



Università  
degli Studi  
di Palermo



# Richiami di Teoria della Probabilità

CORSO DI BIG DATA – MODULO ANALISI PER I BIG DATA  
a.a. 2023/2024

Prof. Roberto Pirrone



# Sommario

- Probabilità di variabili discrete
- Probabilità marginale, condizionale e congiunta di variabili discrete
- Variabili continue
- Densità di probabilità e distribuzione di probabilità
- Probabilità marginale, condizionale e congiunta di variabili continue
- Indipendenza statistica e condizionale
- Media varianza e covarianza
- Correlazione
- Regola di Bayes
- Principali distribuzioni di probabilità
- Cenni di teoria dell'informazione

# Probabilità di variabili discrete

- Una variabile casuale discreta  $X$  assume valori casuali appartenenti ad un insieme discreto  $\mathcal{X} = \{x_i, i = 1, 2, \dots\}$
- Se, su  $N$  osservazioni di  $X$ , registriamo che  $c_i$  volte  $X=x_i$  :

$$p(X = x_i) = \frac{c_i}{N}$$

$$\forall x_i, 0 \leq p(X = x_i) \leq 1$$

$$\sum_{x_i} p(X = x_i) = 1$$

# Probabilità di variabili discrete

- La legge che esprime la probabilità  $P(x) = p(X = x)$ ,  $x \in \mathcal{X}$  per una certa variabile casuale discreta  $X$ , si chiama *distribuzione discreta di probabilità*
- Formalmente la distribuzione discreta di probabilità è realizzata da una *funzione massa* (Probability Mass Function) che esprime il fatto che una certa probabilità è *concentrata* su ogni elemento di  $\mathcal{X}$ .

# Probabilità marginale

- Sia  $n_{ij}$  il numero di osservazioni di un'altra variabile  $Y=y_j$  quando  $X=x_i$ , allora:

$$c_i = \sum_j n_{ij}$$

$$p(X = x_i) = \sum_j p(X = x_i, Y = y_j)$$

*Regola della somma*

# Probabilità condizionale e congiunta

- La probabilità che  $Y=y_j$  posto che  $X=x_i$   $p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$
- Calcoliamo la probabilità congiunta che  $Y=y_j$  e  $X=x_i$

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} * \frac{c_i}{N} =$$
$$p(Y = y_j | X = x_i) * p(X = x_i)$$

*Regola del prodotto*

# Variabili continue

- Se  $x$  è una variabile continua in  $\mathbb{R}$ , le sommatorie divengono integrali
- La probabilità che  $x \in [a, b]$  è:

$$p(x \in [a, b]) = \int_a^b p(x) dx$$

# Densità di probabilità e distribuzione di probabilità

- La legge che esprime la probabilità  $P(x) = p(x = x)$ ,  $x \in \mathbb{R}$  per una certa variabile casuale continua  $x$ , si chiama *distribuzione di probabilità* di cui la funzione densità di probabilità è la realizzazione

N.B.  $x \rightarrow$  la variabile,  $x \rightarrow$  il singolo valore che essa può assumere

