Probabilità: il minimo indispensabile

Marco Baroni

7 ottobre 2004

1 Introduzione

In questi appunti, cerco di fornire in maniera informale un'infarinatura minima di teoria della probabilità che dovrebbe bastare per capire (o almeno avere un'intuizione su) gran parte dei materiali trattati nel mio corso di Linguistica Computazionale che si rifanno a tale teoria, e a fare le letture.

2 Universi di eventi, probabilità e distribuzioni

Quando parliamo di probabilità nella vita di tutti i giorni seguiamo, implicitamente, le stesse regole formalizzate nella teoria della probabilità. Acquisire le basi di teoria della probabilità consiste dunque in gran parte in una riflessione esplicita su cose che, fondamentalmente, sappiamo già, ma normalmente sono nascoste nelle acque torbide delle lingue naturali.

Fortebraccio e Serafina stanno giocando a scacchi. Un esperto ben informato potrebbe dire: Fortebraccio ha una probabilità di vincere del 60 per cento.[1]

Quando facciamo un'affermazione del genere, stiamo obbedendo a certe "regole".

Primo, esprimiamo la probabilità di un evento in termini di un numero che assegniamo a quell'evento: in questo caso, assegniamo il numero *60 per cento* all'evento che Fortebraccio vinca la partita.

Secondo, la probabilità di un evento ha senso soltanto in relazione ad un particolare "universo" di eventi elementari. Nel nostro esempio, stiamo probabilmente considerando un universo con due eventi elementari: l'evento che vinca Fortebraccio e l'evento che vinca Serafina.

Terzo, la probabilità ha sempre un valore tra 0 e 1: 60 per cento vuol dire 60/100, cioè .6.[2] Il valore massimo è 100 per cento, cioè 100/100, cioè 1. Il valore minimo è 0 per cento, cioè 0/100, cioè 0.[3]

[1] Un esempio più naturale potrebbe riguardare la probabilità di sopravvivenza di una persona con una certa malattia, ma poi va a finire che toccate ferro quando mi incontrate per la strada e simili.

[2] Seguo convenzioni anglosassoni nella notazione dei numeri. Dunque, 100,000 vuol dire centomila, 1.233 vuol dire uno e duecentotrentatre millesimi e .6 vuol dire sei decimi.

[3] Quando uno dice di essere sicuro al 110 per cento o che le possibilità di vincere di For-

Quarto, la somma delle probabilità di tutti gli eventi elementari che stiamo considerando è sempre 1. Usiamo, implicitamente, questa regola quando dal fatto che Fortebraccio ha una probabilità del 60 per cento di vincere deduciamo che Serafina ha una probabilità del 40 per cento di vincere. Infatti, abbiamo un universo con due eventi (vince Fortebraccio vs. vince Serafina). Il primo evento ha una probabilità di .6. Siccome la somma delle probabilità di tutti gli eventi deve essere 1, per calcolare la probabilità che vinca Serafina facciamo $1 - .6$ e otteniamo .4, che espresso in percentuale è uguale al 40 per cento.

Notate che il fatto che la probabilità complessiva sia sempre 1 (ossia, 100 per cento) ha senso. La probabilità totale altro non è che la probabilità che almeno uno degli eventi nel nostro universo abbia luogo, e almeno uno di tali eventi *deve* aver luogo!

Nel nostro caso, la probabilità complessiva è la probabilità che o Fortebraccio vinca o Serafina vinca: è ovvio che uno di questi due eventi debba avvenire, e dunque la probabilità complessiva è di 1 (100%).

Sapere quanti e quali siano gli eventi nell'universo di cui stiamo parlando è fondamentale. Se io non sapessi che a scacchi ci si gioca in due, non potrei dedurre che Serafina ha una probabilità di vincere del 40 per cento dal fatto che Fortebraccio ha una probabilità di vincere del 60 per cento.

Inoltre, finora abbiamo assunto un universo senza pareggi: se consideriamo tre eventi elementari (vince Fortebraccio, vince Serafina, i due pareggiano), di nuovo non basta sapere che la vittoria di Fortebraccio ha probabilità .6 per dedurre che la vittoria di Serafina ha probabilità .4.

Sorprendentemente, in questa discussione abbiamo già enunciato le leggi fondamentali della teoria della probabilità, che possiamo riassumere come segue:

Le leggi fondamentali della probabilità

- Dato un *universo* di eventi elementari, le probabilità sono dei valori numerici che assegniamo a ciascun evento.
- Ciascuna probabilità ha un valore tra 0 e 1.
- La somma delle probabilità di tutti gli eventi elementari nell'universo che stiamo considerando è 1.

Quest'ultima condizione si può esprimere così:

$$\sum_i P(i) = 1$$

La parte sinistra di questa formula si legge così: somma le probabilità di tutti gli eventi i nell'universo che stai considerando.

tebraccio sono del meno 100 per cento sta facendo lo spiritoso. Una cosa più importante: ricordatevi che *50 per cento*, *1/2* e *.5* sono tre modi diversi di riferirsi allo stesso numero. Nella teoria della probabilità la forma percentuale non viene usata quasi mai, ma se all'inizio vi fa comodo pensare in questi termini potete farlo.

Il simbolo di sommatoria, nella mia esperienza, intimorisce chi ha poca familiarità con la matematica, ma, come questo esempio dovrebbe mostrare, dal punto di vista matematico corrisponde in realtà ad un processo elementare – la somma di un certo numero di valori.

Nella letteratura, si trovano diverse variazioni notazionali, ma l'idea di base è sempre la medesima: la sommatoria somma i valori in un range specificato.

