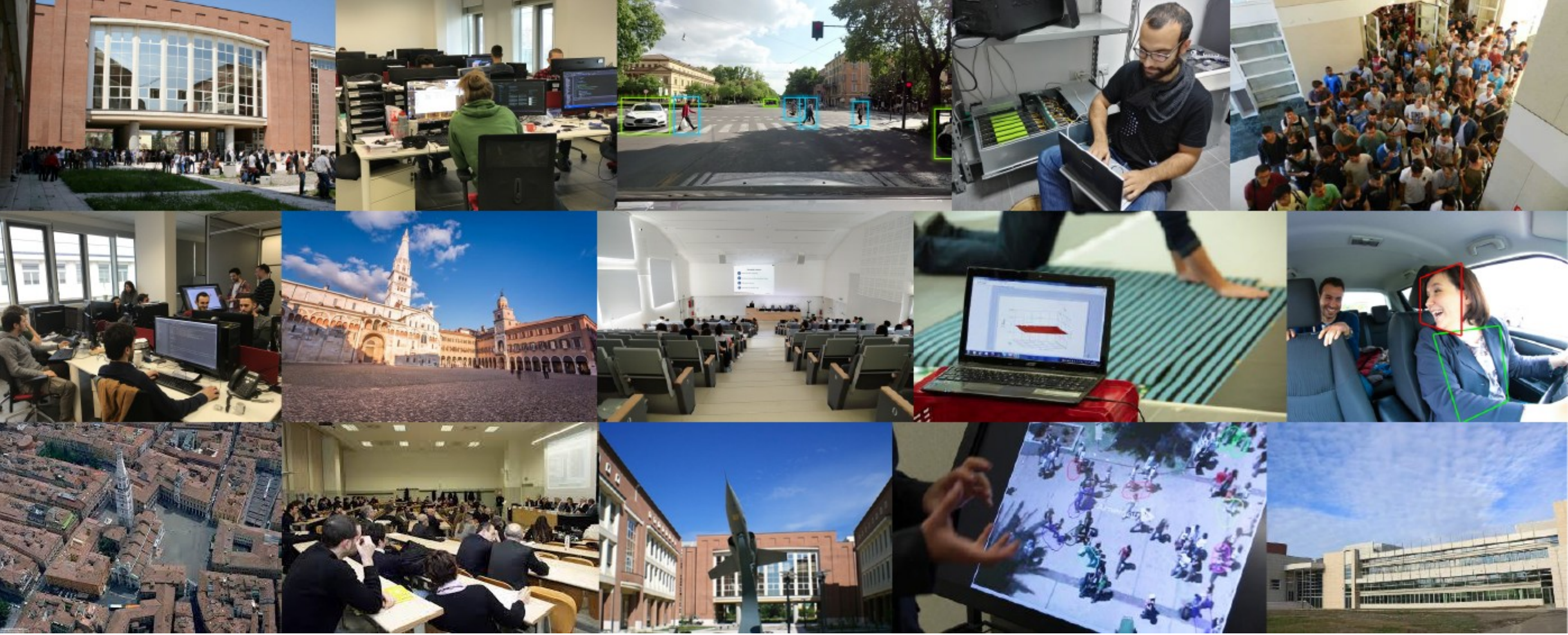


MACHINE LEARNING
Basi Matematiche -

AImage Lab
UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Alcune Basi Matematiche

Faremo una carrellata di concetti matematici probabilmente in gran parte già noti, con lo scopo di ripassare le basi necessarie per le lezioni successive

Gli argomenti **non** verranno trattati in maniera esaustiva

Altri elementi matematici di base verranno rivisti in maniera preliminare ad argomenti specifici di ML

Gli argomenti che tratteremo in questa lezione sono importanti per poter capire il resto del corso e (soprattutto quelli riguardanti il calcolo delle probabilità e il gradiente) possono essere materia d'esame

Elementi di Algebra Lineare

Spazio vettoriale

Definition (over reals)

