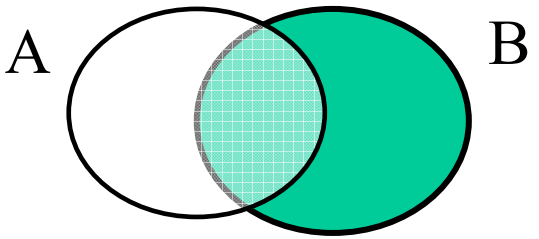
Apprendimento Bayesiano

Metodi di apprendimento di concetti
basati sul calcolo delle probabilità

Richiamo di Concetti di Calcolo delle Probabilità

Assiomi del calcolo delle probabilità

- Spazio di campionamento Ω è l'insieme degli esiti di una prova
- ω è l'esito di una prova (es. il lancio del dado ha esito 2)
- A è un evento (sottoinsieme di Ω)
- Indichiamo con P(A) la **probabilità** (massa di probabilità) di un evento A (es: x=1, o x "pari")
- Per ogni coppia di *eventi* A e B:



$$0 \le P(A) \le 1$$

$$P(true) = 1$$

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B) = P(A \cup B)$$

$$dove\ \ P(A \wedge B) = P(A \cap B) = 0\ \ se\ \ P(A), P(B)\ \ mutuamente\ \ esclusive$$

$$es.\ (lanci\ \ dado)\ \ A = \{1,3,5\}\ \ B = \{2,4,6\}$$

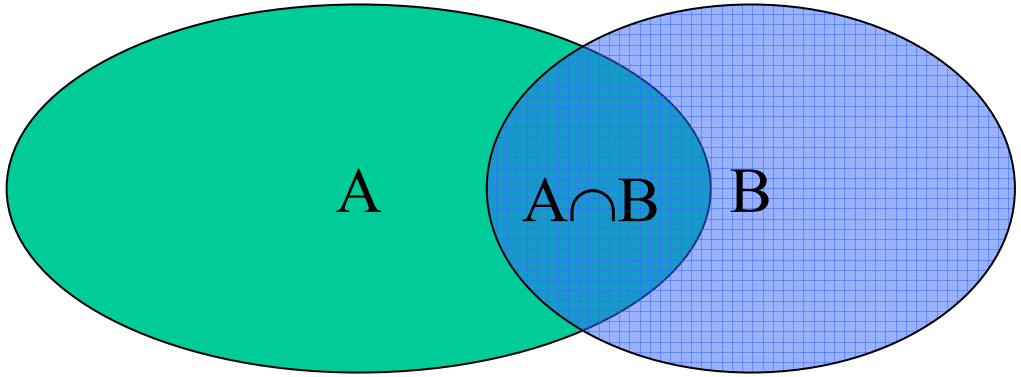
Probabilità condizionata P(A|B)

probabilità di un evento A supponendo verificato l'evento B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad P(B) \neq 0$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \qquad P(A) \neq 0$$

$$\Rightarrow P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$



Es.:
A = Pr(studente segue IUM)
B = Pr(studente segue AA)

Teorema di Bayes

- Una formulazione alternativa della regola vista prima:

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

Esempio lancio di un dado:

– A numeri pari, P(A)=1/2
– B numero 2, P(B)=1/6

– A ∩ B={2}, P(A∩B)=1/6

$$P(B \mid A) = P(A \cap B) / P(A) = \frac{1/6}{1/2} = 1/3$$

Proprietà derivate:

Se due eventi A e B sono disgiunti ($A \cap B = \emptyset$)
segue $P(B|A)=0$ e $P(A|B)=0$ poiché il verificarsi di un evento esclude il verificarsi dell'altro.

Se $b_1, b_2, \ldots b_m$ sono mutuamente disgiunti ed esaustivi:

$$P(A) = \sum_i P(A \mid b_i) P(b_i)$$

Es:

$$P(A) = P(A \mid B)P(B) + P(A \mid \neg B)P(\neg B)$$

$$\Pr(\textit{promosso}) = \Pr(\textit{promosso} \mid \textit{studiare}) \Pr(\textit{studiare}) +$$
$$\Pr(\textit{promosso} \mid \textit{non\_studiare}) \Pr(\textit{non\_studiare}) +$$
$$\Pr(\textit{promosso} \mid \textit{sfortuna}) \Pr(\textit{sfortuna})$$

Variabili aleatorie e probabilità

- Una **variabile aleatoria** X descrive un evento non predicibile in anticipo (lancio di un dado, infatti *alea*=dado in latino)
- Lo **spazio di campionamento** (*sample space*) Ω di X è l'insieme dei possibili esiti della variabile (per il dado, Ω = {1,2,3,4,5,6})
- Un **evento** è un sottoinsieme di Ω, es.: e_1 ={1}, e_2 = {2,4,6}
- La **massa di probabilità** è definita come P(X=x) o P(x) o P_x
- La **distribuzione di probabilità** per una variabile discreta è la lista di probabilità associate a tutti i possibili eventi di S
- **ASSIOMA 1** : Se X è una variabile discreta, allora

