Cenni di Teoria della Probabilità

Marco Baroni

21 febbraio 2005

1 Le leggi fondamentali della probabilità

1. Dato un *universo* di eventi elementari, le probabilità sono dei valori numerici che assegniamo a ciascun evento.
2. Ciascuna probabilità ha un valore tra 0 e 1.
3. La somma delle probabilità di tutti gli eventi elementari nell'universo che stiamo considerando è 1 ($\sum_i P(i) = 1$).

- Implicitamente, facciamo riferimento a queste leggi nella vita quotidiana.
- Per esempio, in una partita di calcio, gli eventi dell'universo in considerazione possono essere:
  1. La squadra A vince.
  2. La squadra B vince.
  3. Le due squadre pareggiano.
- La probabilità che A vinca non può essere minore di 0 ne maggiore di 1: diciamo, e.g., che la squadra ha una probabilità del 50% di vincere, ma non ha senso dire che ha una probabilità del -50% di vincere, o del 150%.
- Dal terzo principio, ricaviamo che se la squadra A ha una probabilità del 50% di vincere e la squadra B ha una probabilità del 30% di vincere, la probabilità di pareggio è del: ...
- Qual'è la probabilità che esca un 3 se hai appena tirato un dado non truccato?
- Sei eventi, con la stessa probabilità (il dado non è truccato):

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6)$$

- Per la terza legge, la somma di queste sei probabilità identiche deve essere 1, dunque ciascuna delle probabilità (inclusa quella che esca un 3) deve essere: ...

2 Probabilità di insiemi di eventi

La probabilità di un insieme di eventi è data dalla somma delle probabilità degli eventi nell'insieme.

- Che probabilità c'è che esca un numero dispari nel tiro di un dado non truccato?
- "Dispari" è etichetta che identifica insieme di 3 eventi elelmentari: che esca un 1, che esca un 3, che esca un 5.
- Ciascuno di questi eventi ha probabilità 1/6.
- Sommando, probabilità che esca numero dispari è 3/6 (50%), come secondo intuizione.

3 Probabilità di eventi indipendenti

La probabilità di due (o più) eventi *indipendenti* è data dal *prodotto* della probabilità degli eventi in questione.

- Tiriamo dado non truccato due volte. Che probabilità c'è che il numero che esce sia dispari in entrambi i casi?
- Probabilità che primo tiro sia dispari è del 50%.
- *All'interno* di questo 50%, probabilità che secondo tiro sia dispari è anche del 50%.
- Dunque, probabilità che entrambi i lanci siano dispari è 50% di 50%, ovvero 25%.
- Matematicamente, dire "x di y" equivale a *moltiplicare* x per y; infatti: $1/2 \times 1/2 = 1/4 = 25\%$.
- Questo vale *solo* per eventi indipendenti.
- La probabilità che il numero uscito in un lancio sia 3 e che sia un numero dispari *non* è $1/6 \times 1/2 = 1/12$, poiché i due eventi *non* sono indipendenti.

4 Probabilità condizionali

$$P(A|B) = \frac{P(A,B)}{P(B)} \tag{1}$$

- Notazione:
  1. $P(A|B)$: probabilità di A dato B.
  2. $P(A,B)$: probabiltà di A e B.
- In parole: la probabilità di A *dato* B è uguale alla probabilità di A e B divisa per la probabilità di B.
- Probabilità che il numero uscito in un lancio sia 3 dato che si tratta di un lancio risultato in numero dispari:
  - $P(\texttt{dispari}) = 1/2$
  - $P(\texttt{3}, \texttt{dispari}) = 1/6$
  - $P(\texttt{3}|\texttt{dispari}) = \frac{1/6}{1/2} = 1/3$
- Intuizione: proporzione di probabilità di A e B rispetto alla "massa" totale di probabilità di B (ci torniamo su in sezione 6.1).
- Calcolate:
  - Probabilità che il numero risultante dal lancio di un dado non truccato sia dispari e minore di quattro;
  - Probabilità che il numero sia dispari *dato che* è minore di quattro.

5 La legge di Bayes e altre formule utili

- Partiamo da equazione (1).
- Moltiplicando entrambi i lati per $P(B)$ (e invertendo destra e sinistra) otteniamo:
$$P(A,B) = P(A|B)P(B) \tag{2}$$
- Con questa formula possiamo calcolare la probabilità che A e B capitino insieme se conosciamo la probabilità di A dato B e la probabilità di B.
- Allo stesso modo possiamo derivare:
$$P(B,A) = P(B|A)P(A)$$
- $P(A,B)$ e $P(B,A)$ sono la stessa cosa, dunque:
$$P(A,B) = P(B|A)P(A) \tag{3}$$

- L'espressione a sinistra dell'equazione (2) e quella a sinistra dell'equazione (3) sono uguali, dunque:

$$P(A|B)P(B) = P(B|A)P(A)$$

- Dividendo entrambi i lati per $P(B)$ otteniamo la *legge di Bayes*:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{4}$$