A set $\mathcal{X}$ is called a *vector space* over $\mathbb{R}$ if addition and scalar multiplication are defined and satisfy for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ and $\lambda, \mu \in \mathbb{R}$:

- Addition:

  associative $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$
  commutative $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$
  identity element $\exists \mathbf{0} \in \mathcal{X} : \mathbf{x} + \mathbf{0} = \mathbf{x}$
  inverse element $\forall \mathbf{x} \in \mathcal{X}\ \exists \mathbf{x}' \in \mathcal{X} : \mathbf{x} + \mathbf{x}' = \mathbf{0}$

- Scalar multiplication:

  distributive over elements $\lambda(\mathbf{x} + \mathbf{y}) = \lambda\mathbf{x} + \lambda\mathbf{y}$
  distributive over scalars $(\lambda + \mu)\mathbf{x} = \lambda\mathbf{x} + \mu\mathbf{x}$
  associative over scalars $\lambda(\mu\mathbf{x}) = (\lambda\mu)\mathbf{x}$
  identity element $\exists 1 \in \mathbb{R} : 1\mathbf{x} = \mathbf{x}$

Prodotto scalare

- “dot product” o “scalar product” o “inner product”

$$\sum_{i=1}^{n} x_i y_i$$

- $\mathbf{x}^{T}\mathbf{y} = \langle \mathbf{x},$ [illegible]

Combinazione lineare di vettori

linear combination given $\lambda_i \in \mathbb{R}, \mathbf{x}_i \in \mathcal{X}$

$$\sum_{i=1}^{n} \lambda_i \mathbf{x}_i$$

Indipendenza lineare

AImage Lab
UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Linear independency

A set of vectors $\mathbf{x}_i$ is *linearly independent* if none of them can be written as a linear combination of the others

Norma vettoriale

Norm

A function $||\cdot|| : \mathcal{X} \rightarrow \mathbb{R}_0^+$ is a *norm* if for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}, \lambda \in \mathbb{R}$:

- $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$
- $||\lambda \mathbf{x}|| = |\lambda| \, ||\mathbf{x}||$
- $||\mathbf{x}|| > 0$ *if* $\mathbf{x} \neq 0$

Norma L_2

$$||\boldsymbol{x}||_2 = \sqrt{\boldsymbol{x}^T \boldsymbol{x}}$$

Norma L_1

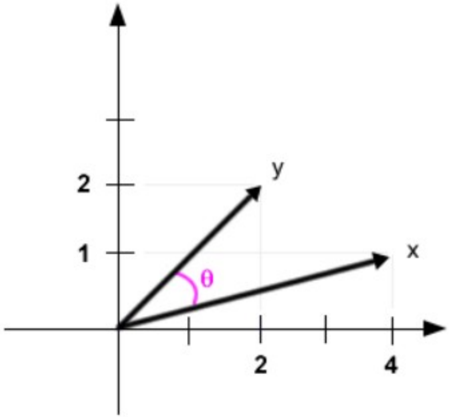
$$||\boldsymbol{x}||_1 = \sum_{i=1}^{n} |\boldsymbol{x}_i|$$

Prodotto scalare: proprietà geometriche

angle The angle θ between two vectors is defined as:

$$cos\theta = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{||\mathbf{x}||\ ||\mathbf{y}||}$$

orthogonal Two vectors are *orthogonal* if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$



Matrici

AImage Lab
UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

$$M \in \mathbb{R}^{m \times n} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \vdots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix}$$

Prodotto tra matrici

Se $X \in \mathbb{R}^{n \times p}$ e $Y \in \mathbb{R}^{p \times m}$, allora $XY = Z \in \mathbb{R}^{n \times m}$ e:
$Z_{i,j} = \sum_k X_{i,k} Y_{k,j}$

Proprietà ed operazioni tra matrici

transpose Matrix obtained exchanging rows with columns (indicated with M^T). Properties:

$$(MN)^T = N^T M^T$$

trace Sum of diagonal elements of a matrix

$$tr(M) = \sum_{i=1}^{n} M_{ii}$$

inverse The matrix which multiplied with the original matrix gives the identity

$$MM^{-1} = I$$

Proprietà ed operazioni tra matrici

Il rango ("rank") di una matrice A $n \times m$ è il massimo numero di righe (o colonne) linearmente indipendenti di A