$$0 \le P(x) \le 1, \forall x \in S, \;\; e \sum_{x \in S} P(x) = 1$$

Funzione densità di probabilità

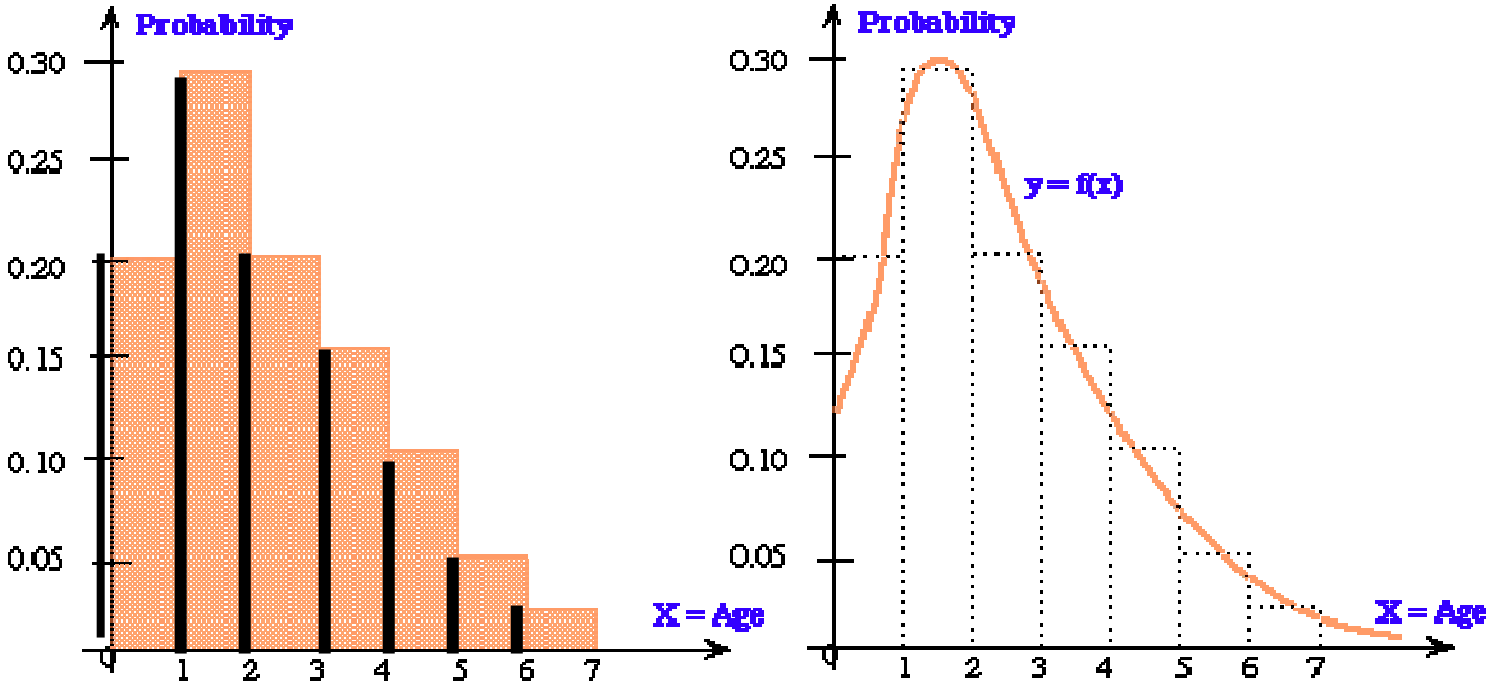
- Se X può assumere infiniti valori (X variabile continua), la somma di questi valori non può essere 1
- Si definisce la funzione **densità di probabilità** come:

$$p(x_0) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} P(x_0 \le X \le x_0 + \varepsilon)$$

- $p(x)$ è il **limite** per $\varepsilon \to 0$ di $1/\varepsilon$ volte la probabilità che X assuma un valore nell'intervallo $[x_0, x_0+\varepsilon]$
- In tal modo si ha, per una variabile aleatoria continua:

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$

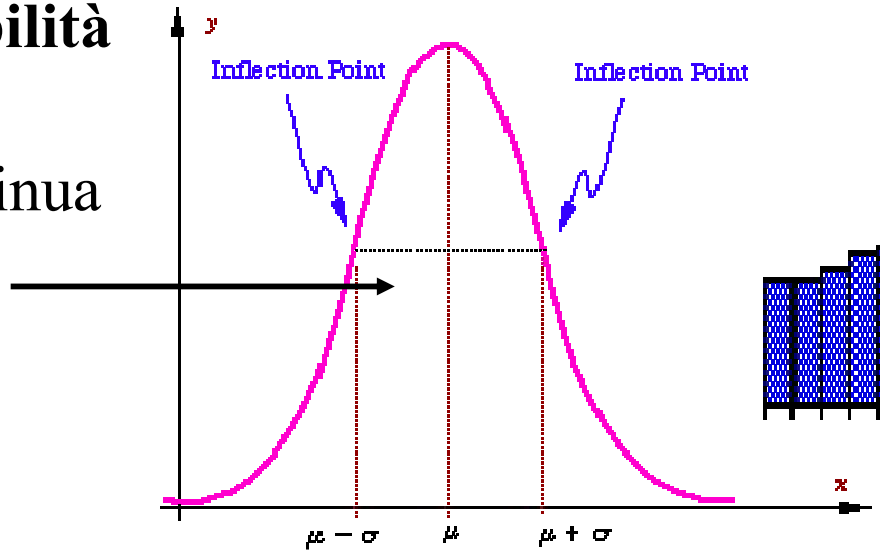
Densità della probabilità di affitto di un'automobile in funzione dell'età



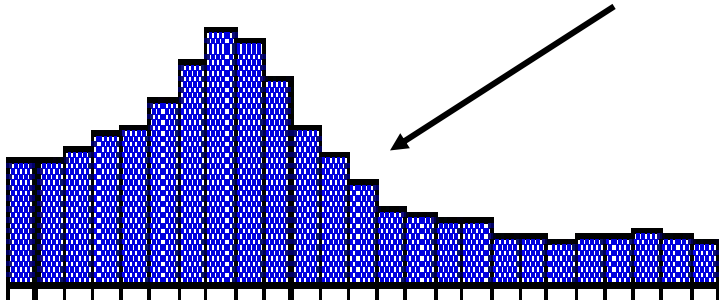
Densità e Distribuzione

Cosí come un oggetto non omogeneo è più o meno denso in regioni differenti del suo volume complessivo, così la densità di probabilità mostra **su quali valori della variabile aleatoria si concentra la probabilità**