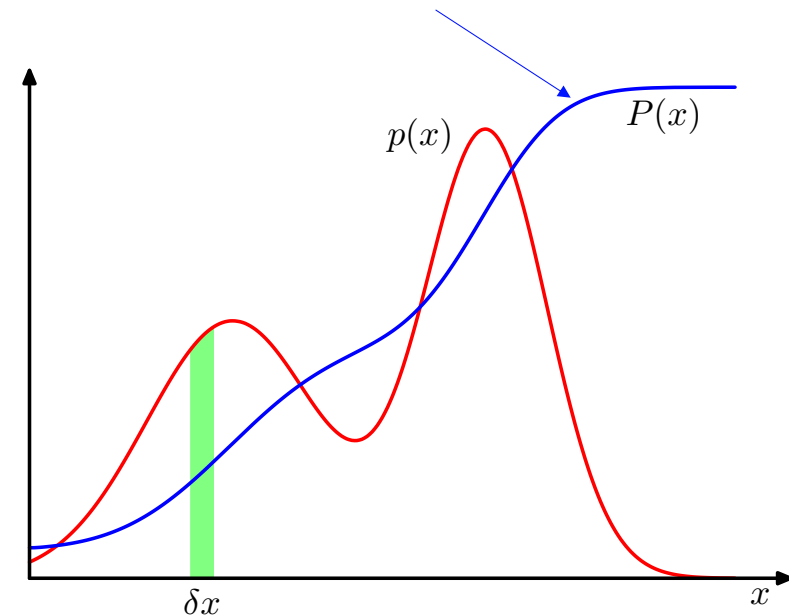


# Densità di probabilità e distribuzione di probabilità

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

$$P(z) = \int_{-\infty}^z p(x) dx \equiv p(x \leq z)$$

Distribuzione cumulativa di probabilità



# Probabilità marginale e condizionale di variabili continue

$$p(x) = \int p(x, y) dy$$

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

$$P(x^{(1)}, \dots, x^{(n)}) = P(x^{(1)}) \prod_{i=2}^n P(x^{(i)} | x^{(1)}, \dots, x^{(i-1)})$$

# Indipendenza statistica e condizionale

- Indipendenza statistica

$$\forall x \in \mathbf{x}, y \in \mathbf{y}, p(\mathbf{x} = x, \mathbf{y} = y) = p(\mathbf{x} = x)p(\mathbf{y} = y)$$

- Indipendenza condizionale

$$\begin{aligned} \forall x \in \mathbf{x}, y \in \mathbf{y}, z \in \mathbf{z}, p(\mathbf{x} = x, \mathbf{y} = y | \mathbf{z} = z) \\ = p(\mathbf{x} = x | \mathbf{z} = z)p(\mathbf{y} = y | \mathbf{z} = z) \end{aligned}$$

# Media varianza e covarianza

- La media o valore atteso, è un operatore lineare

Se  $x$  è una variabile aleatoria  
qualunque variabile  $y = f(x)$  è ancora  
Una variabile aleatoria!!

$$\mathbb{E}_{x \sim P}[f(x)] = \sum_x P(x) f(x)$$

$$\mathbb{E}_{x \sim p}[f(x)] = \int p(x) f(x) dx$$

$$\mathbb{E}_x[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_x[f(x)] + \beta \mathbb{E}_x[g(x)]$$

# Media varianza e covarianza

- La varianza è il valore atteso della differenza tra  $f(x)$  ed il quadrato del suo valore atteso

$$\text{Var}(f(x)) = \mathbb{E} [(f(x) - \mathbb{E}[f(x)])^2] = \mathbb{E} [f(x)^2] - \mathbb{E} [f(x)]^2$$

# Media varianza e covarianza

- Proprietà della varianza (sotto ipotesi di indipendenza statistica)

$$\begin{aligned}\text{Var}(\alpha \cdot f(x)) &= \mathbb{E} [\alpha^2 \cdot f(x)^2] - \mathbb{E} [\alpha \cdot f(x)]^2 = \\ &\alpha^2 \cdot \mathbb{E} [f(x)^2] - \alpha^2 \cdot \mathbb{E} [f(x)]^2 = \\ &\alpha^2 \cdot \left( \mathbb{E} [f(x)^2] - \mathbb{E} [f(x)]^2 \right) \equiv \alpha^2 \cdot \text{Var}(f(x))\end{aligned}$$

# Media varianza e covarianza

- Proprietà della varianza (sotto ipotesi di indipendenza statistica)

$$\begin{aligned}\text{Var}(f(x) + g(y)) &= \mathbb{E} [(f(x) + g(y))^2] - \mathbb{E} [(f(x) + g(y))]^2 = \\ &= \text{Var}(f(x)) + \text{Var}(g(y))\end{aligned}$$

# Media varianza e covarianza

- La covarianza esprime quanto due variabili aleatorie siano legate linearmente l'una all'altra

$$\text{Cov}(f(x), g(y)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])] = \\ \mathbb{E}[f(x)g(y)] - \mathbb{E}[f(x)]\mathbb{E}[g(y)]$$