Dato un universo, la lista di probabilità assegnate a ciascun evento si chiama la *distribuzione* di probabilità per l'universo in questione.

Per esempio, nell'universo di cui sopra abbiamo la seguente distribuzione:

- Probabilità che vinca Fortebraccio: .6
- Probabilità che vinca Serafina: .4

In quasi tutte le trattazioni elementari di teoria della probabilità prima o poi compaiono i dadi, e io non sarò da meno...

Qual è la probabilità che esca un 3 se hai appena tirato un dado non truccato?

Allora, qui gli eventi sono sei: che esca 1, che esca 2, che esca 3, che esca 4, che esca 5 e che esca 6.

Il fatto che il dado non sia truccato significa che ciascun evento ha la stessa probabilità:

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

E sappiamo che la somma di queste probabilità è 1:

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

Mettendo queste due condizioni assieme,[4] otteniamo:

$$P(3) + P(3) + P(3) + P(3) + P(3) + P(3) = 1$$

O, più in breve:

$$6P(3) = 1$$

Dunque:

$$P(3) = \frac{1}{6} \approx .166$$

E, più in generale:

$$P(n) = \frac{1}{6} \approx .166$$

Quindi, la probabilità che esca un 3 (o un 4, o un 1...) è più o meno di .166.

[4]Notate che la prima condizione ci permette di rimpiazzare qualsiasi $P(n)$ con $P(3)$.

Questo dovrebbe essere intuitivo, e lo possiamo generalizzare: se ci sono N eventi che potrebbero capitare, e ciascuno di essi ha la stessa probabilità, la probabilità che uno di essi capiti è data da $1/N$.

Chiamiamo una distribuzione di questo genere, dove tutti gli eventi di base hanno la stessa probabilità e dunque la probabilità di un evento è data da $1/N$, una *distribuzione uniforme*.

3 Probabilità di insiemi di eventi

Molto spesso, piuttosto che alla probabilità degli eventi elementari, siamo interessati alla probabilità di insiemi di eventi.

Per esempio, è ovvio che la probabilità che esca un numero dispari con un dado non truccato è di .5 (50%). Ma come spieghiamo questo, nel nostro modello?

"Dispari" è semplicemente un'etichetta che assegniamo a un insieme di eventi elementari. Nel nostro universo, ci sono tre eventi che classifichiamo come dispari: che esca un 1, che esca un 3 e che esca un 5.

La probabilità che il tiro sia dispari è la probabilità che si verifichi uno dei tre eventi nell'insieme degli eventi dispari, ossia la probabilità che esca un 1 o un 3 o un 5.

Siccome ciascuno di questi eventi elementari ha una probabilità di 1/6, la probabilità complessiva che il tiro sia dispari sarà di $1/6 + 1/6 + 1/6$, ossia 3/6, ossia .5, che è quello che già sapevamo.

Generalizzando, se conosciamo la distribuzione di probabilità per gli eventi elementari del nostro universo, possiamo assegnare una probabilità anche ad un *insieme* di eventi elementari (ossia una classe di eventi elementari che decidiamo di raggruppare, di solito perché hanno una certa proprietà in comune). Tale probabilità sarà uguale alla somma delle probabilità degli eventi elementari che rientrano nell'insieme.[5]

La probabilità che capiti un evento qualsiasi nell'insieme di eventi $\{a, b, c\}$ è data dalla somma della probabilità che capiti a più la probabilità che capiti b più la probabilità che capiti c.

Si tratta di un fatto abbastanza interessante da metterlo in un box:

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

In realtà, avevamo già introdotto questo concetto quando abbiamo detto che la probabilità complessiva di tutti gli eventi elementari (che è sempre 1 e che rappresenta la probabilità che *qualcosa* capiti) è la somma delle probabilità di

[5] Attenti: in italiano "probabilità di un insieme di eventi" può essere interpretato in due modi: la probabilità che si verifichi un evento nell'insieme in questione – questo è il senso giusto in questo contesto – e la probabilità che si verifichino tutti gli eventi in un certo insieme. Questo secondo senso non è rilevante qui, visto che per ora stiamo parlando di eventi tra loro mutuamente esclusivi (se esce un 1 non può uscire un 3).

tutti gli eventi di base – in questo caso, calcoliamo la probabilità dell'insieme che contiene tutti gli eventi elementari.

4 Probabilità di eventi indipendenti

Spesso siamo interessati alla probabilità di due eventi indipendenti tra loro.

Per esempio, potremmo essere interessati a cosa succede quando lancio un dado non truccato due volte. Qual è la probabilità che il numero che esce sia dispari in entrambi i casi?

Abbiamo già visto che la probabilità che esca un numero dispari nel primo lancio è di .5. *All'interno* di questo .5 di probabilità che il primo lancio risulti in un numero dispari, avremo un .5 di probabilità che il secondo lancio sia pari.

Siccome .5 di .5 è uguale a .25, la probabilità che entrambi i lanci risultino in numeri dispari è di .25.[6]

Abbiamo appena usato l'espressione .5 *di* .5. Matematicamente, dire x *di* y significa moltiplicare x per y (un mezzo di venti significa un mezzo per venti, il doppio di 10 significa due per 20, ecc.) Infatti, abbiamo concluso che la probabilità di due tiri dispari è .25, cioè .5 per .5.

Ecco dunque un altro concetto che si merita il box:

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

Nella sezione 6 deriveremo matematicamente questa regola da altre proprietà delle probabilità di eventi indipendenti.

