Machine Learning per la fisica applicata e la fisica delle alte energie

Lezione 2 : Richiami di probabilità

"Probability theory is nothing but common sense reduced to calculation"
[Pierre Laplace, 1812]

DEFINIZIONE FREQUENTISTA

La probabilità di un evento è il rapporto tra il numero di casi favorevoli ed il numero di casi possibili.

DEFINIZIONE SOGGETTIVA

La probabilità di un evento è il prezzo che un individuo ritiene equo pagare per ricevere 1 se l'evento si verifica o 0 altrimenti.

DEFINIZIONE BAYESIANA

La probabilità di un evento è l'<u>incertezza</u> con cui l'evento si verifica

DEFINIZIONE ASSIOMATICA (Kolmogorov, 1933)

Viene dato un fondamento logico al concetto di probabilità mediante assiomi.

TIPI d'INCERTEZZA:

- aleatoria (DATA UNCERTAINTY)
- epistemica (MODEL UNCERTAINTY)

PROBABILITÀ di UN EVENTO

$$\boxed{Pr(A)}$$

2.1 PROPRIETÀ

$$Pr(A \wedge B) = Pr(A) \cdot Pr(B) = Pr(A, B)$$ eventi indipendenti — JOINT PROBABILITY

$$Pr(A \vee B) = Pr(A) + Pr(B) - Pr(A \wedge B)$$ UNION PROBABILITY

$$Pr(B/A) = \frac{Pr(A,B)}{Pr(A)}$$ CONDITIONAL PROBABILITY

$$Pr(A, B/C) = Pr(A/C) \cdot Pr(B/C)$$ CONDITIONAL INDEPENDENCE

X rappresenta una quantità di cui non si conosce il valore; è detta variabile casuale (RANDOM VARIABLE). Il set dei possibili valori di X è detto spazio di sampling (SAMPLING SPACE) $\mathcal{S}$. Un evento è un set di risultati dato uno spazio di sampling definito.

- sampling space finito $\Longleftrightarrow$ variabile casuale DISCRETA $\rightarrow$ PMF
- sampling space $\mathbb{R}$ $\Longleftrightarrow$ variabile casuale CONTINUA $\rightarrow$ CDF

PMF (PROBABILITY MASS FUNCTION)

$$p(x) \stackrel{\Delta}{=} Pr(X = x)$$

CDF (CUMULATIVE DISTRIBUTION FUNCTION)

$$P(x) \stackrel{\Delta}{=} Pr(X \leq x)$$

PDF (PROBABILITY DENSITY FUNCTION)

$$p(x) \stackrel{\Delta}{=} \frac{d}{dx} P(x)$$

segue che

$$Pr(a < X < b) = \int_a^b dx\, p(x) = P(b) - P(a) \qquad Pr(x < X < x+dx) \approx p(x)\,dx$$

Se la CDF è monotonica crescente, allora ha un'inversa che si chiama quantile. Se P è la CDF di X, allora $P^{-1}(q)$ è il valore x_q tale che $Pr(X \leq x_q)$ (il quantile q di P).

Si considerino due variabili casuali X e Y. La JOINT DISTRIBUTION è $p(x,y) = p(X = x, Y = y)$ $\forall X, Y$. Esempio

$p(X,Y)$	$Y=0$	$Y=1$
$X=0$	0,2	0,3
$X=1$	0,3	0,2

Se le variabili sono indipendenti e hanno cardinalità finite, allora

$$p(X = x) = \sum_y p(X = x, Y = y)$$

DISTRIBUZIONE MARGINALE

Definisco invece CONDITIONAL DISTRIBUTION

$$p(Y=y/X=x) = \frac{p(X=x, Y=y)}{p(X=x)} \quad \text{cioè} \quad p(x,y)=p(x)p(y/x)$$

segue la chain rule:

$$p(\bar{x}_1^2, D) = p(x_1)\,p(x_2/x_1)\,p(x_3/x_1, x_2) \dots p(x_D | x_{1:D-1})$$

