[Exercise 1.]

Calculate the correlation coefficient for the two random variables, X and Y, represented by the following data:

$$X = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, Y = \begin{pmatrix} 2 \\ 3 \\ 1 \\ 4 \\ 5 \end{pmatrix}$$

Once you have calculated the correlation coefficient, interpret the result. Does it suggest a strong, moderate, weak, or no linear relationship between X and Y?
Provide a brief explanation supporting your interpretation.

$$\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{5}(1+2+3+4+5) = 3$$

$$\mu_y = \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{1}{5}(2+3+1+4+5) = 3$$

$$Cov(X,Y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y) = \frac{1}{5}\left[(1-3)(2-3) + (2-3)(3-3) + (3-3)(1-3) + (4-3)(4-3) + (5-3)(5-3)\right] = 1,4$$

$$Var(X) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)^2 = \frac{1}{5}\left[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2\right] = 2$$

$$Var(Y) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu_y)^2 = \frac{1}{5}\left[(2-3)^2 + (3-3)^2 + (1-3)^2 + (4-3)^2 + (5-3)^2\right] = 2$$

$$\sigma(X) = \sqrt{Var(X)} = 1,414 \quad ; \quad \sigma(Y) = \sqrt{Var(Y)} = 1,414$$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\,\sigma(Y)} = \frac{1,4}{\sqrt{2}\cdot\sqrt{2}} = \frac{1,4}{2} = 0,7 \rightarrow$$

RELAZIONE LINEARE POSITIVA (FORTE) tra X, Y.

[Exercise 1.]

Consider two random variables X and Y, for which the following data is provided:

$$X = \{2, 3, 4, 5, 6\}$$

$$Y = \{3, 4, 5, 6, 7\}$$

1. Calculate the covariance matrix $\mathrm{Cov}(X, Y)$.
2. Compute the correlation coefficient $\rho(X, Y)$.
3. Interpret and comment on the results obtained from the covariance matrix and the correlation coefficient. Discuss the strength and direction of the relationship between the two random variables based on the calculated values.

$$\mu_X = \frac{1}{5}(2+3+4+5+6) = 4$$

$$\mu_Y = \frac{1}{5}(3+4+5+6+7) = 5$$

MATRICE COVARIANZA:

$$\begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

$$\mathrm{Cov}(X,Y) = \frac{1}{5}[(2-4)(3-5) + (3-4)(4-5) + (4-4)(5-5) + (5-4)(6-5) + (6-4)(7-5)] = 2$$

$$\mathrm{Var}(X) = \frac{1}{5}[(2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2] = 2$$

$$\mathrm{Var}(Y) = \frac{1}{5}[(3-5)^2 + (4-5)^2 + (5-5)^2 + (6-5)^2 + (7-5)^2] = 2$$

$$\sigma(X) = \sqrt{2} = 1{,}414 \quad ; \quad \sigma(Y) = \sqrt{2} = 1{,}414$$

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\,\sigma(Y)} = \frac{2}{2} = 1 \rightarrow \text{RELAZIONE LINEARE POSITIVA FORTE.}$$

(e perfetta)

{assenza di rumori o errori}

[Exercise 1.] Given the following sets of data:

Data Pair	x	y	s	t
1	2	3	2	3
2	4	7	4	5
3	6	5	6	7
4	8	10	8	9
5	10	12	10	11

1. Compute the correlation coefficient for x and y.
2. Compute the correlation coefficient for s and t.
3. Comment on the results, discussing the strength and direction of the relationships between the variables in each set.

$$\mu_x = \frac{1}{5}(2+4+6+8+10) = 6$$

$$\mu_y = \frac{1}{5}(3+7+5+10+12) = 7,4$$

$$Cov(X,Y) = \frac{1}{5}\left[(2-6)(3-7,4) + (4-6)(7-7,4) + (6-6)(5-7,4) + (8-6)(10-7,4) + (10-6)(12-7,4)\right] = 8,4$$

$$Var(X) = \frac{1}{5}\left[(2-6)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2\right] = 8$$

$$Var(Y) = \frac{1}{5}\left[(3-7,4)^2 + (7-7,4)^2 + (5-7,4)^2 + (10-7,4)^2 + (12-7,4)^2\right] = 10,64$$

$$\sigma(X) = \sqrt{8} = 2,83 \qquad \sigma(Y) = \sqrt{10,64} = 3,26$$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\,\sigma(Y)} = \frac{8,4}{9,23} = 0,91 \rightarrow \text{RELAZIONE LINEARE POSITIVA FORTE}$$

$$\mu_S = \frac{1}{5}(2+4+6+8+10) = 6$$

$$\mu_t = \frac{1}{5}(3+5+7+9+11) = 7$$

$$Var(S) = \frac{1}{5}\left[(2-6)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2\right] = 4,8$$