Funzioni lineari

Una funzione $f(x_1, \ldots, x_n)$ è *lineare* se è un polinomio di grado 1 o 0:

$$f(x_1, \ldots, x_n) = a_1 x_1 + a_2 x_2 \ldots a_n x_n + a_{n+1}$$

Esempio: $f(x,y) = 6.5\,x + y + 12$

Il grado di un polinomio è dato dal termine (monomio) con esponente massimo, dove l'esponente di un termine è la somma degli esponenti delle variabili di quel termine. Esempi:

- $4x^2 + 5y$ ha grado 2
- $x + 3xy$ ha grado 2 (!!)

Per cui $f(x,y) = x + 3xy$ non è una funzione lineare

Anche $f(x,y) = log(x) + 7y$ non è una funzione lineare (perchè non è un polinomio)

x^{-1} non è lineare perchè non è un monomio (gli esponenti di un monomio devono essere numeri naturali)

Sistema di equazioni lineari

Un sistema lineare con *n* incognite (x_1, ..., x_n) e stesso numero (*n*) di equazioni *lineari* può essere espresso come:

$$\begin{aligned} a_{11}x_1 + ... + a_{1n}x_n &= b_1 \\ ... &= ... \\ a_{n1}x_1 + ... + a_{nn}x_n &= b_n \end{aligned}$$

che in forma matriciale diventa: $A\mathbf{x} = \mathbf{b}$

Sistema di equazioni lineari

Se A è invertibile (ovvero il rango di A è n), allora la soluzione unica del sistema è:

$$\boldsymbol{x} = A^{-1}\boldsymbol{b}$$

Elementi di Calcolo delle Probabilità e Statistica

Variabili aleatorie discrete

Probability mass function

Given a discrete random variable X taking values in $\mathcal{X} = \{v_1, \ldots, v_m\}$, its *probability mass function* $P : \mathcal{X} \rightarrow [0, 1]$ is defined as:

$$P(v_i) = \Pr[X = v_i]$$

and satisfies the following conditions:

- $P(x) \geq 0$
- $\sum_{x \in \mathcal{X}} P(x) = 1$

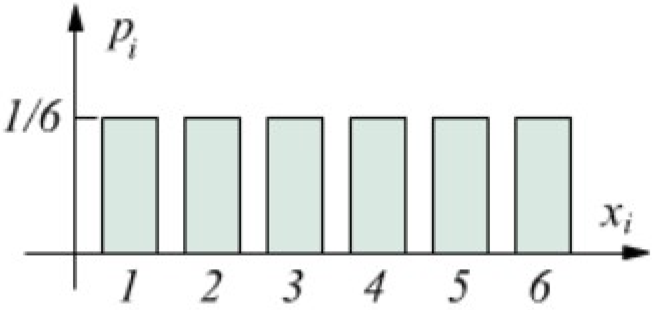
Distribuzione di probabilità discreta

$$P(A) = \sum_{x \in A} P(x), A \in 2^{\mathcal{X}}$$

Nella terminologia del ML spesso si usa il termine "distribuzione di probabilità discreta" anche quando sarebbe più appropriato parlare di "funzione di massa di probabilità"

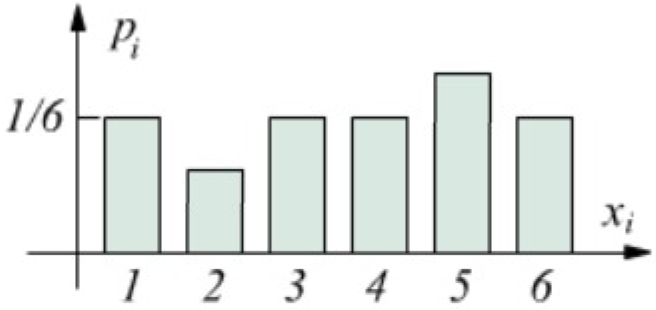
Ciò avviene anche perché raramente saremo interessati a stimare *P(A)* per insiemi di valori *A* che abbiano più di un elemento...