Due variabili casuali sono MARGINALMENTE INDIPENDENTI se

$$X \perp Y \Leftrightarrow p(X,Y) = p(X)\,p(Y)$$

Due variabili casuali sono CONDIZIONALMENTE INDIPENDENTI se

$$X \perp Y \Leftrightarrow p(X, Y/Z) = p(X/Z)\,p(Y/Z)$$

2.3 MOMENTI DI UNA DISTRIBUZIONE

- MEDIA $\mathbb{E}[X] \triangleq \int_{\mathcal{X}} x\,p(x)\,dx \quad \left(= \sum_{x\in\mathcal{X}} x\,p(x) \right)$ se la variabile è discreta

La media è lineare: $\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$

- VARIANZA $\mathbb{V}[X] \triangleq \mathbb{E}[(X-\mu)^2] = \int_{\mathcal{X}} (x-\mu)^2 p(x)\,dx$

$$= \int x^2 p(x)\,dx + \mu^2 \underbrace{\int p(x)\,dx}_{1} - 2\mu \underbrace{\int x\,p(x)\,dx}_{\mu} = \mathbb{E}[X^2] - \mu^2.$$

$$\mathbb{V}[aX+b] = a^2\,\mathbb{V}[X] \qquad \mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{V}[X_i] \text{ (per variabili indipendenti)}$$

- MODA

$x^* = \operatorname{argmax}_x p(x)$ (valore più probabile; può non essere UNICO)

Tutti questi stimatori non danno conto di tutta l'informazione contenuta nella pdf. Esempio: Anscombe's quartet e Datasaurus set.

2.3 TEOREMA DI BAYES

Data una quantità non nota H e dei dati noti $Y = y$:

$$p(H=h/Y=y) = \frac{p(H=h)\,p(Y=y/H=h)}{p(Y=y)} \qquad [\text{posterior} \propto \text{prior} \times \text{likelihood}]$$

Il teorema segue dall'identità $p(h/y)\,p(y) = p(h)\,p(y/h) = p(h,y)$

$p(h)$ si chiama PRIOR (ciò che conosciamo/assumiamo per H prima di fare qualunque misura); $p(y/h)$ è la DISTRIBUZIONE OSSERVATA; $p(y/h)$ è la likelihood (che non è una distribuzione di probabilità); $p(y)$ è la LIKELIHOOD MARGINALE.

$$p(Y=y) = \sum_{h' \in \mathcal{H}} p(H=h')\,p(Y=y/H=h') = \sum_{h' \in \mathcal{H}} p(H=h', Y=y)$$

$p(h/y)$ è detta POSTERIOR DISTRIBUTION.

<u>Esempio:</u> The Monty Hall problem

Tre porte. Un premio dietro una porta. Il concorrente sceglie una porta. Il presentatore ne apre un'altra. Che cosa conviene fare al concorrente?

Supponiamo che il concorrente scelga inizialmente la porta 1

$$p(H_1) = p(H=1) = \frac{1}{3} \qquad \text{probabilità di scegliere una porta}$$

$$p(Y=2/H_1) = \frac{1}{2} \ \left(\text{se il premio è dietro la porta 1}\right) \quad p(Y=2/H_2) = 0 \quad p(Y=2/H_3) = 1$$

$$p(Y=3/H_1) = \frac{1}{2} \ \left(\text{se il premio è dietro la porta 1}\right) \quad p(Y=3/H_2) = 1 \quad p(Y=3/H_3) = 0$$

SI APRE LA PORTA 3

$$p(Y=3) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2} \quad (\text{ottenuta marginalizzando})$$

$$p(H_1/Y=3) = \frac{1/3 \cdot 1/2}{1/2} = \frac{1}{3} \qquad p(H_2/Y=3) = \frac{1/3 \cdot 1}{1/2} = \frac{2}{3} \qquad p(H_3/Y=3) = 0$$

