

Teoria dell'Informazione e Inferenza: note delle lezioni

Alessandro Verri

Queste note raccolgono il materiale utilizzato per le ventiquattro lezioni di *Teoria dell'Informazione e Inferenza*. Le lezioni dalla 1 alla 9, raggruppate nel capitolo 1, illustrano i rudimenti della **Teoria della Probabilità**. Le quattro lezioni successive, riunite nel capitolo 2, coprono argomenti più avanzati basati sulla proprietà dei valori attesi. Molto del materiale di queste lezioni è tratto dal *Ross* [1], una piccola parte, invece, dal *Motwani & Raghavan* [2]. Le lezioni dalla 14 alla 19 formano il capitolo 3 e introducono i principali concetti della **Teoria dell'Informazione** con alcune incursioni nella **Teoria dei Codici**. La principale fonte è il *McKay* [3] con occasionali puntate al *Khinchin* [4] e al *Reza* [5]. Il capitolo 4, che comprende le ultime cinque lezioni, apre all'**Inferenza** attingendo a piene mani da diverse fonti le più importanti delle quali sono il *Duda e Hart* [6], le note di un corso a NYU di *Miranda Holmes-Cefron* [7] e il *Brémaud* [8]. Tutti sono benvenuti a inviare all'indirizzo `alessandro.verri@unige.it` rilievi e correzioni.

Bibliografia

- [1] S. Ross. *A First Course on Probability*. Prentice Hall, 2010.
- [2] R. Motwani & P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.
- [3] D.J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- [4] A.I. Khinchin. *Fondamenti Matematici della Teoria dell'Informazione*. Cremonese, s.l., 1978.
- [5] F.M. Reza. *An introduction to Information Theory*. Dover, 1994.
- [6] R.O. Duda e P.E.Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.
- [7] M. Holmes-Cerfon. *Applied Stochastic Analysis: lecture notes* Courant Institute of Mathematical Sciences, 2019.
- [8] P. Brémaud. *Markov Chains* Springer, 1999

Indice

1	Elementi di Teoria della Probabilità	7
1.1	Impariamo a contare	7
1.2	Definizione assiomatica di probabilità	10
1.3	Probabilità condizionata	13
1.4	Teorema di Bayes	16
1.5	Variabili casuali discrete	19
1.6	Distribuzioni discrete di probabilità	22
1.7	Variabili casuali continue	25
1.8	Distribuzioni continue di probabilità	28
1.9	Distribuzioni congiunte e indipendenza	31
2	Valori attesi	35
2.10	Somme di variabili casuali	35
2.11	Risultati asintotici	38
2.12	Problemi di occupazione	41
2.13	Grandi deviazioni	44
3	Elementi di Teoria dell'Informazione	47
3.14	Informazione di Shannon e codifica dell'informazione	47
3.15	Entropia di Shannon	52
3.16	Lo stretto indispensabile sulla teoria dei codici	57
3.17	Codifiche in assenza di rumore	62
3.18	Codifica aritmetica	67
3.19	Codifiche in presenza di rumore	71
4	Elementi di Inferenza	77
4.20	Inferenza frequentista	77
4.21	Inferenza Bayesiana	82
4.22	Metodi Monte Carlo	87
4.23	Catene di Markov	92
4.24	Catene di Markov <i>Monte Carlo</i>	97

Capitolo 1

Elementi di Teoria della Probabilità

1.1 Impariamo a contare

Introduciamo gli elementi principali del calcolo combinatorio che sono utili per calcolare la probabilità associata a eventi nel caso in cui vale l'ipotesi di equiprobabilità.

Principio base

Nel seguito fare uso della nozione di *esperimento*, ovvero di un'azione il cui *risultato* è un elemento di un insieme costituito da tutti i *possibili risultati* di quell'azione. Tutto quello che vedremo è ottenibile mediante l'applicazione di un principio molto semplice.

Principio 1.1.1. *Uno alla volta*

Se un esperimento fornisce m possibili risultati e se per ciascuno di essi un secondo esperimento fornisce n possibili risultati, allora i due esperimenti forniscono $m \times n$ possibili risultati. \square

Due le vere difficoltà: definire in modo univoco ogni esperimento e individuare il numero dei possibili risultati di ogni esperimento definito.

Esercizio 1.1.1. *Tutte le targhe*

Se una targa è formata da 4 lettere e 3 cifre, quante sono le targhe possibili?

Soluzione

Assumiamo che le lettere possibili siano 26 e le cifre 10. La scelta di ogni elemento della targa può essere visto come un esperimento, l'elemento scelto uno dei possibili risultati.

Ripetizioni ammesse : 26 risultati per la prima, 26 per la seconda, 26 per la terza e 26 per la quarta lettera, 10 risultati per la prima, 10 per la seconda e 10 per la terza cifra, in tutto

$$26^4 \cdot 10^3 = 456,976,000$$

Ripetizioni non ammesse : 26 risultati per la prima, 25 per la seconda, 24 per la terza e 23 per la quarta lettera, 10 risultati per la prima, 9 per la seconda e 8 per la terza cifra, in tutto

$$26 \cdot 25 \cdot 24 \cdot 23 \cdot 10 \cdot 9 \cdot 8 = 258,336,000$$

Consideriamo ora tre casi importanti: permutazioni, disposizioni e combinazioni.

Permutazioni

Una *permutazione* è un particolare ordinamento di n oggetti. Applicando il principio base per contare le permutazioni possibili, otteniamo $n!$ poichè abbiamo n scelte per il primo oggetto, $n - 1$ per il secondo e così via fino alla scelta obbligata dell' n -esimo oggetto, ultimo rimasto. In una permutazione tutti gli n oggetti sono distinguibili.

Esercizio 1.1.2. Tutti gli ordinamenti

In quanti modi è possibile ordinare su uno scaffale 2 libri di chimica, 3 di fisica, 4 di matematica e 5 di informatica in modo che i libri di una stessa materia siano in un unico gruppo di libri consecutivi?

Soluzione

Principio base applicato alle materie : $4! = 24$

Principio base ai libri di ogni materia : $2! \cdot 3! \cdot 4! \cdot 5! = 34,560$

Combinando i due risultati otteniamo in tutto

$$24 \cdot 34,560 = 829,440$$

Notiamo che in assenza di qualunque vincolo le possibili permutazioni sono molte di più, ovvero

$$14! = 87,178,291,200$$

Disposizioni

Una *disposizione* è un particolare ordinamento di i oggetti scelti da n oggetti con $i \leq n$. Se applichiamo il principio base per contare le disposizioni possibili, otteniamo

$$n(n-1) \dots (n-i+1)$$

perchè abbiamo n scelte per il primo, $n - 1$ per il secondo e così via fino alla scelta dell' i -esimo oggetto tra gli $n - i + 1$ oggetti rimasti.

Osservazione 1.1.1. Pensando in termini di permutazioni

Dall'identità

$$n(n-1) \dots (n-i+1) = \frac{n!}{(n-i)!}$$

segue che le disposizioni possibili di i oggetti scelti tra n possono essere ottenute anche ragionando in modo diverso. Ovvero considerando distinguibili gli i oggetti scelti e indistinguibili gli $n-i$ oggetti della cui disposizione non ci curiamo. Delle $n!$ permutazioni possibili di n oggetti, consideriamo equivalenti quelle in cui gli i oggetti scelti si trovano nelle stesse posizioni. Queste permutazioni sono $(n-i)!$ ovvero tante quante le possibili permutazioni degli $n - i$ oggetti non scelti che consideriamo indistinguibili.

Esercizio 1.1.3. Tutti gli anagrammi

Gli anagrammi di *CINEMA* (la maggior parte dei quali non fornisce parole di senso compiuto) sono $6! = 720$. Quanti sono, invece, gli anagrammi di *ERRORE*?

Soluzione

Dividendo per $3!$ (per le 3 *R*) e per $2!$ (per le 2 *E*) i possibili $6!$ anagrammi di *ERRORE* otteniamo

$$\frac{6!}{3!2!} = \frac{6 \cdot 5 \cdot 4}{2} = 120$$

Combinazioni

Una *combinazione* è una scelta di i oggetti da n oggetti con $i \leq n$. Se applichiamo il principio base per contare le combinazioni possibili e teniamo presente che l'ordine, questa volta, è irrilevante sia per gli i oggetti scelti sia per gli $n - i$ oggetti non scelti, otteniamo

$$\frac{n!}{i!(n-i)!} = \binom{n}{i}$$

Esercizio 1.1.4. Tutti i comitati

Quanti comitati di tre persone possiamo formare partendo da un gruppo di 20 persone?

Soluzione

$$\binom{20}{3} = \frac{20!}{17!3!} = \frac{20 \cdot 19 \cdot 18}{6} = 1140$$

Le nozioni di permutazione, disposizione e combinazione sono illustrate in un caso particolare nel diagramma di Venn di figura 1.1.

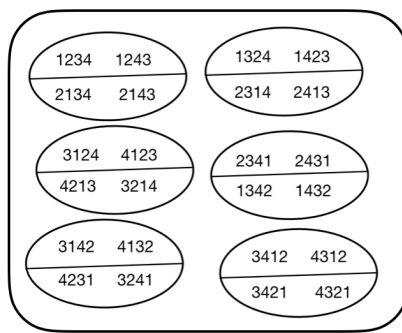


Figura 1.1: Le $4! = 24$ permutazioni di 1, 2, 3 e 4 sono rappresentate come tutte le sequenze possibili delle quattro cifre. Le $4 \times 3 = 12$ disposizioni di 1 e 2 sono le 12 coppie all'interno della parte all'alta e della parte bassa delle 6 ellissi. Poiché nelle disposizioni le posizioni degli oggetti scelti contano, in ognuna delle 12 coppie la posizione di 1 e 2 non cambia. Infine le $\binom{4}{2} = 6$ combinazioni sono rappresentate dalle ellissi. Per le quattro coppie in ogni ellisse sia le posizioni di 1 e 2 sia le posizioni di 3 e 4 possono essere scambiate.

Esercizi non risolti

1. Cerca la definizione di *Codice Fiscale* e determina il numero di codici fiscali possibili
2. La serie A è composta da venti squadre. Calcola quanti mini-campionati diversi si potrebbero disputare composti da dieci squadre.
3. Includendo anche le parole senza senso quanti sono gli anagrammi di *ANILINA*?
4. Hai tre maglioni, quattro camicie e due paia di pantaloni. In quanti ordinamenti diversi puoi trovarli senza mescolarli?

1.2 Definizione assiomatica di probabilità

Definiamo ora la probabilità in modo assiomatico discutendo brevemente alcune proprietà nel caso semplice, ma importante, in cui sia possibile individuare eventi equiprobabili. Ci limitiamo al caso discreto in modo da non dover affrontare il tema di quali sono gli eventi misurabili, ovvero cui è possibile associare una probabilità, e quelli che non lo sono.

Nozioni fondamentali

Spazio campionario: l'insieme S dei possibili risultati di un esperimento (testa e croce o il numero di persona in coda alla cassa).

Evento: un qualunque sottoinsieme E di S che si *realizza* se il risultato dell'esperimento appartiene a E (testa nel lancio di una moneta o quattro persona in coda).

Nel caso discreto un evento E è un qualunque sottoinsieme di S , ovvero un elemento dell'insieme delle parti di S . Indichiamo con $E \cup F$ l'*unione* degli eventi E e F e con EF la loro *intersezione*. Inoltre, due eventi E e F tali che $EF = \emptyset$ sono *mutuamente esclusivi*. L'evento E^c tale che $E \cup E^c = S$ è il *complementare* di E . L'uso dei concetti di base della teoria degli insiemi e la loro illustrazione tramite diagrammi di Venn consentono di trattare in modo preciso e intuitivo diverse proprietà delle probabilità.

Assiomi

Una probabilità $P(\cdot)$ risulta ben definita sugli eventi di uno spazio campionario S se

$$A1: 0 \leq P(E) \leq 1 \quad \forall E \subseteq S$$

$$A2: P(S) = 1$$

A3: Se gli eventi E_i , con $i = 1, 2, \dots$ sono mutuamente esclusivi, allora

$$P\left(\bigcup_i E_i\right) = \sum_i P(E_i)$$

Sono conseguenze immediate di A1, A2 e A3

$$(i) \quad \forall E, P(E^c) = 1 - P(E)$$

$$(ii) \quad \forall E \text{ e } F, \text{ se } E \subseteq F \text{ allora } P(E) \leq P(F)$$

$$(iii) \quad \forall E \text{ e } F, P(E \cup F) = P(E) + P(F) - P(EF)$$

Dimostrazione: la prima segue dal fatto che $E^c E = \emptyset$ e $E^c \cup E = S$, per cui $P(E) + P(E^c) = 1$. La seconda si dimostra notando che $F = E \cup E^c F$ con E e $E^c F$ mutuamente esclusivi. Per la terza e ultima conviene scrivere $E \cup F$ e EF come unione di insiemi disgiunti, ovvero $E \cup F = E \cup E^c F$ e $F = EF \cup E^c F$ e applicare l'assioma A3 (o utilizzare i diagrammi di Venn). ■

Esercizio 1.2.1. Chi canta e chi suona

Quanti sono i componenti di un complesso in cui 3 cantano, 3 suonano la chitarra e 2 entrambe le cose?

Soluzione

Se $E = \{i \text{ tre cantanti}\}$ ed $F = \{i \text{ tre chitarristi}\}$ abbiamo che $EF = \{i \text{ due cantanti chitarristi}\}$.

Dalla proprietà (iii), pertanto, segue che

$$3 + 3 - 2 = 4$$

Eventi equiprobabili

Supponiamo che S sia costituito da un insieme finito di N risultati che indichiamo con i primi N numeri naturali, ovvero $S = \{1, 2, \dots, N\}$. Se le probabilità $P(i)$ sono tutte uguali allora $P(i) = 1/N$. La probabilità di E , in questo caso, si calcola come frazione del numero di risultati in E , $\#E$, sul numero di risultati in S , $\#S$.

Esercizio 1.2.2. Lancio di 2 dadi

Calcola la probabilità P di ottenere 7 lanciando 2 dadi.

Soluzione

Le coppie di risultati ottenibili dal lancio di due dadi sono 36. I casi favorevoli sono sei in tutto: (1,6), (2,5), (3,4), (4,3), (5,2) e (6,1), per cui $P = 6/36 = 1/6$.

Osservazione 1.2.1. Attenzione all'ordinamento!

Le coppie sono ordinate: il primo elemento è il risultato del lancio del primo dado, il secondo elemento il risultato del lancio del secondo dado.

Esercizio 1.2.3. Palline bianche e rosse

Calcola la probabilità P di estrarre 1 pallina bianca e 2 palline rosse da un'urna con 6 palline bianche e 5 rosse.

Soluzione con le combinazioni

Se consideriamo l'insieme delle palline estratte come non ordinato, i casi possibili sono le combinazioni di 3 palline scelte tra 11, i favorevoli quelle di 1 pallina bianca scelta tra 6 e 2 nere scelte tra 5, ovvero

$$\#S = \binom{11}{3} = 165 \quad \text{e} \quad \#E = \binom{6}{1} \binom{5}{2} = 60$$

da cui segue $P = 60/165 = 4/11$.

