Data Mining
Exercises
François Bachoc
Université Paul Sabatier

The lecture notes can be found here: `https://www.math.univ-toulouse.fr/~xgendre/ens/m2se/DataMining.pdf`.

Exercise 1

1) *Consider 10 cars that are for sale with prices (in k euros)* 10, 6, 7, 6, 22, 43, 33, 7, 8, 11*. Consider the case of uniform weights. Compute the mean price. Compute the median price.*

2) *Consider the following regions with population (in millions) and unemployment rates (in percent) given by the pairs* $(4, 8), (4, 4), (6, 10), (3, 9), (2, 6), (7, 21), (6, 11), (4, 7), (5, 2), (8, 8)$ *of the form* $(population, unemployment)$*. Each region is weighted by its population. Compute the normalized weights* $w_1, \ldots, w_n$ *for these data. With these normalized weights, compute the mean and the standard deviation of the unemployment rate. You can use the formula* $\sigma^2(x) = \overline{x^2} - \bar{x}^2$*. Compute the quantile* $q_{0.25}(x)$ *for these data* x*. You can use a calculator.*

3) *Consider data* $x_1, \ldots, x_n$ *with* $n = 100$ *and* $x_i = i$ *for* $i = 1, \ldots, 100$*. Consider uniform weights. Compute the mean and median of* x*. Consider that a data point* $x_{n+1} = 10^6$ *is added. Compute the new mean and median (still with uniform weights). Interpret the results.*

Exercise 2

Prove that the histogram function $h_{x,\lambda} : \mathbb{R} \to \mathbb{R}$ *on page 4 of the lecture notes has integral* 1.

Exercise 3

Prove that $|\rho(x, y)| = 1$ *if and only if* $(x_1, y_1), \ldots, (x_n, y_n)$ *are all distributed on a straight line (as written on page 6 of the lecture notes).*

Exercise 4

Consider 7 *elementary school pupils with ages and weights given by the pairs*

$$(4, 25), (5, 28), (6, 31), (7, 33), (8, 32), (9, 39), (10, 43)$$

(of the form $(age, weight)$*). Compute the covariance between age and weight (uniform weights). You can use the formula* $\sigma(x, y) = \overline{xy} - \bar{x}\bar{y}$ *Interpret the result.*

Exercise 5

Prove the expression of $\hat{a}$ and $\hat{b}$ on page 6 of the lecture notes. You can use the formulas

$$\sum_{i=1}^n w_i x_i y_i = \sigma(x,y) + \bar{x}\bar{y}, \quad \sum_{i=1}^n w_i x_i^2 = \sigma(x)^2 + \bar{x}^2, \quad \sum_{i=1}^n w_i y_i^2 = \sigma(y)^2 + \bar{y}^2.$$

Exercise 6

We consider 6 *companies and their numbers of employees (in thousands), annual growth (percent) and age (years). The data for these companies, of the form (employees,growth,age) are*

$$(4,5,22),(6,7,11),(6,8,2),(3,8,54),(8,2,34),(4,5,5).$$

Construct the data matrix X and the centered data matrix $\bar{X}$ associated to these data (uniform weights).

Exercise 7

Show that ${}^t u M u = {}^t u[(M + {}^t M)/2]u$ as stated on page 8 of the lecture notes.

Exercise 8

Consider the two-dimensional linear subspace E of $\mathbb{R}^3$ spanned by the two vectors $(1,1,1)$ and $(1,0,0)$. Consider the diagonal matrix M with diagonal elements $(1,1/2,1/2)$ to define the inner product $\langle \cdot,\cdot \rangle_M$.

1) *Show that $(1,0,0)$ and $(0,1,1)$ constitute an M-orthonormal basis of E.*

2) *Compute the M-orthogonal projection of $(1,1,-1)$ on E.*

Exercise 9

The goal is to prove that taking the d first eigenvectors of ΣM (that are M-orthonormal) maximizes the inertia (page 10 of the lecture notes). We let the p eigenvalues of ΣM be $\lambda_1 > \cdots > \lambda_p > 0$. For any M-orthonormal vectors $z^1,\dots,z^d$ in $\mathbb{R}^p$ the intertia is $\sum_{j=1}^d \langle \Sigma M z^j, z^j \rangle_M$ (page 10 of the lecture notes). Assume $d < p$.