X continua



X discreta

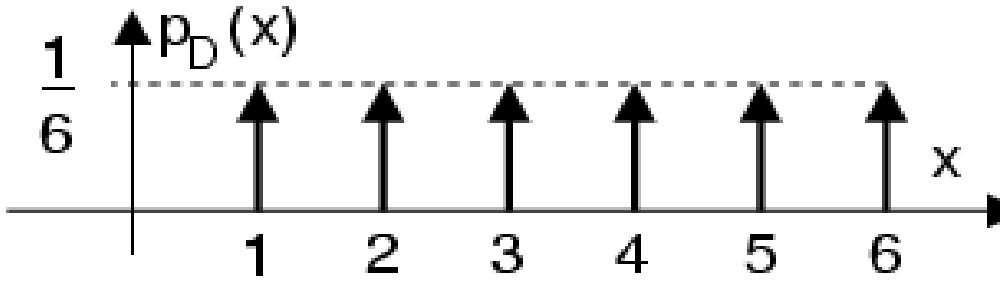


Mentre la funzione **distribuzione di probabilità** per la v.a. **X** è definita come:

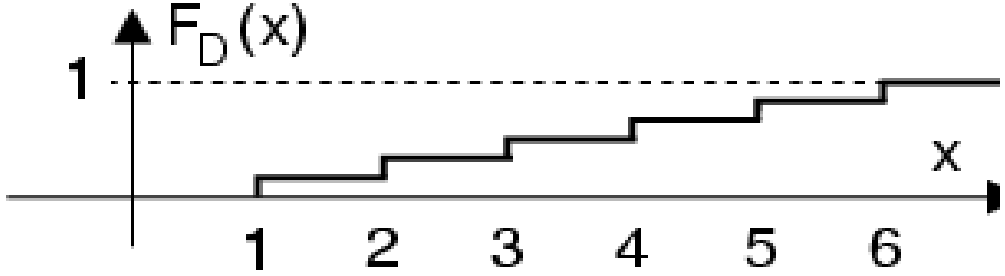
$$F_X(x) = \int_{-\infty}^{x} p_X(\xi)d\xi = Pr\{X \le x\}$$

Esempi

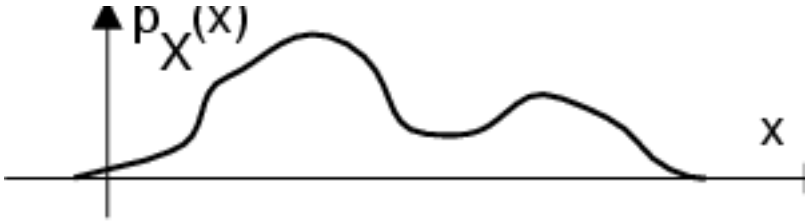
Massa di prob. per X discreta (es. lancio del dado)



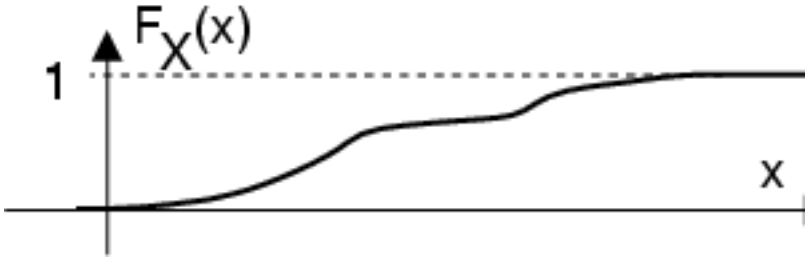
Distribuzione di probabilità per p_D



Densità di prob. per X continua



Distribuzione di prob. per p_X



Medie e statistiche

- La media statistica (o valor medio, o valore atteso) di una v.a. aleatoria X è la **somma** dei suoi possibili valori pesati per le rispettive probabilità. E' indicata con μ_X, o E(X) o E[X] o μ
- Per variabili **discrete**:

$$E(X) = \sum_{i=1}^{k} x_i P(X = x_i)$$

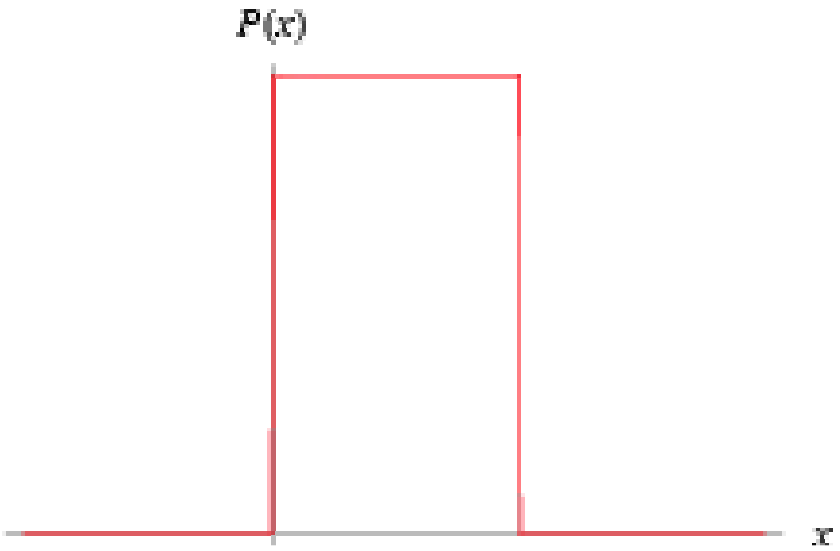
- Per variabili **continue**:

$$E(X) = \int_{-\infty}^{+\infty} xp(x)dx$$

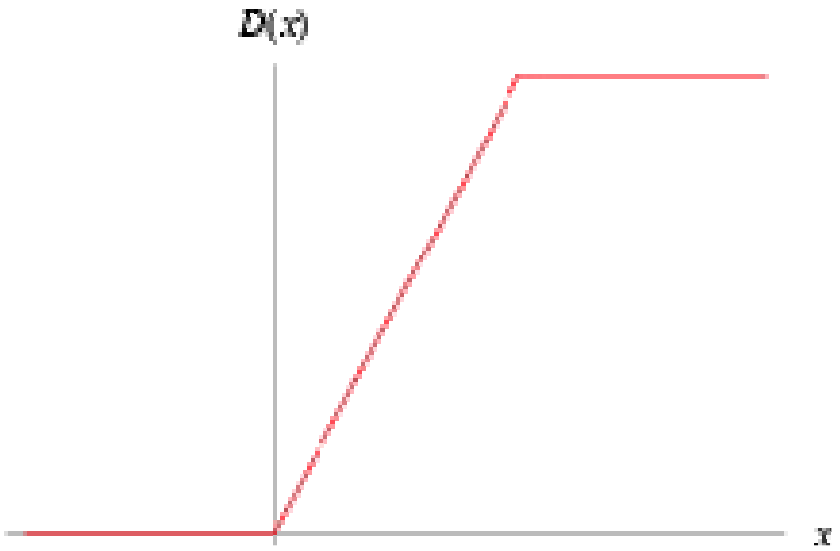
Esempi

- Se X è **uniformemente distribuita** in [a,b], E(X)=(b+a)/2

Massa di probabilità



Distribuzione di probabilità



Esempio (continua)