$$Var(T) = \frac{1}{5}\left[(3-7)^2 + (5-7)^2 + (7-7)^2 + (9-7)^2 + (11-7)^2\right] = 4,8$$

$$\sigma(S) = \sqrt{Var(S)} = 2,19 = \sigma(t) = \sqrt{Var(T)}$$

$$Cov(S,T) = \frac{1}{5}\left[(2-6)(3-7) + (4-6)(5-7) + (6-6)(7-7) + (8-6)(9-7) + (10-6)(11-7)\right] = 4,8$$

$$\rho(S,T) = \frac{Cov(S,T)}{\sigma(S)\,\sigma(T)} = 1 \rightarrow \text{RELAZIONE LINEARE PERFETTA}$$

[Exercise 1.] Consider the following two datasets, each containing 4 elements:

$$X = \{2, 4, 6, 8\}$$

$$Y = \{1, 3, 5, 7\}$$

Calculate the covariance matrix and the correlation coefficient between the datasets X and Y.

Comment on the results appropriately.

$$\mu_X = \frac{1}{4}(2+4+6+8) = 5$$

$$\mu_Y = \frac{1}{4}(1+3+5+7) = 4$$

$$\text{Var}(X) = \frac{1}{4}\left[(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2\right] = 5$$

$$\sigma(X) = \sqrt{5}$$

$$\text{Var}(Y) = \frac{1}{4}\left[(1-4)^2 + (3-4)^2 + (5-4)^2 + (7-4)^2\right] = 5$$

$$\sigma(Y) = \sqrt{5}$$

$$\text{Cov}(X, Y) = 5$$

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X)\,\sigma(Y)} = 1 \rightarrow \text{RELAZIONE LINEARE PERFETTA}$$

[Exercise 1.]

Apply the graphical method to solve the following Linear Programming problem:

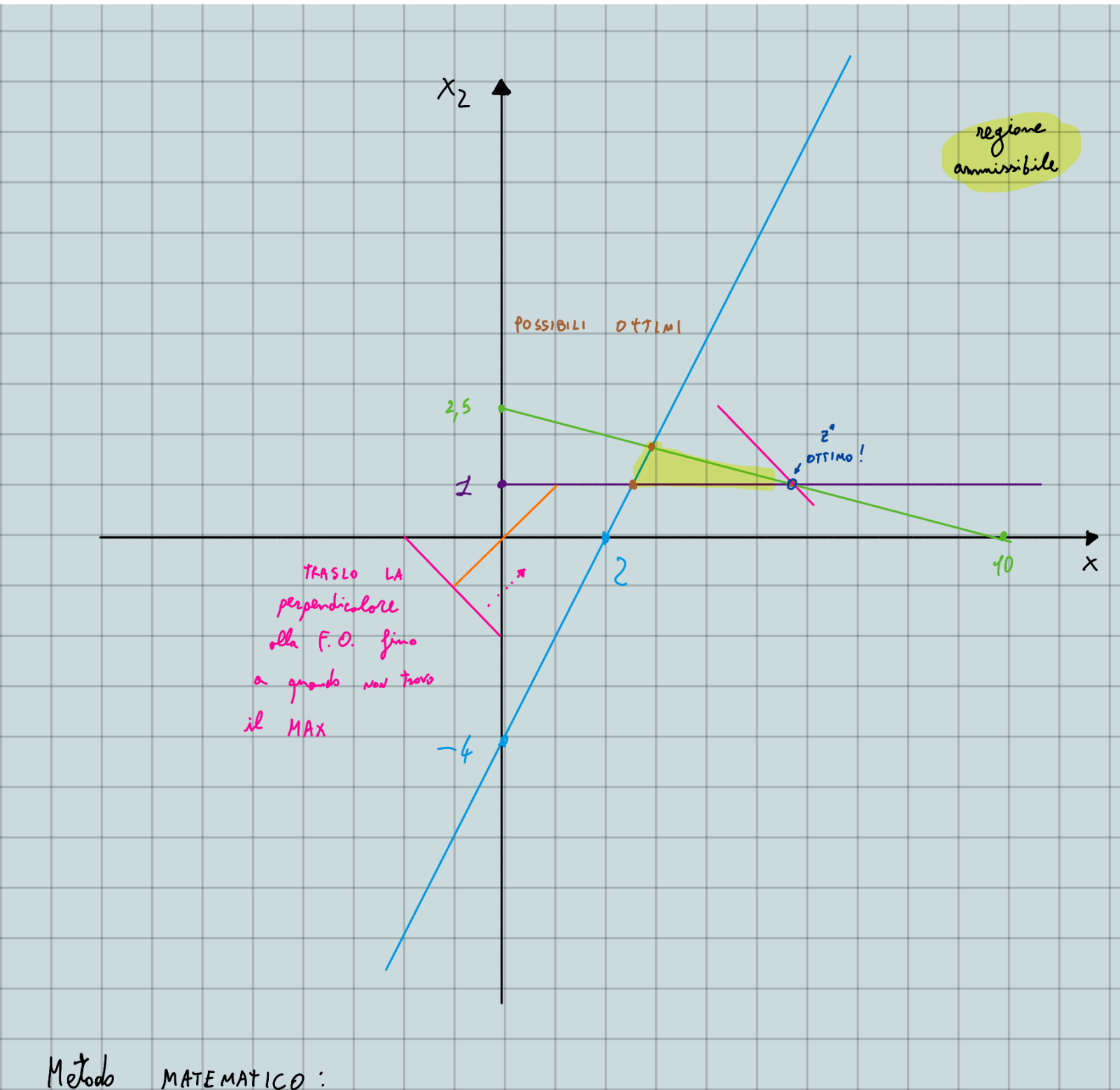
$$\max Z = x_1 + x_2$$

$$2x_1 - x_2 \geq 4$$

$$x_1 + 4x_2 \leq 10$$

$$x_2 \geq 1$$

$$x_1, x_2 \geq 0$$



Metodo MATEMATICO:

$$1^\circ \begin{cases} x_2 = 1 \\ 2x_1 - x_2 = 4 \end{cases} \quad \begin{cases} x_2 = 1 \\ x_1 = 5/2 = 2,5 \end{cases} \qquad z^* = 1 + 2,5 = 3,5$$

$$2^\circ \begin{cases} 2x_1 - x_2 = 4 \\ x_1 + 4x_2 = 10 \end{cases} \quad \begin{cases} x_2 = 2x_1 - 4 \\ x_1 + 8x_1 - 16 = 10 \end{cases} \quad \begin{cases} x_2 = 5,78 - 4 = 1,78 \\ x_1 = \frac{26}{9} = 2,8 \end{cases} \qquad z^* = 4,6$$

$$3^\circ \begin{cases} x_2 = 1 \\ x_1 + 4x_2 = 10 \end{cases} \quad \begin{cases} x_2 = 1 \\ x_1 = 6 \end{cases} \rightarrow \quad z^* = 1 + 6 = 7$$

SOLUZIONE OTTIMA!

[Exercise 1.]

Write the formula for the gradient descent method.

Use 5 iterations of the gradient descent method to minimize the function:

$$f(x, y) = x^2 + y^2 + 2xy$$

with starting point $(x_0, y_0) = (1, 1)$ and a learning rate $\alpha = 0.1$

Comment briefly on the convergence behavior you observed:

1. Does the sequence (x_n, y_n) appear to approach the minimum?

2. Evaluate the values of $f(x_n, y_n)$: Do they decrease as the iteration proceeds?

3. Is the convergence fast or slow?

4. In general, how does the choice of learning rate α affect the results?

$\frac{\partial f}{\partial x} = 2x + 2y$ $\quad$ $\frac{\partial f}{\partial y} = 2y + 2x$

$\nabla f(x,y) = \begin{pmatrix} \partial f/\partial x \\ \partial f/\partial y \end{pmatrix}$ $\quad$ $(x_0, y_0) = (1, 1)$ $\quad$ $\alpha = 0,1$

$x_1 = x_0 - \alpha \frac{\partial f(x_0, y_0)}{\partial x} = 1 - 0,1\,(2 + 2) = 0,6$

$y_1 = y_0 - \alpha \frac{\partial f(x_0, y_0)}{\partial y} = 1 - 0,1\,(2 + 2) = 0,6$

$$x_2 = x_1 - \alpha \frac{\partial f(x_1, y_1)}{\partial x} = 0,6 - 0,1 \left(1,2 + 1,2\right) = 0,36$$

$$y_2 = y_1 - \alpha \frac{\partial f(x_2, y_2)}{\partial y} = 0,6 - 0,1 \left(1,2 + 1,2\right) = 0,36$$

$$x_3 = x_2 - \alpha \frac{\partial f(x_2, y_2)}{\partial x} = 0,36 - 0,1 \left(0,72 + 0,72\right) = 0,216$$

$$y_3 = y_2 - \alpha \frac{\partial f(x_2, y_2)}{\partial y} = 0,36 - 0,1 \left(0,72 + 0,72\right) = 0,216$$

$$x_4 = x_3 - \alpha \frac{\partial f(x_3, y_3)}{\partial x} = 0,13$$

$$y_4 = y_3 - \alpha \frac{\partial f(x_3, y_3)}{\partial y} = 0,13$$

$$x_5 = x_4 - \alpha \frac{\partial f(x_4, y_4)}{\partial x} = 0,078$$

$$y_5 = y_4 - \alpha \frac{\partial f(x_4, y_4)}{\partial x} = 0,078$$

1) La sequenza si avvicina al minimo, che sarà ottenuto per valori di $y = -x = 0$.
$f(x,y) = x^2 + y^2 + 2xy = (x+y)^2$ → IL MINIMO É ZERO!