Soluzione con le disposizioni

Consideriamo ora invece rilevante l'ordine col quale estraiamo le palline. I casi possibili sono le disposizioni di 3 palline scelte tra 11, ovvero $11 \cdot 10 \cdot 9 = 990$. Dividiamo i casi favorevoli in 3 gruppi. Nel primo gruppo la pallina bianca è estratta per prima, $6 \cdot 5 \cdot 4$ casi favorevoli, nel secondo per seconda, $5 \cdot 6 \cdot 4$ casi favorevoli, e nel terzo per terza, $5 \cdot 4 \cdot 6$ casi favorevoli. Otteniamo ancora $P = 3 \cdot 120/990 = 4/11$.

Esercizio 1.2.4. Paradosso del compleanno (o delle collisioni nell'hashing)

Calcola la probabilità P_n che n individui festeggino il compleanno in n giorni diversi.

Soluzione

Consideriamo l'evento complementare ovvero che nessuno tra n individui sia nato lo stesso giorno. Perché un secondo individuo non sia nato nel giorno del primo i casi favorevoli sono $(365-1) = 364$, per un terzo $365-2=363$ e per l' n -esimo $365 - n + 1$. Applicando il principio base nel caso di eventi equiprobabili, la probabilità che tutti gli n individui siano nati in giorni diversi è allora

$$P_n = \frac{(365-1)(365-2) \dots (365-n+1)}{365^{n-1}}$$

Osservazione 1.2.2. Scommettiamo?

Al crescere di n la probabilità P_n diminuisce velocemente. Per $n = 100$ abbiamo che $P_{100} < 0.000001$ (esperimento in classe). L'ultimo valore di n per cui $P_n < 1/2$ è 23. Una spiegazione intuitiva di questo risultato apparentemente paradossale è che le coppie possibili di 23 individui sono $23 \cdot 22/2 = 253$ (ben più della metà dei 365 giorni in un anno).

Probabilità soggettiva

Capita di associare il concetto di probabilità a eventi incerti non ripetibili (probabilità di pioggia a Milano o di vittoria in un incontro di schermo). In questi casi parliamo di *probabilità soggettiva*.

Esercizio 1.2.5. *Piove o non piove?*

La probabilità dell'evento *oggi pioverà* è del 40% e che *domani pioverà* è del 30%. Se la probabilità che *oggi o domani pioverà* è del 60% dimostra che non è possibile che la probabilità che *oggi e domani pioverà* sia del 20%.

Soluzione

Se E è l'evento *oggi pioverà* e F *domani pioverà* sappiamo che $P(E \cup F) = P(E) + P(F) - P(EF)$. Nel nostro caso, invece, abbiamo che

$$40\% + 30\% - 20\% = 50\% \neq 60\%$$

Esercizi non risolti

1. Calcola la probabilità di pescare una pallina rossa e due palline bianche da un'urna che contiene due palline rosse, due palline bianche e tre palline azzurre.
2. Peschi due carte da un mazzo. Calcola la probabilità che siano due A, un A e un K e due carte dal 10 al 2.
3. Lanci un dado tre volte. A quale somma dei tre risultati corrisponde la probabilità più alta?

1.3 Probabilità condizionata

Introduciamo ora il concetto forse più importante della teoria della probabilità, la probabilità condizionata, ovvero la *probabilità di un evento una volta che si venga a conoscenza della realizzazione di un altro evento casuale*.

Esercizio 1.3.1. Probabilità che cambia

Lancia una volta un dado onesto. Qual è la probabilità di ottenere 1 se sai che il risultato del lancio è tra 1 e 3 o se invece sai che il risultato è tra 4 e 6?

Soluzione

Se l'evento E è di ottenere 1 e l'evento F è di ottenere $\{1, 2, 3\}$, abbiamo

$$P(F) = \frac{3}{6} \text{ e } P(EF) = P(E) + P(F) - P(E \cup F) = \frac{1}{6} + \frac{3}{6} - \frac{3}{6}$$

Conseguentemente la probabilità dell'evento E condizionata alla realizzazione dell'evento F è pari a $1/3$. Ragionando in modo analogo nel caso in cui $F = \{4, 5, 6\}$ otterremmo una probabilità pari a 0. In conclusione la probabilità di ottenere 1, inizialmente uguale a $1/6$, si modifica in funzione di informazioni aggiuntive che potrebbero rendersi disponibili. \square

In generale, dati due eventi E e F , siamo interessati a calcolare la probabilità di E quando sappiamo che si è realizzato F . La probabilità di E condizionata a F , indicata con $P(E|F)$, è definita come

$$P(E|F) = \frac{P(EF)}{P(F)}$$

Osservazione 1.3.1. Riduzione dello spazio campionario

Se si realizza anche E dopo che si è realizzato F , significa che il risultato appartiene all'intersezione EF . Lo spazio campionario per la realizzazione di E successiva alla realizzazione di F si è quindi ridotto da S a F e misura $P(F)$.

Esercizio 1.3.2. Sempre una probabilità

Dimostra che $P(\cdot|F)$ è una probabilità e quindi

A1 $\forall E \ 0 \leq P(E|F) \leq 1$

A2 $P(S|F) = 1$

A3 Se gli eventi E_i con $i = 1, \dots$ sono mutuamente esclusivi allora

$$P((\cup_i E_i) | F) = \sum_i P(E_i | F)$$

Dimostrazione: l'assioma **A1** è soddisfatto in quanto poiché $\forall E \ EF \subseteq F$ e quindi $P(EF) \leq P(F)$, l'assioma **A2** in quanto $P(SF) = P(F)$. Per la verifica dell'assioma **A3**, invece, ricordiamo che per via della mutua esclusività degli E_i abbiamo

$$(\cup_i E_i)F = \cup_i E_i F \text{ e } P(\cup_i E_i F) = \sum_i P(E_i F)$$

da cui otteniamo

$$P((\cup_i E_i) | F) = \frac{P((\cup_i E_i) F)}{P(F)} = \frac{P(\cup_i E_i F)}{P(F)} = \frac{\sum_i P(E_i F)}{P(F)} = \sum_i P(E_i | F)$$

■

Osservazione 1.3.2. Regola della moltiplicazione

Generalizziamo l'identità $P(EF) = P(F)P(E|F)$ al caso dell'intersezione di n eventi. Abbiamo che

$$P(E_1 E_2 \dots E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1 E_2) \dots P(E_n|E_1 E_2 \dots E_{n-1})$$

Per dimostrare la validità di questa uguaglianza è sufficiente utilizzare la definizione di probabilità condizionata in modo iterativo su ognuno degli $n - 1$ fattori. Ad esempio, per tre insiemi E, F, G , si ha

$$P(EFG) = \frac{P(EFG)}{P(EF)} \frac{P(EF)}{P(E)} P(E) = P(G|EF)P(F|E)P(E)$$

Esercizio 1.3.3. Assi equidistribuiti

In un mazzo di cinquantadue carte che contiene quattro assi qual è la probabilità che dividendo le carte in quattro pile da tredici ogni pila contenga un asso?

Soluzione

Poniamo

E_1 : l'asso di picche è in uno qualunque delle quattro pile

E_2 : l'asso di picche e l'asso di cuori sono in due pile diverse

E_3 : l'asso di picche, l'asso di cuori e l'asso di quadri sono in tre pile diverse

E_4 : l'asso di picche, l'asso di cuori, l'asso di quadri e l'asso di fiori sono in quattro pile diverse

Dobbiamo calcolare $P(E_1 E_2 E_3 E_4)$. Per la regola della moltiplicazione

$$P(E_1 E_2 E_3 E_4) = P(E_1)P(E_2|E_1)P(E_3|E_1 E_2)P(E_4|E_1 E_2 E_3)$$

Ora, $P(E_1) = 1$, mentre $P(E_2|E_1) = 39/51$, perché dobbiamo sottrarre a 52 l'asso di picche al denominatore e le tredici carte della pila che lo contiene al numeratore, $P(E_3|E_1 E_2) = 26/50$ perché dobbiamo sottrarre l'asso di cuori al denominatore e le altre tredici carte della pila che lo contiene al denominatore e $P(E_4|E_1 E_2 E_3) = 13/49$ perché dobbiamo sottrarre l'asso di quadri al denominatore e le altre tredici carte della pila che lo contiene al numeratore. Mettendo tutto assieme otteniamo

$$P(E_1 E_2 E_3 E_4) = \frac{39}{51} \frac{26}{50} \frac{13}{49} = \frac{13 \cdot 26 \cdot 39}{49 \cdot 50 \cdot 51} \approx 0.105$$

Esercizio 1.3.4. Prima una e poi l'altra

Un'urna contiene otto palline rosse e quattro bianche. Se estraiano due palline qual è la probabilità che entrambe siano rosse?

Soluzione

Se R_1 è l'evento che la prima pallina è rossa e R_2 l'evento che la seconda pallina è rossa avremo $P(R_1) = 2/3$ e $P(R_2|R_1) = 7/11$. Pertanto

$$P(R_1 R_2) = P(R_1)P(R_2|R_1) = \frac{8}{12} \frac{7}{11} = \frac{14}{33}$$

Otteniamo lo stesso risultato mediante le combinazioni: i casi favorevoli sono contati come la scelta di due palline rosse tra le otto, contro tutti i casi possibili, due palline tra le dodici, ovvero

$$P(R_1 R_2) = \frac{\binom{8}{2}}{\binom{12}{2}} = \frac{28}{66} = \frac{14}{33}$$

Esercizio 1.3.5. Un risultato sorprendente

Lancia due volte una moneta onesta e calcola la probabilità di ottenere 2 volte testa, T , se (i) il risultato del primo lancio è T , o se (ii) il risultato di uno dei due lanci è T .

Soluzione

Nel caso (i) gli eventi sono $E = \{T, T\}$ e $F = \{\{T, C\}, \{T, T\}\}$, e poichè

$$P(E) = P(\{T, T\}) \text{ e } P(F) = P(\{\{T, C\}, \{T, T\}\})$$

allora

$$P(E|F) = \frac{P(EF)}{P(F)} = \frac{1/4}{2/4} = \frac{1}{2}$$

Nel caso (ii) gli eventi sono $E = \{T, T\}$ e $F = \{\{T, C\}, \{T, T\}, \{C, T\}\}$ e, poichè

$$P(E) = P(\{T, T\}) \text{ e } P(F) = P(\{\{T, C\}, \{T, T\}, \{C, T\}\})$$

allora

$$P(E|F) = \frac{P(EF)}{P(F)} = \frac{1/4}{3/4} = \frac{1}{3}$$

La differenza tra (i) e (ii) è da imputarsi al diverso numero di risultati possibili nei due casi!

Osservazione 1.3.3. Notazione pesante ma necessaria per capire e non cadere in tentazione

L'uso delle parentesi graffe appesantisce la notazione ma chiarisce che la probabilità di un evento è la misura di un sottoinsieme dello spazio campionario (in questo caso le coppie di possibili risultati ottenibili lanciando due volte una moneta onesta). Lasciarsi guidare dall'intuizione con la probabilità è raramente una buona idea!

Esercizi non risolti

1. Lancia un dado due volte. Con che probabilità ottieni esattamente un 3? Qualè la probabilità che uno dei risultati sia 3 se la somma è 8?
2. Un mazzo di carte è costituito da quattro Assi, quattro King e due Queen. Se peschiamo due carte qual è la probabilità che siano due Assi? E quale che siano due Queen?
3. Con lo stesso mazzo calcola la probabilità che la seconda carta sia un Asso se la prima è un Asso.

1.4 Teorema di Bayes

Appena dietro la nozione di probabilità condizionata arriva il Teorema di Bayes, risultato di fondamentale importanza in moltissime applicazioni.

Per ogni coppia di eventi E e F , applicando la definizione di probabilità condizionata, possiamo riscrivere $P(F|E)$ in termini di $P(E|F)$ (o viceversa). Ovvero

$$P(F|E) = \frac{P(E|F)P(F)}{P(E)} \quad (1.1)$$

o

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

Supponiamo ora che E sia l'unione dei due (o più) eventi mutuamente esclusivi, EF e EF^c . In questo modo, applicando la definizione di probabilità condizionata, la probabilità $P(E)$ può essere scritta come probabilità totale, ovvero

$$P(E) = P(EF) + P(EF^c) = P(F)P(E|F) + P(F^c)P(E|F^c) \quad (1.2)$$

dove nel primo passaggio abbiamo utilizzato la mutua esclusività mentre nel secondo la definizione di probabilità condizionata.

Sostituendo la formula (1.2) nella (1.1) otteniamo infine la formula nota come *Teorema di Bayes*

$$P(F|E) = \frac{P(F)P(E|F)}{P(F)P(E|F) + P(F^c)P(E|F^c)}$$

Vediamone alcune applicazioni.

Esercizio 1.4.1. Probabilità totale

Hai due monete uguali A e B . La probabilità di *testa* per A è $1/2$, per B $1/10$. Qual è la probabilità di ottenere *testa* lanciando una moneta a caso?

Soluzione

Abbiamo $P(A) = P(B) = 1/2$, $P(\text{testa}|A) = 1/2$ e $P(\text{testa}|B) = 1/10$. Pertanto

$$P(\text{testa}) = P(\text{testa}|A)P(A) + P(\text{testa}|B)P(B) = \frac{1}{2} \frac{1}{2} + \frac{1}{10} \frac{1}{2} = \frac{3}{10}$$

Esercizio 1.4.2. Da quale urna?

Hai tre urne uguali, A , B e C . L'urna A contiene 3 palline rosse e 1 bianca, l'urna B 3 palline bianche e 1 rossa e l'urna C 4 palline rosse. Se peschi una pallina rossa, calcola la probabilità che provenga da A , B o C .

Soluzione

Se indichiamo r per *rossa*, dobbiamo calcolare $P(A|r)$, $P(B|r)$ e $P(C|r)$. Poiché le tre urne sono uguali abbiamo che $P(A) = P(B) = P(C) = 1/3$. Inoltre $P(r|A) = 3/4$, $P(r|B) = 1/4$ e $P(r|C) = 1$. Per la probabilità totale $P(r)$ abbiamo quindi

$$P(r) = P(r|A)P(A) + P(r|B)P(B) + P(r|C)P(C) = \left(\frac{3}{4} + \frac{1}{4} + 1 \right) \frac{1}{3} = \frac{2}{3}$$

Applicando il teorema di Bayes otteniamo

$$\begin{aligned} P(A|r) &= \frac{P(r|A)P(A)}{P(r)} = \frac{3}{4} \frac{1}{3} \frac{1}{2} = \frac{3}{8} \\ P(B|r) &= \frac{P(r|B)P(B)}{P(r)} = \frac{1}{4} \frac{1}{3} \frac{1}{2} = \frac{1}{8} \\ P(C|r) &= \frac{P(r|C)P(C)}{P(r)} = \frac{1}{3} \frac{1}{3} \frac{1}{2} = \frac{1}{2} \end{aligned}$$

Esercizio 1.4.3. Paradosso delle tre carte

Ci sono tre carte A , B e C . La carta A , è rossa sul lato 1 e sul lato 2, la carta B su rossa sull'1 e bianca sul 2, mentre la carta C è bianca su entrambi i lati. Ponendo su un tavolo una delle tre carte, scelta a caso, ottengo che il lato visibile è di colore rosso. Qual è la probabilità che anche il lato non visibile sia di colore rosso?