2.4. LA DISTRIBUZIONE DI GAUSS

Tra le distribuzioni più utilizzate nel Machine Learning (e non solo) vi è la distribuzione gaussiana così definita

CDF $P(y) \triangleq Pr(Y \leq y)$ $\qquad Pr(a < Y \leq b) = P(b) - P(a)$

$$\Phi(y; \mu, \sigma^2) \triangleq \int_{-\infty}^{y} \mathcal{N}(z/\mu, \sigma^2)\, dz = \frac{1}{2}\left[1 + \text{erf}(z/\sqrt{2})\right], \; z = \frac{y-\mu}{\sigma}$$

$$\text{erf}(x) \triangleq \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \qquad \mathcal{N}(y/\mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

(PDF)

$$\mathbb{E}[Y] \triangleq \int_y y\, p(y)\, dy = \mu \qquad \mathbb{V}[Y] \triangleq \int (y-\mu)^2 p(y)\, dy = \sigma^2$$

$\hookrightarrow Y \rightarrow (Y-\mu)^2$

Ragioni della popolarità della distribuzione

- dipende da due soli parametri di facile interpretazione
- il teorema del limite centrale ci dice che, nel limite $N \rightarrow \infty$, ciascuna distribuzione può essere approssimata dalla distribuzione gaussiana; inoltre essa rappresenta la somma di variabili casuali indipendenti
- il numero di assunzioni è minimo.

2.5 ALTRE DISTRIBUZIONI NOTEVOLI

t-Student $\quad \mathcal{T}(y/\mu, \sigma^2, \nu) \propto \left[1 + \frac{1}{\nu}\left(\frac{y-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}}$

μ mean
σ scale param
ν degree of normality

Lorentz (Cauchy) $\quad C(x/\mu, \gamma) = \frac{1}{\gamma\pi}\left[1 + \left(\frac{x-\mu}{\gamma}\right)^2\right]^{-1}$

Laplace $\quad L(y/\mu, b) = \frac{1}{2b} \exp\left(-\frac{|y-\mu|}{b}\right)$

μ mean
μ mode
$2b^2$ variance

Beta $\quad \text{Beta}(x/a, b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$

$a/(a+b)$ mean
$(a-1)(a+b-2)$ mode
$ab/(a+b)^2(a+b+1)$ variance

Gamma $\quad Ga(x/a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}$ $\quad$ a shape, b rate $\quad$ 6/

Exponential $\quad Exp(x/\lambda) = Ga(x/a=1, b=\lambda)$

χ^2 $\quad \chi^2_\nu(x) = Ga\left(x/a=\frac{\nu}{2}, b=\frac{1}{2}\right)$

2.6 ALCUNE OSSERVAZIONI

Si supponga di calcolare la PDF della somma di due variabili casuali continue. Se si tratta di distribuzioni gaussiane, avrò:

$$x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \qquad x_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

se $y = x_1 + x_2$ allora

$$p(y) = \mathcal{N}(x_1/\mu_1, \sigma_1^2) \otimes \mathcal{N}(x_2/\mu_2, \sigma_2^2) = \mathcal{N}(y/\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$$

cioè la convoluzione di due gaussiane è una gaussiana

Si supponga che x sia una variabile casuale e $y=f(x)$ una funzione di essa. Talora è difficile calcolare analiticamente $p(y)$. Si supponga per esempio che $x \sim Unif(-1,1)$ e $y=f(x) = x^2$. Possiamo approssimare $p(y)$ campionando $p(x)$ usando un generatore (uniforme) di numeri casuali, facendone il quadrato e prendendo la distribuzione empirica

$$p_S(y) = \frac{1}{N_S} \sum_{x=1}^{N_S} \delta(y - y_s) \qquad \text{METODO MONTE CARLO}$$