2) Sì i valori di $f(x,y)$ diminuiscono ad ogni iterazione:
$f(x_0,y_0) = 4$; $f(x_1,y_1) = 1,44$; $f(x_2,y_2) = 0,52$; $f(x_3,y_3) = 0,19$; $f(x_4,y_4) = 0,07$; $f(x_5,y_5) = 0,02$

3) La convergenza sembra veloce!

4) Se α é troppo piccolo si converge al minimo ma in tempi troppo lunghi.
Se α é troppo alto potrebbe consure una divergenza.

[Exercise 2.]
Given two sets of discrete probability distributions P and Q as follows:

$$P = \{0.2, 0.3, 0.5\}$$

$$Q = \{0.4, 0.4, 0.2\}$$

Calculate the Kullback-Leibler (KL) divergence from P to Q and from Q to P.

Comment on the results.

La DKL misura l'errore tra una distribuzione "vera" ed una sua approssimazione.

FORMULAZIONE GENERALE:

$$D_{KL}(f \| g) = \begin{cases} \int f(x) \ln \frac{f(x)}{g(x)} \, dx & [\text{CONTINUO}] \\ \sum_{x \in X} f(x) \ln \frac{f(x)}{g(x)} & [\text{DISCRETO}] \end{cases}$$

(Valore più basso è meglio)

$$D_{KL}(P \| Q) = 0,2 \ln\left(\frac{0,2}{0,4}\right) + 0,3 \ln\left(\frac{0,3}{0,4}\right) + 0,5 \ln\left(\frac{0,5}{0,2}\right) \simeq 0,233$$

$$D_{KL}(Q \| P) = 0,4 \ln\left(\frac{0,4}{0,2}\right) + 0,4 \ln\left(\frac{0,4}{0,3}\right) + 0,2 \ln\left(\frac{0,2}{0,5}\right) \simeq 0,209$$

Notiamo come la DKL non è SIMMETRICA. In questo caso deduciamo che P è una migliore approssimazione di Q rispetto a quanto Q lo sia per P.

[Exercise 2.]
Given two sets of discrete probability distributions P and Q as follows:

$$P = \{0.3, 0.2, 0.1, 0.4\}$$

$$Q = \{0.1, 0.4, 0.2, 0.3\}$$

Calculate the Kullback-Leibler (KL) divergence of distribution Q from distribution P $D_{\mathrm{KL}}(Q||P)$.

Comment on the result.

$$D_{KL}(Q\|P) = 0,1 \ln\left(\frac{0,1}{0,3}\right) + 0,4 \ln\left(\frac{0,4}{0,2}\right) + 0,2 \ln\left(\frac{0,2}{0,1}\right) + 0,3 \ln\left(\frac{0,3}{0,4}\right) =$$

$$\simeq 0,22$$

Il valore assunto dalla D_{KL} è abbastanza basso, quindi P è una buona approssimazione per Q.

[Exercise 2.] Consider two distributions P and Q such that

Distribution	$X = 1$	$X = 2$	$X = 3$
$P(X)$	0.2	0.5	0.3
$Q(X)$	0.1	0.6	0.3

Calculate the Kullback-Leibler (KL) divergence $D_{\text{KL}}(P\|Q)$ and comment on the result.

$$D_{KL}(P\|Q) = 0{,}2 \ln\left(\frac{0{,}2}{0{,}1}\right) + 0{,}5 \ln\left(\frac{0{,}5}{0{,}6}\right) + 0{,}3 \ln\left(\frac{0{,}3}{0{,}3}\right) = 0{,}047$$

Q é un ottima approssimazione per P.

[**Exercise 2.**] Given two discrete probability distributions P and Q:

$$P = \begin{pmatrix} 0.2 & 0.5 & 0.3 \end{pmatrix}, \quad Q = \begin{pmatrix} 0.1 & 0.6 & 0.3 \end{pmatrix},$$

compute the Kullback-Leibler (KL) divergence $D_{KL}(P \parallel Q)$ and comment on the results.

What does the KL divergence indicate about the relationship between these two distributions?

$$D_{KL}(P\|Q) = 0{,}2 \ln\left(\frac{0{,}2}{0{,}1}\right) + 0{,}5 \ln\left(\frac{0{,}5}{0{,}6}\right) + 0{,}3 \ln\left(\frac{0{,}3}{0{,}3}\right) = 0{,}047$$

La divergenza Kullback-Leibler è usata per quantificare quanto bene una distribuzione predittiva si adatta ad una distribuzione osservata (di riferimento). [Un valore basso è desiderabile].