1) *Consider M-orthonormal vectors $z^1,\dots,z^d$ in $\mathbb{R}^p$ and write for $j = 1,\dots,d$ $z^j = \sum_{i=1}^p A_{i,j} v^i$, where $v^1,\dots,v^p$ are p M-orthonormal eigenvectors of ΣM associated to the eigenvalues $\lambda_1,\dots,\lambda_p$ and A is a $p \times d$ matrix. Show that ${}^t A A = I_d$.*

2) *For the same vectors $z^1,\dots,z^d$ as before, show that*

$$\sum_{j=1}^d \langle \Sigma M z^j, z^j \rangle_M = \sum_{j=1}^d \sum_{a=1}^p \lambda_a A_{a,j}^2.$$

3) *For $a = 1,\dots,p$, let $\beta_a = \sum_{j=1}^d A_{a,j}^2$. Show that $\beta_a \leq 1$. Hint: you can add columns to A to obtain a $p \times p$ orthonormal matrix $\tilde{A}$ such that ${}^t\tilde{A}\tilde{A} = I_p$ (this is always possible since the d columns of A are I_p-orthonormal).*

4) *Show that $\sum_{a=1}^p \beta_a = d$.*

5) *Show that the maximum of $\sum_{a=1}^p \lambda_a \beta_a$ under the constraints $0 \leq \beta_a \leq 1$ for $a = 1,\dots,p$ and $\sum_{a=1}^p \beta_a = d$ is $\sum_{a=1}^d \lambda_a$.*

6) *Conclude by showing that the d first eigenvectors of ΣM maximize the inertia.*

Exercise 10

Prove the equation ${}^tV(M\Sigma M)V = \Lambda$ *on page 16 of the lecture notes.*

Exercise 11

Show that if a d*-dimensional subspace* E_d *maximizes the intertia of the projected observations* $I_M(x, E_d)$ *(page 10 of the lecture notes) then it minimizes the intertia of the projection errors* $I_M(x - \pi_{E_d}(x)) = \sum_{i=1}^n w_i ||\tilde{x}_i - \pi_{E_d}(\tilde{x}_i)||_M^2$.

Exercise 12

1) *In the context of page 20 of the lecture notes, show that for* $i, j \in \{1, \dots, p\}$,

$$\langle \tilde{x}^j, u^i \rangle_W = \sqrt{\lambda_i} v_j^i.$$

2) *In the context of page 20 of the lecture notes, show that, for* $j \in \{1, \dots, p\}$,

$$\sum_{k=1}^p \rho(x^j, c^k)^2 = 1.$$

Exercise 13

1) *Show that for* $j = 1, \dots, p$, *the new vector of observations* c^j *has mean zero (context of page 15 of the lecture notes).*

2) *For* $d = 1, \dots, p$, *let* E_d *be the two-dimensional subset of* $\mathbb{R}^n$ *spanned by the* d *first* W*-orthonormal eigenvectors* $u^1, \dots, u^d$ *of* $\bar{X}M^t\bar{X}W$ *(context of page 20 of the lecture notes). Show that, for* $j = 1, \dots, p$,

$$\frac{||\pi_{E_d}(x^j)||_W^2}{||x^j||_W^2} = \sum_{k=1}^d \rho(x^j, c^k)^2.$$

Exercise 14

The goal is to carry out the computations of PCA on simple simulated data. Note that you are not expected to interpret the results, since the data are simulated arbitrarily and do not come from a real data set. Consider the data matrix

$$X = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 2 & 2 & 1 \\ 0 & -4 & -4 & 0 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \end{pmatrix}.$$

1) *Consider uniform weights and the matrix* $M = I_4$ *to compute the distances on the space of individuals. Compute the covariance matrix.*

2) *Show that two eigenvalues of* Σ *are* 0 *and that the two first eigenvectors are*

$$v^1 = \begin{pmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{pmatrix} \quad \text{and} \quad v^2 = \begin{pmatrix} 1/\sqrt{2} \\ 0 \\ 0 \\ 1/\sqrt{2} \end{pmatrix}$$

with eigenvalues $44/5$ *and* $8/5$.

3) *Compute* $c^1, c^2, \rho(x^1, c^1), \rho(x^1, c^2), \ldots, \rho(x^4, c^1), \rho(x^4, c^2)$.

4) *Draw a biplot (Section 1.2.4 of the lecture notes).*

Exercise 15

We consider n *individuals, where each of them has the two qualitative variables employment and education. Employment takes the* $p = 2$ *values "employed" (E) and "unemployed" (U). Education takes the* 3 *values "Up to High school" (HS), "Undergraduate degree" (U) and "Graduate degree" (G). The* n *individuals are given by the data matrix*

$$\begin{pmatrix} E & HS \\ E & U \\ U & HS \\ U & G \\ E & G \\ E & G \\ U & HS \\ U & HS \\ E & U \\ U & U \\ U & HS \end{pmatrix},$$

with column 1 *for employment and column 2 for education.*

1) *Construct the contingency table, with the marginal totals and grand total (page 26 of the lecture notes).*

2) *Compute the line profiles matrix* P_1 *and the corresponding center of gravity* g_1

3) *Compute the column profiles matrix* P_2 *and the corresponding center of gravity* g_2

Exercise 16

1) *In the context of page 30 of the lecture notes, show that* $||g_1||^2_{M_1} = 1$.