- Nel discreto, supponiamo che X assuma i valori 1, 2, 3, 4, 5 e 6 con probabilità uniforme, P(x)=1/6 $\forall$x
- Allora:

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} =$$

$$= \frac{1}{6} \sum_{i=1}^{6} i = \frac{21}{6} = \frac{7}{2} = \frac{1+6}{2} = \frac{a+b}{2}$$

$$\sum_{i=1}^{n} i = \frac{n}{2}(n+1)$$

Varianza

Varianza di una distribuzione di probabilità $p_X(x)$:

v.a. discreta $\sigma_X^2 = E((X-\mu_X)^2) = \sum_i (x_i - \mu_X)^2 p_X(x_i)$

v.a. continua $\sigma_X^2 = E((X-\mu_X)^2) = \int_{-\infty}^{+\infty} (x-\mu_X)^2 p_X(x) dx$

La varianza indica la **dispersione** della v.a. rispetto al suo valore medio

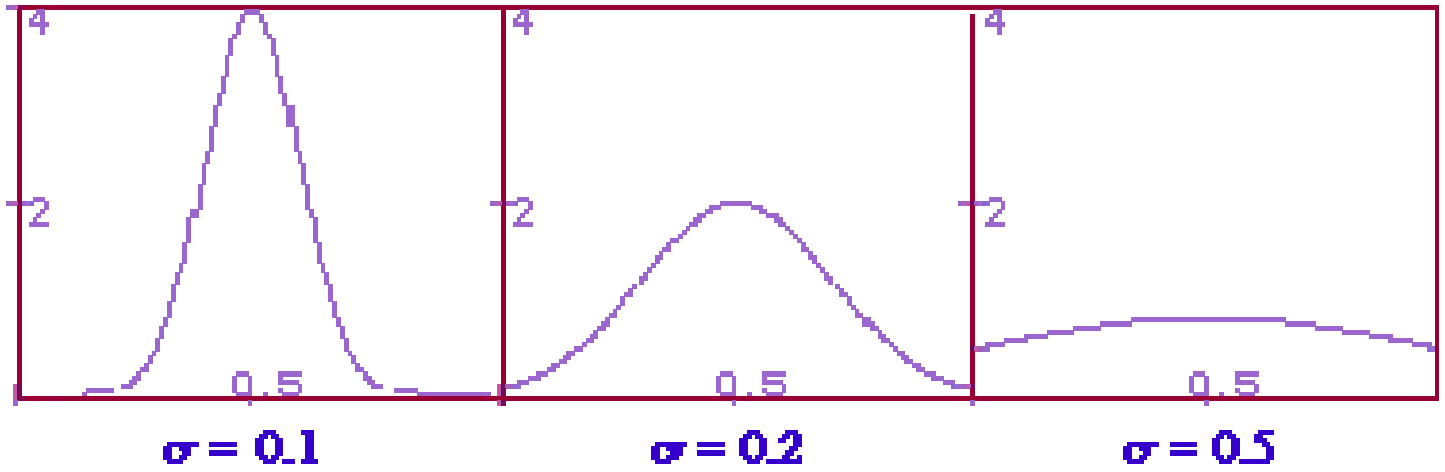
– Esempio: X può assumere i due valori -1 e 1 con uguale probabilità $E[X]=\mu_X=0$, $p(-1)=p(+1)=0,5$

$$\sigma^2 = (-1-0)^2 0,5 + (+1-0)^2 0,5 = 1$$

Lo **scarto quadratico medio** o **deviazione standard** è definito come:

$$\sqrt{\sigma}$$

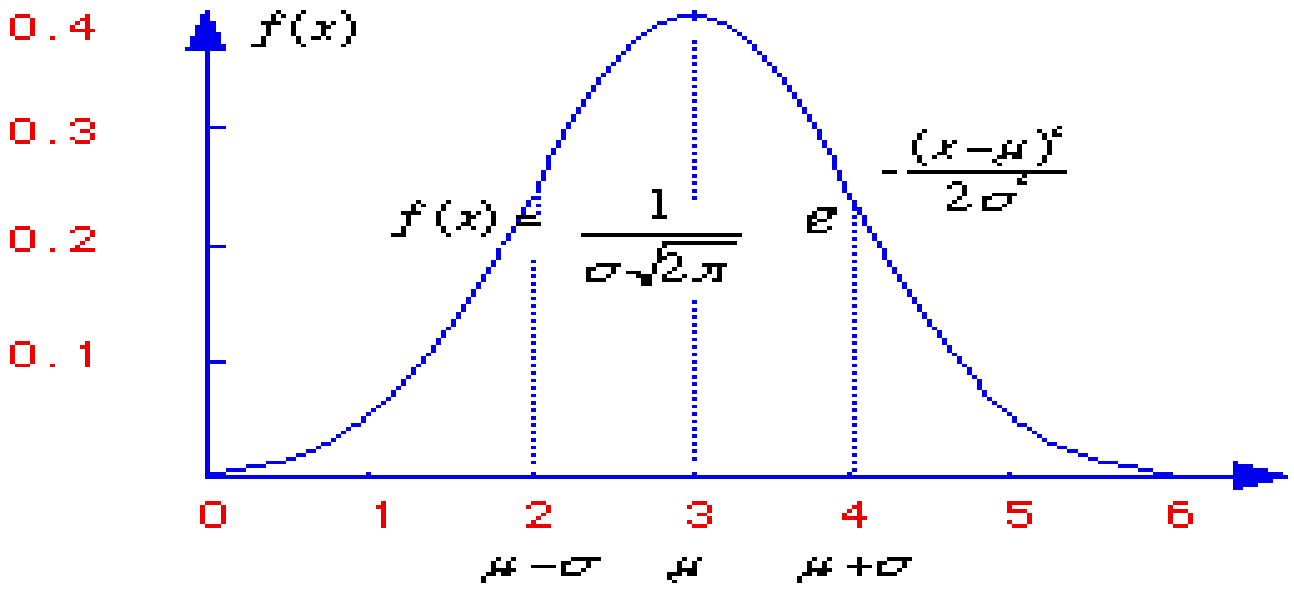
Esempi



Stesso valore medio, ma scarto quadratico assai diverso!!!