Esempio: distribuzione uniforme



Nel caso di un dado la distribuzione di probabilità discreta dei suoi eventi singoli è uniforme

Esempio: distribuzione non uniforme



Distribuzione di probabilità di un dado truccato

Media e varianza (caso discreto)

Expected value

- The *expected value*, *mean* or *average* of a random variable x is:

$$\mathrm{E}[x] = \mu = \sum_{x \in \mathcal{X}} xP(x) = \sum_{i=1}^{m} v_i P(v_i)$$

- The *expectation* operator is linear:

$$\mathrm{E}[\lambda x + \lambda' y] = \lambda \mathrm{E}[x] + \lambda' \mathrm{E}[y]$$

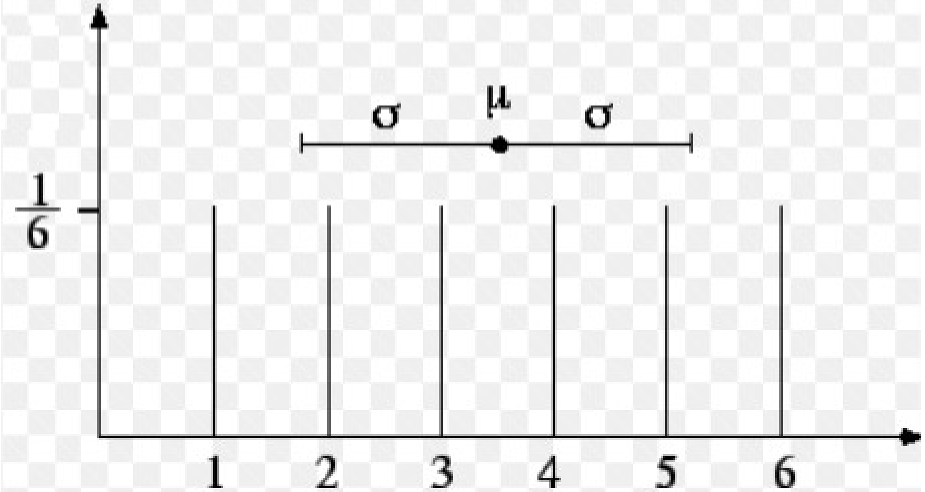
Variance

- The *variance* of a random variable is the moment of inertia of its probability mass function:

$$\mathrm{Var}[x] = \sigma^2 = \mathrm{E}[(x-\mu)^2] = \sum_{x \in \mathcal{X}} (x-\mu)^2 P(x)$$

- The *standard deviation* σ indicates the typical amount of deviation from the mean one should expect for a randomly drawn value for x.

Esempio: dado non truccato



Media e deviazione standard di un dado non truccato

Coppia di variabili aleatorie discrete

AImage Lab
UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Probability mass function

Given a pair of discrete random variables X and Y taking values $\mathcal{X} = \{v_1, \ldots, v_m\}$ $\mathcal{Y} = \{w_1, \ldots, w_n\}$, the *joint probability mass function* is defined as:

$$P(v_i, w_j) = \Pr[X = v_i, Y = w_j]$$

with properties:

- $P(x, y) \geq 0$
- $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) = 1$

Esempio: probabilità congiunta del lancio di un dado e di una moneta

	1	2	3	4	5	6
Heads						
Tails						

Variabili aleatorie continue

Distribuzione di probabilità continua:

$$P(A) = \int_a^b f(x)dx,\ A = [a, b] \in \mathbb{R}$$

f(x) è detta funzione di densità di probabilità:

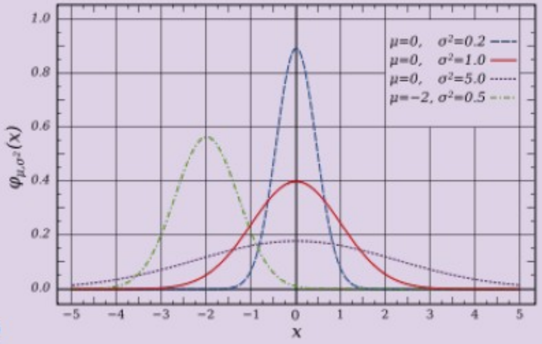
$$f(x) \geq 0,$$
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Distribuzione Gaussiana (o “Normale”)

- Bell-shaped curve.
- Parameters: μ mean, σ^2 variance.
- Probability density function:

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{(x-\mu)^2}{2\sigma^2}$$

- $\mathrm{E}[x] = \mu$
- $\mathrm{Var}[x] = \sigma^2$

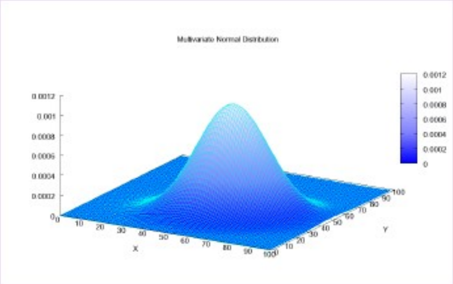


Distribuzione Gaussiana

- Si indica con $\mathcal{N}(\mu, \sigma^2)$
- Distribuzione Normale Standard: $\mathcal{N}(0, 1)$
- In statistica è molto importante perchè descrive una serie di fenomeni fisici: la somma di processi stocastici indipendenti
- Ad esempio, il risultato della somma di errori indipendenti può essere descritto tramite una Gaussiana

Distribuzione Gaussiana multidimensionale (o "multivariata")

- normal distribution for d-dimensional vectorial data.
- Parameters: μ mean vector, Σ covariance matrix.
- Probability density function:



$$p(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp -\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)$$

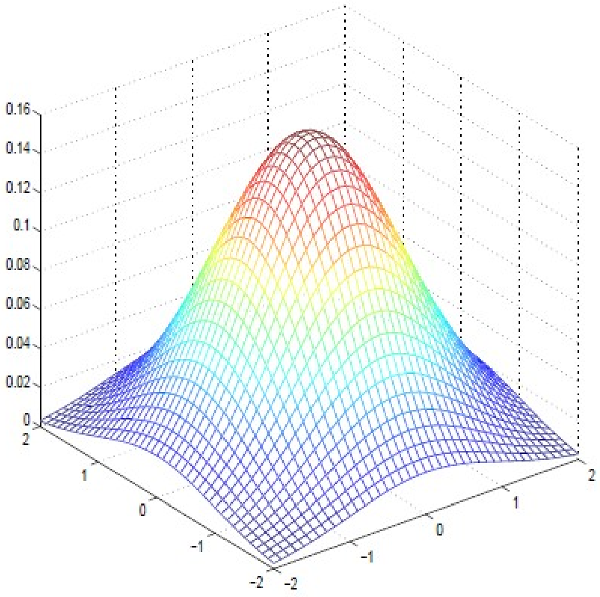
- $\mathrm{E}[x] = \mu$
- $\mathrm{Var}[x] = \Sigma$

Matrice di Covarianza

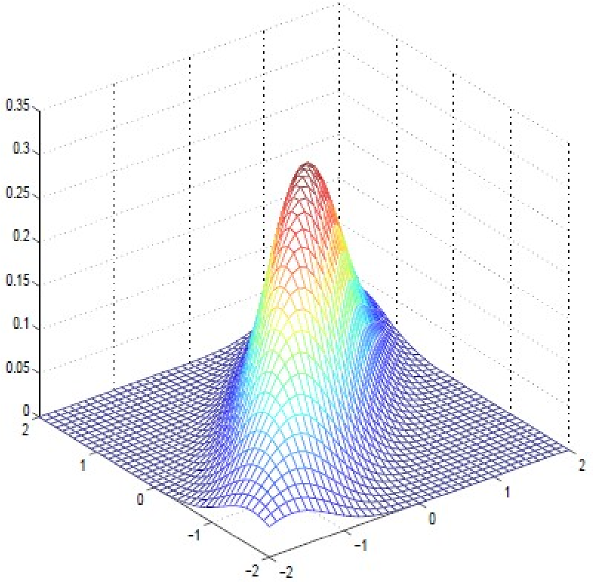
$$\text{cov}(\mathbf{X}) = \begin{bmatrix} \sigma(X_1)^2 & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \sigma(X_2)^2 & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \sigma(X_n)^2 \end{bmatrix}$$