Prima soluzione

Calcoliamo i casi favorevoli e tutti i casi possibili. Estrahendo una carta e posandola sul tavolo si possono verificare i sei casi equiprobabili in figura 1.2.













	vis nasc		vis nasc	
A				
B				
C				

Figura 1.2: Vedi testo.

Escludendo gli ultimi tre casi col lato visibile bianco, rimangono tre casi col lato visibile rosso, due dei quali nascondono un lato anch'esso rosso. La probabilità è quindi pari a $2/3$.

Seconda soluzione

Utilizzando il teorema di Bayes e tenendo conto che la carta C non ha lati rossi, otteniamo

$$P(A|\text{vis rosso}) = \frac{P(\text{vis rosso}|A)P(A)}{P(\text{vis rosso})} = \frac{P(\text{vis rosso}|A)P(A)}{P(A)P(\text{vis rosso}|A) + P(B)P(\text{vis rosso}|B)}$$

Ora, $P(A) = P(B) = 1/3$. Inoltre, la carta A ha due lati rossi, per cui $P(\text{vis rosso}|A) = 1$, mentre la carta B uno solo per cui $P(\text{vis rosso}|B) = 1/2$. Mettendo tutto assieme otteniamo

$$\begin{aligned} P(\text{vis rosso}) &= P(A)P(\text{vis rosso}|A) + P(B)P(\text{vis rosso}|B) \\ &= \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{2} \end{aligned}$$

per cui

$$P(A|\text{vis rosso}) = \frac{1 \cdot 1/3}{1/2} = \frac{2}{3}$$

Esercizio 1.4.4. Monty Hall

Dietro una delle tre porte a , b e c c'è una macchina, dietro ognuna delle altre due una capra. Scopo del gioco è indovinare la porta dietro la quale si trova la macchina. Scegli la porta a e Monty, che vede che cosa si nasconde dietro le tre porte, scopre la porta c mostrandoti una capra. È meglio mantenere la scelta e chiedere di aprire la porta a o cambiare scelta e chiedere di aprire la porta b ?

Soluzione

Sia R_c l'evento *Monty sceglie di aprire la porta c*. Indichiamo con X l'evento *dietro la porta x c'è una macchina*. Valutiamo prima di tutto $P(R_c|A)$, $P(R_c|B)$ e $P(R_c|C)$. Se la macchina è dietro la porta a , Monty può aprire le porte b e c con uguale probabilità, per cui $P(R_c|A) = 1/2$. Se la macchina è dietro

la porta b , Monty può aprire solo la porta c , per cui $P(R_c|B) = 1$. Se la macchina è dietro la porta c , Monty non può aprire la porta c , per cui $P(R_c|C) = 0$. Pertanto

$$P(A|R_c) = \frac{P(A)P(R_c|A)}{P(A)P(R_c|A) + P(B)P(R_c|B) + P(C)P(R_c|C)} = \frac{\frac{1}{3}\frac{1}{2}}{\frac{1}{3}\frac{1}{2} + \frac{1}{3}1 + \frac{1}{3}0} = \frac{1}{3}$$

$$P(B|R_c) = \frac{P(B)P(R_c|B)}{P(A)P(R_c|A) + P(B)P(R_c|B) + P(C)P(R_c|C)} = \frac{\frac{1}{3}1}{\frac{1}{3}\frac{1}{2} + \frac{1}{3}1 + \frac{1}{3}0} = \frac{2}{3}$$

Se questo risultato ci lascia perplessi proviamo a pensare di ripetere il gioco con 1,000,000 porte, una macchina e 999,999 capre e con Monty che, dopo la nostra scelta, apre 999,998 porte mostrandoci altrettante capre...

Eventi indipendenti

Due eventi sono *indipendenti* se la realizzazione di uno non modifica la probabilità dell'altro. Se E e F sono indipendenti allora

$$P(EF) = P(E)P(F)$$

ovvero $P(E|F) = P(E)$ e $P(F|E) = P(F)$.

Esercizio 1.4.5. Ancora un doppio lancio

Lancia un dado onesto due volte. Verifica che l'evento E_1 la somma dei due risultati è 7 e l'evento F il primo risultato è 4 sono indipendenti, mentre l'evento E_2 la somma dei due risultati è 6 e F non lo sono.

Soluzione

Chiaramente abbiamo che $P(F) = 1/6$ e $P(E_1F) = P(E_2F) = 1/36$. Per quanto riguarda E_1 abbiamo 6 casi favorevoli su 36 possibili, per cui $P(E_1) = 1/6$. Per E_2 , invece, i casi favorevoli sono 5 e, quindi, $P(E_2) = 5/36$.

Osservazione 1.4.1. Strano ma vero

Il motivo per cui le cose cambiano è dovuto al fatto che mentre la probabilità di ottenere 7 con due lanci non dipende dal primo risultato, la probabilità di ottenere 6 richiede che il primo risultato non sia 6.

Esercizi non risolti

1. Hai due monete A e B . La probabilità di *testa* per A è $1/4$, per B $1/8$. Se ottieni *testa* con che probabilità hai lanciato la moneta A e con quale la moneta B ?
2. Hai dieci monete di tipo A e due monete di tipo B . Se ottieni *testa* con che probabilità hai lanciato una moneta di tipo A e con quale una moneta di tipo B ?
3. Risolvi *Monty Hall* nel caso di 1,000,000 di porte con una macchina e 999,999 capre.

1.5 Variabili casuali discrete

Con l'introduzione delle variabili casuali nel caso discreto e delle nozioni essenziali di valore atteso e varianza entriamo nel vivo della nostra breve incursione nella teoria della probabilità.

Variabili casuali

Molto spesso le quantità di interesse in un esperimento non sono i risultati ma una qualche funzione del risultato nota come *variabile casuale*. **Una variabile casuale è una funzione a valori reali definita sullo spazio campionario.** Nel caso di valori discreti, come la somma di due dadi, il numero di teste in n lanci o il genere alla nascita, una variabile casuale X è completamente definita in termini della probabilità con la quale assume ognuno dei suoi possibili valori.

Esercizio 1.5.1. Numero di teste in tre lanci di una moneta

Lancia 3 volte una moneta. Il numero X di teste ottenute è una variabile casuale i cui possibili valori sono 0, 1, 2 e 3. Fissa lo spazio campionario e determina la variabile casuale come funzione dallo spazio campionario ai reali.

Soluzione

Lo spazio campionario è l'insieme delle otto possibili triple

$$S = \{TTT, TTC, TCT, TCC, CTT, CTC, CCT, CCC\}$$

Valutiamo ora la probabilità con la quale X assume i valori 0, 1, 2 e 3 nell'assunzione che la moneta sia onesta. Poiché tutte le 8 triple sono ugualmente probabili abbiamo

$$\begin{aligned} P(X = 0) &= P(\{CCC\}) = 1/8 & P(X = 1) &= P(\{TCC, CTC, CCT\}) = 3/8 \\ P(X = 2) &= P(\{CTT, TCT, TTC\}) = 3/8 & P(X = 3) &= P(\{TTT\}) = 1/8 \end{aligned}$$

Esercizio 1.5.2. Palline numerate

Estrai casualmente 3 palline senza reinserimento tra 20 palline numerate da 1 a 20. Il numero estratto più grande X è una variabile casuale i cui possibili valori sono 3, 4, ..., 20. I risultati possibili dell'esperimento, ovvero lo spazio campionario, sono le combinazioni di 3 numeri scelti tra 1, 2, ..., 20. Valuta la probabilità con la quale X assume il valore i (con $3 \leq i \leq 20$) nell'assunzione di equiprobabilità.

Soluzione

Se i è il numero estratto più grande, le altre due palline sono numerate con una delle possibili coppie di numeri diversi compresi tra 1 e $i - 1$. Avremo pertanto

$$P(X = i) = p(i) = \binom{i-1}{2} / \binom{20}{3}, \text{ per } i = 3, 4, \dots, 20$$

che fatti i calcoli fornisce

i	$p(i)$	i	$p(i)$	i	$p(i)$	i	$p(i)$	i	$p(i)$	i	$p(i)$
3	1/1140	4	3/1140	5	6/1140	6	10/1140	7	15/1140	8	21/1140
9	28/1140	10	36/1140	11	45/1140	12	55/1140	13	66/1140	14	78/1140
15	91/1140	16	105/1140	17	120/1140	18	136/1140	19	153/1140	20	171/1140

Funzione di probabilità di massa

Nel caso di una variabile casuale X a valori discreti x_i con $i = 1, 2, \dots$, la *funzione di probabilità di massa* $p(\cdot)$ definita sulla retta reale, o *pmf* o anche solo *funzione di probabilità*, contiene tutta l'informazione necessaria per descrivere completamente X . Si ha che $p(x_i) = P(X = x_i) \geq 0$ con $\sum_i p(x_i) = 1$. La *pmf* per un dado onesto è illustrata a sinistra nella figura 1.3.

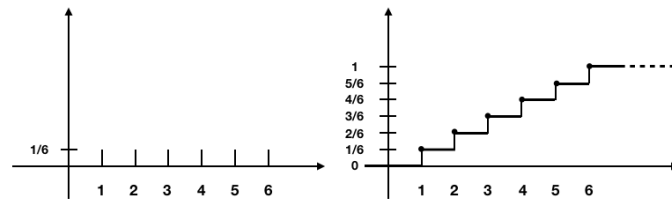


Figura 1.3: Funzione di probabilità di massa e funzione di probabilità cumulata per un dado onesto.

Funzione di probabilità cumulata

Ordiniamo i valori x_i in modo tale che $x_1 < x_2 < \dots < x_i < \dots$ e introduciamo la *funzione di probabilità cumulata* $F(a)$, o *cdf*, definita come

$$F(a) = \sum_{x_i \leq a} p(x_i)$$

È facile verificare che la funzione F è continua da destra e *crescente* da 0 a 1. Una *cdf* di una *pmf* è una funzione a gradini. Se i valori sono in numero finito, la *cdf* vale 0 a sinistra del valore più piccolo e 1 dal valore più grande in poi. L' i -esimo gradino è localizzato nel punto x_i e il salto corrispondente vale $p(x_i)$. La somma di tutti i gradini, ovviamente, è sempre 1. La *cdf* per un dado onesto è illustrata a destra nella figura 1.3.

Valore atteso

Introduciamo una delle nozioni centrali dell'intera teoria: il valore atteso di una variabile casuale. Dovremo attendere la legge dei grandi numeri per apprezzarne appieno l'importanza.

Il *valore atteso* μ di una variabile casuale X , indicato con $\mathbb{E}[X]$ e che non deve essere confuso con la media empirica, è la media pesata dei valori x_i che può assumere X . Ogni x_i è pesato con la sua probabilità $p(x_i)$ e quindi si ha

$$\mu = \mathbb{E}[X] = \sum_i x_i p(x_i)$$

Esercizio 1.5.3. *Valore atteso del numero di teste in tre lanci di una moneta onesta*

Valuta il valore atteso della variabile casuale dell'esercizio **1.5.1**.

Soluzione

$$\mu = 0 \cdot 1/8 + 1 \cdot 3/8 + 2 \cdot 3/8 + 3 \cdot 1/8 = 1.5$$

Valore atteso di una funzione di variabile casuale

Per calcolare il valore atteso di una funzione g di una variabile casuale discreta X possiamo determinare la *pmf* della variabile casuale discreta $g(X)$, oppure calcolare il valore atteso come media pesata.

Esercizio 1.5.4. *Due modi diversi di calcolare il valore atteso*

Sia $Y = X^2$. Calcola $\mathbb{E}[X^2]$ per una variabile casuale X con

$$P(X = -1) = 0.2, \quad P(X = 0) = 0.5 \quad \text{e} \quad P(X = 1) = 0.3$$

Soluzione

Poiché $P(Y = 1) = P(X = -1) + P(X = 1) = 0.5$ e $P(Y = 0) = 0.5$ otteniamo

$$\mathbb{E}[X^2] = 1 \cdot 0.5 + 0 \cdot 0.5 = 0.5$$

Calcoliamo ora il valore atteso come $\mathbb{E}[g(X)] = \sum_i g(x_i)p(x_i)$. In questo caso scriviamo

$$\mathbb{E}[X^2] = 1 \cdot 0.2 + 1 \cdot 0.3 = 0.5$$

Esercizio 1.5.5. Linearità

Valuta $\mathbb{E}[aX + b]$ con a e $b \in \mathbb{R}$ in funzione di $\mathbb{E}[X] = \sum_i x_i p(x_i)$.

Soluzione

Coerentemente con la struttura lineare del valore atteso, abbiamo

$$\mathbb{E}[aX + b] = \sum_i (ax_i + b)p(x_i) = a \sum_i x_i p(x_i) + b = a\mathbb{E}[X] + b$$

Varianza

Una seconda quantità che cattura proprietà importanti di una variabile casuale X è la *varianza* $Var(X)$ definita come $Var(X) = E[(X - \mu)^2]$.

Osservazione 1.5.1. Quadrato è meglio

Il quadrato nella definizione di varianza è fondamentale per ovviare al fatto che le differenze dal valore atteso hanno valore atteso nullo per qualunque X . Infatti

$$\mathbb{E}[(X - \mu)] = \mathbb{E}[X] - \mathbb{E}[\mu] = \mu - \mu = 0$$

Esercizio 1.5.6. Una formula utile

Dimostra che $\mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

Dimostrazione

$$\mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 + \mu^2 - 2\mu X] = \mathbb{E}[X^2] + (\mathbb{E}[X])^2 - 2(\mathbb{E}[X])^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Esercizio 1.5.7. Nonlinearità della varianza

Dimostra che $Var(aX + b) = a^2 Var(X)$.

Dimostrazione

$$Var(aX + b) = \mathbb{E}[(aX + b - a\mathbb{E}[X] - b)^2] = \mathbb{E}[(aX - a\mathbb{E}[X])^2] = a^2 Var(X)$$

Osservazione 1.5.2. Deviazione standard

Una quantità molto usata è la radice quadrata della varianza, nota come *deviazione standard*, o

$$SD(X) = \sqrt{Var(X)}$$

Esercizi non risolti

1. Calcola il valore atteso e la varianza della variabile casuale X se $p(1) = 1/2$, $p(2) = 1/4$, $p(3) = 1/8$ e $p(4) = 1/8$. Disegna e commenta il grafico della *cdf*.
2. Sia X la variabile casuale che conta il numero di 6 ottenuti lanciando un dado onesto tre volte. Determina la *pmf* e la *cdf* di X . Con che probabilità ottieni almeno un 6? E con che probabilità un numero pari di 6?
3. Calcola il valore atteso e la varianza della variabile casuale X dell'esercizio precedente.

1.6 Distribuzioni discrete di probabilità

Prendiamo ora in considerazione importanti distribuzioni di probabilità nel caso discreto.

Bernoulli

Una variabile casuale di *Bernoulli* X assume due soli valori, 0 e 1 (talvolta associati al fallimento e al successo di un esperimento), con funzione di probabilità di massa $P(X = 0) = 1 - p$ e $P(X = 1) = p$ con $0 < p < 1$.

Esercizio 1.6.1. Calcoliamo il valore atteso e la varianza di una variabile casuale di Bernoulli.