[**Exercise 2.**] Consider the following two discrete probability distributions P and Q over the same set of outcomes $\{1, 2, 3\}$:

$$P = \{0.7, 0.2, 0.1\}$$

$$Q = \{0.1, 0.1, 0.8\}$$

Calculate the KL divergence $D_{\mathrm{KL}}(P \parallel Q)$ (use the natural logarithm or logarithm base 2).

Comment on the value obtained. What does the magnitude of the KL divergence suggest about how different the two distributions are?

$$D_{KL}(P||Q) = 0,7 \ln\left(\frac{0,7}{0,1}\right) + 0,2 \ln\left(\frac{0,2}{0,1}\right) + 0,1 \ln\left(\frac{0,1}{0,8}\right) \simeq 1,29$$

Indica quanta un'informazione è persa quando una distribuzione Q viene usata per approssimare un'altra distribuzione P.

In questo caso P e Q sono molto diverse (Q è una scarsa approssimazione per P → → l'informazione contenuta in P è persa se si tenta di rappresentarla usando Q).

[**Exercise 2.**] Let P and Q be two discrete probability distributions over the set of three possible events $X = \{x_1, x_2, x_3\}$, defined as follows:

$$P(x_1) = 0.5, \quad P(x_2) = 0.3, \quad P(x_3) = 0.2$$
$$Q(x_1) = 0.4, \quad Q(x_2) = 0.4, \quad Q(x_3) = 0.2$$

1. Write the definition of the Kullback-Leibler (KL) divergence.
2. Calculate the Kullback-Leibler divergence between P and Q, i.e., compute $D_{KL}(P \parallel Q)$
3. Comment on the result

La $D_{KL}(f||g)$ misura l'errore presente tra una funzione di distribuzione di probabilità "esatta" f ed una sua approssimazione g.

$$D_{KL}(P||Q) = 0,5 \ln\left(\frac{0,5}{0,4}\right) + 0,3 \ln\left(\frac{0,3}{0,4}\right) + 0,2 \ln\left(\frac{0,2}{0,2}\right) \simeq 0,253$$

Abbiamo ottenuto un valore prossimo a zero (piccolo e positivo), quindi Q è una buona approssimazione di P.

[Se le due distribuzioni fossero identiche avremmo ottenuto esattamente 0].

[Exercise 3.]
Give a brief description of the gradient descent method.

Provide a clear and concise explanation of what the learning rate coefficient represents in the gradient descent algorithm.

Describe how the choice of learning rate affects the convergence and discuss the trade-offs involved in selecting a high or low learning rate.

Il metodo della discesa del gradiente é una tecnica di ottimizzazione usata per trovare il minimo di una funzione. Il learning rate "α" rappresenta la quantitá di tempo trascorso tra due misurazioni (la dimensione del passo con cui l'algoritmo si muove lungo la direzione del gradiente ad ogni iterazione). La scelta di α é cruciale perché influisce sulla convergenza del metodo.

Se il learning rate é troppo basso → discesa lenta → numero alto di iterazioni → tempo di calcolo eccessivo.

Se α é troppo grande → si potrebbe superare il minimo
- DIVERGENZA → il valore di $f(x,y)$ aumenta anziché diminuire dopo una certa iterazione.
- OSCILLAZIONI intorno al min senza convergere
- BLOCCO IN un punto di sella.

Bisogna trovare il giusto compromesso tra velocitá di convergenza e stabilitá dell'algoritmo.

Criteri di stop: $[\varepsilon \simeq 10^{-5}]$

$\|\nabla f(x_{i+1})\| \leq \varepsilon$ → NORMA del gradiente minore di epsilon.

$\|x_{i+1} - x_i\| < \varepsilon$ → NORMA della differenza dei parametri tra un'iterazione e l'altra é al di sotto di una certa soglia

$i \geq i_{max}$ → raggiungimento NUMERO MAX iterazioni

[Exercise 3.]

Describe briefly how you would choose the reduced dimension in techniques such as Singular Value Decomposition (SVD) or Principal Component Analysis (PCA). Discuss any considerations or strategies involved in making this decision. Additionally, explain how the choice of reduced dimension impacts the performance and efficiency of these dimensionality reduction methods.

La riduzione della dimensionalità ha l'obiettivo di produrre una rappresentazione più compatta e facilmente interpretabile dei dati, concentrando l'attenzione sulle variabili più rilevanti. (Un algoritmo di Machine Learning può degradare con troppe variabili di INPUT).

