Richiami di Teoria della Probabilità

Corso di Big Data – Modulo Analisi per i Big Data

a.a. 2022/2023

Prof. Roberto Pirrone

Laboratorio di Interazione Uomo-Macchina
CHILAB

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 $\mathcal{X} = \{x_i, \ i = 1, 2, \ldots\}$

- Se, su *N* osservazioni di *X*, registriamo che c_i volte $X = x_i$:

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

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

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

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 $p\left(Y = y_j | X = x_i\right) = \frac{n_{ij}}{c_i}$

- Calcoliamo la probabilità congiunta che Y=y_j e X=x_i

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

$$p\left(Y = y_j | X = x_i\right) * p\left(X = x_i\right)$$

Regola del prodotto

Variabili continue

- Se *x* è una variabile continua in $\mathbb{R}$, le sommatorie divengono integrali

- La probabilità che x ∈ [*a*,*b*] è:

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

Densità di probabilità e distribuzione di probabilità

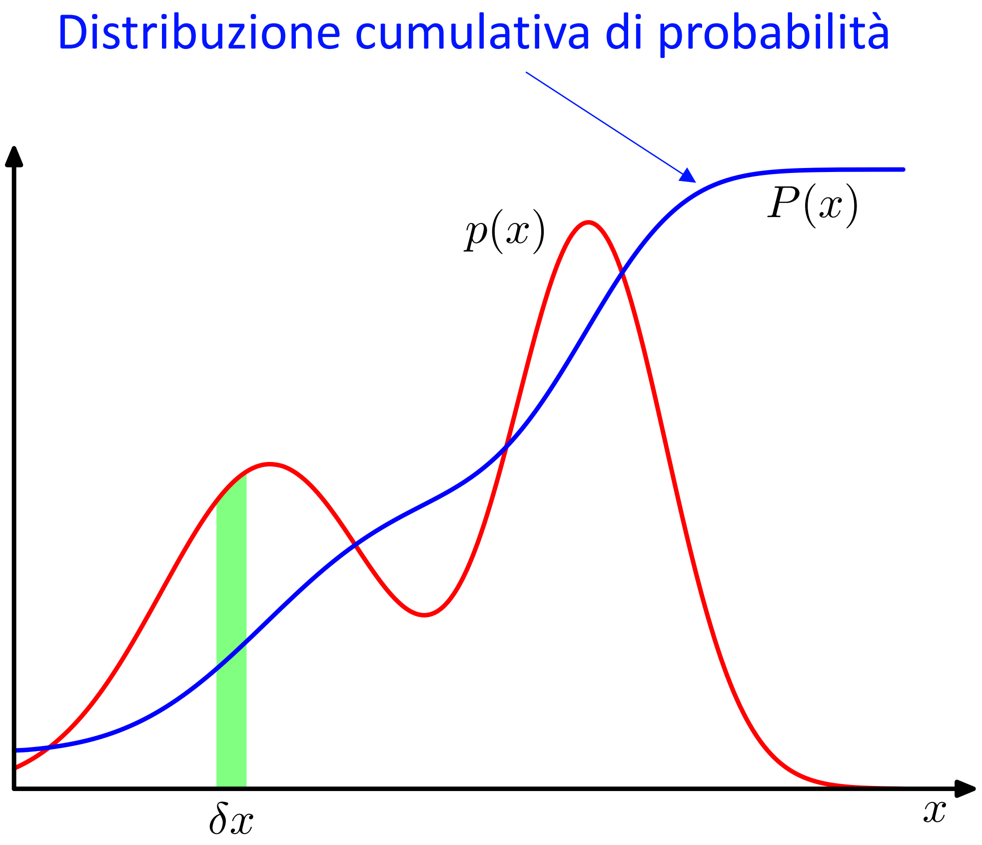
- La legge che esprime la probabilità $P(x) = p$ (x = x), $x \in \mathbb{R}$ per una certa variabile casuale continua x, si chiama *distribuzione di probabilità* di cui la funzione densità di probabilità è la realizzazione

N.B. x → la variabile, *x* → il singolo valore che essa può assumere

Densità di probabilità e distribuzione di probabilità

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

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



Probabilità marginale e condizionale di variabili continue

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

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

$$P\left(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}\right) = P\left(\mathrm{x}^{(1)}\right) \Pi_{i=2}^{n} P\left(\mathrm{x}^{(i)} | \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)}\right)$$

Indipendenza statistica e condizionale

- Indipendenza statistica

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

- Indipendenza condizionale

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

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!!

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

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

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

Media varianza e covarianza

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

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

Media varianza e covarianza

- Proprietà della varianza (sotto ipotesi di indipendenza statistica)

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

Media varianza e covarianza

- Proprietà della varianza (sotto ipotesi di indipendenza statistica)

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

Media varianza e covarianza

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

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

Media varianza e covarianza

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

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

- La covarianza di **x** con sé stesso è:

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

Media varianza e covarianza

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

$$C = \frac{D^T D}{n} - \bar{\mu}^T \bar{\mu}, \;\; \bar{\mu} = \left\{\mathbb{E}\left[x_k^1\right], \ldots, \mathbb{E}\left[x_k^d\right]\right\}$$

- I suoi elementi si possono esprimere come

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

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

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*

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

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

- Il coefficiente di correlazione si definisce come

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

Usate $a^2 - b^2 = (a+b)(a-b)$ sulla formula del determinante

Indica il segno della correlazione!!