Da una diretta applicazione delle definizioni di valore atteso e varianza si ha,

$$\mathbb{E}[X] = p \cdot 1 + (1 - p) \cdot 0 = p \quad \text{e} \quad \text{Var}(X) = p(1 - p)^2 + (1 - p)p^2 = p(1 - p)$$

Binomiale

La variabile casuale *binomiale* X conta i successi in una sequenza di n realizzazioni indipendenti di una variabile casuale di Bernoulli con $p(1) = p$. La sua funzione di probabilità di massa si scrive come

$$p(i) = \binom{n}{i} p^i (1 - p)^{n-i} \quad i = 0, 1, \dots, n$$

Il coefficiente binomiale $\binom{n}{i}$ conta in quanti modi diversi si possono realizzare i successi in una sequenza di n realizzazioni indipendenti.

Esercizio 1.6.2. Somma sempre uguale a 1

Verifica che $\sum_i p(i) = 1$.

Soluzione

Riconoscendo nella scrittura della somma il binomio di Newton si ha

$$\sum_{i=0}^n \binom{n}{i} p^i (1 - p)^{n-i} = (p + (1 - p))^n = 1$$

Osservazione 1.6.1. Valore atteso e varianza della binomiale

Per il valore atteso di una binomiale abbiamo

$$\mathbb{E}[X] = np$$

mentre per la varianza

$$\text{Var}(X) = np(1 - p)$$

La derivazione di questi due risultati richiede passaggi algebrici un po' noiosi. Tra qualche lezione vedremo che entrambi discendono prontamente da proprietà di base dei valori attesi. Il risultato sul valore atteso dalle proprietà di linearità del valore atteso applicato alla somma di n variabili casuali di Bernoulli, quello sulla varianza dal fatto che le n variabili sono indipendenti.

Geometrica

La variabile casuale *geometrica* X vale n se si ottiene un successo dopo $n - 1$ fallimenti in una sequenza di n realizzazioni indipendenti di una variabile casuale di Bernoulli. La pmf, vedi figura 1.4, è

$$P(X = n) = (1 - p)^{n-1} p \quad \text{per } n = 1, 2, \dots$$

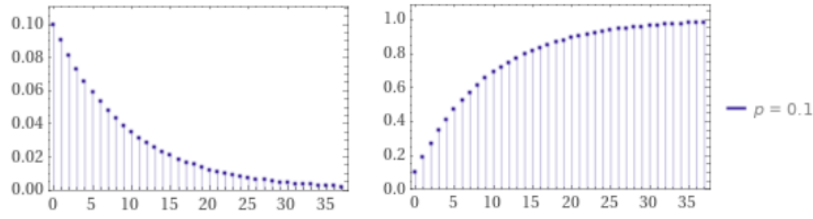


Figura 1.4: *Pmf e cdf per una distribuzione geometrica con $p = 1/10$.*

Esercizio 1.6.3. *Verifica della somma a 1*

Verifica che $\sum_{i=1}^{\infty} (1-p)^{i-1} p = 1$.

Soluzione

Per $0 < p < 1$ abbiamo che

$$\sum_{k=0}^{\infty} p^k = \frac{1}{1-p} \quad \text{per } p \rightarrow 1-p \text{ fornisce } \sum_{k=0}^{\infty} (1-p)^k = \frac{1}{1-(1-p)} = \frac{1}{p}$$

Pertanto abbiamo

$$\sum_{i=1}^{\infty} (1-p)^{i-1} p = \sum_{i=0}^{\infty} (1-p)^{i-1} p = \frac{1}{p} \cdot p = 1$$

Esercizio 1.6.4. *Valore atteso di una geometrica*

Decomponiamo la somma in due addendi

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} i(1-p)^{i-1} p = \sum_{i=1}^{\infty} (i-1+1)(1-p)^{i-1} p = \sum_{i=1}^{\infty} (i-1)(1-p)^{i-1} p + \sum_{i=1}^{\infty} (1-p)^{i-1} p$$

Ponendo $j = i - 1$ per la seconda serie otteniamo $\sum_{i=1}^{\infty} (1-p)^{i-1} p = \sum_{j=0}^{\infty} (1-p)^j p$. Inoltre, sempre ponendo $j = i - 1$ si ha

$$\sum_{i=1}^{\infty} (i-1)(1-p)^{i-1} p = \sum_{j=0}^{\infty} j(1-p)^j p = \sum_{j=1}^{\infty} j(1-p)^j p = (1-p) \sum_{j=1}^{\infty} j(1-p)^{j-1} p = (1-p) \mathbb{E}[X]$$

dove abbiamo escluso il termine nullo, raccolto $1-p$, e infine ricordato l'espressione del valore atteso. Combinando queste equazioni si ha che

$$\mathbb{E}[X] = (1-p)\mathbb{E}[X] + 1$$

da cui abbiamo che $\mathbb{E}[X] = 1/p$. Questo risultato ci riconcilia con l'intuizione che se la probabilità di ottenere *testa* con una moneta truccata in cui la probabilità p di testa è $1/10$ ci aspettiamo di ottenere *testa* una volta ogni 10 lanci. Dal grafico della *cdf* di figura 1.4 notiamo che la probabilità di ottenere almeno una volta testa con 10 lanci è intorno al %70.

Osservazione 1.6.2. *Ancora sulla varianza*

In modo simile si ottiene $\text{Var}(X) = (1-p)/p^2$.

Poisson

Una variabile casuale di *Poisson* è definita dalla *pmf*

$$P(X = i) = \frac{\mu^i}{i!} e^{-\mu} \quad \text{con } i = 0, 1, 2, \dots$$

con

$$\sum_{i=0}^{\infty} \frac{\mu^i}{i!} e^{-\mu} = e^{-\mu} \sum_{i=0}^{\infty} \frac{\mu^i}{i!} = e^{-\mu} e^{\mu} = 1 \quad \text{in quanto} \quad e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \quad \forall x \in \mathbb{R}$$

Il numero di errori di stampa per pagina, il numero di ultracentenari di una comunità, il numero di numeri di telefono sbagliati da un centralino, il numero di pacchi di biscotti venduti in un giorno, il numero di clienti in un ufficio postale sono esempi di variabili casuali di Poisson. La figura 1.5 mostra tre distribuzioni di Poisson al variare di μ .

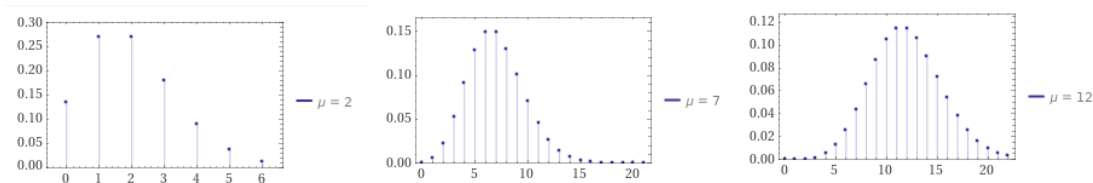


Figura 1.5: Vedi testo.

Esercizio 1.6.5. Valore atteso di una Poissoniana

Calcola il valore atteso di una variabile casuale di Poisson.

Soluzione

Si ha che

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} i e^{-\mu} \frac{\mu^i}{i!} = \sum_{i=1}^{\infty} i e^{-\mu} \frac{\mu^i}{i!} = \mu \sum_{i=1}^{\infty} e^{-\mu} \frac{\mu^{i-1}}{(i-1)!} = \mu \sum_{j=0}^{\infty} e^{-\mu} \frac{\mu^j}{j!} = \mu$$

dove abbiamo escluso il termine nullo, semplificato i , raccolto μ e, infine, rinominato gli indici.

Osservazione 1.6.3. Varianza di una Poissoniana

La varianza di una variabile casuale di Poisson è ancora μ .

Osservazione 1.6.4. Poisson e binomiale

Per n grande, p piccolo e $\mu = np \sim 1$, la pmf della binomiale tende a una Poissoniana con $\mathbb{E}[X] = \mu$. Infatti possiamo scrivere

$$\begin{aligned} P(X = i) &= \binom{n}{i} p^i (1-p)^{n-i} \sim \frac{n!}{(n-i)! i!} \left(\frac{\mu}{n}\right)^i \left(1 - \frac{\mu}{n}\right)^{n-i} \\ &= \frac{n(n-1) \dots (n-i+1)}{n^i} \frac{\mu^i}{i!} \left(1 - \frac{\mu}{n}\right)^{-i} \left(1 - \frac{\mu}{n}\right)^n \\ &\sim 1 \cdot \frac{\mu^i}{i!} \cdot 1 \cdot e^{-\mu} = e^{-\mu} \frac{\mu^i}{i!} \end{aligned}$$

Esercizi non risolti

1. Che tipo di variabile casuale è la X che conta il numero di 6 ottenuti lanciando un dado onesto tre volte? Calcola il valore atteso e la varianza di X applicando le formule appropriate.
2. Se un algoritmo restituisce la risposta corretta il 50% delle volte quante volte devi lanciarlo per ottenere il risultato corretto con probabilità superiore al 99,9%?
3. Un libro contiene in media un errore di stampa ogni pagina. Con che probabilità contiene due errori in una stessa pagina?

1.7 Variabili casuali continue

Estendiamo il nostro studio al caso di variabili casuali che assumono valori nel continuo.

Funzione densità di probabilità

L'insieme dei valori che può assumere una variabile casuale spesso non è finito o numerabile (pensiamo al tempo di vita di un componente, all'ora d'arrivo di un treno o al tempo di percorrenza di un viaggio in auto). Una variabile casuale X è continua se esiste una funzione $f : \mathbb{R} \rightarrow \mathbb{R}^+$ tale che

$$P(X \in B) = \int_B f(x) dx \quad (1.3)$$

su ogni sottoinsieme misurabile $B \subset \mathbb{R}$ (la misurabilità è una condizione tecnica che, per i nostri scopi, non ha conseguenze rilevanti in quanto tutti i sottoinsiemi di nostro interesse sono misurabili). La funzione f è la *densità di probabilità*, o *pdf*.

Osservazione 1.7.1. Diversamente dal caso discreto

La probabilità che una variabile casuale continua X assuma un determinato valore x è sempre 0. Se $B = (x - \epsilon/2, x + \epsilon/2)$ la probabilità 1.3 diventa

$$P(x - \epsilon/2 < X < x + \epsilon/2) = \int_{x-\epsilon/2}^{x+\epsilon/2} f(t) dt \approx \epsilon \cdot f(x)$$

Fissato un intervallo di ampiezza ϵ , pertanto, la *pdf* esprime la probabilità che il valore assunto da X sia vicino a x .

Esercizio 1.7.1. Una questione di normalizzazione

Se $f(x) = C(4x - 2x^2)$ per $0 < x < 2$ e 0 altrimenti, calcola il valore di C per cui la funzione f è una densità di probabilità e valuta $P(X > 1)$.

Soluzione

La costante C si ottiene imponendo la condizione di normalizzazione $P(X \in S) = 1$ che, in questo caso, diventa

$$\frac{1}{C} = \int_0^2 (4x - 2x^2) dx = \left(2x^2 - \frac{2}{3}x^3 \right) \Big|_0^2 = \frac{8}{3}$$

da cui ricaviamo $C = 3/8$. Calcolando l'integrale definito tra 1 e 2 della *pdf* normalizzata otteniamo $P(X > 1) = 1/2$. \square

Di seguito vediamo come le definizioni introdotte per le variabili casuali discrete si estendono al caso continuo, considerando le *pdf* al posto delle *pmf* e integrali invece che sommatorie.

Funzione di distribuzione cumulata

La *funzione di distribuzione cumulata* $F : \mathbb{R} \rightarrow [0, 1]$, o *cdf*, è definita $\forall a \in \mathbb{R}$ come

$$F(x) = \int_{-\infty}^x f(t) dt$$

Come nel caso discreto la *cdf* è una funzione crescente non negativa compresa tra 0 e 1. La *pdf* e la *cdf* forniscono due caratterizzazioni equivalenti delle variabili casuali continue.

Osservazione 1.7.2. Per il Teorema fondamentale del calcolo integrale

Applicando la prima parte di questo teorema troviamo che se $f : (a, b) \rightarrow \mathbb{R}$ e

$$F(x) = \int_{-\infty}^x f(t)dt$$

allora

$$\frac{dF}{dx}(x) = f(x)$$

Osservazione 1.7.3. Che succede con la discontinuità

Se la densità presenta punti di discontinuità, la *cdf* è sempre continua ma non è più ovunque derivabile. La funzione densità è ottenibile come la derivata della *cdf* escludendo i punti angolosi (vedi figura 1.6).

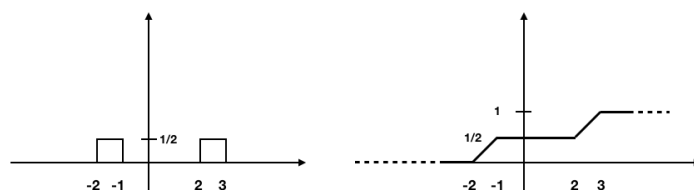


Figura 1.6: Funzione densità con punti di discontinuità e corrispondente *cdf*. La *cdf*, l'integrale definito della funzione densità da $-\infty$ a x , è sempre una funzione continua. I punti di discontinuità della funzione densità creano spigoli (punti di non derivabilità) della *cdf*.

Valore atteso e varianza

In piena analogia con il caso discreto definiamo il valore atteso μ e la varianza $Var(X)$ di una variabile casuale X continua come

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x) dx \quad \text{e} \quad Var(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

Esercizio 1.7.2. Un semplice calcolo

Mostra che $\mathbb{E}[X] = 2/3$ per la variabile casuale X con densità di probabilità

$$f(x) = \begin{cases} 2x & \text{per } 0 < x < 1 \\ 0 & \text{altrimenti} \end{cases}$$

Soluzione

Per il valore atteso abbiamo

$$\mathbb{E}[X] = \int_0^1 2x^2 dx = \frac{2}{3}$$

Esercizio 1.7.3. Uno un po' più difficile

Calcola $\mathbb{E}[e^X]$ per la variabile casuale X con densità di probabilità

$$f(x) = \begin{cases} 1 & \text{per } 0 \leq x \leq 1 \\ 0 & \text{altrimenti} \end{cases}$$

Soluzione

In analogia col caso discreto scriviamo $\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f(x) dx$. Quindi,

$$\mathbb{E}[e^X] = \int_{-\infty}^{+\infty} e^x f(x) dx = \int_0^1 e^x dx = e - 1$$

Distribuzione di una funzione di variabile casuale

Data una variabile casuale X di distribuzione nota vogliamo trovare la distribuzione di $g(X)$ per una funzione g data. Nel caso di g monotona è sufficiente un po' di attenzione.