Nel caso di due eventi indipendenti, la probabilità che entrambi gli eventi capitino è dunque sempre minore o uguale alla probabilità che soltanto uno di essi capiti. Questo è vero perché moltiplichiamo due numeri che sono per definizione tra 0 e 1, ma ha anche senso dal punto di vista intuitivo: se due eventi non hanno nulla a che fare l'uno con l'altro, la probabilità che capitino entrambi non può essere maggiore della probabilità che ne capiti uno!

Dal punto di vista metodologico, è interessante osservare che regole come quella appena esposta ci permettono di calcolare probabilità in un universo più complesso sulla base di probabilità calcolate in un universo più semplice.

Per esempio, l'universo in cui ha senso l'affermazione che la probabilità di due tiri dispari è uguale a .25 è un universo dato da 36 eventi elementari (6 possibili esiti per il primo lancio, e, per ciascuno di questi, 6 possibili esiti per il secondo lancio: $6 \times 6 = 36$): 1 e 1, 1 e 2, 1 e 3... 6 e 4, 6 e 5, 6 e 6.

Infatti, avremmo potuto calcolare la stessa probabilità osservando che in questo universo ci sono 9 eventi in cui entrambi i numeri sono dispari (1 e 1, 1 e 3, 1 e 5, 2 e 1, 2 e 3, 2 e 5, 5 e 1, 5 e 3, 5 e 5). Siccome ciascun evento in questo

[6] Quanto appena detto magari risulta più chiaro parlando in percentuali: il primo lancio ha una probabilità del 50% di risultare in un numero dispari e, all'interno di questo 50%, la probabilità di un secondo lancio dispari è del 50%. Il 50% di 50% è uguale a 25%, e dunque la probabilità complessiva di due lanci dispari è del 25%.

universo ha una probabilità di 1/36 (questo è chiaro, vero?), per calcolare la probabilità dell'insieme dei lanci interamente dispari facciamo:

$$9 \times \frac{1}{36} = \frac{9}{36} = .25$$

Lo stesso risultato ottenuto di sopra combinando, tramite moltiplicazione, probabilità calcolate assumendo due universi di 6 eventi ciascuno.

Man mano che l'universo analizzato diventa più complesso, diventa vieppiù importante essere in grado di scomporre il problema in termini di universi più semplici.

5 Probabilità condizionali

Tiriamo il solito dado. Viene un numero dispari. Qual è la probabilità che tale numero sia 3?

Ovviamente, una volta che sappiamo che il numero è dispari la probabilità che il numero sia 3 aumenta – l'informazione che il numero è dispari riduce lo spazio di eventi possibili da 6 a 3, e dunque la probabilità che venga un 3 passerà da 1/6 a 1/3, un bel guadagno!

Parliamo in questo caso di una *probabilità condizionale* – in questo caso, la probabilità che esca un 3 *a condizione che* esca un numero dispari, o, in altre parole, la probabilità che esca un 3 *dato che* sappiamo che è uscito (o deve uscire) un numero dispari.

Per calcolare la probabilità di A dato che B, calcoliamo la probabilità di A e B, e poi dividiamo per la probabilità di B.

Nel caso di una distribuzione uniforme, la probabilità di A e B si ottiene dividendo il numero di eventi che sono sia A che B per il numero complessivo di eventi.

Per esempio, la probabilità che esca un 3 e che esca un numero dispari è data dal numero di eventi (tiri del dado) nei quali esce un 3 ed esce un numero dispari (un solo evento) diviso per il numero di eventi totale (sei eventi):

$$P(3, dispari) = \frac{1}{6} \approx .166$$

La probabilità di B, cioè nel nostro caso la probabilità che il numero sia dispari è .5, come ben sappiamo:

$$P(dispari) = .5$$

Dunque, la probabilità condizionale sarà data da:

$$P(3|dispari) = \frac{P(3, dispari)}{P(dispari)} \approx \frac{.166}{.5} = .332$$

Questo valore corrisponde più o meno un terzo, e cioè a quello che avevamo calcolato sopra.

È bene fermarsi un attimo sulle formule che ho appena presentato.

Prima di tutto, ci sono questioni di notazione: per la probabilità che capitino sia A che B, abbiamo introdotto la notazione $P(A, B)$ (come in $P(3, dispari)$), ma a volte si trovano anche le notazioni $P(AB)$ e $P(A\&B)$.

Invece, la notazione standard per la probabilità condizionale è $P(A|B)$, dove è importante ricordarsi che si tratta della probabilità dell'evento *a sinistra* della barra dato l'evento a destra.[7]

Dal punto di vista sostanziale, è invece importante convincersi del perché la formula per le probabilità condizionali abbia senso:

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

Mi pare che questo concetto venga spiegato molto chiaramente da Goldsmith nel suo *Probability for Linguists*, che dunque qui cito in blocco:[8]

> Let the universe of outcomes be the 52 cards of a standard playing card deck. The probability of drawing any particular card is 1/52 (that's a uniform distribution). What if we restrict our attention to red cards? It might be the case, for example, that of the card drawn, we know it is red, and that's all we know about it; what is the probability now that it is the Queen of Hearts?
>
> The sub-universe consisting of the red cards has probability mass 0.5, and the Queen of Hearts lies within that sub-universe. So if we restrict our attention to the 26 outcomes that comprise the "red card sub-universe," we see that the sum total of the probability mass is only 0.5 (the sum of 26 red cards, each with 1/52 probability). In order to consider the sub-universe as having a distribution on it, we must divide each of the 1/52 in it by 0.5, the total probability of the sub-universe in the larger, complete universe. Hence the probability of the Queen of Hearts, given the Red Card sub-Universe (given means with the knowledge that the event that occurs is in that sub-universe), is 1/52 divided by 1/2, or 1/26.