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)

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

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* (ciò che possiamo stimare *dopo* aver osservato i dati) $P(\mathbf{w}|\mathcal{D})$

Teorema di Bayes

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

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

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

Teorema di Bayes

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

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

- In genere è da ritenersi una costante perché è l'evidenza del data set a prescindere dalla scelta dei parametri

Principali distribuzioni di probabilità

- Distribuzione di Bernoulli (variabili binarie)

$$P(\mathrm{x}=1)=\phi$$
$$P(\mathrm{x}=0)=1-\phi$$
$$P(\mathrm{x}=x)=\phi^x(1-\phi)^{1-x}$$
$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}]=\phi$$
$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x})=\phi(1-\phi)$$

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

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

Principali distribuzioni di probabilità

- Distribuzione Multinoulli (distribuzione categorica) x è una variabile che può assumere uno tra *K* stati diversi

$$\mathrm{x} = (0_1, 0_2, \ldots, 1_k, \ldots, 0_K)^T$$

One-hot encoding

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

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

Principali distribuzioni di probabilità

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

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

Principali distribuzioni di probabilità

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

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

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

Principali distribuzioni di probabilità

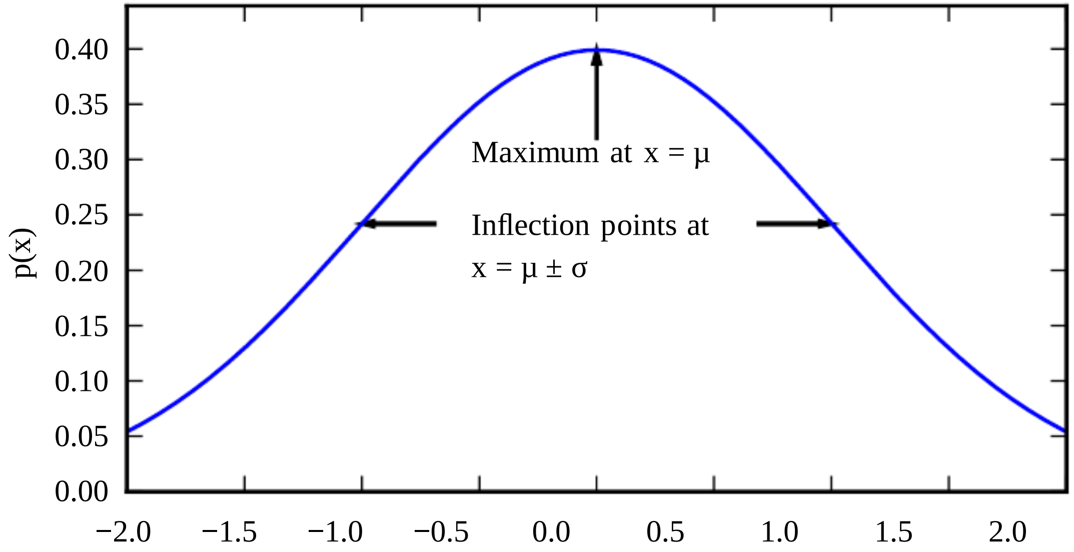
- Distribuzione empirica
  - Rappresenta la distribuzione di una variabile continua di cui si osservano *m* campioni che sono ovviamente equiprobabili

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

Principali distribuzioni di probabilità

- Distribuzione gaussiana

  o «Normale»



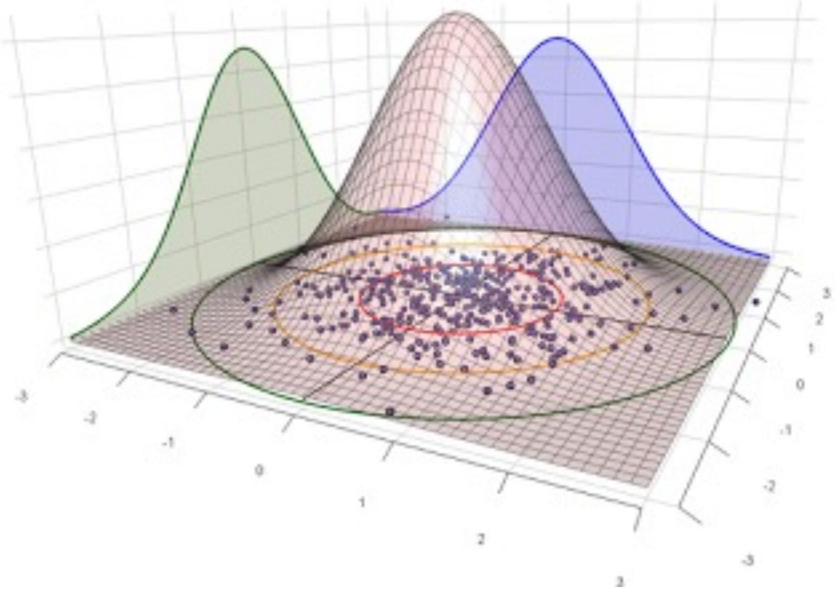
$$\mathcal{N}\left(x;\mu,\sigma^2\right)=\sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Principali distribuzioni di probabilità