Esempio 1.7.1. *Prima le cose facili*

Sia X distribuita uniformemente tra 0 e 1. Abbiamo quindi $f(x) = 1$ e $F(x) = x$ tra 0 e 1. Se $Y = X^n$ allora

$$F_Y(y) = P(Y \leq y) = P(X^n \leq y) = P(X \leq y^{1/n}) = F(y^{1/n}) = y^{1/n}$$

Pertanto, per $0 \leq y \leq 1$, derivando $F_Y(y)$ otteniamo

$$f_Y(y) = \frac{y^{(1-n)/n}}{n}$$

Esempio 1.7.2. *Poi quelle un po' più difficili*

Data $f(x)$ per X , troviamo f_Y per $Y = X^2$. Per $y \geq 0$ abbiamo

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F(\sqrt{y}) - F(-\sqrt{y})$$

la cui derivata per $y \geq 0$ fornisce

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}$$

Esercizio 1.7.4. *Una seconda soluzione per l'esercizio 1.7.3*

Per $0 \leq x \leq 1$ abbiamo che $1 \leq e^x \leq e$. Ora, per $1 \leq a \leq e$,

$$F_Y(a) = P\{Y \leq a\} = P\{e^X \leq a\} = P\{X \leq \ln a\} = \int_0^{\ln a} f(x) dx = \int_0^{\ln a} 1 dx = \ln a$$

Segue che

$$f_Y(a) = \frac{dF_Y(a)}{da} = \frac{1}{a} \quad \text{per } 1 \leq a \leq e$$

e 0 altrimenti. Pertanto,

$$\mathbb{E}[e^X] = \mathbb{E}[Y] = \int_1^e x \left(\frac{1}{x}\right) dx = e - 1$$

Esercizi non risolti

1. Per quale valore della costante C la funzione

$$f(x) = \begin{cases} Cx & \text{se } 0 \leq x < 1 \\ C & \text{altrimenti} \end{cases}$$

per x nell'intervallo $[0, 2]$ è una *pdf*?

2. Se la *pdf* di una variabile casuale X è definita come

$$f(x) = \begin{cases} 1 & \text{se } 0 \leq x < 1 \\ 0 & \text{altrimenti} \end{cases}$$

calcola la *cdf*, $\mathbb{E}[X]$ e $Var(X)$.

3. Se $Y = X^n$, dove X è la variabile casuale dell'esercizio precedente, calcola $\mathbb{E}[Y]$ e commenta il risultato che ottieni per n grande.

1.8 Distribuzioni continue di probabilità

Anche nel caso continuo vale la pena soffermarsi su alcune funzioni di distribuzioni importanti.

Digressione sull'integrazione per parti

Siano f e g due funzioni continue e derivabili. La derivata del prodotto delle due funzioni é

$$\frac{d}{dx}(f(x)g(x)) = \frac{df(x)}{dx}g(x) + f(x)\frac{dg(x)}{dx} = f'(x)g(x) + f(x)g'(x)$$

Considerando l'integrale di entrambi i membri, applicando il teorema fondamentale del calcolo integrale e riordinando i termini, otteniamo

$$\int f'(x)g(x)dx = f(x)g(x) - \int f(x)g'(x)dx + C$$

dove C è una costante arbitraria. La forza di questo metodo risiede nella capacità di individuare, quale tra le funzioni f e g sia più facilmente derivabile o integrabile in modo da poter semplificare l'integrale. Nel caso di integrale definito su un intervallo $[a, b]$ la costante C scompare e si ottiene

$$\int_a^b f'(x)g(x)dx = f(x)g(x)\Big|_a^b - \int_a^b f(x)g'(x)dx$$

Distribuzione normale (o Gaussiana)

Una variabile casuale normale $X = \mathcal{N}(\mu, \sigma^2)$, con μ e $\sigma^2 > 0$ entrambi parametri reali fissati, ha come pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$$

Osservazione 1.8.1. Normalizzazione garantita

Omettiamo la verifica che l'integrale di f vale 1 perché troppo laboriosa. Ci limitiamo a osservare che il valore di $f(0)$ è inversamente proporzionale alla costante σ . Tre distribuzioni normali per diversi valori di μ e σ sono mostrate in figura 1.7.

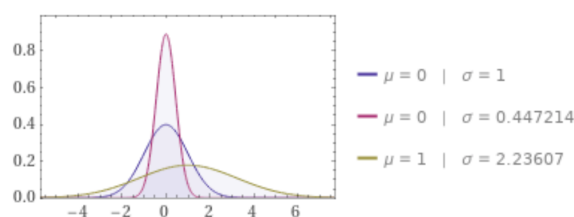


Figura 1.7: Vedi testo.

Esercizio 1.8.1. Valore atteso e varianza per la normale standard $Z = \mathcal{N}(0, 1)$

Determina $\mathbb{E}[Z]$ e $\text{Var}(Z)$.

Soluzione

Abbiamo

$$\mathbb{E}[Z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x e^{-x^2/2} dx = 0$$

perché la funzione integranda è dispari. Per la varianza otteniamo

$$Var(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-x^2/2} dx = 1$$

Infatti, ponendo

$$f(x) = \frac{x}{\sqrt{2\pi}} \implies f'(x) = \frac{1}{\sqrt{2\pi}}$$

e

$$g(x) = -e^{-\frac{x^2}{2}} \implies g'(x) = xe^{-\frac{x^2}{2}}$$

la formula di integrazione per parti fornisce

$$\int_{-\infty}^{+\infty} \frac{x^2 e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = -\frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}} \Big|_{-\infty}^{+\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = 0 + 1$$

Osservazione 1.8.2. Una gaussiana è sempre una gaussiana

Se $X = \mathcal{N}(\mu, \sigma^2)$ e $Y = aX + b$ con a e b reali qualunque, abbiamo che la variabile casuale Y è gaussiana con $Y = \mathcal{N}(a\mu + b, a^2\sigma^2)$. Infatti,

$$F_Y(x) = P(Y \leq x) = P(aX + b \leq x) = P\left(X \leq \frac{x-b}{a}\right) = F_X\left(\frac{x-b}{a}\right) \quad e$$

$$f_Y(x) = \frac{dF_Y(x)}{dx} = \frac{f_X\left(\frac{x-b}{a}\right)}{a} = \frac{1}{a\sigma\sqrt{2\pi}} e^{-((x-b)/a-\mu)^2/2\sigma^2} = \frac{1}{a\sigma\sqrt{2\pi}} e^{-(x-a\mu-b)^2/2a^2\sigma^2}.$$

In particolare, ponendo $a = 1/\sigma$ e $b = -\mu/\sigma$ otteniamo

$$Y = (X - \mu)/\sigma = \mathcal{N}(0, 1) = Z$$

ovvero una normale standard! Invertendo la relazione, poichè $X = \sigma Z + \mu$, abbiamo che per una variabile normale $X = \mathcal{N}(\mu, \sigma^2)$

$$\mathbb{E}[X] = \mu \quad e \quad Var(X) = \sigma^2$$

Distribuzione esponenziale

La densità di probabilità di una variabile casuale esponenziale è $f(x) = \lambda e^{-\lambda x}$ per $x \geq 0$ e 0 altrimenti. Tre distribuzioni esponenziali per diversi valori di λ sono mostrate in figura 1.8.

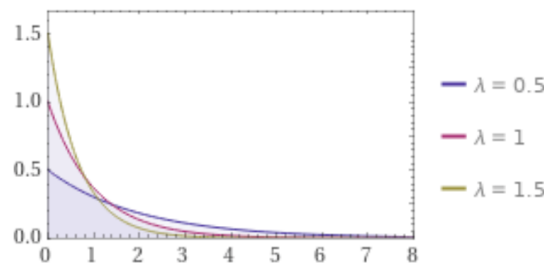


Figura 1.8: Vedi testo.

Osservazione 1.8.3. *La solita proprietà di normalizzazione*

Verifica che

$$\int_0^{+\infty} \lambda e^{-\lambda x} dx = 1$$

Soluzione

$$\int_0^{+\infty} e^{-\lambda x} dx = -\frac{1}{\lambda} e^{-\lambda x} \Big|_0^{+\infty} = 0 - \left(-\frac{1}{\lambda}\right) = \frac{1}{\lambda}$$

Osservazione 1.8.4. *Funzione di probabilità cumulata*

$$F(x) = P\{X \leq x\} = -\lambda \frac{1}{\lambda} e^{-\lambda x} \Big|_0^x = 1 - e^{-\lambda x} \quad \text{per } x \geq 0$$

Osservazione 1.8.5. *Una distribuzione smemorata*

Una distribuzione per la quale $P(X > s+t | X > t) = P(X > s)$ per tutti gli $s, t \geq 0$, è *senza memoria*. L'esponenziale è senza memoria, infatti,

$$P(X > s+t | X > t) = \frac{P(X > s+t, X > t)}{P(X > t)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} = P(X > s)$$

Esercizio 1.8.2. $\mathbb{E}[X] = 1/\lambda$

Integrando per parti si ha

$$\mathbb{E}[X] = \int_0^{+\infty} x \lambda e^{-\lambda x} dx = -x e^{-\lambda x} \Big|_0^{+\infty} - \left(-\int_0^{+\infty} e^{-\lambda x} dx\right) = 0 + 1/\lambda$$

Esercizio 1.8.3. $\text{Var}(X) = 1/\lambda^2$

Poiché $\text{Var}(X) = E[X^2] - (\mathbb{E}[X])^2$ è sufficiente calcolare $\mathbb{E}[X^2]$. Integrando per parti si ha

$$\mathbb{E}[X^2] = \int_0^{+\infty} x^2 \lambda e^{-\lambda x} dx = -x^2 e^{-\lambda x} \Big|_0^{+\infty} - \left(-\int_0^{+\infty} 2x e^{-\lambda x} dx\right) = 2 \int_0^{+\infty} x e^{-\lambda x} dx$$

Moltiplicando e dividendo per λ si ottiene

$$\mathbb{E}[X^2] = \frac{2}{\lambda} \mathbb{E}[X] = \frac{2}{\lambda^2}$$

per cui

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

1.9 Distribuzioni congiunte e indipendenza

Estendiamo quanto visto precedentemente al caso di più variabili casuali.

Caso discreto

La distribuzione di una coppia di variabili congiunte (X, Y) con valori in $\{x_1, \dots, x_N\}$ e $\{y_1, \dots, y_M\}$ è data da una *pmf*

$$P(X = x_i, Y = y_j) = p(x_i, y_j) \quad \text{per } i = 1, \dots, N, \text{ e } j = 1, \dots, M$$

Distribuzioni marginali A partire dalla *pmf congiunta* di due variabili casuali è immediato determinare le *pmf* per le singole variabili, note come *marginali*, sommando su tutti i possibili valori assunti dall'altra variabile casuale. Ovvero

$$p_X(x_i) = P(X = x_i) = \sum_{j=1}^M p(x_i, y_j) \quad \text{e} \quad p_Y(y_j) = P(Y = y_j) = \sum_{i=1}^N p(x_i, y_j)$$

Esercizio 1.9.1. Palline di tre colori

Estrai 3 palline a caso da un'urna che ne contiene 3 rosse, 4 bianche e 5 blu. Se X conta le rosse estratte e Y le bianche, calcola la distribuzione congiunta di probabilità di massa e le probabilità marginali.

Soluzione

Abbiamo $\binom{12}{3} = 220$ casi possibili. Per i casi favorevoli otteniamo

$$(0, 0) \rightarrow \binom{5}{3} = 10 \quad (0, 1) \rightarrow \binom{4}{1} \binom{5}{2} = 40 \quad (0, 2) \rightarrow \binom{4}{2} \binom{5}{1} = 30 \quad (0, 3) \rightarrow \binom{4}{3} = 4$$

$$(1, 0) \rightarrow \binom{3}{1} \binom{5}{2} = 30 \quad (1, 1) \rightarrow \binom{3}{1} \binom{4}{1} \binom{5}{1} = 60 \quad (1, 2) \rightarrow \binom{3}{1} \binom{4}{2} = 18$$

$$(2, 0) \rightarrow \binom{3}{2} \binom{5}{1} = 15 \quad (2, 1) \rightarrow \binom{3}{2} \binom{4}{1} = 12$$

$$(3, 0) \rightarrow \binom{3}{3} = 1$$

Le probabilità congiunte e le marginali, che si chiamano così perché collocate nel margine destro per la X e nel margine inferiore per la Y , sono

$p(0, 0) = 10/220$	$p(0, 1) = 40/220$	$p(0, 2) = 30/220$	$p(0, 3) = 4/220$	$p_X(0) = 84/220$
$p(1, 0) = 30/220$	$p(1, 1) = 60/220$	$p(1, 2) = 18/220$		$p_X(1) = 108/220$
$p(2, 0) = 15/220$	$p(2, 1) = 12/220$			$p_X(2) = 27/220$
$p(3, 0) = 1/220$				$p_X(3) = 1/220$

$$p_Y(0) = 56/220 \quad p_Y(1) = 112/220 \quad p_Y(2) = 48/220 \quad p_Y(3) = 4/220$$

È appena il caso di rimarcare che $p(i, j) \neq p_X(i)p_Y(j)$.

Esercizio 1.9.2. Da 0 a 3 figli

In un paese il 15% delle famiglie non ha figli, il 20% uno, il 35% due e il 30% tre. Se la probabilità di essere maschio (M) o femmina (F) per un figlio è la stessa, determina la distribuzione di probabilità congiunta per $P(M = i, F = j)$ e le marginali. $i, j = 0, \dots, 3$.

Soluzione

Abbiamo che $P(0, 0) = P(0 \text{ figli}) = 0.15$. Dal fatto che $P(0, 1) = P(1 \text{ figlio})P(F|1 \text{ figlio})$ segue che $P(0, 1) = 0.2 \times 0.5 = 0.1$ e lo stesso per $P(1, 0)$. Per $P(2, 0)$, e similmente per $P(0, 2)$, abbiamo $P(2 \text{ figli})P(2 F|2 \text{ figli}) = 0.35 \times 0.25 = 0.0875$. Poiché $P(2 \text{ figli}) = 0.35$ abbiamo che $P(1, 1) = 0.175$ (la figlia femmina può essere la prima o la seconda). Ragionando in modo analogo otteniamo le restanti probabilità riassunte nella tabella

	$M = 0$	$M = 1$	$M = 2$	$M = 3$	
$F = 0$	0.15	0.10	0.0875	0.0375	0.3750
$F = 1$	0.10	0.175	0.1125	0	0.3875
$F = 2$	0.0875	0.1125	0	0	0.20
$F = 3$	0.0375	0	0	0	0.0375
	0.3750	0.3875	0.20	0.0375	

Funzione di distribuzione cumulata congiunta È data da

$$F(a, b) = \sum_{i: x_i \leq a} \sum_{j: y_j \leq b} p(x_i, y_j)$$

Come nel caso di singola variabile casuale abbiamo che $0 \leq F(a, b) \leq 1$.