Funzione densità normale o gaussiana

$$p(X)=\frac{1}{\sqrt{2\pi\sigma^2}}\quad e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$



Riassunto dei concetti esposti

- Spazio di campionamento, esiti, eventi
- Probabilità condizionata, somma i probabilità, probabilità congiunte, proprietà fondamentali
- Teorema di Bayes
- Definizione di v.a. discreta e continua
- Definizione di massa di probabilità, densità di probabilità (per v.a. continue) e distribuzione di probabilità
- Media, scarto quadratico e varianza
- Gaussiana

Apprendimento Bayesiano

Caratteristiche dell'Apprendimento Bayesiano

- Ogni esempio di addestramento progressivamente **decrementa o incrementa la probabilità stimata** che un'ipotesi sia corretta
- La **conoscenza pregressa** può essere **combinata con i dati osservati** per determinare la probabilità finale di un'ipotesi
- I metodi Bayesiani possono **fornire predizioni probabilistiche** (es. questo paziente ha il 93% di possibilità di guarire)
- Nuove istanze possono essere **classificate combinando le predizioni di ipotesi multiple**, pesate con le loro probabilità
- Anche quando i metodi Bayesiano sono intrattabili computazionalmente, possono **fornire uno standard di decisione ottimale** rispetto al quale misurare metodi più pratici

Il teorema di Bayes nell'Apprendimento Automatico

- Sia h un'ipotesi in H e D sia l'insieme dei dati di apprendimento (x_i, d_i):

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

dove:

P(h) è la probabilità **a priori** dell'ipotesi h (precedente all'apprendimento)

P(D) è la probabilità **a priori** di D (la probabilità di estrarre un campione D dallo spazio delle istanze X)

P(D|h) è la probabiltà di osservare i dati D dato un mondo in cui vale l'ipotesi h

P(h|D) è la probabilità **a posteriori** di h

Obiettivo: scegliere l'ipotesi h più probabile (ovvero che massimizzi P(h|D)

Maximum A Posteriori hypothesis (MAP)

- Scegli l'ipotesi

P(D) è costante

$$h_{MAP} = \arg\max_{h \in H} P(h \mid D) = \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)} = \arg\max_{h \in H} P(D \mid h)P(h)$$

- Richiede la stima di P(D|h) per ogni h di H

Il Teorema di Bayes e l'Apprendimento di Concetti

- Progettiamo un semplice algoritmo di apprendimento di concetti per emettere l'ipotesi MAP (maximum a posteriori), basata sul teorema di Bayes:
- **Algoritmo di MAP learning con forza bruta**:
  - Emetti l'ipotesi h_{MAP} con la massima probabilità a posteriori:

$$h_{MAP} = \arg\max_{h \in H} P(h \mid D) = \arg\max_{h \in H} P(D \mid h) P(h)$$

  - Richiede molti calcoli, perché deve calcolare P(h|D) per ogni ipotesi in H (non fattibile per H grande)
  - Come scegliere P(h) e P(D|h)?

MAP Learning (1)

- Facciamo le seguenti assunzioni:
  1. L'insieme di addestramento D è libero da rumore
  2. Il concetto obiettivo c è contenuto nello spazio delle ipotesi H (c è consistente con H)
  3. Non abbiamo ragioni a priori di credere che un'ipotesi sia più probabile di un'altra
- Che valore scegliere per P(h)? Per la (3) assegniamo la stessa probabilità a priori a ogni ipotesi h in H:

$$P(h) = \frac{1}{|H|} \qquad \forall h \in H$$

MAP Learning (2)

- Che valore scegliere per P(D|h)?
- P(D|h) è la probabilità di osservare i valori d_1, ..., d_m per l'insieme di istanze x_1, ..., x_m, dato un mondo in cui *h* è la corretta descrizione del concetto *c*
- Per la (1), la probabilità di osservare gli esempi (x_i, d_i) dato *h* è 1 se $d_i = h(x_i)$ per ogni esempio x_i e 0 altrimenti:

$$P(D \mid h) = \begin{cases} 1 & \text{se } d_i = h(x_i) \text{ per ogni } d_i \in D \\ 0 & \textit{altrimenti} \end{cases}$$

MAP Learning (3)

- Che valore scegliere per P(D)?

$$P(D) = \sum_{h_i \in H} P(D \mid h_i) P(h_i) = \sum_{h_i \in VS_{H,D}} 1 \cdot \frac{1}{|H|} + \sum_{h_i \notin VS_{H,D}} 0 \cdot \frac{1}{|H|}$$

$$= \sum_{h_i \in VS_{H,D}} 1 \cdot \frac{1}{|H|} = \frac{|VS_{H,D}|}{|H|}$$

- dove $VS_{H,D}$ è lo spazio delle versioni di H rispetto a D (cioè il sottoinsieme delle ipotesi di H consistenti con D)

MAP Learning (4)

- Quindi, se h è inconsistente con D:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} = \frac{0 \cdot P(h)}{P(D)} = 0$$

- Se h è consistente con D:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} = \frac{1 \cdot \frac{1}{|H|}}{P(D)} = \frac{1 \cdot \frac{1}{|H|}}{\frac{|VS_{H,D}|}{|H|}} = \frac{1}{|VS_{H,D}|}$$

MAP Learning (5)

- Per concludere, il teorema di Bayes implica che la probabilità a posteriori P(h|D) date le nostre assunzioni di distribuzione uniforme delle ipotesi h su H (per calcolare P(h)) e assenza d'errore (per P(D|h)) è:

$$P(h \mid D) = \begin{cases} \dfrac{1}{|VS_{H,D}|} & \text{se } h \text{ è consistente con D} \\ 0 & \text{altrimenti} \end{cases}$$

- Ovvero qualsiasi ipotesi consistente h appresa da un apprendista ha probabilità a posteriori $1/|VS_{H,D}|$, ovvero è una ipotesi MAP
- Il teorema di Bayes **ci aiuta a caratterizzare le assunzioni implicite in un modello di apprendimento** (ad es. Version Space), sotto le quali il modello si comporta in maniera ottima.

