*.

The PCA analysis resulted in the following eigen values and corresponding proportions of explained variances:

	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	3.313	47.327	47.327
2	2.616	37.369	84.696
3	.575	8.209	92.905
4	.240	3.427	96.332
5	.134	1.921	98.252
6	9.E-02	1.221	99.473
7	4.E-02	.527	100.000

Extraction Method: Principal Component Analysis.

How many principal components would you choose to better represent the data? Justify your answer.

>>> The first two or three PCA's **due to the explanation of the data scatter > 84%**

Question 3 – SVD problem [2.5 + 1.0 points]

The next table presents the preferences of seven individuals about five movies. The preference scores are expressed in a qualitative scale between 0 and 5, with value '0' indicating '*without opinion'*. There are two "concepts" underlying the movies characterising their gender: '*science-fiction*' and '*romance'*.

	Matrix	Allen	Star Wars	Casablanca	Titanic
Espetador_1	1	1	1	0	0
Espetador_2	3	3	3	0	0
Espetador_3	4	4	4	0	0
Espetador_4	5	5	5	0	0
Espetador_5	0	2	0	4	4
Espetador_6	0	0	0	5	5
Espetador_7	0	1	0	2	2

One had applied singular value decomposition analysis to this matrix leading to the decomposition $U \circ \Sigma \circ V^T$:

$$\begin{bmatrix} .13 & .02 & -.01 \\ .41 & .07 & -.03 \\ .55 & .09 & -.04 \\ .68 & .11 & -.05 \\ .15 & -.59 & .65 \\ .07 & -.73 & -.67 \\ .07 & -.29 & .32 \end{bmatrix} \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \\ .40 & -.80 & .40 & .09 & .09 \end{bmatrix}$$

3.1 Interpret the values of matrices U, Σ e V^T in the context of the problem. For that, choose one mathematical greatness of each matrix.

We may consider the hidden factor as 'Gender of Movies', that can be: Fiction or Romance
- Then matrix U connects people to concepts. For example, Expetactor_1, likes mostly the concept science fiction.
The value 0.14 in the first row and first column of U is smaller than some of the other entries in that column, because while Expectator_1 watches only science fiction, he doesn't rate those movies highly. The second column of the first row of U is close to 0, because Expectator_1 rates romance movies very low.
The matrix V relates movies to concepts. The 0.56 in each of the first three columns of the first row of V^T indicates that the first three movies – The Matrix, Alien, and Star Wars – each are of the science-fiction genre, while the .09's in the last two columns of the first row say that these movies do not partake of the concept romance at all. Likewise, the second row of V^T tells us that the
Finally, matrix Σ gives the strength of each of the concepts. In the example, the strength of the science-fiction concept is 12.4, while the strength of the romance concept is 9.5. Intuitively, the science-fiction concept is stronger because the data provides more information about the movies of that genre and the people who like them.

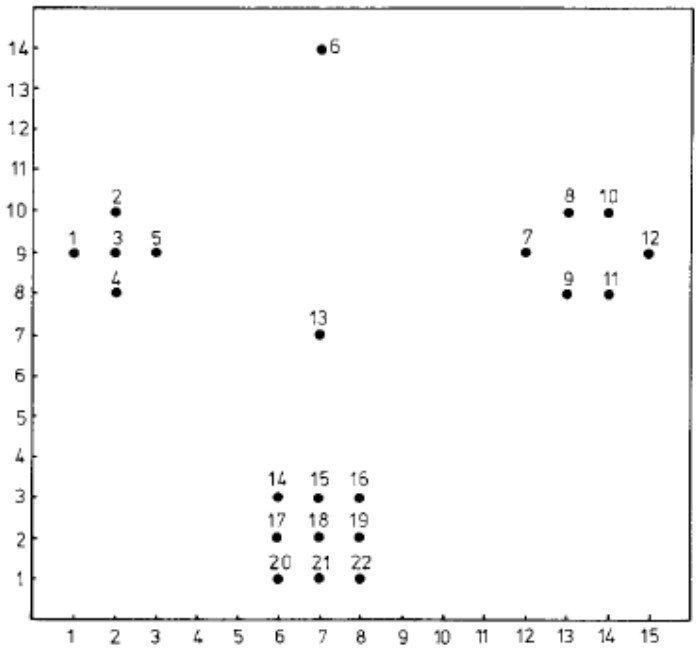
3.2 A new individual is considered. He only watched the *Matrix* movie, with a preference rate score of 4. So, the new individual is represented by the preference vector q= [4, 0, 0, 0, 0].

To characterize the individual in the *space of concepts* one has to calculate qV. Calculate this vector and explain its meaning in the framework of the problem.

We find qV= [2.32,0]^3 . This is to say, Quincy is high in science-fiction interest, and not at all interested in romance.