Cumulate marginali A partire dalla *cdf congiunta* $F(a, b)$ di due variabili casuali è immediato definire le *cdf* per le singole variabili, $F_X(a)$ e $F_Y(b)$, note come *cdf marginali*

$$F_X(a) = P(X \leq a) = P(X \leq a, Y < +\infty) = F(a, +\infty)$$

$$F_Y(b) = P(Y \leq b) = P(X < +\infty, Y \leq b) = F(+\infty, b)$$

Esercizio 1.9.3. Spesso assieme

Dimostra che $P(X > a, Y > b) = 1 + F(a, b) - F_X(a) - F_Y(b)$.

$$\begin{aligned} P(X > a, Y > b) &= 1 - P((X > a, Y > b)^c) = 1 - P((X > a)^c \cup (Y > b)^c) \\ &= 1 - P((X \leq a) \cup (Y \leq b)) \\ &= 1 - (P(X \leq a) + P(Y \leq b) - P(X \leq a, Y \leq b)) \\ &= 1 + F(a, b) - F_X(a) - F_Y(b). \end{aligned}$$

Caso continuo

Il caso continuo è in completa analogia col discreto, ma richiede l'uso di integrali doppi. Per valutare la probabilità che la coppia X, Y assuma valori nel dominio C dobbiamo saper calcolare

$$P((X, Y) \in C) = \int \int_{(x, y) \in C} f(x, y) dx dy = P(X \in A, Y \in B) = \int_B \left(\int_A f(x, y) dx \right) dy$$

con

$$F(a, b) = \int_{-\infty}^b \left(\int_{-\infty}^a f(x, y) dx \right) dy \quad \text{e} \quad f(a, b) = \frac{\partial^2 F(a, b)}{\partial a \partial b}$$

Inoltre, per le probabilità marginali, se

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy \quad \text{e} \quad f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx$$

abbiamo

$$P\{X \in A\} = P\{X \in A, Y < +\infty\} = \int_A \left(\int_{-\infty}^{+\infty} f(x, y) dy \right) dx = \int_A f_X(x) dx$$

e

$$P\{X \in B\} = P\{X < +\infty, Y \in B\} = \int_B \left(\int_{-\infty}^{+\infty} f(x, y) dx \right) dy = \int_B f_Y(y) dy$$

Variabili casuali indipendenti

Due variabili casuali sono indipendenti se per ogni coppia di insiemi A e B

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

Una definizione equivalente richiede che per ogni coppia a e b di numeri reali

$$P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq b)$$

Se X e Y sono indipendenti, allora

$$F(a, b) = F_X(a)F_Y(b)$$

per le funzioni di distribuzione cumulate e

$$f(x, y) = f_X(x)f_Y(y) \quad \text{e} \quad p(x, y) = p_X(x)p_Y(y)$$

rispettivamente per le *funzioni densità* nel caso continuo e le *pmf* nel caso discreto.

Esercizi non risolti

1. Lanci una moneta con $P(\text{testa}) = 1/10$ e un dado con $P(1) = P(2) = P(3) = P(4) = 1/8$ e $P(5) = P(6) = 1/4$. Scrivi la tabella delle probabilità congiunte e verifica l'uguaglianza con il prodotto delle corrispondenti probabilità marginali.
2. Calcola la *cdf* per l'esercizio precedente.
3. Calcola la *cdf* per l'esercizio **1.9.2**.

Capitolo 2

Valori attesi

2.10 Somme di variabili casuali

La capitale importanza della nozione di valore atteso merita un'attenzione speciale. Partiamo dal ruolo giocato dal valore atteso nel caso di somme di variabili casuali.

Funzione di variabili casuali

Il valore atteso di $g(X, Y)$, una variabile casuale funzione di X e Y nel caso discreto può essere calcolato come

$$\mathbb{E}[g(X, Y)] = \sum_x \sum_y g(x, y)p(x, y)$$

Se $g(X, Y) = X + Y$ abbiamo

$$\mathbb{E}[X + Y] = \sum_x \sum_y (x + y)p(x, y) = \sum_x \sum_y xp(x, y) + \sum_y \sum_x yp(x, y) = \mathbb{E}[X] + \mathbb{E}[Y]$$

Osservazione 2.10.1. *Valore atteso di una variabile casuale binomiale*

Poiché $X = \sum_i X_i$ è la somma di variabili casuali di Bernoulli con $\mathbb{E}[X_i] = p$ per tutti gli i , abbiamo

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^n X_i\right] = np$$

Media e varianza campionaria

Se le X_i per $i = 1, \dots, n$ sono variabili casuali identicamente e indipendentemente distribuite con valore atteso μ e varianza σ^2 definiamo la *media campionaria* μ_n e la *varianza campionaria* σ_n^2 come

$$\mu_n = \frac{\sum_i X_i}{n} \quad \text{e} \quad \sigma_n^2 = \frac{\sum_i (X_i - \mu_n)^2}{n - 1}$$

Calcola $\mathbb{E}[\mu_n]$, $Var(\mu_n)$ e $\mathbb{E}[\sigma_n^2]$.

Soluzione

$$\begin{aligned}\mathbb{E}[\mu_n] &= \mathbb{E}\left[\frac{\sum_i X_i}{n}\right] = \frac{\mathbb{E}[\sum_i X_i]}{n} = \frac{\sum_i \mathbb{E}[X_i]}{n} = \mu \\ Var(\mu_n) &= \frac{1}{n^2} Var\left(\sum_i X_i\right) = \frac{1}{n^2} \sum_i Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \\ \mathbb{E}[\sigma_n^2] &= \mathbb{E}\left[\frac{\sum_i (X_i - \mu_n)^2}{n-1}\right] = \frac{\mathbb{E}[\sum_i (X_i - \mu + \mu - \mu_n)^2]}{n-1} \\ &= \frac{\mathbb{E}[\sum_i (X_i - \mu)^2] + \mathbb{E}[\sum_i (\mu - \mu_n)^2] - 2\mathbb{E}[\sum_i (\mu - X_i) \sum_i (\mu - \mu_n)]}{n-1} \\ &= \frac{n\sigma^2 + nVar(\mu_n) - 2nVar(\mu_n)}{n-1} = \frac{n\sigma^2 - nVar(\mu_n)}{n-1} = \frac{(n-1)\sigma^2}{n-1} = \sigma^2\end{aligned}$$

Covarianza e varianza di somme

Se X e Y sono indipendenti, $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$. Infatti

$$\begin{aligned}\mathbb{E}[g(X)h(Y)] &= \sum_x \sum_y g(x)h(y)p(x, y) = \sum_x \sum_y g(x)h(y)p(x)p(y) \\ &= \sum_x g(x)p(x) \sum_y h(y)p(y) = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]\end{aligned}\quad (2.1)$$

La covarianza di due variabili casuali X e Y è definita come

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

Esercizio 2.10.1. Come per la varianza

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

Soluzione

$$\begin{aligned}Cov(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[(XY - \mathbb{E}[X]Y - \mathbb{E}[Y]X + \mathbb{E}[X]\mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

Osservazione 2.10.2. Indipendenza e covarianza

Se X e Y sono indipendenti, allora $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ e, quindi, $Cov(X, Y) = 0$.

Osservazione 2.10.3. Varianza di una variabile casuale binomiale

Poiché $X = \sum_i X_i$ è la somma di variabili casuali di Bernoulli indipendenti e identicamente distribuite con $Var(X_i) = p(1-p)$ per tutti gli i , abbiamo che

$$Var(X) = Var\left(\sum_{i=1}^n X_i\right) = np(1-p)$$

Esercizio 2.10.2. Varianza di una somma

Esprimi la varianza di una somma di due variabili casuali in termini delle varianze delle singole variabili e della loro covarianza.

$$\begin{aligned}Var(X + Y) &= \mathbb{E}[(X + Y - (\mathbb{E}[X + Y]))^2] = \mathbb{E}[(X - \mathbb{E}[X] + Y - \mathbb{E}[Y])^2] \\ &= \mathbb{E}[(X - \mathbb{E}[X])^2] + \mathbb{E}[(Y - \mathbb{E}[Y])^2] + 2\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= Var(X) + Var(Y) + 2Cov(X, Y)\end{aligned}$$

Esercizio 2.10.3. Moltiplicazione per una costante

Dimostra che se a è una costante, allora $Cov(aX, Y) = aCov(X, Y)$.

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

Correlazione

La correlazione $\rho(X, Y)$ di due variabili casuali X e Y è definita come

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Esercizio 2.10.4. Quanto può valere una correlazione

Dimostra che $-1 \leq \rho(X, Y) \leq 1$.

Soluzione

Siano $Var(X) = \sigma^2$ e $Var(Y) = \tau^2$. Abbiamo

$$0 \leq Var\left(\frac{X}{\sigma} + \frac{Y}{\tau}\right) = \frac{Var(X)}{\sigma^2} + \frac{Var(Y)}{\tau^2} + 2\frac{Cov(X, Y)}{\sigma\tau} = 2(1 + \rho(X, Y))$$

e

$$0 \leq Var\left(\frac{X}{\sigma} - \frac{Y}{\tau}\right) = \frac{Var(X)}{\sigma^2} + \frac{Var(Y)}{\tau^2} - 2\frac{Cov(X, Y)}{\sigma\tau} = 2(1 - \rho(X, Y))$$

Osservazione 2.10.4. Correlazione e legame lineare

Per $\rho(X, Y) \approx 1$ la relazione tra X e Y è ben approssimata da una retta con coefficiente angolare positivo (negativo per $\rho(X, Y) \approx -1$).

2.11 Risultati asintotici

Enunciamo ora due risultati fondamentali della Teoria della Probabilità: la legge dei grandi numeri e il teorema centrale del limite che legano il valore atteso alla media empirica.

Disuguaglianze fondamentali

Disuguaglianza di Markov Sia X una variabile casuale che assume valori non negativi ed $f(x)$ la sua densità di probabilità. Allora per ogni $a > 0$ vale la disuguaglianza

$$P\{X \geq a\} \leq \frac{\mathbb{E}[X]}{a} \quad (2.2)$$

Infatti, dato che $xf(x) \geq 0$ abbiamo

$$\mathbb{E}[X] = \int_0^{+\infty} xf(x)dx \geq \int_a^{+\infty} xf(x)dx$$

Moltiplicando e dividendo per a e notando che $x > a$ e quindi $x/a > 1$ si ottiene

$$\mathbb{E}[X] \geq a \int_a^{+\infty} \frac{x}{a} f(x)dx \geq a \int_a^{+\infty} f(x)dx$$

La disuguaglianza segue allora dalla definizione di funzione di probabilità cumulata

$$P\{X \geq a\} = \int_a^{+\infty} f(x)dx$$

Disuguaglianza di Chebyshev Sia X una variabile casuale con valore atteso μ e varianza σ^2 finiti. Allora per tutti gli $\epsilon > 0$ vale la disuguaglianza

$$P\{|X - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{\epsilon^2} \quad (2.3)$$

La disuguaglianza di Chebyshev si ottiene dalla disuguaglianza di Markov applicata alla variabile casuale non negativa $(X - \mu)^2$ con $a = \epsilon^2$ e notando che

$$\forall \epsilon > 0, P\{|X - \mu| \geq \epsilon\} = P\{|X - \mu|^2 \geq \epsilon^2\}$$

Osservazione 2.11.1. Nessuna ipotesi

La portata fondamentale delle disuguaglianze di Markov e Chebyshev è dovuta al fatto che valgono per qualunque distribuzione di probabilità.

Esercizio 2.11.1. Il vantaggio di saperne di più

Sia X uniformemente distribuita tra 0 e 10. Quanto vale la probabilità che X si scosti di 4 dal suo valore medio? Confronta il risultato ottenuto utilizzando la disuguaglianza di Chebyshev con quello ottenuto sfruttando il fatto che la distribuzione è uniforme.

Soluzione

Abbiamo che $\mathbb{E}[X] = 5$ e $\sigma^2 = 25/3$. Se applichiamo la disuguaglianza di Chebyshev con $\epsilon = 4$ otteniamo $P\{|X - 5| \geq 4\} \leq 25/48 \approx 0.52$. Utilizzando l'informazione sulla forma della distribuzione di X otteniamo

$$P\{|X - 5| \geq 4\} = \frac{2}{10} \int_0^1 dx = \frac{1}{5} = 0.2$$

Legge dei grandi numeri

Questo risultato chiarisce in che senso la frequenza o la media empirica converge a un valore atteso ed è alla base di tutti i metodi che stimano una quantità ignota a partire da un numero finito di osservazioni.

Teorema 2.11.1. *Legge (debole) dei grandi numeri*

Siano X_i con $i = 1, 2, \dots, n$ variabili casuali indipendenti e identicamente distribuite con $\mathbb{E}[X_i] = \mu$. Allora

$$\forall \epsilon > 0 \lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_i X_i - \mu \right| \geq \epsilon \right\} = 0$$

Con l'ipotesi aggiuntiva (e non necessaria) che la varianza σ^2 sia finita, poichè

$$\mathbb{E} \left[\frac{1}{n} \sum_i X_i \right] = \mu \quad e \quad Var \left(\frac{1}{n} \sum_i X_i \right) = \frac{\sigma^2}{n}$$

applicando la disuguaglianza di Chebyshev per $k = \epsilon$ otteniamo infine

$$P \left\{ \left| \frac{1}{n} \sum_i X_i - \mu \right| \geq \epsilon \right\} \leq \frac{\sigma^2}{n\epsilon^2}$$

■

Per il teorema 2.11.1, quindi, al crescere di n la probabilità che la media empirica differisca dal valore atteso tende a 0. Differenze significative per n sufficientemente grande possono essere rilevate, ma non frequentemente.

Esercizio 2.11.2. *Passeggiata dell'ubriaco*

Un ubriaco si muove con passi di lunghezza unitaria, indipendenti e in una direzione Θ uniformemente distribuita tra 0 e 2π . Dopo n passi a quale distanza D si troverà dal punto di partenza?

Soluzione

Se θ_i è la direzione del passo i -esimo, dopo i passi l'ubriaco si troverà nella posizione $(\sum_i X_i, \sum_i Y_i)$ con $(X_i, Y_i) = (\cos \theta_i, \sin \theta_i)$. Pertanto, dopo n passi avremo

$$\begin{aligned} D^2 &= \left(\sum_i X_i \right)^2 + \left(\sum_i Y_i \right)^2 \\ &= \sum_i (\cos^2 \theta_i + \sin^2 \theta_i) + \left(\sum_i \cos \theta_i \sum_{j \neq i} \cos \theta_j \right) + \left(\sum_i \sin \theta_i \sum_{j \neq i} \sin \theta_j \right) \\ &= n + \left(\sum_i \cos \theta_i \sum_{j \neq i} \cos \theta_j \right) + \left(\sum_i \sin \theta_i \sum_{j \neq i} \sin \theta_j \right) \end{aligned}$$

Poichè $\mathbb{E}[\cos \Theta] = \mathbb{E}[\sin \Theta] = 0$ per n grande avremo $\sum_i \cos \theta_i \approx \sum_i \sin \theta_i \approx 0$ e, pertanto, $D \approx \sqrt{n}$.

Compito 2.11.1. *Verifica empirica della legge dei grandi numeri*

Simula un dado truccato con

$$p_1 = 0.4, \quad p_2 = p_3 = 0.2, \quad p_4 = 0.1 \quad e \quad p_5 = p_6 = 0.05$$

campionando u da una distribuzione uniforme in $[0, 1]$.

Calcola $(\#_n i)/n$, ovvero la frequenza con la quale ottieni la faccia i su n lanci. Poiché ogni faccia i è una variabile casuale di Bernoulli con $\mu_i = p_i$ e $\sigma_i^2 = p_i(1 - p_i)$, ponendo $\epsilon = 10^{-2}$ per il teorema 2.11.1 si ha

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{(\#_n i)}{n} - p_i \right| \geq 10^{-2} \right\} \leq \frac{p_i(1 - p_i)}{n10^{-4}}$$

Ripeti per $m = 1000$ volte $n = 10^5$ lanci e verifica che, per ogni faccia i , $f_i \leq p_i(1 - p_i)/10$ approssimando

$$P \left\{ \left| \frac{(\#_n i)}{n} - p_i \right| \geq 10^{-2} \right\} \quad \text{con} \quad f_i = \frac{1}{m} \#_m \left(\left| \frac{(\#_n i)}{n} - p_i \right| \geq 10^{-2} \text{ è vera} \right)$$

Teorema centrale del limite

Presentiamo ora uno dei risultati più importanti della matematica.