PCA (Principal Component Analysis) le componenti principali (Principal Components) [PC: calcolate come combinazione lineare delle variabili originali] sono ordinate in base alla varianza spiegata (dalla più grande alla più piccola). L'obiettivo è mantenere una percentuale di varianza sufficientemente alta [85% - 95%].

compresa tra l'85% ed il 95%

La varianza spiegata cumulativa si calcola come:

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{m} \lambda_i}$$

Dove:

- "m" è il numero totale di componenti
- "K" è il numero di componenti scelte per ridurre la dimensione → SI SCEGLIE OPPORTUNAMENTE
- "λ" gli autovalori delle componenti principali

SVD (Singular Value Decomposition) permette di scomporre una matrice A $(m \times m)$ in tre matrici $U \Sigma V^T$ di dimensioni $(m \times m)$, $(m \times m)$, $(m \times m)$.

Σ contiene i valori singolari "σ_i" ordinati in modo decrescente. Per ridurre la dimensionalità scegliamo i primi K valori singolari (i più grandi). In questo caso l'obiettivo è mantenere un energia conservata compresa tra l'85% ed il 90%, che si calcola come:

$$\frac{\sum_{i=1}^{K} \sigma_i^2}{\sum_{i=1}^{m} \sigma_i^2}$$

Le due tecniche sono strettamente correlate e spesso usate in modo interscambiabile, perché $\lambda_i = \sigma_i^2$.

Bisogna però trovare il giusto trade-off tra compressione (che porta ad una minore occupazione della memoria → velocità di calcolo più elevata) e la perdita di informazione.

[Exercise 3.]

1. Define the Maximum Likelihood Estimate (MLE). Explain the contexts in which it is used.

2. Given the following data representing the number of successes in a series of Bernoulli trials:

$$\{1, 0, 1, 1, 0, 1, 0, 1, 1, 0\}$$

where 1 represents a success and 0 represents a failure, calculate the Maximum Likelihood Estimate (MLE) of the probability of success, p.

3. Comment on the result, discussing its implications for the given data.

1

Il Maximum Likelihood Estimate é un metodo fondamentale per la stima dei parametri di un modello statistico. L'obiettivo é trovare i valori dei parametri del modello che massimizzano la probabilitá di (VEROSIMIGLIANZA) aver osservato i dati a disposizione (l'MLE cerca i parametri che rendono i dati osservati "piú verosimili" al modello scelto).

L'MLE é ampiamente utilizzato in vari contesti di modellazione statistica e ML, per stimare i parametri di diverse distribuzioni e modelli:

- STIMA PARAMETRI DISTRIBUZIONI: Bernulli; Binomiale; Gaussiana; Poisson; Esponenziale; Geometrica.
- REGRESSIONE; CLASSIFICAZIONE.

2

Distribuzione di Bernulli:

$$L(\rho) = \prod_{i=1}^{n} \rho^{x_i} (1-\rho)^{1-x_i}$$

$$\hat{\rho} = \arg\max_{\rho} \left[\ln \prod_{i=1}^{n} \rho^{x_i} (1-\rho)^{1-x_i} \right] = \arg\max_{\rho} \sum_{i=1}^{n} \ln \rho^{x_i} (1-\rho)^{1-x_i} = \arg\max_{\rho} \sum_{i=1}^{n} \left(\ln \rho^{x_i} + \ln (1-\rho)^{1-x_i} \right)$$

Ponendo la derivata uguale a zero:

$$\sum_{i=1}^{n} \frac{x_i}{\rho} - \sum_{i=1}^{n} \frac{(1-x_i)}{1-\rho} = \frac{1}{\rho} n \bar{x}_n - \frac{n}{1-\rho} + \frac{n x_n}{1-\rho} = 0 \rightarrow \frac{n\bar{x}_n - \rho n \bar{x}_n - n\rho + n\rho\bar{x}_n}{\rho(1-\rho)} = 0$$

$$\rightarrow n\bar{x}_n - n\rho = 0 \rightarrow \hat{\rho} = \bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i =$$

$$= \frac{1}{10}\left(1+0+1+1+0+1+0+1+1+0\right) = \frac{6}{10} = 0,6$$

3]

Quindi la probabilità di successo è del 60%.
Se il processo che ha generato i dati continua nello stesso modo, ci aspettiamo una propensione maggiore verso il successo (60% delle future osservazioni saranno 1).

Avendo solo 10 osservazioni però l'incertezza sulla stima è elevata. Con più dati l'MLE diventerebbe più affidabile.

[Exercise 3.]
Consider the function

$$f(x, y) = x^2 + y^2 + 4xy.$$

Apply two iterations of the gradient descent method with an initial point $(x_0, y_0) = (1, 2)$ and a learning rate $\alpha = 0.1$.