Eviteremo di definirla esattamente nel caso continuo perché è più utile la sua versione con variabili discrete, che è molto simile alla matrice di correlazione (la vedremo tra poco...)

Distribuzione Gaussiana multivariata: Esempio 1



Distribuzione Gaussiana multivariata: Esempio 2



Probabilità condizionale

AImage Lab
UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

conditional probability probability of x once y is observed

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

statistical independence variables X and Y are statistical independent iff

$$P(x,y) = P(x)P(y)$$

implying:

$$P(x|y) = P(x) \qquad P(y|x) = P(y)$$

Teorema di Bayes

product rule conditional probability definition implies that

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

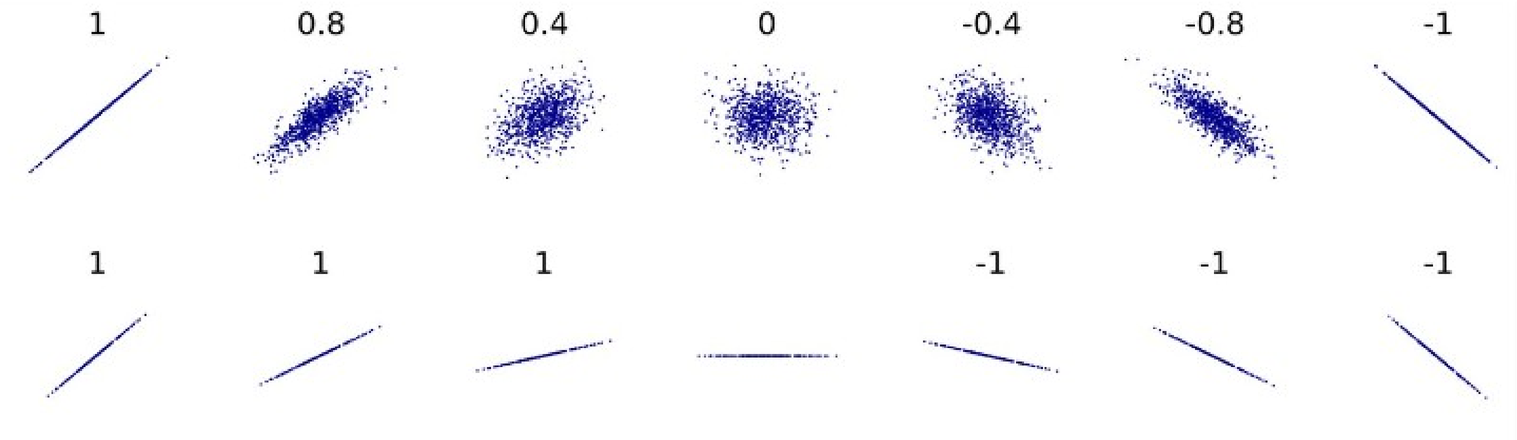
Bayes' rule

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

$$posterior = \frac{likelihood \times prior}{evidence}$$

Correlazione

La correlazione tra due variabili aleatorie X ed Y esprime il loro grado di dipendenza *lineare*



Correlazione: definizione

$$-1 \leq \rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \leq +1$$

Caso *X, Y* discrete:

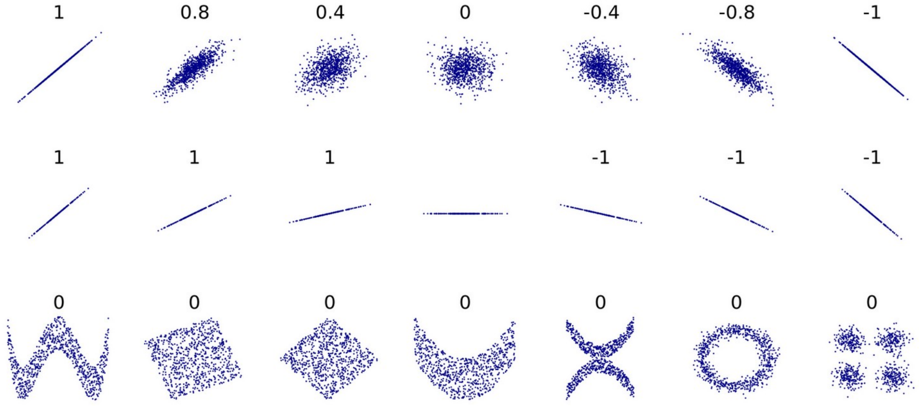
$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}}$$

Correlazione: significato intuitivo

- Correlazione positiva: le variazioni di X (rispetto al suo valore medio) corrispondono in maniera diretta alle variazioni di Y. Per esempio, c'è una correlazione positiva tra l'altezza e il peso delle persone.
- Correlazione negativa: alle variazioni di X corrispondono, in senso contrario, variazioni di Y. Ad esempio, ad una maggior produzione di grano corrisponde un prezzo minore.

Correlazione e dipendenza statistica

Attenzione: la dipendenza *lineare* è solo un caso particolare di dipendenza statistica!



$$X, Y \text{ independent} \quad \Rightarrow \quad \rho_{X,Y} = 0 \quad (X, Y \text{ uncorrelated})$$
$$\rho_{X,Y} = 0 \quad (X, Y \text{ uncorrelated}) \quad \not\Rightarrow \quad X, Y \text{ independent}$$

Matrice di Correlazione

Dato un vettore di *n* variabili ale $\mathbf{X} = (X_1, X_2, \ldots, X_n)^{\mathrm{T}}$

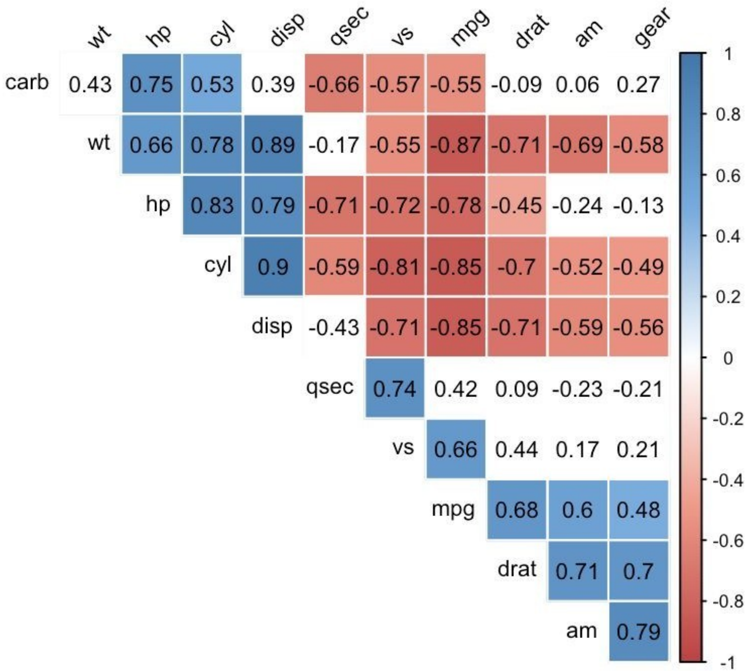
la matrice di correlazione tra tutte le coppie di tali variabili è definita come segue:

$$\operatorname{corr}(\mathbf{X}) = \begin{bmatrix} 1 & \frac{\mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)]}{\sigma(X_1)\sigma(X_2)} & \cdots & \frac{\mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)]}{\sigma(X_1)\sigma(X_n)} \\ \frac{\mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)]}{\sigma(X_2)\sigma(X_1)} & 1 & \cdots & \frac{\mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)]}{\sigma(X_2)\sigma(X_n)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)]}{\sigma(X_n)\sigma(X_1)} & \frac{\mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)]}{\sigma(X_n)\sigma(X_2)} & \cdots & 1 \end{bmatrix}$$