2) *In the context of page 30 of the lecture notes, show that* $\Sigma_1 M_1$ *and* ${}^tTD_1TD_2$ *have the same eigenvalues apart from the one associated to* g_1.

Exercise 17

1) *In the context of page 31 of the lecture notes, show that the data from the matrix* $C^{(1)}$ *have* $\kappa \times 1$ *mean vector* 0*, with the weight matrix* W_1.

Exercise 18

We consider $n = 8$ voters, where each of them has the two qualitative variables work and preference. Work takes the $p = 3$ values "Dentist" (Den), "Teacher" (T) and "Developper" (Dev). Preference takes the 3 values "Left" (L), "Center" (C) and "Right" (R). The n individuals are given by the data matrix

$$\begin{pmatrix} Den & R \\ T & L \\ T & L \\ Dev & C \\ Dev & C \\ Den & R \\ T & L \\ Dev & C \end{pmatrix},$$

with column 1 for work and column 2 for preference. In this exercize, you have the choice between giving exact expressions of the results (using fractions, square roots,...) or giving (approximate) numerical results.

1) *Construct the contingency table, with the marginal totals and grand total (page 26 of the lecture notes).*

2) *Compute the lines profile matrix P_1, the center of gravity g_1, the matrix D_1, the matrix D_2, the weight matrix W_1 and the distance matrix M_1.*

3) *To perform the PCA on the line profiles, provide the matrix that has the $\kappa = \min(p, q) - 1$ non trivial eigenvalues $\lambda_1 \geq \cdots \geq \lambda_\kappa \geq 0$. Compute $\kappa, \lambda_1, \ldots, \lambda_\kappa$. Compute also the $q \times \kappa$ matrix V_1 which columns are the M_1-orthogonal eigenvectors corresponding to these non-trivial eigenvalues (the choice of these κ eigenvectors will not be unique, so you can take any choice you want that is valid).*

4) *Compute the principal component matrix $C^{(1)}$.*

5) *Using the transition formulae, compute the the principal component matrix $C^{(2)}$.*

6) *Plot the individuals of the two principal component matrix $C^{(1)}$ and $C^{(2)}$, with one color for each of the matrices, together with category name of each individual. Interpret the plot.*

Exercise 19

1) *Prove that for $\ell, \ell' \in \{1, \ldots, m\}$, $(\bar{W})_{\ell,\ell'} = \bar{w}_\ell$ if $\ell = \ell'$ and $(\bar{W})_{\ell,\ell'} = 0$ if $\ell \neq \ell'$ (context of page 39 of the lecture notes).*

2) *Prove that*

$$\bar{W}^{-1}\,{}^t TWX = \begin{pmatrix} {}^t g_1 \\ \vdots \\ {}^t g_m \end{pmatrix}$$

(context of page 40 of the lecture notes).

3) *Prove that ${}^t\bar{G}\bar{W}\bar{G} = {}^t\bar{X}_b W \bar{X}_b$ (context of page 41 of the lecture notes).*

Exercise 20

1) *Here, you can use the concept of rank from linear algebra. For a matrix K of size $a \times b$, its rank* $\mathrm{rank}(K)$ *satisfies* $\mathrm{rank}(K) \leq \min(a, b)$. *Also, if ${}^t v K = 0$ for a non-zero $a \times 1$ vector v, then* $\mathrm{rank}(K) \leq a - 1$. *Also, for two (rectangular) matrices A, B we have* $\mathrm{rank}(AB) \leq \min(\mathrm{rank}(A), \mathrm{rank}(B))$. *Finally, the number of non-zero eigenvalues of a matrix is smaller than its rank.*

Assume that Σ is invertible. Consider the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ of $\Sigma_b \Sigma^{-1}$ (context of page 41 of the lecture notes). Show that at most $\min(p, \kappa - 1)$ *elements of $(\lambda_1, \ldots, \lambda_p)$ are non-zero.*

2) *Here we will show that the eigenvalues of $\Sigma_b \Sigma^{-1}$ are in $[0, 1]$.* **a)** *Show that the eigenvalues of $\Sigma_b \Sigma^{-1}$ are eigenvalues of $\Sigma^{-1/2} \Sigma_b \Sigma^{-1/2}$.* **b)** *Show that these eigenvalues are positive.* **c)** *Then, you can use the following result from linear algebra: the largest eigenvalue of $\Sigma^{-1/2} \Sigma_b \Sigma^{-1/2}$ is*

$$\max_{||x||=1} {}^t x \Sigma^{-1/2} \Sigma_b \Sigma^{-1/2} x.$$

Show that this largest eigenvalue is smaller than 1.