Maximum Likelihood learning

- Supponiamo di trovarci in una situazione più complessa, e ***più realistica:***
  - Dobbiamo apprendere una funzione obiettivo c che assume valori in $\Re$, nel continuo
  - Il campione di addestramento **produce errori**, cioè: D = { (x_i,d_i)} $d_i=c(x_i)+e_i$, dove e_i è una variabile aleatoria estratta indipendentemente per ogni x_i secondo una distribuzione *gaussiana* con media zero (**errore**)
- Quale è **l'ipotesi massimamente probabile** (ML)?
- Questa situazione è tipica di molti metodi di apprendimento, come i metodi basati su *reti neurali, regressioni lineari, interpolazione di polinomi* (metodi algebrici)

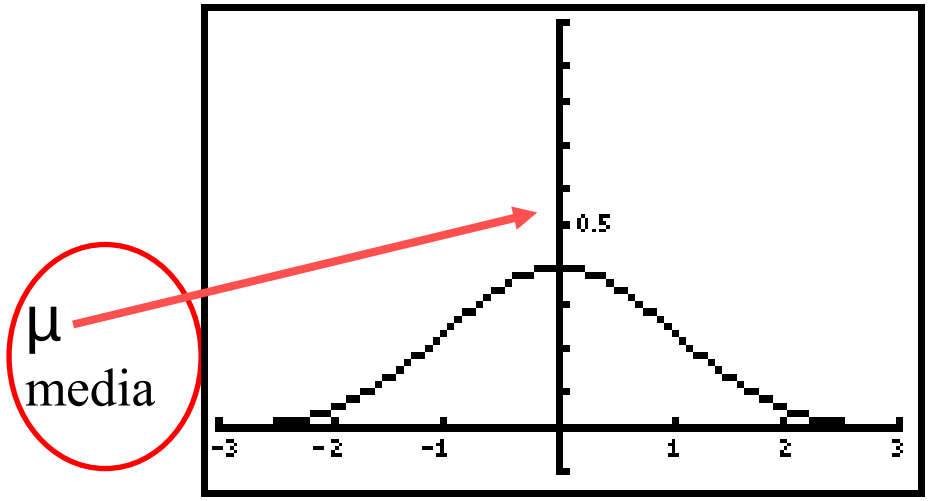
Distribuzione gaussiana dell'errore

- L'errore agisce sulla funzione di classificazione generalmente casuale
- La distribuzione gaussiana (introdotta precedentemente) ci aiuta a rappresentare la densità di probabilità della variabile aleatoria *e*

Cosa è una distribuzione gaussiana?

- Quando molti fattori *casuali ed indipendenti* agiscono in modo **additivo** per creare fattori di variabilità, i dati seguono un andamento "a campana" chiamato **distribuzione gaussiana**, o anche distribuzione **normale**. Molti dati seguono una distribuzione che approssima la distribuzione Gaussiana (e le sue proprietà matematiche)

$$p(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$



Ipotesi ML con rumore Gaussiano

densità di probabilità (variabile aleatoria continua!!)

$$h_{ML} = \arg\max_{h \in H} p(D \mid h) = \arg\max_{h \in H} \prod_{i=1}^{m} p(d_i \mid h) =$$

$$\arg\max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2}$$

$\mu = c(x_i)$

L'errore segue una distribuzione gaussiana, quindi anche i valori d_i si distribuiscono secondo una gaussiana, con scostamenti centrati attorno al valore "reale" $c(x_i)$ $\quad d_i = c(x_i) + e_i$

Dato che gli esempi sono estratti in modo indipendente, $p(di \wedge dj) = p(di)p(dj)$

Poiché stiamo esprimendo la probabilità di di condizionata all'essere h(x) una ipotesi corretta, avremo $\mu = c(x) = h(x)$

ML (2)

- Anziché massimizzare l'espressione precedente, massimizzianone il logaritmo

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \left(\sum_{i=1}^{m} \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2} \right) \right) = \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2 \right) =$$

$$= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} -\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2 =$$

$$= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} -\left(d_i - h(x_i)\right)^2 = \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{m} \left(d_i - h(x_i)\right)^2$$

questi fattori sono uguali per tutte le h_i non influenza argmax

ML (3)