Provate a calcolare la probabilità che il numero risultante dal lancio di un dado non truccato sia dispari e minore di quattro, e la probabilità che il numero sia dispari *dato che* è minore di quattro.

[7] Un modo per ricordarsi ciò è pensare che l'ordine in cui i due eventi si scrivono è quello in cui si leggerebbero: scriviamo $A|B$ e diciamo: "A dato B".

[8] In questo passo Goldsmith usa un concetto piuttosto utile che non avevo ancora introdotto, e cioè quello di "probability mass", *massa di probabilità*. L'idea è quella che si può pensare alla probabilità complessiva condivisa dagli eventi di un universo come ad una "massa", una sostanza. Per rendere le cose ancora più concrete, Goldsmith suggerisce di pensare a questa massa come ad una sostanza collosa di un kilo, che viene divisa tra un evento e l'altro.

6 La legge di Bayes e altre formule utili

Ricaviamo un po' di formule utili da quella che abbiamo appena visto e che qui ripeto:

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

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.

Dovrebbe essere ovvio che con lo stesso ragionamento usato per derivare 2 (sostituendo A con B e B con A nelle due formule sopra) possiamo derivare:

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

Dovrebbe essere anche ovvio che $P(A,B)$ e $P(B,A)$ sono la stessa cosa, e che dunque possiamo riscrivere l'equazione precedente come:

$$P(A,B) = P(B|A)P(A) \tag{3}$$

Siccome l'espressione a sinistra di 2 e quella a sinistra di 3 sono uguali, devono esserlo anche quelle a destra. Dunque, ricaviamo l'equazione:

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

Da cui, dividendo entrambi i lati per $P(B)$, otteniamo:

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

Questa è un'equazione di importanza fondamentale nella teoria della probabilità, e nella storia della probabilità e della statistica, dove è nota come *legge di Bayes*.

L'equazione in 4 ci permette di mettere in relazione $P(A|B)$ e $P(B|A)$ se conosciamo le probabilità indipendenti $P(A)$ e $P(B)$. Notate che questo è un risultato semplice da derivare, estremamente utile e non particolarmente intuitivo (prima di studiare probabilità, non avrei mai detto che dalla probabilità che nevichi se fanno 0 gradi si potesse in qualche modo derivare la probabilità che facciano 0 gradi se nevica).

Se approfondirete studi di linguistica computazionale e aree affini, vi troverete di fronte alla legge di Bayes in tutte le salse.

Concludiamo confermando matematicamente quanto intuito nella sezione 4 sulle probabilità di due eventi indipendenti.

Prima di tutto, osservate che se A è indipendente da B la probabilità di A non cambia se B è dato, ovvero:

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

Questo dovrebbe essere ovvio: per definizione, se l'evento A è indipendente dall'evento B, la probabilità di A è indipendente da B, e dunque il fatto che B abbia o non abbia luogo non ha effetti sulla probabilità di A.

Per fare un esempio concreto, mettiamo che tu sappia che la probabilità che il lancio di un certo dado risulti in un numero pari sia di .5. Se io ti dico che ho tirato una moneta ed è venuto testa (e tu sei una persona sensata), continui a pensare che la probabilità che il lancio del dado risulti in un numero pari sia di .5. Ovvero: $P(pari_{dado}) = P(pari_{dado}|testa_{moneta})$.

Bene. L'equazione 5 ci dice che, per due eventi indipendenti, ovunque vediamo $P(A|B)$ possiamo rimpiazzarlo con $P(A)$. In particolare, 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 4!

Tra l'altro, l'equazione 6 ci fornisce un test per verificare se due eventi sono indipendenti o no – se i due eventi sono indipendenti, la probabilità che i due eventi capitino insieme deve essere uguale al prodotto delle probabilità dei due eventi. Nella sezione 8.1 parleremo di un'applicazione linguistica di questo test.

7 Stima delle probabilità

Fino a qui, abbiamo considerato casi dove per qualche ragione conosciamo già la distribuzione delle probabilità degli eventi elementari (la partita a scacchi), o possiamo dedurla da principi generali. Per esempio, nel caso del dado non truccato gli eventi sono 6 e per definizione hanno tutti la stessa probabilità (è un dado non truccato!) Dunque, la probabilità di ciascun evento deve essere di 1/6.

Quando analizziamo fenomeni più complessi, come per esempio il linguaggio, è raro il caso in cui possiamo derivare le probabilità da principi generali e semplici definizioni di questo genere.

Considerate per esempio il seguente problema: qual è la parola più probabile in inizio di frase in italiano?

In questo caso, il nostro universo è dato dall'insieme degli eventi che una parola w dal lessico dell'italiano capiti all'inizio di una frase qualsiasi.

Sorvoliamo qui sul problema che il lessico dell'italiano in realtà è infinito[9] e

[9]Basta per esempio applicare la regola della formazione del diminutivo per generare un numero infinito di parole italiane: `Fortebraccio, Fortebraccino, Fortebraccinino, Fortebraccininino, Fortrebraccinininino` ecc.

consideriamolo come costituito dalle 880,111 parole distinte (*types*) che capitano nel corpus Repubblica/SSLMIT.[10]

Le parole che capitano in inizio di frase sono come le facce di un dado non truccato? Se lo fossero, e cioè se la distribuzione di probabilità nell'universo che stiamo considerando fosse uniforme, avremmo che ciascuna parola ha una probabilità di 1/880111 di capitare in inizio di frase, e dunque che ciascuna parola ha la stessa probabilità di capitare in inizio di frase.