- Distribuzione gaussiana

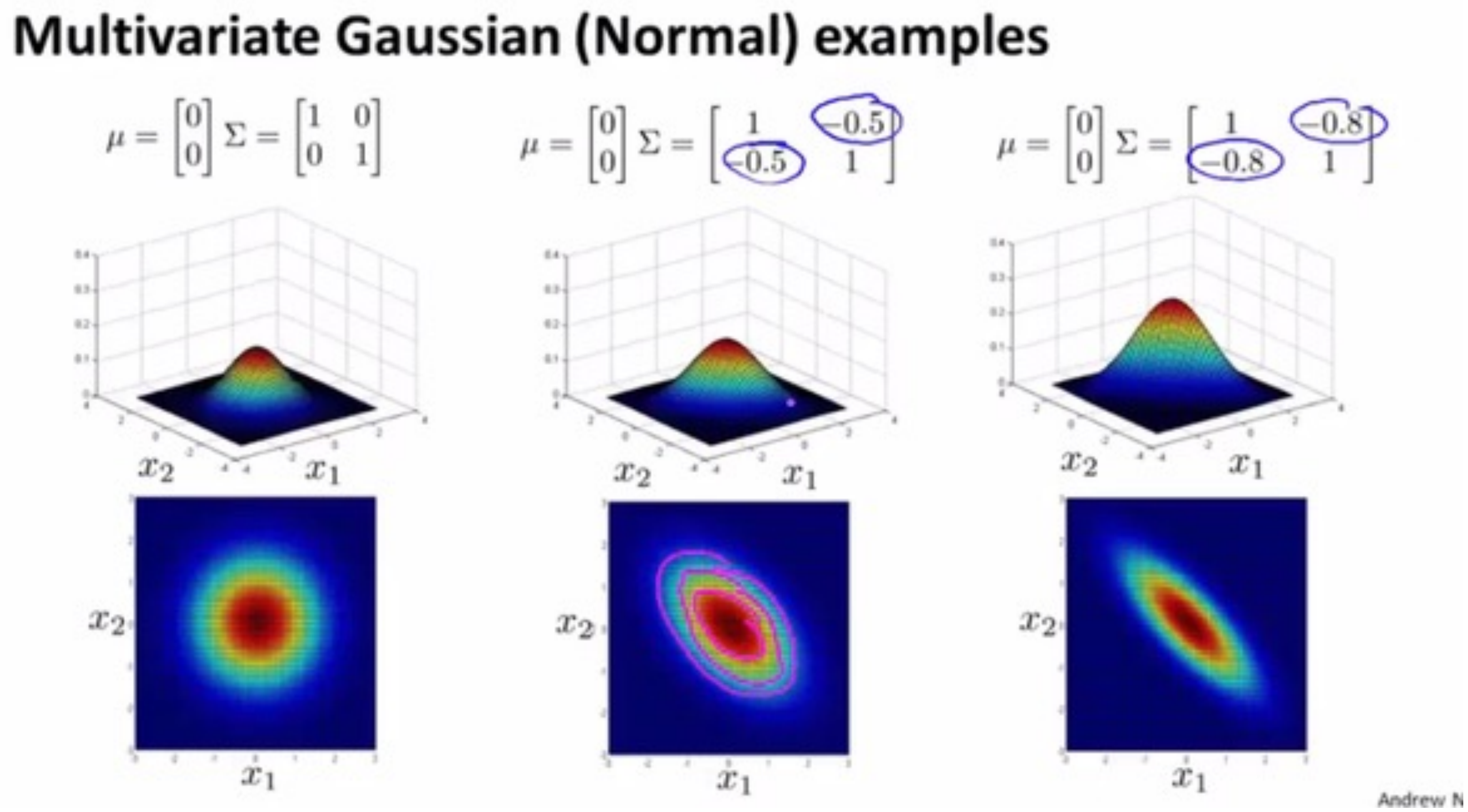
  o «Normale» multivariata

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



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

Principali distribuzioni di probabilità



Misture di distribuzioni

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

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

- *P*(*c*) è la distribuzione multinoulli di appartenenza all'i-esima distribuzione componente
- *c* : *variabile latente* → 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

$$I(\mathrm{x} = x) = -\log P(x)$$

- $\log_2$ → *bit* $\log_e$ → *nat*

Cenni di teoria dell’informazione

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

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

Cenni di teoria dell’informazione

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

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

$$\mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - \log Q(x)\right]$$

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

Cenni di teoria dell’informazione

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

$$H(P,Q) = H(P) + D_{KL}(P \| Q) = -\mathbb{E}_{\mathrm{x}\sim P}[\log Q(x)]$$
$$= -\sum_{\mathrm{x}\sim P} Q(x) \log Q(x)$$

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

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