From Statistics to Data Mining — Written exam
Master 1 COSI / CPS^2 — Saint-Étienne, France

Preliminary remarks

Two-hour "From Statistics to Data Mining" written exam. No documents are allowed. The exam is scored out of 20 points but it is possible to obtain up to 24 points, so choose to answer the questions that seem easy to you. The 3 exercises are independent of each other. The written exam corresponds to 70% of the final mark.

Exercise 1 — 8 points

Let A be a dataset described by two features, x et y :

x_i	-10	-8	-6	-4	-2	1	3	5	7	9
y_i	0	7	13	17	19	20	18	15	10	4

2 pts **1.1.** Plot the graphical representation of y against x.

2 pts **1.2.** Calculate the correlation coefficient between x and y.

As a reminder, the correlation coefficient ρ_{XY} between two variables X and Y is $\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$, the covariance $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)$, $\sigma_X^2 = \mathrm{E}(X^2) - \mathrm{E}^2(X)$ and $\sigma_Y^2 = \mathrm{E}(Y^2) - \mathrm{E}^2(Y)$.

Does the correlation coefficient ρ_{xy} seem to you to indicate a significant linear relationship between the variables x and y ?

2 pts **1.3.** Propose a simple linear regression model on the dataset. The regression line (or least squares line) is given by :

- $y = mx + b$
- with $m = \frac{\overline{x.y} - \bar{x}.\bar{y}}{\overline{x^2} - (\bar{x})^2}$
- and $b = \bar{y} - m.\bar{x}$

2 pts **1.4.** Try a change of variable to find a linear relationship between x and y. We want to find $y = \Phi(x)$ by using a polynomial function such as $y = mx^q + b$. Choose the value of q which would seem to you the most suitable for a linear approximation between this transformation of the predictive variable x and the variable y we want to predict, for example $y = m\sqrt{|x|} + b$ with $q = 1/2$ or $y = mx^2 + b$ with a polynomial of order 2.

What are the coefficients and the general shape of this new linear regression model $y = \Phi(x)$? (In other words, find the 3 parameters q, m and b.)

Indicate what are the values of y predicted with this model if $x = -9$, $x - 3$, $x = 5$, and $x = 15$.

Exercise 2 — 8 points

4 pts **2.1.** Let B a dataset described by 3 variables whose covariance matrix is $\Sigma_B = \begin{pmatrix} 6 & 3 & 3 \\ 3 & 5 & 4 \\ 3 & 4 & 5 \end{pmatrix}$

Deduce the eigenvalues λ_1, λ_2 et λ_3 thanks to the resolution of the characteristic equation : $\det(\Sigma_B - \lambda I) = 0$

You will be asked to solve an equation with a polynomial of degree 3 of the form $x^3 + P \times x^2 + Q \times x + R = 0$, i.e., find the roots of the equation λ_1, λ_2 and λ_3 such that $(x - \lambda_1)(x - \lambda_2)(x - \lambda_3) = 0$. A fast solution of an equation of a polynomial of degree 3 is possible if the roots λ_1, λ_2 and λ_3 are integers : for that, just find the divisors (integers) of R because $R = -\lambda_1 \times \lambda_2 \times \lambda_3$.

4 pts **2.2.** Sort (largest to smallest) the eigenvalues ($\lambda_1 \geq \lambda_2 \geq \lambda_3$). From these eigenvalues, it is possible to obtain a PCA (by looking for the eigenvectors and the eigenspaces associated with the eigenvalues).

By reducing the dimension and representing the initial dataset in 3 dimensions in a 2-dimensional space (i.e., a graph drawn in a usual plane), how much of the variance will be explained by the two larger eigenvalues (λ_1 and λ_2) used for the projection of the data ? Will the quality of this representation in 2 dimensions along the two factorial axes obtained with this PCA be good compared to the 3 initial dimensions ?

Same question by making a reduction of dimension keeping only the first component on the 3 initial dimensions.

Exercise 3 — 8 points

Let C by a dataset of 20 observations (ω_1 to ω_{20}) described by 3 variables (X_1, X_2 et X_3). The values of C are the following :

ω_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X_1	1	2	3	4	5	8	9	10	11	12	13	14	15	16	17	19	20	21	22	23
X_2	11	9	10	13	9	1	3	2	5	3	4	2	3	1	3	16	14	15	16	15
X_3	2	1	5	3	4	19	20	18	16	17	7	5	6	3	4	2	4	1	3	5

4 pts **3.1.** Make the following 3 graphical representations from the dataset C : X_2 as a function of X_1, X_3 as a function of X_2 and X_3 as a function of X_1. Title and caption your graphics. Tip : Use the length of a square on your exam sheet (= 0.5 cm) as the unit of your graph.

In your opinion, from a simple observation of these three graphs, how many clusters could we obtain if we made a classification on the data of C ?

4 pts **3.2.** The distance matrix (squared Euclidean distances) between the 20 observations of the dataset C is the following :

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	ω_{11}	ω_{12}	ω_{13}	ω_{14}	ω_{15}	ω_{16}	ω_{17}	ω_{18}	ω_{19}
ω_2	6																		
ω_3	14	18																	
ω_4	14	24	14																
ω_5	24	18	6	18															
ω_6	438	424	302	416	298														
ω_7	452	446	310	414	308	6													
ω_8	418	402	282	382	270	6	6												
ω_9	332	322	210	282	196	34	24	14											
ω_{10}	410	392	274	360	254	24	18	6	6										
ω_{11}	218	182	140	178	98	178	186	134	86	102									
ω_{12}	259	209	185	225	131	233	251	185	139	149	9								
ω_{13}	276	230	194	230	140	222	232	170	120	130	6	3							
ω_{14}	326	264	254	288	186	320	342	262	210	216	34	9	14						
ω_{15}	324	270	246	270	180	310	320	246	184	194	26	11	8	6					
ω_{16}	349	339	301	235	249	635	593	533	381	443	205	230	201	235	177				
ω_{17}	374	358	306	258	250	538	498	440	306	354	158	181	150	186	130	9			
ω_{18}	417	397	365	297	301	689	649	579	425	481	221	234	205	225	169	6	11		
ω_{19}	467	453	401	333	339	677	627	565	411	465	241	264	227	261	195	10	9	6	
ω_{20}	509	493	425	369	361	617	565	507	365	409	225	250	209	249	181	26	11	20	6

From this distance matrix, perform an agglomerative hierarchical clustering by using the single link algorithm for the agglomeration criterion (the dissimilarity between two clusters is equal to the minimum of the distances between individuals of the 2 clusters). Draw the dendrogram associated with this hierarchical clustering.

How many clusters do you think it would be appropriate to find in this clustering ? How can we find the right number of clusters ?