Question 4 – Partitional Clustering [1.5 + 0.5 + 1.5 + 1.5 points]

Consider the following set of 22 points on a bi-dimensional space



i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
x_i1	1	2	2	2	3	7	12	13	13	14	14	15	7	6	7	8	6	7	8	6	7	8
x_i2	9	10	9	8	9	14	9	10	8	10	8	9	7	3	3	3	2	2	2	1	1	1

The fuzzy *c*-means algorithm was run looking for *c*= 3 clusters with parameter *m*=2.0 and ε=0.001.

The next table presents the fuzzy membership values of the fuzzy *3*-parttion obtained by the FCM algorithm.

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
u_1i	0.97	0.99	1.00	0.96	0.99	0.50	0.02	0.01	0.01	0.01	0.01	0.02	0.37	0.03	0.01	0.02	0.01	0.00	0.01	0.03	0.02	0.02
u_2i	0.02	0.01	0.00	0.03	0.01	0.16	0.02	0.01	0.02	0.01	0.02	0.02	0.41	0.95	0.98	0.95	0.98	1.00	0.98	0.95	0.97	0.96
u_3i	0.01	0.01	0.00	0.01	0.00	0.35	0.96	0.98	0.97	0.97	0.97	0.96	0.22	0.02	0.01	0.03	0.01	0.00	0.01	0.02	0.01	0.02

4.1 Define the membership values for points $\mathbf{x}_5$=(3,9), $\mathbf{x}_{12}$=(15,9) and $\mathbf{x}_{13}$=(7,7). Justify presenting all necessary calculations.

We can "estimate" the approximate membership values by the symmetry of the data points, and by the definition of Fuzzy Membership: weight proportion to the Euclidean distance.

>>> To train, calculate analytically the membership values

4.2 By applying the maximum membership transformation, the former fuzzy membership matrix is transformed into a crisp (hard) one. Define the crisp membership vectors of points $\mathbf{x}_6$ e $\mathbf{x}_{13}$..

4.3 Give an example of a real world 'problem' where the concept of fuzzy partition has advantages over hard (i.e. *crisp*) partition. To illustrate your problem present a fuzzy *3*-membership matrix with five entities.

>>> Choose a real world example pointing out:

- What are the groups;
- Why & how the problem is better modelled with the concept of grades of membership (fuzzy membership function) instead of crisp (0/1) membership values (characteristic membership function.
- Point out the meaning and importance of cluster prototype to the problem

4.4 Outline the main properties of the *iterative anomalous pattern (IAP)* clustering algorithm. Comment on its advantage to initialize the *fuzzy c*-means algorithm.

Concentrate on

- Iterative architecture of the algorithm
- "Bi-partition" and the criterion
- Discuss the elementary conditions of the stop condition and why this allows to get "number of clusters"

Question 5 – Spectral Clustering [0.5 + 1.0 + 1.5 + 1.5 points]

5.1 Suppose you want to split a graph *G* into two subgraphs. Let *L* be G's Laplacian matrix. Choose the sentences that are *True* in order to find a good split.

(i) *The eigenvector corresponding to the second-largest eigenvalue of L.*

(ii) *The eigenvector corresponding to the second-smallest eigenvalue of L.*

(iii) The left singular vector corresponding to the *second-largest singular value of L.*

(iv) The left singular vector corresponding to the second-smallest singular value of L.

>>> (ii) and (iv)

5.2 Consider the data set shown in Fig. 1 formed by a ring and a heap in the center generated from a Gaussian distribution $N(0,1)$, with 100 points for the ring and 100 points for the heap. This data set shows two intuitively obvious clusters that are difficult for conventional clustering algorithms.

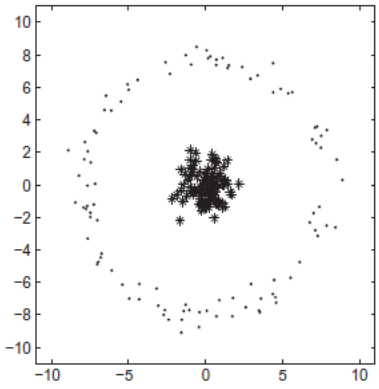


Fig. 1. Two intuitively obvious clusters: stars in the middle and dots in the ring.

The data are first transformed into a 200×200 similarity matrix (using the Gaussian kernel function), on which it is then applied a variant of the Laplacian transformation

The next table shows: (i) five points randomly selected from the original data set on the left: two from the heap and three from the ring; (ii) the affinity similarities of them in the middle; and (iii) the Laplacian transformed similarities on the right.