3) *In the context of page 41 of the lecture notes, show that if $\lambda_1 = 0$, then the m centers of gravity of the m groups are equal.*

Exercise 21

1) *Consider the setting of pages 50, 51 of the lecture notes. Consider $n = 20$ individuals with corresponding 20 values of a qualitative variable, taking values in $\{\tau_1, \tau_2, \tau_3\}$, given by*

$$t = {}^t \left(\tau_1, \tau_2, \tau_3, \tau_2, \tau_1, \tau_1, \tau_2, \tau_2, \tau_3, \tau_3, \tau_1, \tau_3, \tau_2, \tau_1, \tau_1, \tau_2, \tau_3, \tau_2, \tau_2, \tau_1 \right).$$

Consider uniform weights. Assume that a partition of the input space has been obtained and that its first region R_1 contains the individuals $x_2, x_4, x_5, x_7, x_{10}, x_{12}, x_{15}, x_{16}, x_{20}$ and only these individuals. Compute the frequencies $\hat{p}_{11}, \hat{p}_{12}, \hat{p}_{13}$. Which value of the qualitative variable should be attributed to R_1, if a classifier is built from this partition of the input space? Compute the value of the Gini index $\mathcal{G}_1$ of R_1.

2) *Consider the data matrix*

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

and the corresponding values of the qualitative variable

$$t = \begin{pmatrix} \tau_1 \\ \tau_2 \\ \tau_1 \\ \tau_2 \end{pmatrix}.$$

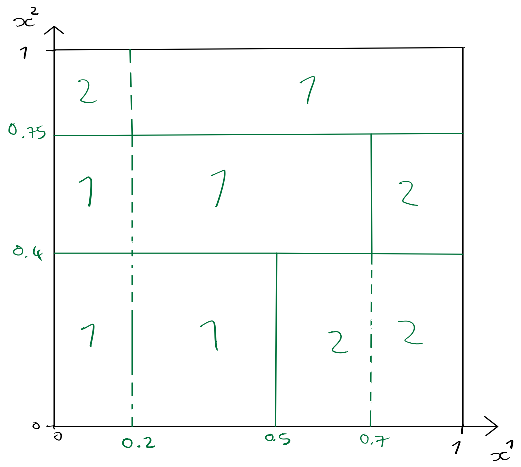
Consider uniform weights. Find $j^\star \in \{1, 2\}$ and $s^\star \in \mathbb{R}$ that minimize

$$\bar{w}_1(j, s) \mathcal{G}_1(j, s) + \bar{w}_2(j, s) \mathcal{G}_2(j, s)$$

(context of page 51 of the lecture notes).

Exercise 22

The following plot is a partition of $[0, 1]^2$ obtained from a classification tree.



Plot a classification tree leading to this partition. Ignore the problem of determining how input points located on the green straight and dashed lines (equality cases in the classification tree) are classified.

Exercise 23

Consider two qualitative variables x^1 with values in $\{1,2\}$ and x^2 with values in $\{1,2\}$. Consider the corresponding data matrix

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 2 & 1 \\ 2 & 1 \\ 2 & 2 \end{pmatrix}.$$

Compute $d_{\chi^2}(1,2)$ (context of page 58 of the lecture notes).

Exercise 24

Consider $n=6$ individuals with corresponding dissimilarity matrix

$$X = \begin{pmatrix} 0 & 1.2 & 3.1 & 4 & 3 & 1 \\ 1.2 & 0 & 3 & 4 & 3 & 1.2 \\ 3.1 & 3 & 0 & 3 & 3 & 2 \\ 4 & 4 & 3 & 0 & 2 & 2 \\ 3 & 3 & 3 & 2 & 0 & 4 \\ 1 & 1.2 & 2 & 2 & 4 & 0 \end{pmatrix}.$$

Consider the hierarchical clustering procedure based on the complete linkage (context of page 65 of the lecture notes). Carry out by hand all the computations of the hierarhical clustering algorithm. Plot the resulting dendogram. What is the clustering obtained if the number of groups is chosen to be 3?