Chiaramente, questo non è un modello molto plausibile dell'italiano – una parola come `Il` ha probabilità molto più alte di capitare in inizio di frase che la parola `Adelfo`.

Su cosa basiamo questa intuizione? Sul fatto che nei dati a nostra disposizione (tutte le frasi in italiano che abbiamo sentito) la parola `Il` capita in inizio di frase molto più di *frequente* che la parola `Adelfo`.[11]

Dunque, intuitivamente, stimiamo la probabilità degli eventi sulla base della loro frequenza relativa nei dati a nostra disposizione:

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

In questa equazione, N è il numero di istanze di cui i nostri dati sono composti, e $fq(x)$ è il numero di istanze in cui l'evento x si è verificato.

Siccome nessuno di noi tiene traccia dei dati che ci sarebbero utili per analisi linguistiche di questo genere (tipo quante volte abbiamo sentito una frase che iniziava con *Il* in vita nostra), useremo un corpus per ricavare stime delle probabilità degli eventi di interesse.

Nel corpus la Repubblica/SSLMIT ci sono 6,606,903 frasi, dunque $N = 6606903$. Di queste frasi, 412,062 iniziano con la parola `Il`, e soltanto una inizia con la parola `Adelfo`.

Dunque, usando il metodo di stima in 7, otteniamo:

$$P(Il) = \frac{fq(Il)}{N} = \frac{412062}{6606903} = .06236$$

$$P(Adelfo) = \frac{fq(Adelfo)}{N} = \frac{1}{6606903} = .0000001$$

Come da intuizione, la probabilità di `Il` in inizio frase è molto più alta della probabilità di `Adelfo` nel medesimo contesto.

[10] Un corpus è semplicemente una collezione di dati linguistici in un formato che ne renda possibile l'analisi computerizzata: per esempio, il corpus la Repubblica/SSLMIT, nella versione da cui ho estratto le frequenze usate in queste note, era una collezione di tutti gli articoli usciti sul quotidiano la Repubblica dal 1985 al 1992. Il corpus, che nel frattempo è cresciuto fino a contenere tutti gli articoli fino al 2000, è codificato in un formato elettronico che ne permette la rapida gestione tramite il programma di elaborazione e consultazione di corpora IMS Corpus WorkBench. Complessivamente, il corpus, nella versione usata qui, conteneva 175,239,348 parole (*tokens*). Nella versione attuale, il corpus contiene più o meno 380,000,000 parole.

[11] Naturalmente, alla stessa maniera, se notassimo che su un milione di lanci del nostro dado $900,000$ risultassero in un 6, saremmo inclini ad abbandonare il modello in cui il dado non è truccato e la probabilità degli eventi è uniforme.

Notate come questo metodo di stima della probabilità garantisce che tutte le probabilità siano valori tra 0 e 1 (se una parola non capita nel corpus, avrà una probabilità di $0/N$, cioè 0; se il corpus contiene una sola parola, questa parola avrà una probabilità di N/N, cioè 1), e che la loro somma sia 1 (perché la somma delle frequenze di tutte le parole è uguale a N).

Tuttavia, la stima delle probabilità basata sulle frequenze relative in un corpus ha i suoi problemi. Il problema maggiore è che, usando la formula in 7, se una parola non capita nel nostro corpus dobbiamo considerarla come una parola a probabilità zero. Questo ha effetti devastanti sul calcolo di probabilità composte: ovunque le probabilità vengono moltiplicate, basta uno zero perché il risultato complessivo sia zero!

Il problema si fa poi drammatico se consideriamo le probabilità di *sequenze* di parole. Le combinazioni possibili sono così tante che anche in un corpus molto esteso una buona proporzione di sequenze del tutto plausibili non comparirà mai.

Per esempio, una sequenza perfettamente sensata come `vedere Silvia` non capita mai nel corpus la Repubblica/SSLMIT, e dovremmo dunque considerarla come a probabilità zero.

In un modello in cui la probabilità di una frase si calcola moltiplicando le probabilità delle sequenze di due parole che capitano in essa (e questo è un modello tipico), la probabilità di qualsiasi frase che contiene `vedere Silvia` sarà dunque zero. E se una di queste frasi capita in un testo più lungo, la probabilità dell'intero testo rischia anche di essere zero!

Questo è un problema così serio che è stato visto da alcuni (come Chomsky) come la dimostrazione dell'impossibilità di usare la nozione di probabilità nell'analisi linguistica.

Il desiderio di evitare gli zeri spiega in parte perché i linguisti computazionali ricorrano a corpora sempre più estesi (di recente la world wide web è una scelta popolare), e perché ci sia molto interesse in metodi per stimare la probabilità di eventi (parole, sequenze di parole, strutture) che non capitano mai nel nostro corpus di riferimento.[12]

8 Probabilità e linguistica

Nell'ultima sezione abbiamo iniziato ad intravedere l'uso che si può fare della probabilità quando studiamo le lingue naturali. Nozioni relate alla probabilità saranno sullo sfondo in molti dei temi trattati nel corso, ma per finire questi appunti su una nota più concreta, consideriamo qui in breve due esempi di come la teoria della probabilità possa venire applicata in linguistica.

Si tratta, sottolineo, di esempi alquanto semplificati.

[12] Se la cosa vi incuriosisce, una lettura relativamente facile che spiega una tecnica piuttosto sofisticata per stimare queste probabilità è: "Good Turing frequency estimation without tears" di Geoffrey Sampson e William Gale, articolo ristampato nel libro *Empirical Linguistics* di Sampson, disponibile alla Ruffilli (io ho fotocopie del capitolo in questione).