Verify that the value of the function has indeed decreased after each iteration.

$$\frac{\partial f}{\partial x} = 2x + 4y$$

$$\frac{\partial f}{\partial y} = 2y + 4x$$

$$\nabla f(x,y) = \left(\frac{\partial f}{\partial x} , \frac{\partial f}{\partial y} \right)$$

ITERAZIONI:

0: $(x_0, y_0) = (1, 2)$

$f(x_0, y_0) = 1 + 4 + 8 = 13$

1:
$$\begin{cases} x_1 = x_0 - \alpha \frac{\partial f}{\partial x}(x_0, y_0) = 1 - 0,1\,(2+8) = 0 \\ y_1 = y_0 - \alpha \frac{\partial f}{\partial y}(x_0, y_0) = 2 - 0,1\,(4+4) = 1,2 \end{cases}$$

$f(x_1, y_1) = 0 + 1,44 + 0 = 1,44$

2:
$$\begin{cases} x_2 = x_1 - \alpha \frac{\partial f}{\partial x}(x_1, y_1) = 0 - 0,1\,(4,8) = -0,48 \\ y_2 = y_1 - \alpha \frac{\partial f}{\partial y}(x_1, y_1) = 1,2 - 0,1\,(2,4) = 0,96 \end{cases}$$

$f(x_2, y_2) = 0,23 + 0,92 - 1,84 = -0,69$

COMMENTO:
la funzione diminuisce ad ogni iterazione, il metodo della discesa del gradiente sta funzionando correttamente.
Il valore di α sembra adeguato.

[**Exercise 3.**] Briefly describe the concept of the Maximum Likelihood Estimate (MLE). In your answer, please include the following:

- A clear mathematical definition of MLE and its derivation.
- Explain how MLE is used to estimate the parameters of a statistical model.
- Provide at least one practical example of applying MLE to real-world data.

- Se g^{G_p} é il GLOBAL MINIMIZER del RISCHIO nella classe G_p. Quindi: $l(g^{G_p}) = \min_{g \in G_p} l(g)$.

Sia $\vartheta^* = \operatorname{argmin}_\vartheta E\, \text{Loss}\left(f(x), g(x|\vartheta)\right) = \operatorname{argmin}_\vartheta \int \left(\ln f(x) - \ln g(x|\vartheta)\right) f(x)\, dx =$

$= \operatorname{argmin}_\vartheta \left[- \int \ln g(x|\vartheta) f(x)\, dx \right] = \operatorname{argmax}_\vartheta E \ln g(x|\vartheta)$.

Allora $g^{G_p} = g(\cdot|\vartheta^*)$; quindi apprendere g^{G_p} equivale a stimare ϑ^*.

Vogliamo minimizzare la perdita durante l'allenamento. $t = x_1, \dots, x_n$ →

$\frac{1}{n} \sum_{i=1}^{n} \text{Loss}\left(f(x_i), g(x_i|\vartheta)\right) = \frac{1}{n} \sum_{i=1}^{n} \left(\ln f(x_i) - \ln g(x_i|\vartheta)\right)$. →

$\hat{\vartheta}_n = \operatorname{argmin}_\vartheta \frac{1}{n} \sum_{i=1}^{n} \left(\ln f(x_i) - \ln g(x_i|\vartheta)\right) = \operatorname{argmax}_\vartheta \frac{1}{n} \sum_{i=1}^{n} \ln g(x_i|\vartheta) = \operatorname{argmax}_\vartheta \ln \prod_{i=1}^{n} g(x_i|\vartheta) =$

$= \operatorname{argmax}_\vartheta \prod_{i=1}^{n} g(x_i|\vartheta)$. → $L(\rho) = \prod_{i=1}^{n} g(x_i|\vartheta)$.

$L(\rho)$ rappresenta la Likelihood dei dati (verosimiglianza).

$\hat{\vartheta}_n$ é l'MLE di ϑ^*.

Quindi la Maximum Likelihood Estimation é usata per stimare i parametri di un modello statistico scelto, che massimizzano la probabilità che i dati a disposizione seguano la distribuzione presa in considerazione

- Si costruisce la funzione di verosimiglianza basata sul modello scelto e la si massimizza per ottenere le stime dei parametri.

- Distribuzione Binomiale: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

$$L(p) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\frac{dL(p)}{dp} = 0 \rightarrow \binom{n}{k} \left[k p^{k-1} (1-p)^{n-k} - (n-k) p^k (1-p)^{n-k-1} \right] = 0 \rightarrow$$

$$\rightarrow \binom{n}{k} p^{k-1} (1-p)^{n-k-1} \left[k(1-p) - (n-k)p \right] = 0 \rightarrow$$

$$\rightarrow \binom{n}{k} p^{k-1} (1-p)^{n-k-1} \left[k - \cancel{kp} - np + \cancel{kp} \right] = 0 \rightarrow k - np = 0$$