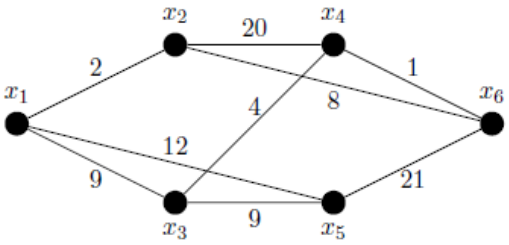
	x-axis	y-axis	2	3	4	5	2	3	4	5
1	-1.1465	0.3274	0.63	0.00	0.02	0.01	0.04	-0.12	-0.14	-0.10
2	0.8956	0.5529		0.00	0.00	0.00		-0.12	-0.15	-0.11
3	0.3086	7.9059			0.00	0.03			0.16	0.40
4	-7.1827	0.0625				0.01				0.46
5	-5.0025	5.8504								

Explain the ability of a spectral clustering algorithm to correctly cluster these five points in case of the: (a) affinity similarity matrix; (b) Laplacian transformed matrix. Specifically, point out the effect of the Laplacian transformation over the affinity similarity.

>>> What is present in the numbers is the effect of the (pseudo-inverse) Laplacian.
Describe what it is from
The Laplacian transformation establishes points {1,2} in a positive relation
points {3,4,5} in a positive relation

It has the effect of establish a "semantic" relation taking positive vs negative values among data points that belong / do not belong to the same cluster.

5.3 Consider the graph in the figure on left and the corresponding similarity matrix on the right. We want to find two communities in this graph using spectral clustering.



$$S = \begin{bmatrix} 0 & 2 & 9 & 0 & 12 & 0 \\ 2 & 0 & 0 & 20 & 0 & 8 \\ 9 & 0 & 0 & 4 & 9 & 0 \\ 0 & 20 & 4 & 0 & 0 & 1 \\ 12 & 0 & 9 & 0 & 0 & 21 \\ 0 & 8 & 0 & 1 & 21 & 0 \end{bmatrix}$$

5.3.1 Calculate the degree matrix, D, and the Laplacian matrix, L.

5.3.2 The two eigenvectors (corresponding to the two smaller eigenvalues) of Laplacian matrix, L, are:

e_1= [0.408, 0.408, 0.408, 0.408, 0.408, 0.408]^T

e_2= [-0.417, 0.520, -0.331, 0.595, -0.295, -0.071]^T

Indicate what are the two clusters obtained from application of K-means. Justify your answer.

C1= {x_1, x_3, x_5, x_6} **C2= {x_2, x_4}**

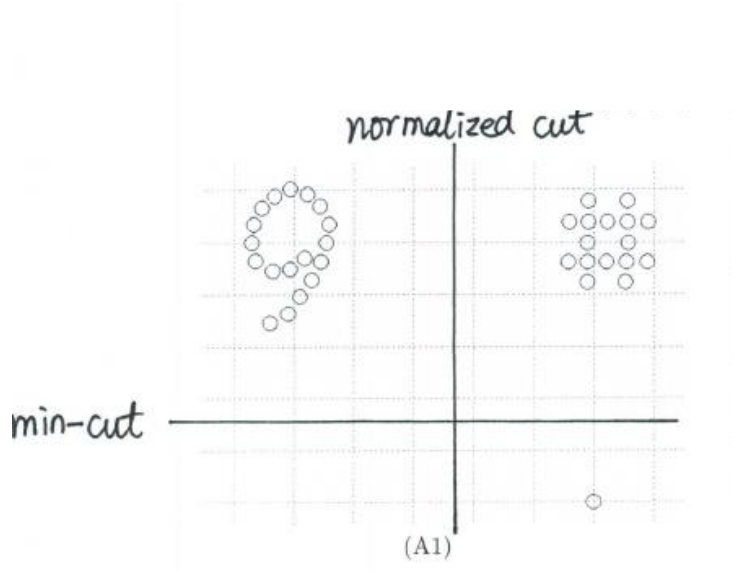
5.3.3 Comment about the obtained partition concerning the graph structure.
>>> The obtained partition is concordant in finding clusters with maximum intra-cluster similarity and minimum inter-cluster similarity

5.4 Consider the next data points plotted in a *1×1* grid. We are going to cluster these data using spectral clustering, in two clusters, denoted by C1 and C2, considering the corresponding *N×N* similarity matrix, W_{ij}, calculated from the Gaussian Kernel. Give the clustering results considering the two clustering criteria, *min-cut* and *normalized cut*, respectively, defined as

$$mincut(A,B) = \sum_{i \in A, j \in B} w_{ij}$$

$$min\, N\, cut(A,B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)}$$

Copy the figure to your answer sheet and mark mark out the clustering results for each of the clustering criteria. Briefly, justify your answer.



The clustering results are justified by the fact that the min-cut only considers the External cluster connections, and does not take into account internal cluster density;

The normalized cut tends to produce balanced clusters (in terms of their cardinality