# Media varianza e covarianza

- La covarianza tra due vettori casuali *colonna*  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  (ogni componente è una variabile aleatoria) è una *matrice*

$$\begin{aligned}\text{Cov}(\mathbf{x}, \mathbf{y}) &= \mathbb{E} \left( \{ \mathbf{x} - \mathbb{E}[\mathbf{x}] \} \{ \mathbf{y}^T - \mathbb{E}[\mathbf{y}^T] \} \right) \\ &= \mathbb{E} [\mathbf{x} \mathbf{y}^T] - \mathbb{E}[\mathbf{x}] \mathbb{E} [\mathbf{y}^T] \in \mathbb{R}^{n \times n}\end{aligned}$$

- La covarianza di  $\mathbf{x}$  con sé stesso è:

$$\begin{aligned}\text{Cov}(\mathbf{x}, \mathbf{x}) &= \mathbb{E} \left( \{ \mathbf{x} - \mathbb{E}[\mathbf{x}] \} \{ \mathbf{x}^T - \mathbb{E}[\mathbf{x}^T] \} \right) \\ &= \mathbb{E} [\mathbf{x}^2] - \mathbb{E}[\mathbf{x}] \mathbb{E} [\mathbf{x}^T] = \mathbb{E} [\mathbf{x}^2] - \mathbb{E} [\mathbf{x}]^2 \equiv \text{Var}(\mathbf{x})\end{aligned}$$

# Media varianza e covarianza

- Matrice di covarianza di una data set  $D$  di  $n$  record  $\mathbf{x}_k = \{x_k^1, \dots, x_k^d\}$ ,  $k = 1, \dots, n$

$$C = \frac{D^T D}{n} - \bar{\mu}^T \bar{\mu}, \quad \bar{\mu} = \{ \mathbb{E} [x_k^1], \dots, \mathbb{E} [x_k^d] \}$$

- I suoi elementi si possono esprimere come

$$c_{ij} = \frac{\sum_{k=1}^n x_k^i x_k^j}{n} - \mu_i \mu_j \quad \forall i, j \in \{1 \dots d\}$$

$$c_{ii} = \frac{\sum_{k=1}^n x_k^{i^2}}{n} - \mu_i^2 \equiv \sigma_i^2$$

# Correlazione

- La matrice di covarianza di una data set  $D$  esprime come le varie componenti dei campioni  $\mathbf{x}_k$  (le singole *feature*) siano più o meno legate linearmente tra loro.
- E' una matrice semidefinita positiva
- Esiste anche un altro indice, derivato dalla covarianza, che viene spesso utilizzato per indicare quanto due feature siano collegate ed è il *coefficiente di correlazione* o *indice di correlazione di Pearson*



# Correlazione

- Si consideri il caso di due variabili aleatorie  $X$  e  $Y$

$$\Sigma_{X,Y} = \begin{bmatrix} \sigma_X^2 & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \sigma_Y^2 \end{bmatrix}$$

$$\det \Sigma_{X,Y} = \sigma_X^2 \sigma_Y^2 - \text{Cov}(X, Y)^2 \geq 0$$

- Il coefficiente di correlazione si definisce come

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \in [-1, 1]$$

*Indica il segno della correlazione!!*



Università  
degli Studi  
di Palermo

**dj** dipartimento  
di ingegneria  
unipa



# Teorema di Bayes

- Dalla definizione di probabilità condizionale, ci rendiamo conto che è possibile esprimere la probabilità congiunta  $P(x,y)$  di due variabili correlate sia a partire da  $P(x|y)$  sia a partire da  $P(y|x)$

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

# Teorema di Bayes

- Assumiamo di avere un dataset  $\mathcal{D}$  dal quale vogliamo apprendere un modello descritto da un vettore di parametri  $\mathbf{w}$
- Avremo appreso il miglior modello quando avremo massimizzato la «probabilità a posteriori» o *posterior*  $P(\mathbf{w} | \mathcal{D})$ 
  - Ciò che possiamo stimare *dopo* aver osservato i dati

# Teorema di Bayes

- Il teorema di Bayes correla il *posterior* con:
  - la conoscenza a priori o *prior* sul nostro modello:  $P(\mathbf{w})$
  - la verosimiglianza o *likelihood* del nostro modello espresso dalla probabilità di predire effettivamente  $\mathcal{D}$  usando i parametri del modello  $\mathbf{w}$ :  $P(\mathcal{D}|\mathbf{w})$

$$P(\mathbf{w}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathbf{w})P(\mathbf{w})}{P(\mathcal{D})}$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