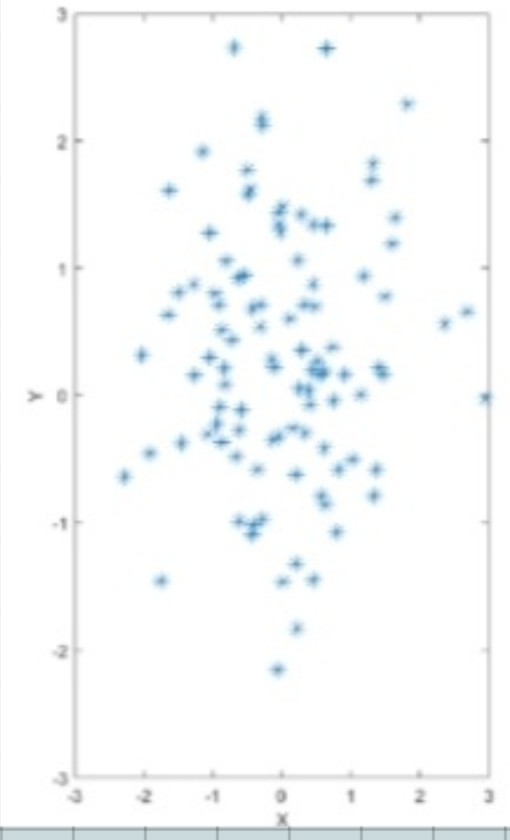
$$\rightarrow \hat{p} = \frac{k}{n}$$ con "k: numero di successi; n: numero totale lanci".

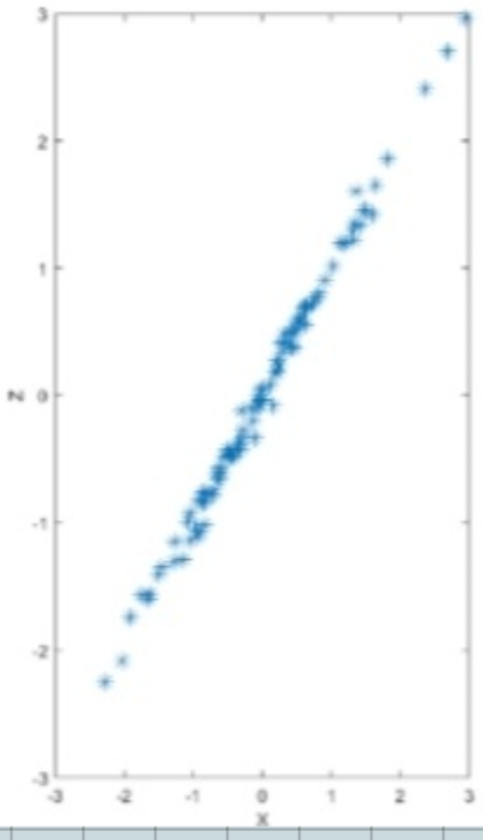
Una moneta è lanciata 100 volte, ottenendo 55 testa. L'MLE per la probabilità di ottenere testa in un singolo lancio è:

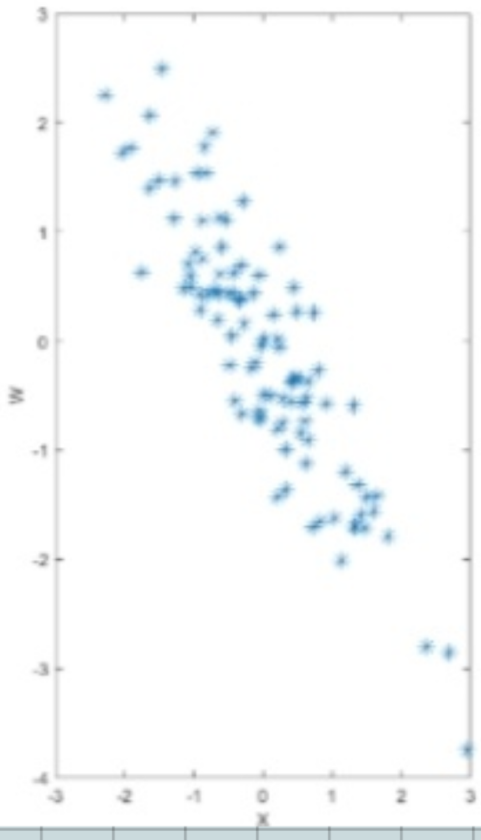
$$\hat{p} = \frac{55}{100} = 0,55.$$

Bonus:

Descrivi la correlazione nei 3 esempi rappresentati qui sotto:







L'immagine a sinistra rappresenta una correlazione nulla tra X ed Y.

L'immagine centrale rappresenta una correlazione positiva forte tra X e Z.

La terza immagine rappresenta una correlazione negativa moderata tra X e W.