8.1 Indipendenza e collocazioni

Abbiamo visto che se due eventi sono indipendenti, la probabilità che co-occorrano è data dall'equazione 6, che qui ripeto:

$$P(A, B) = P(A)P(B)$$

Gli eventi in questione potrebbero essere l'occorrenza di due parole una accanto all'altra. Dunque, se le due parole sono indipendenti, dovremmo avere che:

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

Ovvero:

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

Con un corpus possiamo stimare le tre probabilità che compaiono in 8 usando l'equazione 7. Poi, possiamo dividere il valore di $P(w_1, w_2)$ ottenuto dal corpus per il valore di $P(w_1)P(w_2)$ ottenuto dal corpus. Se il risultato è lontano dall'1 predetto dall'equazione 8, vuol dire che le due parole *non* sono indipendenti.

Consideriamo per esempio le seguenti coppie: `fatto che`, `Hong Kong`, `gelato alla`, `conoscere Andreotti`.

Per ciascuna coppia, possiamo stimare le probabilità sulla base del corpus la Repubblica/SSLMIT usando queste formule:

$$P(w_1, w_2) = \frac{fq(w_1, w_2)}{N}$$

$$P(w_1) = \frac{fq(w_1)}{N}$$

$$P(w_2) = \frac{fq(w_2)}{N}$$

In queste formule $fq(w_1, w_2)$ è il numero di volte che la sequenza $w_1 w_2$ capita nel corpus,[13] e N è il numero di parole nel corpus.[14]

[13] Abbiamo detto nella sezione 6 che $P(A, B)$ e $P(B, A)$ sono due modi di dire la stessa cosa. Infatti quando scriviamo $P(A, B)$ non ci riferiamo all'ordine (spaziale o temporale) in cui gli eventi A e B hanno luogo. Tuttavia, quando in linguistica parliamo della probabilità di co-occorrenza di lettere, parole o sintagmi, di solito intendiamo la probabilità che queste unità hanno di occorrere in *sequenza*, e non in ordine sparso. Dunque spesso quando scriviamo $P(Hong, Kong)$ (o $P(a, z)$) intendiamo: la probabilità che in una sequenza di due parole (o lettere), la prima sia `Hong` (o `a`) e la seconda sia `Kong` (o `z`). Stiamo allora usando la notazione $P(A, B)$ come una "scorciatoia" per $P(E_1 = A, E_2 = B)$, o qualcosa del genere (dove gli indici di E si riferiscono ad una sequenza ortografica o temporale). Una volta che usiamo questa forma più esplicita, possiamo di nuovo invertire l'ordine senza creare confusione: $P(E_1 = A, E_2 = B)$ e $P(E_2 = B, E_1 = A)$ sono ovviamente la stessa cosa. Però la forma esplicita complica notevolmente la notazione, e dunque noi useremo soltanto quella meno esplicita ma più breve. Se tutto ciò vi ha confuso le idee anziché chiarirvele, lasciate pure perdere e continuate a leggere.

[14] A rigor di termini, N nella prima equazione è il numero di *coppie* di parole nel corpus,

Siccome non ci interessano queste probabilità individualmente, ma il risultato ottenuto dividendo la prima per il prodotto delle altre due, osserviamo che:

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

Nel corpus la Repubblica/SSLMIT il valore di N è 175239348 e la seguente tabella contiene gli altri dati che ci interessano per ciascuna coppia:

coppia	$fq(w_1,w_2)$	$fq(w_1)$	$fq(w_2)$
fatto che	27538	184859	3042908
Hong Kong	1656	1762	1797
gelato alla	16	586	559750
conoscere Andreotti	1	8220	26589

Usando la formula in 9 calcoliamo il rapporto $P(w_1,w_2)/P(w_1)P(w_2)$ per ciascuna coppia, ottenendo i seguenti valori (riportati in ordine decrescente):

coppia	$P(w_1,w_2)/P(w_1)P(w_2)$
Hong Kong	91651.16
fatto che	8.58
gelato alla	8.55
conoscere Andreotti	0.8

Ricordiamo che, seguendo il ragionamento fatto sopra, una coppia di parole indipendenti dovrebbe avere un valore vicino ad 1 per il rapporto che abbiamo calcolato.

Delle quattro coppie considerate, soltanto `conoscere Andreotti` ha un valore abbastanza vicino ad 1 per la misura in questione[15]. Infatti, abbiamo probabilmente tutti l'intuizione che non c'è nessuna dipendenza particolare tra le parole `conoscere` e `Andreotti`.

È anche interessante osservare che `Hong Kong` ha un valore molto più alto delle altre due coppie. Di nuovo, questo va d'accordo con la nostra intuizione, che ci dice che le parole `Hong` e `Kong` sono associate in maniera molto forte – se abbiamo appena letto o sentito la parola `Hong` possiamo quasi essere certi che la parola che seguirà sarà `Kong`.

Anche se la misura che abbiamo usato ha dei problemi (che non discuterò qui), essa illustra il tipo di ragionamento che è alla base di molte misure usate

e c'è una coppia in meno poiché la prima parola non forma una coppia con un elemento a sinistra. Tuttavia, in un corpus esteso la differenza è negligibile. Potremmo anche assumere che la prima parola è preceduta da un simbolo speciale che marca l'inizio del corpus e forma una coppia con essa, nel qual caso i due valori di N diventano uguali.