Teorema 2.11.2. *Come si distribuiscono le stime di un valore atteso*

Siano le X_i con $i = 1, 2, \dots, n$ variabili casuali indipendenti e identicamente distribuite con $\mathbb{E}[X_i] = \mu$ e $\text{Var}(X_i) = \sigma^2$. Allora

$$\frac{\sum_i X_i - \mu}{\sigma/\sqrt{n}} \rightarrow \mathcal{N}(0, 1) \quad \text{per } n \rightarrow \infty$$

■

Il teorema **2.11.2** garantisce che le stime di un valore atteso, al crescere di n , si distribuiscono come una normale standard centrata sul valore atteso **indipendentemente** dalla distribuzione sottostante.

Osservazione 2.11.2. *Aguzzate la vista*

La legge dei grandi numeri ci assicura che la stima di un valore atteso al crescere di n , con grande probabilità, è vicina al piacere al valore atteso. Tuttavia, non ci dice nulla su quanto debba essere grande n e non ci consente di stimare la velocità con la quale la stima si avvicina al valore atteso. Il teorema centrale del limite, invece, è un risultato molto più forte perché garantisce che al crescere di n la distribuzione della stime approssima una distribuzione normale centrata sul valore atteso con varianza σ^2/n .

Compito 2.11.2. *Verifica empirica del CLT*

Simula una moneta onesta con $p = 1/2$, con 1 il valore attribuito alla realizzazione dell'evento *testa* e 0 a *croce*. Lancia la moneta n volte e calcola la frequenza t_n con la quale ottieni *testa* come

$$t_n = \frac{\#_n \text{testa}}{n}$$

Sapendo che per una moneta onesta $\mu = p = 0.5$ e $\sigma^2 = p(1 - p) = 1/4$, ripeti per 1000 volte n lanci e confronta il grafico della distribuzione normale standard $\mathcal{N}(0, 1)$ con l'istogramma delle tre distribuzioni empiriche ottenute

$$2\sqrt{n}(t_n - \mu) \quad \text{per } n = 10^2, 10^4 \text{ e } 10^6$$

2.12 Problemi di occupazione

L'analisi del problema del bilanciamento di un carico, ovvero di come distribuire un carico su risorse multiple, è centrale nello studio dell'allocazione dinamica di risorse e nell'*hashing*. Adottiamo come modello il lancio di m palline indistinguibili in n contenitori nell'ipotesi di probabilità uniforme e lanci indipendenti e stimiamo valori attesi importanti.

Formule utili

Maggiorazione del coefficiente binomiale Per ogni numero naturale n e $k \leq n$ abbiamo che

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} = \frac{n \cdot (n-1) \dots (n-(k-1))}{k!} \leq \frac{n^k}{k!} \quad (2.4)$$

Maggiorazione del reciproco del fattoriale Dallo sviluppo in serie della funzione esponenziale otteniamo che, per qualunque $k > 0$ intero,

$$\frac{k^k}{k!} < 1 + \frac{k^1}{1!} + \frac{k^2}{2!} + \dots + \frac{k^k}{k!} + \dots = \sum_{i=0}^{+\infty} \frac{k^i}{i!} = e^k$$

da cui segue che

$$\frac{1}{k!} < \left(\frac{e}{k}\right)^k \quad (2.5)$$

Somma della serie geometrica Sia $S = \sum_{i=k}^n a^i$ per $a > 0$. Poiché

$$aS = \sum_{i=k}^n a^{i+1}$$

abbiamo che

$$aS - S = a^{n+1} - a^k$$

da cui segue

$$S = \frac{a^k - a^{n+1}}{1 - a} \quad (2.6)$$

Disuguaglianza di Boole Dalla definizione di probabilità segue che la probabilità dell'unione di n eventi arbitrari E_i con $i = 1, \dots, n$, non necessariamente indipendenti, non è più grande della somma delle loro probabilità, ovvero

$$\Pr(\cup_{i=1}^n E_i) \leq \sum_{i=1}^n \Pr(E_i) \quad (2.7)$$

Domande

Con quale probabilità due palline finiscono in uno stesso contenitore? Se C_{ij} è la variabile casuale indicatrice di una *collisione* tra le palline i e j , ovvero dell'evento *la pallina i e la pallina j finiscono nello stesso contenitore*, e R_i^k la variabile casuale indicatrice dell'evento *il contenitore k riceve la pallina i* , avremo

$$\Pr(C_{ij}) = \sum_{k=1}^n \Pr(R_i^k) \Pr(R_j^k | R_i^k) = \sum_{k=1}^n \frac{1}{n} \cdot \frac{1}{n} = \frac{1}{n}$$

Quale è il numero atteso di collisioni? Se indichiamo con C la variabile casuale che conta le collisioni, $C = \sum_{i \neq j} C_{ij}$, avremo

$$E[C] = E \left[\sum_{i \neq j} C_{ij} \right] = \sum_{i \neq j} E[C_{ij}] = \sum_{i \neq j} \Pr(C_{ij}) = \binom{m}{2} \frac{1}{n}$$

Problema 2.12.1. *Il paradosso del compleanno*

La numerosità minima per avere un compleanno in comune con probabilità maggiore del 50% è 23. Il più piccolo m per il quale $\frac{1}{365} \binom{m}{2} \geq 1$ è 28. Come spieghi questa differenza?

Qual è il numero atteso di contenitori vuoti? Sia V_j la variabile casuale indicatrice dell'evento *il contenitore j è vuoto* e V la variabile casuale che conta il numero di contenitori vuoti. Dal fatto che la probabilità che una pallina non cada in un particolare contenitore è $1 - 1/n$ e nell'ipotesi $m = n$ otteniamo

$$\Pr(V_j) = \prod_{i=1}^n \left(1 - \frac{1}{n}\right) = \left(1 - \frac{1}{n}\right)^n$$

da cui ricaviamo che il valore atteso di V , per la linearità del valore atteso e per n grande, è

$$E[V] = E \left[\sum_{j=1}^n V_j \right] = \sum_{j=1}^n E[V_j] = \sum_{j=1}^n \Pr(V_j) = \sum_{j=1}^n \left(1 - \frac{1}{n}\right)^n \approx \sum_{j=1}^n \frac{1}{e} = \frac{n}{e}$$

Con quale probabilità un contenitore dato riceve esattamente k palline? Sia RE_j la variabile casuale indicatrice dell'evento *il contenitore j riceve esattamente k palline*. Utilizzando le disuguaglianze (2.4) e (2.5), e sempre nell'ipotesi $m = n$, otteniamo

$$\Pr(RE_j) = \binom{n}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{n-k} \leq \binom{n}{k} \left(\frac{1}{n}\right)^k \leq \frac{n^k}{k!} \left(\frac{1}{n}\right)^k = \frac{1}{k!} < \left(\frac{e}{k}\right)^k$$

Con quale probabilità un contenitore riceve almeno k palline? Sia RA_j la variabile casuale indicatrice dell'evento *il contenitore j riceve almeno k palline*. Usando la disuguaglianza (2.7) e la formula (2.6), e sempre nell'ipotesi $m = n$, abbiamo

$$\Pr(RA_j) \leq \sum_{i=k}^n \left(\frac{e}{i}\right)^i \leq \sum_{i=k}^n \left(\frac{e}{k}\right)^i = \left(\frac{e}{k}\right)^k \frac{1 - (e/k)^{n+1-k}}{1 - e/k} < \left(\frac{e}{k}\right)^k \frac{1}{1 - e/k}$$

Poniamo

$$k^* = \frac{3 \ln n}{\ln(\ln n)}$$

Poiché $e/k^* < 0.5$ abbiamo

$$\begin{aligned} \left(\frac{e}{k^*}\right)^{k^*} \frac{1}{1 - e/k^*} &= \left(\frac{e \ln(\ln n)}{3 \ln n}\right)^{\left(\frac{3 \ln n}{\ln(\ln n)}\right)} \\ &= \exp \left(\frac{3 \ln n}{\ln(\ln n)} (1 + \ln(\ln(\ln n)) - \ln(\ln(n)) - \ln 3) \right) \\ &< \exp \left(-3 \ln n + \frac{3 \ln n \cdot \ln(\ln(\ln n))}{\ln(\ln n)} \right) < \exp(-2 \ln n) = \frac{1}{n^2} \end{aligned}$$

dove l'ultima disuguaglianza vale per n sufficientemente grande.

Usando nuovamente la disuguaglianza di Boole e tenendo presente che il complementare dell'unione di eventi è l'intersezione dei complementari abbiamo

$$\begin{aligned}\Pr(\text{un contenitore qualunque riceve almeno } k^* \text{ palline}) &\leq n \cdot \frac{1}{n^2} = \frac{1}{n} \\ \Pr(\text{tutti i contenitori ricevono al massimo } k^* \text{ palline}) &\geq 1 - \frac{1}{n}\end{aligned}$$

Osservazione 2.12.1. *Caso $m = n$*

In conclusione abbiamo che quando il valore atteso dell'occupazione è pari a 1 ($m = n$), il numero massimo di palline ricevute da un contenitore è dell'ordine di $\ln n / \ln(\ln n)$.

2.13 Grandi deviazioni

Completiamo l'analisi del problema dell'occupazione trattando il caso di $m = n \ln n$. Il risultato principale è che il numero massimo di occupazione è dello stesso ordine, $O(\ln n)$, del valore atteso. Grandi variazioni dal valore atteso, per n grande, sono pertanto improbabili.

Disuguaglianza di Chernoff

Siano X_1, \dots, X_n variabili casuali *indipendenti* di Bernoulli con $\Pr(X_i = 1) = p$ per $i = 1, \dots, n$. Se per la variabile casuale somma $X = \sum_i X_i$ abbiamo

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i] = np = \mu$$

allora per ogni $\epsilon > 0$

$$\Pr(X > (1 + \epsilon)\mu) < \left(\frac{e^\epsilon}{(1 + \epsilon)^{1+\epsilon}} \right)^\mu \quad (2.8)$$

Dimostrazione: per ogni $t > 0$ applicando l'esponenziale è sempre vero che

$$P(X > (1 + \epsilon)\mu) = P(e^{tX} > e^{t(1+\epsilon)\mu})$$

La disuguaglianza di Markov (2.2) per la variabile casuale positiva e^{tX} , tenuto conto che $X = \sum_i X_i$, fornisce

$$P(e^{tX} > e^{t(1+\epsilon)\mu}) < \frac{\mathbb{E}[e^{tX}]}{e^{t(1+\epsilon)\mu}} = \frac{\prod_{i=1}^n \mathbb{E}[e^{tX_i}]}{e^{t(1+\epsilon)\mu}} \quad (2.9)$$

Ora

$$\mathbb{E}[e^{tX_i}] = pe^t + (1 - p) = 1 - p(1 - e^t)$$

e siccome $1 - x < e^x$ ponendo $x = p(1 - e^t)$ possiamo scrivere

$$\mathbb{E}[e^{tX_i}] < e^{p(1-e^t)}$$

Pertanto, poiché $np = \mu$, abbiamo

$$\prod_{i=1}^n \mathbb{E}[e^{tX_i}] < \prod_{i=1}^n e^{p(1-e^t)} = e^{np(1-e^t)} = e^{\mu(1-e^t)}$$

che, sostituito nella disequazione (2.9), fornisce

$$P(e^{tX} > e^{t(1+\epsilon)\mu}) < \frac{e^{\mu(1-e^t)}}{e^{t(1+\epsilon)\mu}} = e^{\mu(1-e^t-t(1+\epsilon))}$$

Poiché questa disuguaglianza è valida per ogni $t > 0$, cerchiamo il valore di t per il quale $e^{\mu(1-e^t-t(1+\epsilon))}$ è minimo. Annullando la derivata di $(e^t - t - t\epsilon - 1)$ rispetto a t troviamo

$$\frac{d}{dt}(e^t - t - t\epsilon - 1) = e^t - 1 - \epsilon = 0 \implies \bar{t} = \ln(1 + \epsilon)$$

In conclusione per $t = \bar{t}$, e sbarazzandoci dell'esponenziale nell'argomento della probabilità, abbiamo

$$P(X > (1 + \epsilon)\mu) < e^{\mu(1+\epsilon-\ln(1+\epsilon)-\epsilon\ln(1+\epsilon)-1)} = e^{\mu(\epsilon-(1+\epsilon)\ln(1+\epsilon))} = \left(\frac{e^\epsilon}{(1 + \epsilon)^{(1+\epsilon)}} \right)^\mu$$

■

Ancora sulla probabilità che un contenitore riceva almeno k palline

Guardiamo a che cosa succede nel caso in cui $m = n \ln n$. Per la linearità del valore atteso in questo caso ricaviamo che il numero atteso di palline ricevute da ogni contenitore è

$$\mu = \Pr \left(\bigcup_{i=1}^{n \ln n} E_i \right) \leq \sum_{i=1}^{n \ln n} \frac{1}{n} = \ln n$$

Vediamo che cosa ci dicono le tre disuguaglianze se ipotizzassimo l'esistenza di un contenitore con $10 \ln n$ palline.

Markov Dalla disuguaglianza (2.2) con $\mu = \ln n$ e $a = 10 \ln n$ ricaviamo

$$\Pr(X > 10 \ln n) = \frac{\ln n}{10 \ln n} = \frac{1}{10}$$

Chebyshev La varianza σ^2 di una distribuzione binomiale con m lanci e probabilità $p = 1/n$ è

$$\sigma^2 = m \frac{1}{n} \left(1 - \frac{1}{n} \right) \leq \frac{m}{n}$$

Pertanto, la disuguaglianza (2.3) per $\mu = m/n = \ln n$ e $k = 9 \ln n$ diventa

$$\Pr\{X \geq \ln n + 9 \ln n\} \leq \frac{\ln n}{81 \ln^2 n} = \frac{1}{81 \ln n}$$

Chernoff Con $\epsilon = 9$ la disuguaglianza (2.8) diventa

$$\Pr(X \geq (1 + 9) \ln n) \leq \left(\frac{e^9}{10^{10}} \right)^{\ln n} < \left(\frac{e^9}{e^{20}} \right)^{\ln n} = \frac{1}{n^{11}}$$

Osservazione 2.13.1. *Non c'è paragone*

La disuguaglianza di Chernoff mostra che la probabilità di grandi variazioni dal valore atteso, almeno per il caso di somma di variabili casuali indipendenti, decresce molto più rapidamente di quanto si possa inferire dalle disuguaglianze di Markov e Chebyshev.

Capitolo 3

Elementi di Teoria dell'Informazione

3.14 Informazione di Shannon e codifica dell'informazione

Introduciamo la nozione di informazione di Shannon e familiarizziamo con le sue principali proprietà.

Misura di informazione

Una misura della quantità di informazione che si acquisisce una volta che si sia realizzato un evento casuale di probabilità $p > 0$ è l'*informazione di Shannon* definita come

$$\log_2 \frac{1}{p}$$

L'unità di misura dell'informazione di Shannon è il *bit*.