# Teorema di Bayes

- La conoscenza a priori di  $P(\mathcal{D})$  può essere marginalizzata ed espressa in termini del numeratore:

$$P(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w}$$

- Non è necessaria:
  - È espressa in termini del numeratore
  - È da ritenersi una costante perché è l'evidenza del data set a prescindere dalla scelta dei parametri



# Prinmcipali distribuzioni di probabilità

- Diverse tipologie di dati
  - Multidimensionali
  - Categorici
  - Serie discrete
  - ...
- *Diverse tipologie di distribuzioni di probabilità che li sottendono*
- Dobbiamo conoscerle perché il Machine Learning è il processo di stima di tali distribuzioni di probabilità a partire dai dati che il modello osserva



Università  
degli Studi  
di Palermo



dipartimento  
di ingegneria  
unipa



# Principali distribuzioni di probabilità

- Distribuzione di Bernoulli (variabili binarie)

$$\begin{aligned}P(x = 1) &= \phi \\P(x = 0) &= 1 - \phi \\P(x = x) &= \phi^x (1 - \phi)^{1-x} \\E_x[x] &= \phi \\Var_x(x) &= \phi(1 - \phi)\end{aligned}$$

$$\text{Bin}(m|N, \phi) = \binom{N}{m} \phi^m (1 - \phi)^{N-m}$$

$m$ : numero di osservazioni di  $x=1$  su  $N$  tentativi

# Principali distribuzioni di probabilità

- Distribuzione Multinoulli (distribuzione categorica) in cui  $x$  è una variabile continua che può assumere uno tra  $K$  stati diversi

$$\mathbf{x} = (0_1, 0_2, \dots, 1_k, \dots, 0_K)^T \quad \text{One-hot encoding}$$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}$$

$$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)^T, \quad \mu_k = p(x_k = 1)$$

# Principali distribuzioni di probabilità

- Distribuzione multinomiale (generalizza la binomiale) per variabili discrete
  - date  $N$  osservazioni descrive la probabilità di osservare  $m_k$  volte lo stato  $x_k=1$  da una distribuzione Multinoulli con media  $\boldsymbol{\mu}$

$$\text{Mult}(m_1, m_2, \dots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \dots m_K} \prod_{k=1}^K \mu_k^{m_k}$$

# Principali distribuzioni di probabilità

- Distribuzione esponenziale e distribuzione di Laplace
  - Si utilizzano quando si vuole concentrare la probabilità rispettivamente vicino a 0 ovvero ad un valore dato

$$p(x; \lambda) = \lambda \mathbf{1}_{x \geq 0} \exp(-\lambda x)$$

$$\text{Laplace}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x - \mu|}{\gamma}\right)$$

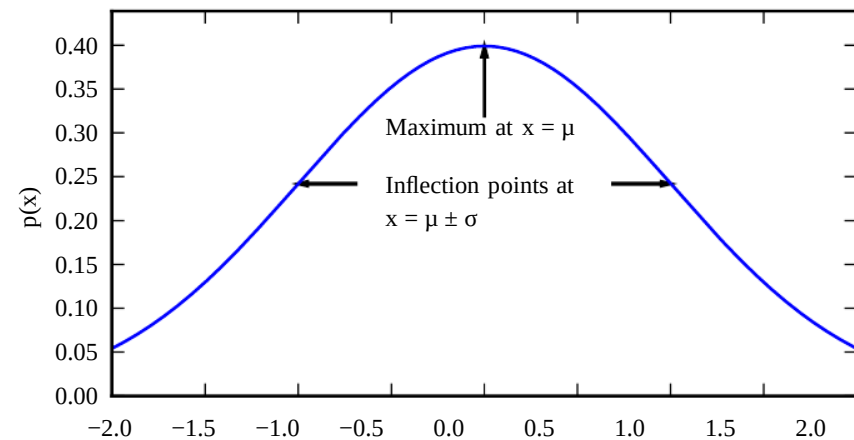
# Principali distribuzioni di probabilità