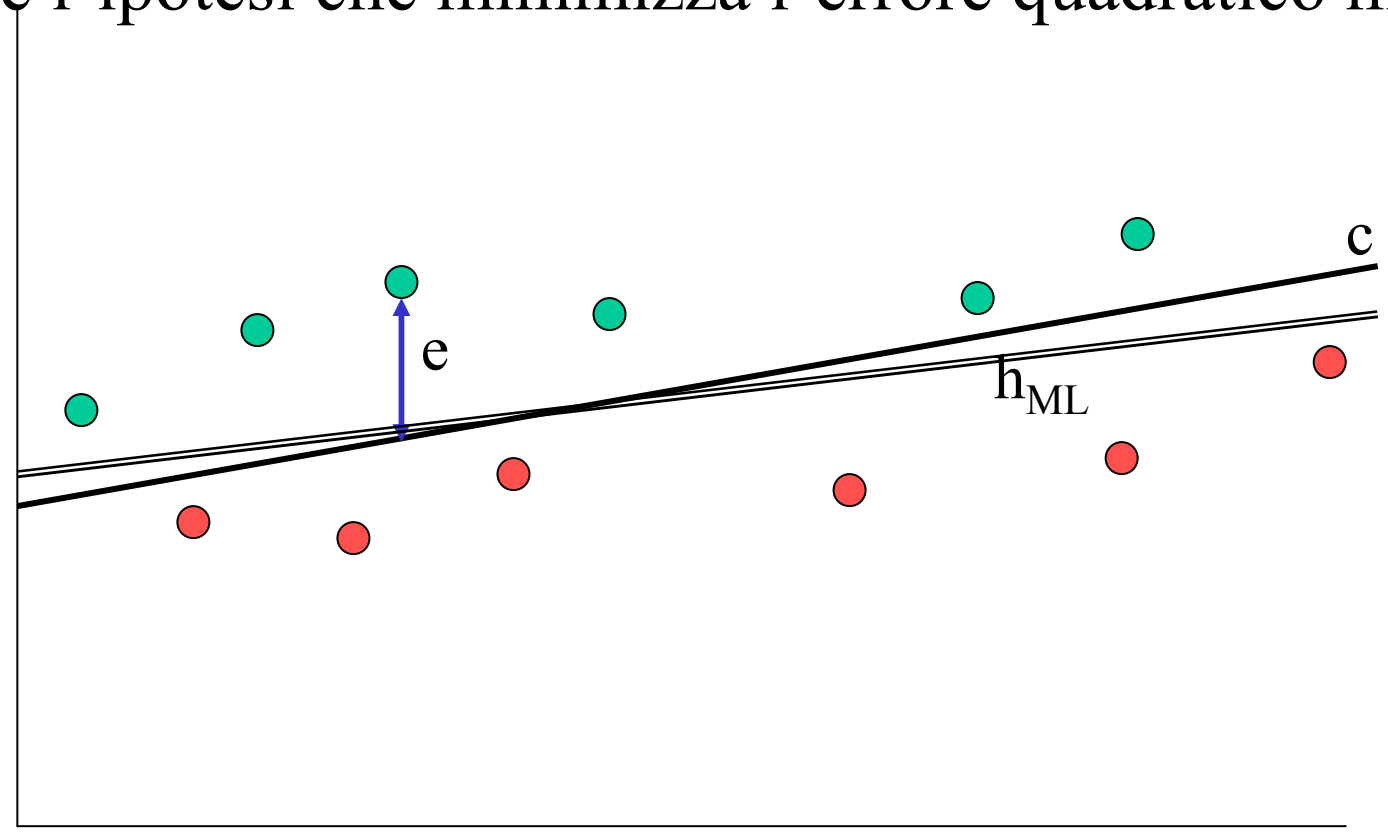
- Dunque, l'ipotesi massimamente probabile h_{ML} è quella che **minimizza la somma degli errori quadratici (dell'ipotesi stessa):** $d_i = c(x_i) + e_i = h(x_i) + e_i$

$$h_{ML} = \arg\min_{h \in H} \prod_{i=1}^{m} (d_i - h(x_i))^2 = \arg\min_{h \in H} \prod_{i=1}^{m} e_i^2$$

- L'ipotesi sottostante è: esempi in D estratti indipendentemente, distribuzione gaussiana dell'errore
- Problema: considera solo errori di classificazione ($c(x_i)$), non errori nei valori degli attributi degli x_i in D

Una spiegazione intuitiva

c(x) è la funzione da apprendere, gli esempi d_i sono rumorosi con distribuzione del rumore e_i gaussiana (uniformemente distribuio attorno al valore reale).
h_{ML} è l'ipotesi che minimizza l'errore quadratico medio



ML e MAP servono a
caratterizzare in termini di
probabilità il problema
dell'apprendimento
di una ipotesi, ovvero:
"Qual è l'ipotesi più probabile dato
l'insieme d'addestramento D?"

Ma qual è la classificazione più probabile?

Supponiamo di avere la seguente situazione:

$$C = \{+,-\},\ H = \{h_1, h_2, h_3\}$$

$$P(h_1 \mid D) = 0{,}4\ \ P(h_2 \mid D) = 0{,}3\ \ P(h_3 \mid D) = 0{,}3$$

(% degli esempi che sono consistenti con h_i)

$h_{ML} = h_1$. Supponiamo che h_1 classifichi il prossimo esempio come positivo.

$$P(+ \mid h_1) = 0{,}6, P(- \mid h_1) = 0{,}4$$

$$P(+ \mid h_2) = 0{,}3, P(- \mid h_2) = 0{,}7$$

$$P(+ \mid h_3) = 0{,}4, P(- \mid h_3) = 0{,}6$$

$$\sum_{h_i \in H} P(+ \mid h_i) P(h_i \mid D) = 0{,}24$$

$$\sum_{h_i \in H} P(- \mid h_i) P(h_i \mid D) = 0{,}76$$

$$\arg\max_{c_j \in C} \sum_{h_i \in H} P(c_j \mid h_i) P(h_i \mid D) = -$$

Se invece consideriamo tutte le ipotesi, pesate secondo le probabilità condizionate, la classe negativa è più probabile!

Optimal Bayes classifier

- Supponiamo che c(x) sia una *funzione obiettivo* discreta ed assuma valori in $C = \{ c_1, c_2, ..., c_m \}$
- Supponiamo che H sia lo spazio corrente delle ipotesi, D sia l'insieme di apprendimento, e $P(h_i|D)$ siano le probabilità **a posteriori** delle h_i dato D (calcolato come % dei casi in cui $h_i(x_j)=c(x_j)$) quindi non si richiede consistenza)
- Supponiamo che x_k sia una nuova istanza. Quale è la classificazione ottima $c_{opt} \in C$ per x_k ?

$$c_{opt} = \arg\max_{c_j \in C} P(c_j \mid D) = \sum_{h_i \in H} P(c_j \mid h_i) P(h_i \mid D)$$

- Si combinano le predizioni di tutte le ipotesi, pesate rispetto alla loro probabilità a posteriori.