2.7 MODELLI MULTIVARIATI

Consideriamo ora due variabili X e Y. La covarianza è:

$$Cov[X,Y] \triangleq \mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

cioè è una matrice D-dimensionale che ha le varianze sulla diagonale se $\vec{x}$ è un vettore D-dimensionale

Si definisce correlazione (di Pearson) tra due variabili X e Y

$$\rho \triangleq \mathrm{corr}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathbb{V}[X]\,\mathbb{V}[Y]}}$$

$$\mathrm{Cov}[X] = \begin{pmatrix} \mathbb{V}[X_1] & \mathrm{Cov}[X_1,X_2] & \cdots & \mathrm{Cov}[X_1,X_D] \\ \mathrm{Cov}[X_2,X_1] & \mathbb{V}[X_2] & \cdots & \mathrm{Cov}[X_2,X_D] \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}[X_D,X_1] & \mathrm{Cov}[X_D,X_2] & \cdots & \mathbb{V}[X_D] \end{pmatrix}$$

$$\mathrm{corr}(x) = \begin{pmatrix} 1 & \frac{\mathbb{E}[(X_1-\mu_1)(X_2-\mu_2)]}{\sigma_1\sigma_2} & \cdots & \frac{\mathbb{E}[(X_1-\mu_1)(X_D-\mu_D)]}{\sigma_1\sigma_D} \\ \frac{\mathbb{E}[(X_2-\mu_2)(X_1-\mu_1)]}{\sigma_2\sigma_1} & 1 & \cdots & \frac{\mathbb{E}[(X_2-\mu_2)(X_D-\mu_D)]}{\sigma_2\sigma_D} \\ \vdots & \vdots & & \vdots \\ \frac{\mathbb{E}[(X_D-\mu_D)(X_1-\mu_1)]}{\sigma_D\sigma_1} & \frac{\mathbb{E}[(X_D-\mu_D)(X_2-\mu_2)]}{\sigma_D\sigma_2} & \cdots & 1 \end{pmatrix}$$

Osservazioni:

1) Il fatto che due variabili siano non correlate non significa che siano indipendenti. Esempio: $X \sim \mathrm{Unif}(-1,1)$ e $Y = X^2$, $\mathrm{corr}[X,Y]=0$
Viceversa, due variabili indipendenti sono necessariamente non correlate

2) Correlazione NON implica causalità.

3) Una correlazione che appare simile in diversi set di dati può sparire (o diventare opposta) se i dati sono combinati. (Simpson's paradox).

1.7 DISTRIBUZIONE GAUSSIANA MULTIVARIATA (MVN)

La definizione di distribuzione gaussiana si estende a più variabili

$$\mathcal{N}(\vec{y}\,|\,\vec{\mu},\Sigma) \triangleq \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{y}-\vec{\mu})^T\Sigma^{-1}(\vec{y}-\vec{\mu})\right\}$$

$\Sigma = \mathrm{Cov}[\vec{y}]$

Supponiamo di avere due variabili aleatorie multidimensionali $\vec{y}_1$ e $\vec{y}_2$. Si può definire la distribuzione "jointly Gaussian" come

$$\vec{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \qquad \Lambda = \Sigma^{-1}$$

Le distribuzioni marginali sono date da

$$p(\vec{y}_1) = \mathcal{N}(\vec{y}_1 | \vec{\mu}_1, \Sigma_{11}) \qquad p(\vec{y}_2) = \mathcal{N}(\vec{y}_2 | \vec{\mu}_2, \Sigma_{22})$$

e le probabilità condizionate è:

$$p(\vec{y}_1 | \vec{y}_2) = \mathcal{N}(\vec{y}_1 | \vec{\mu}_{1/2}, \Sigma_{1/2})$$

con

$$\vec{\mu}_{1/2} = \vec{\mu}_1 + \Sigma_{12} \Sigma_{22}^{-1} (\vec{y}_2 - \vec{\mu}_2) \qquad \Sigma_{1/2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} = \Lambda_{11}^{-1}$$

Sono tutte distribuzioni gaussiane.