- Distribuzione empirica
  - Rappresenta la distribuzione di una variabile continua di cui si osservano  $m$  campioni che sono ovviamente equiprobabili

$$\hat{p}(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m \delta(\mathbf{x} - \mathbf{x}^{(i)})$$

# Principali distribuzioni di probabilità

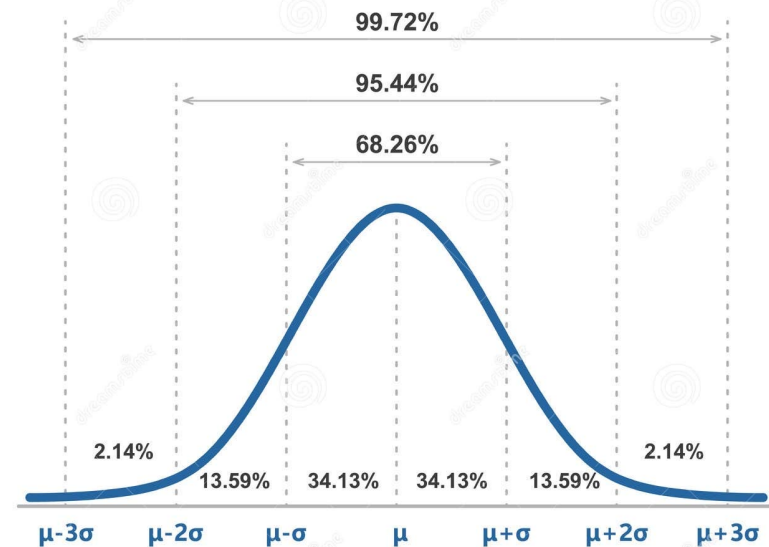
- Distribuzione gaussiana  
o «Normale»



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

# Principali distribuzioni di probabilità

- Distribuzione gaussiana  
o «Normale»



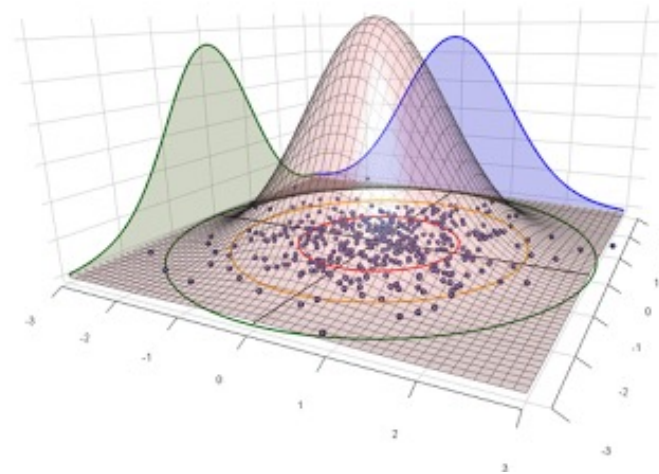
$$\mathcal{N}(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$



# Principali distribuzioni di probabilità

- Distribuzione gaussiana  
o «Normale» multivariata

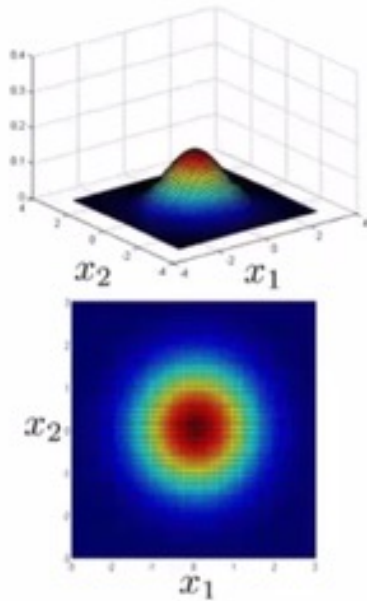
$$\mathbf{x} \in \mathbb{R}^n$$



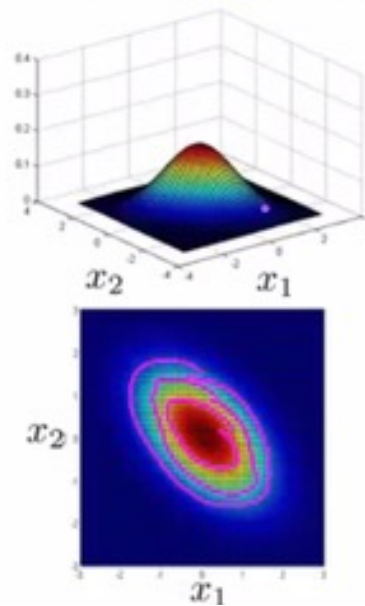
$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