[15] Addirittura un po' inferiore ad 1, ma sorvoliamo su questo fatto

per estrarre automaticamente collocazioni e altre coppie di parole fortemente associate (idiomi, frasi fatte, nomi proprii, ecc.) da corpora.

Più in generale, abbiamo illustrato con un caso estremamente semplice un tipico modo di ragionare in maniera statistica: prima abbiamo derivato da principi interamente teorici un valore per la misura $P(w_1, w_2)/P(w_1)P(w_2)$ se due parole sono indipendenti (il valore 1).

Abbiamo poi calcolato il valore di questa misura sulla base di dati empirici, e abbiamo tratto delle conclusioni (quali coppie sono indipendenti, quali no) sulla base del confronto tra il valore teorico della misura in caso di indipendenza e il suo valore empirico estratto dal corpus.

Allo stesso tempo, quando ho messo in ordine le coppie sulla base dei valori ricavati dal corpus, più ancora che il paragone di ciascun valore empirico con il valore teorico, è risultato interessante il confronto tra i valori ottenuto per le varie coppie.

Questo è caratteristico degli approcci statistici/quantitativi alla linguistica computazionale, dove spesso ci interessa di più l'ordine in cui una misura dispone gli elementi in una lista che decidere, da un punto di vista teorico, per quali elementi il valore della misura è "significativo" e per quali no.[16]

8.2 Probabilità di una stringa di caratteri in italiano

Questo è un esempio più complesso, che introduce, seppure un po' superficialmente, varie nozioni e metodi importanti nell'approccio probabilistico a problemi linguistici.

Prima di tutto, riprendiamo l'equazione 3 (invertendo il primo e il secondo termine dell'espressione a destra):

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

Naturalmente, sia A che B potrebbero essere a loro volta dati dalla cooccorrenza di due eventi. Per esempio al posto di A potremmo avere A_1, A_2:

$$P(A_1, A_1, B) = P(A_1, A_2)P(B|A_1, A_2)$$

Sempre usando l'equazione 3 possiamo espandere $P(A_1, A_2)$:

$$P(A_1, A_1, B) = P(A_1)P(A_2|A_1)P(B|A_1, A_2)$$

Sostituendo B con B_1, B_2 otteniamo:

$$P(A_1, A_1, B_1, B_2) = P(A_1)P(A_2|A_1)P(B_1, B_2|A_1, A_2)$$

Espandendo di nuovo:

[16] Per esempio, a un lessicografo può interessare analizzare una lista di coppie ordinate secondo la nostra misura dall'alto in basso, alla ricerca di collocazioni da inserire in un dizionario.

$$P(A_1, A_1, B_1, B_2) = P(A_1)P(A_2|A_1)P(B_1|A_1, A_2)P(B_2|A_1, A_2, B_2)$$

Dovrebbe essere chiaro che questo processo si può ripetere per stringhe di eventi sempre più lunghe.

Facendo un po' di pulizia nel simbolismo, possiamo esprimere nella seguente maniera l'equazione per calcolare la probabilità che gli eventi $E_1...E_n$ capitino insieme:

$$P(E_1, ..., E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1, E_2)...P(E_n|E_1, ..., E_{n-1}) \qquad (10)$$

Questa equazione è importante per noi perché nell'analisi linguistica, spesso e volentieri dobbiamo calcolare la probabilità di stringhe (stringhe di caratteri, stringhe di parole, stringe di frasi...)

Per esempio, possiamo usare la formula in 10 per confrontare la probabilità della stringa `azione` e la probabilità della stringa `action` in italiano:[17]

$$P(a, z, i, o, n, e) = P(a)P(z|a)P(i|a, z)P(o|a, z, i)P(n|a, z, i, o)P(e|a, z, i, o, n)$$

$$P(a, c, t, i, o, n) = P(a)P(c|a)P(t|a, c)P(i|a, c, t)P(o|a, c, t, i)P(n|a, c, t, i, o)$$

Le varie probabilità possono essere stimate usando il nostro corpus e le equazioni 1 e 7.

Per esempio, la probabilità $P(z|a)$ si stima così:

$$P(z|a) = \frac{P(a, z)}{P(a)}$$

$$P(a, z) = \frac{fq(az)}{fq(..)}$$

$$P(a) = \frac{fq(a)}{fq(.)}$$

Prestiamo attenzione ai denominatori della seconda e terza equazione. Nella seconda equazione il denominatore è il numero totale di *bigrammi* (sequenze di due lettere) nel corpus. In pratica, questa quantità è uguale al numero di lettere nel corpus.[18]

[17] Attenzione: stiamo parlando delle probabilità delle due *stringhe* (sequenze di lettere) `azione` e `action`, e non delle probabilità delle due *parole* `azione` e `action`, che si potrebbero stimare contando il numero di occorrenze delle parole in questione nel corpus. Noto incidentalmente che il nostro modello potrebbe anche basarsi su probabilità condizionate da ciò che *segue*, anziché da ciò che precede (per esempio, potremmo calcolare la probabilità di `z` se è seguita da `ione` anziché la probabilità di `z` se è preceduta da `a`).

[18] Anche in questo caso potremmo assumere che la prima lettera del corpus sia preceduta da un simbolo speciale di inizio corpus: così i due valori saranno uguali anche in teoria.

Nella terza equazione il denominatore è anche dato dal numero totale di lettere nel corpus.