Esempio

- Le matrici di correlazione possono essere usate, ad esempio, per trovare coppie di feature che sono altamente correlate tra di loro o con le variabili dipendenti

- Possono essere usate per decidere se scartare o mantenere gruppi di features (lo vedremo meglio in seguito)



Gradiente di funzioni a più variabili

Gradiente

Data una funzione *f: R^n -> R*, il suo gradiente è definito da:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

dove $\frac{\partial f}{\partial x_i}$ è la derivate parziale di f rispetto a x_i

Esempio

Data la funzione $f(\mathbf{x}) = f(x,y) = 3x^2 + y^2 + 2x + 7$
il suo gradiente è:

$$\nabla f = \begin{bmatrix} 6x + 2 \\ 2y \end{bmatrix}$$

Gradiente

Data una funzione *f*: R^n -> R,
il gradiente, calcolato rispetto al vettore di coordinate $p = (x_1, ..., x_n)$,
è dato da:

$$\nabla f(p) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(p) \\ \vdots \\ \frac{\partial f}{\partial x_n}(p) \end{bmatrix}$$

Esempio

Data la funzione $f(\mathbf{x}) = f(x,y) = 3x^2 + y^2 + 2x + 7$
Il cui gradiente è:

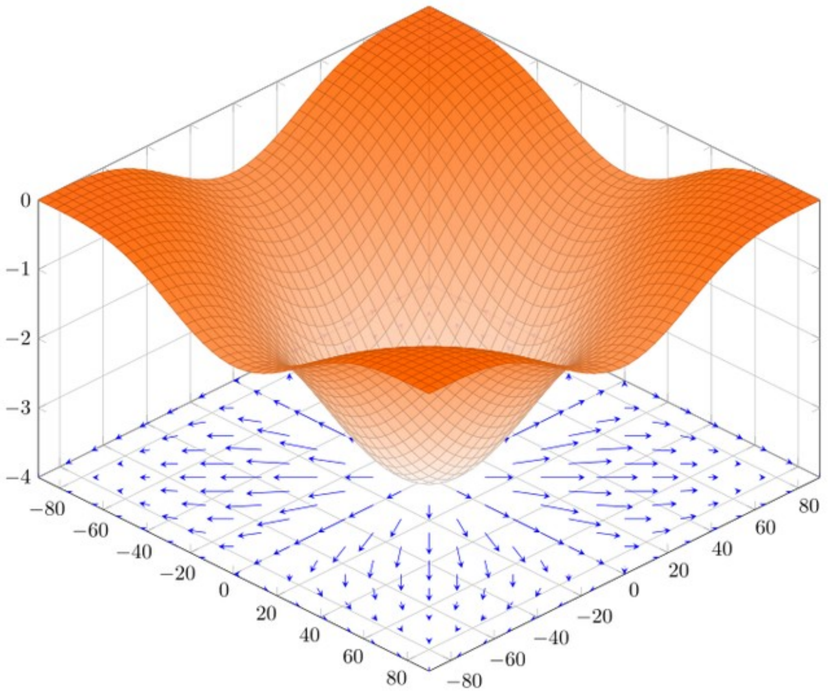
$$\nabla f = \left[\begin{array}{l} 6x + 2 \\ 2y \end{array} \right]$$

Posso ad esempio calcolare:

$$\nabla f(1,1) = \left[\begin{array}{c} 8 \\ 2 \end{array} \right]$$

Proprietà geometriche del gradiente

- Il gradiente calcolato in *p* corrisponde alla direzione di *massima crescita locale* di *f* in *p*



Proprietà geometriche del gradiente

- Il gradiente calcolato in *p* corrisponde alla direzione di *massima crescita locale* di *f* in *p*
- I valori delle derivate parziali in *p* rappresentano le componenti di questo vettore rispetto