# Principali distribuzioni di probabilità

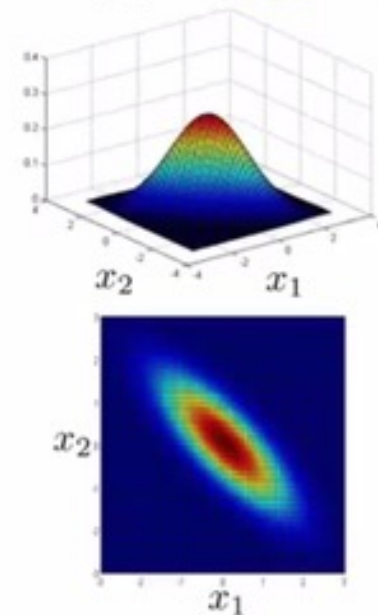
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$



# Mixture di distribuzioni

- Rappresentazione di una distribuzione di probabilità complessa e non nota attraverso un insieme di più *distribuzioni componenti*

$$P(\mathbf{x}) = \sum_i P(c = i) P(\mathbf{x} | c = i)$$

- $P(c)$  è la distribuzione multinoulli di appartenenza all' $i$ -esima distribuzione componente
- $c$  : *variabile latente*  $\rightarrow$  una variabile randomica correlata al processo e non direttamente osservabile
- Mistura di gaussiane: *approssimatore universale* di distribuzioni

# Cenni di teoria dell'informazione

- Intuitivamente il contenuto informativo di un evento raro è maggiore del contenuto informativo di un evento altamente probabile
- Due eventi indipendenti hanno informazione che si somma

$$I(x = x) = -\log P(x)$$

- $\log_2 \rightarrow \textit{bit}$      $\log_e \rightarrow \textit{nat}$

# Cenni di teoria dell'informazione

- Entropia di Shannon
  - Misura l'informazione per un'intera distribuzione come valore atteso su di essa

$$H(x) = \mathbb{E}_{x \sim P} [I(x)] = -\mathbb{E}_{x \sim P} [\log P(x)] \triangleq H(P)$$

# Cenni di teoria dell'informazione

- Divergenza di Kullback-Leibler
  - Quanto due distribuzioni  $P(x)$  e  $Q(x)$  sono dissimili → *non è una distanza!*

$$D_{KL}(P \parallel Q) = \mathbb{E}_{x \sim P} \left[ \log \frac{P(x)}{Q(x)} \right] = \\ \mathbb{E}_{x \sim P} [\log P(x) - \log Q(x)]$$

$$D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$$

# Cenni di teoria dell'informazione

- Cross-entropia
  - Misura l'entropia mutua di P e Q
  - Ha un significato analogo alla  $D_{KL}$

$$\begin{aligned} H(P, Q) &= H(P) + D_{KL}(P \parallel Q) = -\mathbb{E}_{x \sim P} [\log Q(x)] \\ &= - \sum_{x \sim P} P(x) \log Q(x) \end{aligned}$$

# Cenni di teoria dell'informazione

- Mutua Informazione

- Date due variabili aleatorie  $X$  e  $Y$  in qualche modo correlate, aventi distribuzioni rispettivamente  $P_X$  e  $P_Y$ , la  $MI$  misura l'ammontare di informazione che ottengo su una variabile se osservo l'altra
- Si definisce come la  $D_{KL}$  tra la distribuzione congiunta ed il prodotto delle due distribuzioni nel senso dell'indipendenza statistica

$$MI(X, Y) = D_{KL}(P_{X,Y} \parallel P_X \otimes P_Y)$$