Dunque, visto che abbiamo deciso che $fq(..) = fq(.)$, rinominiamo questa quantità N e riscriviamo le equazioni come:

$$P(a, z) = \frac{fq(az)}{N}$$

$$P(a) = \frac{fq(a)}{N}$$

Quindi, $P(z|a)$ si calcolerà come:

$$P(z|a) = \frac{P(a, z)}{P(a)} = \frac{\frac{fq(az)}{N}}{\frac{fq(a)}{N}} = \frac{fq(az)}{N} \times \frac{N}{fq(a)} = \frac{fq(az)}{fq(a)}$$

Questo semplifica notevolmente i nostri calcoli, visto che ci permette di calcolare $P(z|a)$ senza bisogno di calcolare $P(z|a)$ e $P(a)$.

Calcolare probabilità condizionate da stringhe molto lunghe, come per esempio $P(n|actio)$ può essere molto problematico. Per esempio, potrebbe darsi che nel nostro corpus la sequenza `action` non capiti mai, impedendoci di calcolare $fq(action)$ (un valore necessario al calcolo di $P(n|actio)$, come $fq(az)$ è necessario al calcolo di $P(z|a)$).

Inoltre, dover calcolare probabilità per tutte le sequenze di due, tre, quattro, cinque... lettere che capitano in un corpus potrebbe darci problemi dal punto di vista del tempo e dello spazio richiesto per l'operazione.

Infine, è chiaro che la distribuzione di una lettera tende ad essere influenzata maggiormente dalla lettera che la precede che dalla lettera che si trova due caratteri a sinistra, e che l'influenza della lettera che si trova cinque caratteri a sinistra sarà pressoché nulla.

Per esempio, abbiamo tutti una chiara intuizione che in italiano `f` è estremamente improbabile se la lettera immediatamente precedente è `v`, ma non credo che nessuno di noi abbia intuizioni riguardo alla probabilità di `f` se `v` si trova cinque caratteri a sinistra.

Visto tutto ciò, possiamo azzardare una soluzione al problema delle probabilità condizionate da sequenze di molte lettere: sostituiamo ogni probabilità condizionata da una sequenza di più di una lettera con la probabilità condizionata soltanto dall'ultima lettera nella sequenza.

Dunque assumiamo che:

$$P(n|a, c, t, i, o) = P(n|o)$$

Più in generale:

$$P(c_n|c_1, ..., c_{n-1}) = P(c_n|c_{n-1})$$

Le formule per calcolare le probabilità di `azione` e `action` dunque diventano:

$$P(a,z,i,o,n,e) = P(a)P(z|a)P(i|z)P(o|i)P(n|o)P(e|n)$$

$$P(a,c,t,i,o,n) = P(a)P(c|a)P(t|c)P(i|t)P(o|i)P(n|o)$$

Una bella semplificazione, no? Incidentalmente, questo tipo di modello, in cui ciascun evento è condizionato da una finestra fissa e limitata di eventi passati (in questo caso, un evento), si chiama *modello markoviano*, e gioca un ruolo fondamentale in molte aree della linguistica computazionale.

Estraendo le frequenze dal corpus la Repubblica/SSLMIT e calcolando le varie probabilità come mostrato sopra otteniamo:

$$P(a,z,i,o,n,e) = .0000020720$$

$$P(a,c,t,i,o,n) = .0000000034$$

Il nostro semplice modello dell'italiano assegna alla stringa *azione* una probabilità molto maggiore di quella assegnata alla stringa *action* – una cosa molto ragionevole![19]

A cosa può servire un risultato del genere? Beh, per esempio potrebbe essere un metodo per scoprire prestiti non-adattati in un corpus, o, più ambiziosamente, essere alla base di un metodo per scoprire in che lingua è scritto un certo documento.

Semplificando molto, se trattiamo un documento come una stringa, possiamo calcolare la probabilità di questa stringa sulla base di bigrammi (trigrammi, n-grammi) estratti da un corpus italiano, da un corpus tedesco, da un corpus inglese, ecc. Se il documento risulta molto più probabile sulla base del modello estratto dal corpus italiano, è verosimile che il documento sia stato scritto in italiano, se è più probabile sulla base del modello tedesco, è verosimile che sia stato scritto in tedesco e via dicendo.[20]

9 Per saperne di più...

Dopo aver letto queste note, potete provare a leggere il paper di John Goldsmith intitolato *Probability for linguists*, almeno fino alla parte sui logaritmi (esclusa).

Se poi l'argomento vi appassiona (e come non potrebbe? ;-)), potete leggere il paper di Goldsmith nella sua interezza (senza disperarvi se non capite tutto:

[19] 1) Quale probabilità causa la "rovina" di `action`? 2) Notate come, a moltiplicare probabilità (che sono valori tra 0 e 1, per cui il cui prodotto è sempre uguale o inferiore ai moltiplicandi), si raggiungono presto valori estremamente bassi. Questa è una delle ragioni per le quali spesso invece di calcolare la probabilità di una stringa di eventi calcoliamo il *logaritmo* di questa probabilità. Infatti, il logaritmo di un prodotto è uguale alla *somma* dei logaritmi dei termini, per cui i problemi causati dalla moltiplicazione scompaiono.

[20] Sospetto che Google usi un meccanismo di questo genere per decidere in che lingua è stata scritta una pagina.

alcune parti richiedono conoscenze matematiche piuttosto avanzate) nonché il paper di David Magerman intitolato *Everything you always wanted to know about probability theory but were afraid to ask* (entrambi sono disponibili sul sito del corso).

Il libro di Richard Isaac *The pleasures of probability* (Springer, 1995) costituisce un'introduzione abbastanza completa ma estremamente semplice e non-matematica alla teoria della probabilità.

Partendo dalla pagina dei miei links, troverete varie risorse utili per studiare probabilità e statistica (incluso un ottimo libro di testo scaricabile interamente dalla rete).