Bayes Optimal classifier (conclusioni)

- Ottiene le migliori prestazioni ottenibili, assegnati i dati di apprendimento D, e uno spazio di ipotesi H
- **Costoso:** calcola la probabilità a posteriori per ogni ipotesi, e combina le predizioni per classificare ogni nuova istanza

Naïve Bayes Classifier

- Si applica al caso in cui le ipotesi in H sono rappresentabili mediante una **congiunzione di valori di attributi (k-monomi)**, e c(x) può assumere valori da un insieme finito C. Le istanze x in X sono descritte mediante tuple di valori $(a_1, a_2, ..., a_n)$ associati agli n attributi di x
- Il classificatore "naif" si basa sull'assunzione semplificativa che i valori degli attributi siano **condizionalmente indipendenti**, assegnato un valore della funzione obiettivo, cioè, dato un nuovo esempio x da classificare, calcoliamo:

$$c_{NB} = \underset{c_j \in C}{\text{argmax}}\ P(c_j \mid a_1, a_2, ..., a_n) = \underset{c_j \in C}{\text{argmax}}\ \frac{P(a_1, a_2, ..., a_n \mid c_j) P(c_j)}{P(a_1, a_2, ..., a_n)} =$$

$$\underset{c_j \in C}{\text{argmax}}\ P(c_j) \prod_i P(a_i \mid c_j)$$

Stima delle probabilità

- Le probabilità $P(a_i|c_j)$ vengono stimate osservando le frequenze nei dati di addestramento D
- Se D include n_i esempi classificati c_i, e n_{ij} di questi n_i esempi contengono il valore a_j per l'attributo j, allora:

$$P(a_j \mid c_i) = \frac{n_{ij}}{n_i}$$

Naïve Bayes: Esempio

- C = {allergia, raffreddore, in_salute} (**valori c(x))**
- att_1 = starnuti (sì, no) ; att_2 = tosse (sì, no) ; att_3 = febbre (sì, no) (attributi booleani)
- x = (1, 1, 0) ovvero (att_1, att_2, ¬ att_3) come lo classifico?

Dall'insieme D stimo le prob. a priori e condizionate es:

Prob	**in salute**	**raffreddore**	**allergia**
$P(c_i)$	0.9	0.05	0.05
$P(a_1 \mid c_i)$	0.027	1.0	1.0
$P(a_2 \mid c_i)$	0.027	0.5	0.5
$P(a_3 \mid c_i)$	0.027	0.5	0.5

Esempio (continua)

- 40 esempi, 36 classificati “in salute”, 2 raffreddore, 2 allergia
- Per stimare, ad esempio, $P(a_1=1|\text{in-salute})$, contare sui 36 esempi nei quali c(x)= “in-salute” quanti hanno $att_1=1$

  se 1 su 36, $P(att_1=1|\text{in-salute})=1/36=0,027$

Analogamente avrò, ad es.:

- $P(att_1=1|\text{raffreddore})=2/2=1$
- $P(att_1=1|\text{allergia})=2/2=1$
- ecc.

Esempio (continua)

- Devo calcolare il massimo al variare di c di: $P(c_j)\prod_i P(a_i \mid c_j)$
- Quindi ad esempio per c=raffreddore

$$P(raffreddo\ e)\left[P(att_1 = sì \mid raff\ )P(att_2 = sì \mid raff\ )P(att_3 = no \mid raff\ )\right] =$$
$$0{,}05 \times \left[1 \times 0{,}5 \times 0{,}5\right] = 0{,}0125$$

- Analogamente, troverò:

$$P(in-salute)\left[P(att_1 = sì \mid sal)P(att_2 = sì \mid sal)P(att_3 = no \mid sal)\right] =$$
$$0{,}9 \times \left[0.027 \times 0{,}027 \times 0{,}027\right] = 0{,}000017$$
$$P(allergia)\left[P(att_1 = sì \mid all)P(att_2 = sì \mid all)P(att_3 = no \mid all)\right] =$$
$$0{,}05 \times \left[1 \times 0{,}5 \times 0{,}5\right] = 0{,}0125$$

Problemi con Naive Bayes

- Se D è piccolo, le stime sono inaffidabili (nell'esempio precedente alcune stime sono = 1!!!).
- Un valore raro a_k può non capitare mai in D e dunque:
  - $\forall c_j$: $\quad P(a_k \mid c_j) = 0$.
- Analogamente, se ho un solo esempio di una classe c_j,
  - $\forall\ a_k$: $\ P(a_k \mid c_j) = 1$ o $\ P(a_k \mid c_j) = 0$.
- Se a_k capita in un test set T, il risultato è che
  - $\forall c_i$: $\quad P(T \mid c_i) = 0$ and $\forall c_i$: $P(c_i \mid T) = 0$

Smoothing

- Per tener conto di eventi rari, si operano degli aggiustamenti sulle probabilità detti **smoothing**
- *Laplace smoothing con una m-stima* assume che ogni evento a_j (ovvero $att_j = a_j$) abbia una probabilità a priori p, che si assume essere stata osservata in un campione virtuale di dimensione m > del campione reale

$$P(a_j \mid c_i) = \frac{n_{ij} + mp}{n_i + m}$$

- Nell'esempio precedente, ad es.

$$P(att_1 = 0 \mid raff) = \frac{0 + m \times 0{,}05}{2 + m}$$

- m è una costante che determina il peso dello smoothing
- In assenza di altre informazioni, si assume p = 1/k dove k è il numero di valori dell'attributo j in esame

Un esempio

- **Classificazione automatica di documenti**
  - Un documento rappresentato come un elenco di termini t_j (j=1,…,|V|), dove V è il vocabolario
  - Rappresentiamo un documento x con il vettore x = (a_1,a_2…,$a_{|V|}$) dove a_j=1 se il termine t_j è presente (0 altrimenti)
  - D = { (x_i, c(x_i)) } insieme di addestramento di documenti già classificati
  - c : X $\rightarrow$ { sport, politica, scienze, ... }
  - Stimare, sulla base di D, le P(t_j|c_k)