Jean Monnet University

"Data Analysis" Exam
MLDM Master program

December 2022 - **(2 hours)**

Pocket calculators and two 2-sided papers of handwritten notes are allowed.

Exercise 1 (4 points): Likelihood Maximization

The Pareto distribution is a power-law probability distribution that is used in description of social, quality control, scientific, geophysical, actuarial, and many other types of observable phenomena. Its probability density function is described as follows:

$$f(x_i) = \frac{\alpha}{x_i^{\alpha+1}} \qquad \forall x_i \geq 1 \text{ and } \alpha > 0$$

1. Show that $f(x_i)$ is an actual density function.
2. Let $x_1, x_2, ..., x_n$ be a set of n observations of random variables $i.i.d.$ according to a Pareto distribution. Using the **log likelihood**, find an estimate of α.

Exercise 2 (4 points): Convex Optimization

Given the following optimization problem:

$$\min_{x,y} x^2 + 3y^2 + 1$$
$$\text{s.t. } x - 6y = 1$$

1. Show that the objective function $x^2 + 3y^2 + 1$ is convex.
2. Using the method of Lagrange multipliers, find the optimal values x and y.

Exercise 3 (9 points): Linear Discriminant Analysis and Linear Regression

Let $S = \{A = (-1, 0), B = (1, 2), C = (2, 2), D = (6, 4)\}$ be a set 4 points in $\mathbb{R}^2$ where A and B belong to the positive class and C and D belong to the negative class.

PART 1: LDA (6 points)

1. Plot the data.
2. Compute the within-class scatter matrix Σ_W.
3. Compute the in-between-class scatter matrix Σ_B.
4. Compute the eigenvalues from the matrix $\Sigma_W^{-1}\Sigma_B$
5. Compute the eigenvector corresponding to the largest eigenvalue.
6. Compute the new coordinates of A, B, C, D in $\mathbb{R}$ and plot them in the same figure as that of 1.

Note that you won't need a pocket calculator to do this exercise. All calculations can be simply done in your head ;-) !

PART 2 (3 points): Linear Regression (note that this part is independent from PART 1)

In this part, we consider the coordinates (x, y) of A, B, C, D where x is a feature and y is an output to predict by regression. For example, from $A = (-1, 0)$ we get $x_A = -1$ and $y_A = 0$. Using the closed-form solution of linear regression, find the equation $y = \theta_1 x + \theta_0$. Draw the regressor on the same figure as that of PART 1.

Exercise 4 (3 points): Multiple Choice Questions

Circle the letter corresponding to the correct answer (**only one is correct**).
- Each correct answer **adds 1/2**.
- Each incorrect answer **subtracts 1/4**.

1. The determinant of the matrix $A = \begin{pmatrix} 1 & 3 & 2 \\ 2 & 1 & 0 \\ 2 & 1 & 1 \end{pmatrix}$ is:

   a. 2

   b. -5

   c. -1

2. What is the correlation coefficient between X and Y whose joint distribution is defined as follows?

$X\backslash Y$	0	1
0	0	$\frac{1}{2}$
1	$\frac{1}{2}$	0

   a. $\rho_{XY} = 0$

   b. $\rho_{XY} = 1$

   c. $\rho_{XY} = -1$

3. Let X be a binomial variable defined as $X \equiv B(4, \frac{1}{3})$. What is the probability $P(X = 2)$?

   a. $\frac{2}{27}$

   b. $\frac{2}{9}$

   c. $\frac{8}{27}$

4. Let $f(x) = 1$ be a density funcrion $\forall x \in [0, 1]$. The expected value is:

   a. $E(X) = 0$

   b. $E(X) = 1/2$

   c. $E(X) = 1/3$

5. Let X be a Gaussian variable defined as $X \equiv N(8, 2)$. What is the probability $P(X < 6)$?

   a. 0.1587

   b. 0.6431

   c. 0.8413

6. Which of the following methods is non-linear?

   a. PCA

   b. UMAP

   c. LDA