- Equazione di importanza fondamentale, perché ci permette di mettere in relazione $P(A|B)$ e $P(B|A)$ (se conosciamo le probabilità $P(A)$ e $P(B)$).
- Infine, possiamo provare che la probabilità di due eventi indipendenti sia uguale al prodotto delle loro probabilità (come discusso sopra in sezione 3).
- Se A è indipendente da B la probabilità di A non cambia se B è dato, ovvero:

$$P(A|B) = P(A) \tag{5}$$

- Dunque, possiamo rimpiazzare $P(A|B)$ con $P(A)$ nell'equazione (2), ottenendo:

$$P(A, B) = P(A)P(B) \tag{6}$$

- Ma 6 ci dice che la probabilità che si verifichino due eventi indipendenti è uguale al prodotto delle probabilità dei due eventi, e questo era proprio ciò che avevamo concluso, sulla base dell'intuizione, nella sezione 3!

6 Stima delle probabilità

- Salvo nei casi più ovvi (dadi non truccati e simili) non conosciamo la probabilità degli eventi a cui siamo interessati.
- Un metodo intuitivamente motivato di *stimare* probabilità: usare la frequenza relativa:

$$P(x) = \frac{fq(x)}{N} \tag{7}$$

- N è il numero totale di casi che abbiamo a disposizione nell'analisi e $fq(x)$ è il numero di casi in cui l'evento x si è verificato.
- Per es., in inizio di frase è più probabile la parola *Il* o la parola *Adelfo*?
- Usando l'equazione (7):
  - Numero di frasi nel corpus *La Repubblica*: 15,344,780
  - Numero di frasi che iniziano con *Il*: 866,760

  - Numero di frasi che iniziano con *Adelfo*: 1
  - Probabilità di *Il*:

$$P(\texttt{Il}) = \frac{fq(\texttt{Il})}{N} = \frac{866760}{15344780} = .056$$

  - Probabilità di *Adelfo*:

$$P(\texttt{Adelfo}) = \frac{fq(\texttt{Adelfo})}{N} = \frac{1}{15344780} = .00000006$$

- Problema con stima basata su frequenze relative: parole, sequenze di parole che non capitano mai in corpus hanno frequenza 0; probabilità di parole/sequenze di parole con frequenza bassa è poco stabile.
- Problemi non da poco, vista proprietà zipfiana di frequenze!

6.1 Stima di probabilità condizionali

- Abbiamo visto in sezione 4 come calcolare probabilità di A dato B.
- La formula per calcolare questa probabilità diventa più intuitiva quando la interpretiamo in termini di frequenze relative.
- Infatti, con metodo di frequenze relative, probabilità di A dato B diventa:

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{fq(A,B)/N}{fq(B)/N} = \frac{fq(A,B)}{fq(B)}$$

- Questo dovrebbe essere abbastanza intuitivo: la probabilità che capiti A se sappiamo che è capitato B è data dalla proporzione di volte in cui è capitato anche A su tutte le volte che è capitato B.

7 Un'applicazione pratica: indipendenza e collocazioni

- Se due parole sono indipendenti, per via di equazione (6) dobbiamo avere che:

$$P(w_1, w_2) = P(w_1)P(w_2)$$

- Ovvero

$$\frac{P(w_1, w_2)}{P(w_1)P(w_2)} = 1 \tag{8}$$

- Se valore dato da questo rapporto è lontano da 1, vuol dire che parole *non* sono indipendenti (e dunque è probabile che formino frasi fatte, nomi propri, collocazioni, termini...)

- In pratica, calcoliamo:

$$\frac{P(w_1,w_2)}{P(w_1)P(w_2)} = \frac{\frac{fq(w_1,w_2)}{N}}{\frac{fq(w_1)}{N}\frac{fq(w_2)}{N}} = \frac{fq(w_1,w_2)}{N} \times \frac{N^2}{fq(w_1)fq(w_2)} = \frac{fq(w_1,w_2)N}{fq(w_1)fq(w_2)} \quad (9)$$

- Calcolare questa formula per le coppie *fatto che*, *Hong Kong*, *gelato alla*, *conoscere Andreotti* sulla base di dati da *La Repubblica* dalla seguente tabella:

coppia	$fq(w_1,w_2)$	$fq(w_1)$	$fq(w_2)$	N
fatto che	54986	388633	6336856	379415868
Hong Kong	4685	4963	5044	379415868
gelato alla	31	1469	1161521	379415868
conoscere Andreotti	1	17238	38368	379415868

- Risultati:

coppia	eq (9)
fatto che	8.5
Hong Kong	71007.7
gelato alla	6.9
conoscere Andreotti	0.6

- I risultati corrispondono alle vostre intuizioni sul livello di lessicalizzazione di queste sequenze?