Basi matematiche per il Machine Learning

Corso di AA, anno 2017/18, Padova

Fabio Aiolli

04 Ottobre 2017

Probabilità

- Un esperimento si dice casuale/aleatorio quando l'output non è predicibile con certezza.
- Considereremo output discreti (perché più semplici)
- Una possibile interpretazione della probabilità come frequenza: ripetendo un esperimento sotto le stesse condizioni per un numero infinito di volte, la proporzione di volte che l'output è contenuto in E tende ad un valore costante. Questa costante è la probabilità dell'evento E, ovvero $P(E)$.

Assiomi della Probabilità

- $0 \leq P(E) \leq 1$. Se E_1 è evento che non può mai succedere (evento impossibile), allora $P(E_1) = 0$. Se E_2 è un evento certo, allora $P(E_2) = 1$.
- Se S è l'insieme di tutti i possibili output, allora $P(S) = 1$
- Se $E_i, i = 1, \ldots, n$ sono mutuamente esclusivi (non possono occorrere contemporaneamente, $E_i \cap E_j = \emptyset, i \neq j$), allora abbiamo: $P(\cup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$.
  In particolare, vale $P(E^c) = 1 - P(E)$ se E^c è l'evento complementare di E.
- Se l'intersezione di due eventi E e F è non vuota, abbiamo: $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

Probabilità Condizionale

- $P(E|F)$ (o probabilità a posteriori di E dato F) è la probabilità di un'evento E dato che l'evento F si è verificato, ed è tale che: $P(E \cap F) = P(F)P(E|F)$
- Poiché $P(F)P(E|F) = P(E \cap F) = P(E)P(F|E)$ (l'operatore $\cap$ è commutativo), otteniamo la *formula di Bayes*: $P(F|E) = \frac{P(E|F)P(F)}{P(E)}$
- Due eventi E, F si dicono indipendenti quando $P(E|F) = P(E)$. Per eventi indipendenti, vale $P(E \cap F) = P(E)P(F)$

Media e Varianza

La media (o valore atteso) di una variabile aleatoria X, $E[X]$, è il valore medio di X ottenuto su un grande numero di esperimenti:
$E[X] = \sum_i x_i P(x_i)$.

- $E[aX + b] = aE[X] + b$
- $E[X + Y] = E[X] + E[Y]$
- $E[X^n] = \sum_i x_i^n P(x_i)$ momento n-esimo

La varianza misura quanto X varia rispetto al valore atteso nei singoli esperimenti. Sia $\mu = E[X]$, la varianza è definita come:

$$Var(X) = E[(X - \mu)^2] = E[X^2] - \mu^2$$

Tipicamente si usa la notazione σ^2 per denotare la varianza. La deviazione standard è definita come $\sigma(X) = \sqrt{Var(X)}$

Distribuzioni Notevoli

Caso Discreto:

- Bernoulli - Output 1 (successo), 0 (fallimento). p è la probabilità di successo. Allora, $P(X = 1) = p$ e $P(X = 0) = 1 - p$. $E[X] = p$, $Var(X) = p(1 - p)$.
- Binomiale - Se eseguiamo N prove di Bernoulli identiche, la variabile aleatoria X che rappresenta il numero di successi ottenuti in N prove ha una distribuzione del tipo: $P(X = i) = \binom{N}{i} p^i (1 - p)^{N-i}$, $i = 0, \dots N$

Caso Reale:

- Uniforme nell'intervallo $[a, b]$. Allora, $p(x) = \frac{1}{b-a}$ se $a \leq x \leq b$ e $p(x) = 0$ altrimenti. $E[X] = \frac{a+b}{2}$, $Var(X) = \frac{(b-a)^2}{12}$.
- Normale (Gaussiana) di media μ e varianza σ^2:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), -\infty < x < +\infty$$

Vettori

Un vettore n dimensionale $\mathbf{x} \in \mathbb{R}^n$, è una collezione di n valori scalari sistemati in colonna:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Dati due vettori n dimensionali $\mathbf{x}, \mathbf{z}$ la loro somma è ancora un vettore n dimensionale:

$$\begin{pmatrix} x_1 + z_1 \\ x_2 + z_2 \\ \vdots \\ x_n + z_n \end{pmatrix}$$

Prodotto scalare

Dati due vettori n dimensionali $\mathbf{x}, \mathbf{z}$ il loro prodotto scalare è uno scalare:

$$\mathbf{x} \cdot \mathbf{z} = \sum_i x_i z_i$$

La lunghezza di un vettore $\mathbf{x}$ si denota $|\mathbf{x}|$, la lunghezza al quadrato vale:

$$|\mathbf{x}|^2 = \sum_i x_i^2$$

Il prodotto scalare ha una naturale interpretazione geometrica:

$$\mathbf{x} \cdot \mathbf{z} = |\mathbf{x}||\mathbf{z}| \cos(\theta)$$

dove θ è l'angolo formato tra i due vettori. Nota che tale quantità viene massimizzata con $\theta = 0$ e viene annullata quando i vettori sono ortogonali.

Vettori binari e insiemi

Quando abbiamo a che fare con vettori a valori binari $\mathbf{x} \in \{0,1\}^n$ possiamo dare una interpretazione insiemistica al vettore considerando un universo di n elementi e $\mathbf{x}$ l'insieme contenente gli elementi corrispondenti agli 1 nel vettore.

La lunghezza (o norma) di un vettore indicherà il numero di elementi nell'insieme (cardinalità)

Il prodotto scalare tra due vettori indicherà il numero di elementi in comune tra i due insiemi (cardinalità dell'intersezione)

Posso calcolare la proiezione di un vettore $\mathbf{x}$ lungo la direzione di un altro vettore $\mathbf{z}$, come

$$\mathbf{x}_\mathbf{z} = \left(\frac{\mathbf{x} \cdot \mathbf{z}}{\mathbf{z} \cdot \mathbf{z}} \right) \mathbf{z} = \alpha \mathbf{z}$$

Il coefficiente α ha una interessante interpretazione probabilistica, $P(\mathbf{x}|\mathbf{z})$.

Matrici

Una matrice $m \times n$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, è una collezione di valori scalari sistemati in un rettangolo di m righe e n colonne:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

L'elemento i, j della matrice $\mathbf{A}$ può essere scritto $a_{ij} = [\mathbf{A}]_{ij}$

Nota che un vettore $\mathbf{x} \in \mathbb{R}^n$ può essere visto come una matrice $\mathbf{x} \in \mathbb{R}^{n \times 1}$

Addizione e prodotto

Date due matrici $\mathbf{A}$ e $\mathbf{B}$ della stessa dimensione,

$$[\mathbf{A} + \mathbf{B}]_{ij} = [\mathbf{A}]_{ij} + [\mathbf{B}]_{ij} = a_{ij} + b_{ij}$$

Date due matrici $\mathbf{A} \in \mathbb{R}^{m \times k}$ e $\mathbf{B} \in \mathbb{R}^{k \times n}$, il loro prodotto matriciale $\mathbf{AB}$ è la matrice con elementi:

$$[\mathbf{AB}]_{ij} = \sum_{q=1}^{k} [\mathbf{A}]_{iq} [\mathbf{B}]_{qj} = \sum_{q=1}^{k} a_{iq} b_{qj}$$

Nota bene: In generale $\mathbf{AB} \neq \mathbf{BA}$

Trasposta

La trasposta $\mathbf{A}^\top \in \mathbb{R}^{m \times n}$ di una matrice $\mathbf{A} \in \mathbb{R}^{n \times m}$ è definita da:

$$[\mathbf{A}^\top]_{ij} = \mathbf{A}_{ji}$$

Proprietà:

- $(\mathbf{A}^\top)^\top = \mathbf{A}$
- $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$

Se $\mathbf{A} = \mathbf{A}^\top$ si dice che la matrice $\mathbf{A}$ è simmetrica.

Inversa

La matrice identità è una matrice diagonale (necessariamente quadrata) $\mathbf{I} \in \mathbb{R}^{n \times n}$ avente valori uguali a 1 nella diagonale e 0 fuori dalla diagonale.

La matrice inversa di una matrice quadrata $\mathbf{A} \in \mathbb{R}^{n \times n}$ è una matrice $\mathbf{A}^{-1}$ tale che

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} = \mathbf{A}^{-1}\mathbf{A}$$

N.B. Non è sempre possibile trovare tale matrice (solo se il rango è massimo, determinante diverso da 0)

Se la matrice inversa esiste, allora

$$(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

Per matrici rettangolari, se la matrice quadrata $\mathbf{A}\mathbf{A}^{\top}$ è invertibile, allora la matrice $\mathbf{A}^{\dagger} = \mathbf{A}^{\top}(\mathbf{A}\mathbf{A}^{\top})^{-1}$ (detta pseudo-inversa) soddisfa $\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{I}$.

Matrici definite positive

Una matrice simmetrica $\mathbf{A} \in \mathbb{R}^{n \times n}$ con la proprietà che $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$ per ogni vettore $\mathbf{x} \in \mathbb{R}^n$ si dice semidefinita positiva (autovalori ≥ 0).

Una matrice simmetrica $\mathbf{A} \in \mathbb{R}^{n \times n}$ con la proprietà che $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$ per ogni vettore $\mathbf{x} \in \mathbb{R}^n$ si dice definita positiva (autovalori > 0).

Una matrice definita positiva è sempre invertibile!