Esempio 3.14.1. Lancio di una moneta onesta

Il risultato del lancio di una moneta può essere testa, T , o croce, C ; nel caso in cui i due eventi sono equiprobabili, ovvero se $p(T) = p(C) = 1/2$, l'informazione di Shannon che acquisiamo osservando il risultato è sempre

$$\log_2 2 = 1bit$$

Esempio 3.14.2. Lancio di una moneta truccata

Consideriamo ora una moneta in cui la probabilità che esca testa sia molto più grande di croce, tipo $p(T) = 7/8$. In questo caso, l'informazione di Shannon acquisita alla realizzazione dell'evento T è

$$\log_2 \frac{8}{7} \approx 0.19bit$$

mentre nel caso dell'evento C , poiché $p(C) = 1/8$, è

$$\log_2 8 = 3bit$$

In accordo con l'intuizione, la sorpresa provocata dalla realizzazione di un evento casuale e l'entità della conseguente rimozione di incertezza aumentano al decrescere della probabilità dell'evento.

Supponiamo ora di lanciare due volte la moneta truccata. L'informazione di Shannon acquisita dalla realizzazione dell'evento TT è

$$\log_2 \left(\frac{8}{7} \cdot \frac{8}{7} \right) = \log_2 \frac{8}{7} + \log_2 \frac{8}{7} \approx 0.39bit$$

mentre nel caso dell'evento CC è

$$\log_2(8 \cdot 8) = \log_2 8 + \log_2 8 = 6\text{bit}$$

Nei casi TC e CT si ottiene

$$\log_2 \left(8 \cdot \frac{8}{7} \right) = \log_2 \frac{8}{7} + \log_2 8 \approx 3.19\text{bit}$$

Anche in questo caso i risultati sono in accordo con la nostra aspettativa: la dipendenza logaritmica dalla probabilità garantisce che la quantità di informazione acquisita dalla realizzazione di eventi indipendenti sia pari alla somma dell'informazione di Shannon associata a ogni singolo evento.

Unicità

Dimostriamo ora che la forma funzionale $S = S(p)$ dell'informazione di Shannon è univocamente determinata, a meno di un fattore costante arbitrario, da quattro assunzioni.

A1 Non si acquisisce informazione dalla realizzazione di un evento certo, ovvero

$$S(1) = 0$$

A2 Minore la probabilità di un evento, maggiore la quantità di informazione ottenuta dalla sua realizzazione, ovvero

$$\text{se } p < q, \text{ allora } S(p) > S(q)$$

A3 A piccoli cambiamenti di p corrispondono piccoli cambiamenti di $S(p)$, ovvero $S(p)$ è una funzione continua di p .

A4 Siano E e F due eventi indipendenti con probabilità rispettivamente p e q . La quantità di informazione acquisita dalla realizzazione dell'evento E non è modificata dal fatto di conoscere che l'evento F si è realizzato, ovvero

$$S(pq) = S(p) + S(q)$$

Da questi assiomi è possibile determinare univocamente la forma funzionale dell'informazione di Shannon.

Teorema 3.14.1. *Unicità della forma funzionale dell'informazione di Shannon*

Se $S(\cdot)$ soddisfa gli Assiomi **A1**, **A2**, **A3** e **A4** allora

$$S(p) = C \log_2 \frac{1}{p}$$

dove C è un numero positivo arbitrario.

Dimostrazione: da **A4**

$$S(p^2) = S(p) + S(p) = 2S(p)$$

e, per induzione,

$$S(p^m) = mS(p)$$

Inoltre, per ogni intero n

$$S(p) = S(p^{1/n} p^{1/n} \dots p^{1/n}) = nS(p^{1/n})$$

Di conseguenza si ha

$$S(p^{m/n}) = mS(p^{1/n}) = \frac{m}{n}S(p)$$

Per **A3** questa relazione vale non solo per ogni numero razionale positivo m/n , ma anche per ogni numero reale positivo w , ovvero

$$S(p^w) = wS(p)$$

Per ogni $p > 0$, ponendo $w = \log_2(1/p)$ si ha $p = 2^{-w}$ per cui

$$S(p) = S\left(\frac{1}{2^w}\right) = wS\left(\frac{1}{2}\right) = C \log_2 \frac{1}{p}$$

poiché, per **A1** e **A2**, $C = S(1/2) > S(1) = 0$. ■

Due giochi

Vediamo ora come l'informazione di Shannon sia legata al numero di bit necessari a rappresentare uno qualunque tra N numeri attraverso l'analisi di due semplici giochi. In entrambi i casi, assumiamo tacitamente che tutte le scelte del nostro avversario siano ugualmente probabili.

Esempio 3.14.3. Indovina il numero

L'avversario sceglie un numero n compreso tra 0 e 1023; il nostro scopo è indovinare il numero scelto formulando il minimo numero di domande alle quali l'avversario può rispondere solo *si* o *no*. Procediamo per dimezzamenti progressivi dell'intervallo originale. Se pensiamo alla codifica binaria di un qualunque n compreso tra 0 e 1023 ognuno dei 10 bit necessari può essere uguale a 0 o a 1 con probabilità $p = 1/2$, per cui ogni dimezzamento corrisponde all'acquisizione di una quantità di informazione pari a

$$\log_2 \frac{1}{p} = \log_2 2 = 1bit$$

I 10bit di informazione acquisiti dopo 10 domande, alla fine del gioco, ricostruiscono i 10 bit necessari per la codifica di n . Con questa strategia a ogni passo si ottiene sempre 1bit di informazione. Inoltre, l'informazione complessiva acquisita alla fine del gioco è la stessa che avremmo acquisito se avessimo indovinato al primo tentativo.

Esempio 3.14.4. Trova il sottomarino

Scopo del gioco, versione semplificata della battaglia navale, è indovinare in quale delle $N = 8 \times 8 = 64$ caselle l'avversario abbia posizionato un sottomarino. Nuovamente ipotizziamo che il sottomarino possa essere con la stessa probabilità in qualunque casella. Al primo tentativo abbiamo 1 probabilità su 64 di indovinare e 63 su 64 di fallire. Se falliamo, la quantità di informazione acquisita è

$$\log_2 64 - \log_2 63 \approx 0.02bit$$

Se indoviniamo, invece,

$$\log_2 64 = 6bit$$

In questo secondo caso il gioco finisce e abbiamo acquisito tutta l'informazione di Shannon che era effettivamente disponibile in un solo colpo. Nell'ipotesi di aver fallito, il gioco prosegue. Calcoliamo come aumenta l'informazione di Shannon a ogni passo. Al secondo tentativo abbiamo 1 probabilità su 63 di indovinare e 62 su 63 di fallire. Nel caso in cui falliamo acquisiamo una quantità di informazione pari a

$$\log_2 64 - \log_2 63 + (\log_2 63 - \log_2 62) = \log_2 64 - \log_2 62 \approx 0.05bit$$

La quantità di informazione acquisita è ancora piuttosto scarsa e non troppo diversa da prima (solo 2 tentativi invece che 1 per 64 caselle). Se indoviniamo, invece, si ha

$$S = \log_2 64 - \log_2 63 + \log_2 63 = 6\text{bit}$$

È un caso? No: in effetti, in qualunque momento finisca il gioco, l'informazione di Shannon acquisita è sempre uguale a 6bit . Anche per questo gioco, quindi, l'informazione di Shannon complessiva è uguale al numero di bit, 6, richiesti per identificare ciascuna delle 64 caselle.

Che cosa succede se fallissimo per 32 volte? L'informazione accumulata sarebbe

$$S = \log_2 64 - \log_2 32 = 6 - 5 = 1\text{bit}$$

Dopo 32 tentativi a vuoto, infatti, abbiamo confinato il sottomarino in metà delle caselle della tabella originale, guadagnando 1bit di informazione!

Risoluzione di problemi guidata dall'informazione di Shannon

Verifichiamo infine l'ottimalità di una strategia che massimizza l'informazione di Shannon. Per semplicità, assumiamo nuovamente che tutti i risultati siano equiprobabili.

Problema 3.14.1. *Quante pesate?*

Supponiamo di avere 12 palline uguali per forma e colore, una delle quali è di peso diverso dalle altre 11. Avendo a disposizione una bilancia con due piatti, quante pesate sono necessarie per individuare la pallina diversa da tutte le altre e se sia più leggera o più pesante? I casi possibili sono 24: ognuna delle 12 palline, infatti, può essere più leggera o più pesante delle altre 11. A ogni pesata la bilancia restituisce uno di 3 possibili risultati: preso come riferimento il peso del piatto sinistro, si può osservare un peso minore, uguale o maggiore sul piatto destro. È evidente che 2 pesate non sono sufficienti per determinare l'alternativa corretta ($3^2 = 9 < 24$) mentre 3 pesate potrebbero bastare ($3^3 = 27 > 24$). Come conviene procedere? Se scegliamo di mettere 6 palline su un piatto e 6 sull'altro, uno dei 3 possibili risultati dell'esperimento (stesso peso sui due piatti) è addirittura impossibile. Una scelta che esclude risultati possibili genera inefficienza. Dopo la pesata, infatti, saremmo ancora alle prese con 12 possibili soluzioni (ognuna delle 6 palline sul piatto più basso potrebbe essere quella più pesante e ognuna delle 6 palline sul piatto più alto quella più leggera). È evidente che altre 2 pesate non sarebbero sufficienti per determinare l'alternativa corretta ($3^2 = 9 < 12$).

Soluzione: disponiamo 4 palline su un piatto e 4 sull'altro. Per ognuno dei 3 possibili risultati le alternative sono ridotte a 8, alternative che possono essere analizzate esaustivamente con 2 ulteriori pesate ($3^2 = 9 > 8$). Vediamo il dettaglio considerando prima il caso in cui i piatti sono in equilibrio.

1: in equilibrio Sappiamo che il peso delle 8 palline è quello standard e che la pallina, più leggera o più pesante, è una delle 4 che non abbiamo pesato (8 alternative in tutto). Pesiamo allora 3 delle 4 palline non ancora utilizzate e 3 di quelle già utilizzate. Anche in questa seconda pesata il risultato dell'esperimento è incerto. Consideriamo le alternative possibili.

2: in equilibrio La pallina è l'unica non ancora utilizzata e con una terza pesata possiamo determinare se sia più leggera o più pesante confrontandola con una delle altre palline.

2: non in equilibrio Sappiamo che la pallina è una delle 3 pesate una sola volta e sappiamo anche se è più leggera o più pesante dal confronto col piatto caricato con le palline di peso standard. Basta allora confrontarne 2 a caso delle 3 per individuare quella giusta.

1: non in equilibrio Supponiamo ora che il piatto sinistro sia più basso (nel caso sia più alto basta scambiare destra con sinistra nel seguito). Sappiamo allora che il peso delle 4 palline non pesate è quello standard e che, se la pallina è più pesante, è una delle 4 sul piatto sinistro o, se è più leggera, è una delle 4 sul piatto destro. Nuovamente, 8 alternative. Mettiamo allora 2 delle palline del piatto sinistro e 1 delle palline del piatto destro su ciascun piatto. Anche in questo caso il risultato dell'esperimento è incerto. Consideriamo le alternative possibili.

2: in equilibrio La pallina è una delle 2 non utilizzate e sappiamo che è più leggera. Basta allora confrontare queste 2 palline tra loro per trovare la soluzione.

non in equilibrio Supponiamo che il piatto sinistro sia più basso (se più alto occorre scambiare destra con sinistra nel seguito). Sappiamo ora che la pallina è una di 2 sul piatto sinistro, se più pesante, o 1 sul piatto destro, se più leggera. Confrontando le 2 palline del piatto sinistro tra loro con una terza pesata possiamo trovare quale delle tre palline risolva il problema.

Osservazione 3.14.1. *La potenza dell'incertezza*

A ogni pesata la possibilità di ottenere tutti e tre i risultati, ovvero la massimizzazione dell'incertezza sull'esito dell'esperimento, gioca un ruolo fondamentale per eliminare alternative in modo bilanciato e determinare una soluzione nel minor numero possibile di pesate. Nel primo passo garantisce la riduzione delle alternative 24 a 8, nei secondi passi da 8 a 3 (in alcuni casi fortunati a 2) e nei terzi passi di poter sempre trovare l'alternativa corretta.

3.15 Entropia di Shannon

Introduciamo l'entropia di Shannon e discutiamo le sue proprietà fondamentali.

Entropia come valore atteso

Sia $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$ l'insieme dei valori che possono essere assunti da una variabile casuale X con probabilità $p_X(x_1), p_X(x_2), \dots, p_X(x_N)$ e $\sum_i p_X(x_i) = 1$. Chiaramente $|\mathcal{X}| = N$.

Definizione 3.15.1. Entropia di una variabile casuale

L'entropia di X è il valore atteso dell'informazione di Shannon

$$H(X) = \sum_{i=1}^N p_X(x_i) \log_2 \frac{1}{p_X(x_i)}$$

dove, se $p_X(x_i) = 0$, poniamo $p_X(x_i) \log_2(1/p_X(x_i)) = 0$.

Nel caso in cui X assuma due soli valori, rispettivamente con probabilità p e $1-p$, l'entropia diventa

$$H_2(X) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$$

Uguagliando a 0 la derivata di $H_2(X)$ rispetto a p otteniamo

$$\log_2 \frac{1}{p} - \log_2 \frac{1}{1-p} = 0$$

da cui segue che il massimo di H_2 si ottiene per $p = 1/2$ e vale 1.

Dalla concavità della funzione logaritmica discende che questo risultato è vero più in generale: nel caso di eventi ugualmente probabili, per qualunque N , l'entropia è sempre massima. Infatti, per ogni funzione convessa f vale la disuguaglianza di Jensen, ovvero

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

Nel nostro caso, poichè il logaritmo è concavo dobbiamo invertire la disuguaglianza e abbiamo

$$H(X) = \mathbb{E} \left[\log_2 \frac{1}{p_X(x_i)} \right] \leq \log_2 \mathbb{E} \left[\frac{1}{p_X(x_i)} \right] = \log_2 \left(\sum_{i=1}^N \frac{p_X(x_i)}{p_X(x_i)} \right) = \log_2 \sum_{i=1}^N 1 = \log_2 N$$

Esempio 3.15.1. Dado equo a otto facce

Consideriamo il caso di un dado equo con otto facce. Ogni faccia i ha probabilità $p_i = 1/8$ e per l'entropia H_e abbiamo

$$H_e = \sum_{i=1}^8 p_i \log_2 \frac{1}{p_i} = \sum_{i=1}^8 \frac{1}{8} \log_2 8 = \log_2 8 = 3$$

Esempio 3.15.2. Dado iniquo (sempre a 8 facce)

Consideriamo il caso di un dado iniquo a otto facce in cui le probabilità sono

$$p(1) = p(2) = p(3) = p(4) = \frac{1}{16}, \quad p(5) = p(6) = \frac{1}{8}, \quad \text{e} \quad p(7) = p(8) = \frac{1}{4}$$

Ci aspettiamo che in questo caso l'entropia del dado iniquo H_i sia minore di H_e , infatti

$$H_i = 4 \frac{1}{16} \log_2 16 + 2 \frac{1}{8} \log_2 8 + 2 \frac{1}{4} \log_2 4 = \frac{1}{4} 4 + \frac{1}{4} 3 + \frac{1}{2} 2 = \frac{11}{4} = 2.75 < 3 = H_e$$

Entropia congiunta e condizionata

Sia $\mathcal{Y} = \{y_1, y_2, \dots, y_M\}$ l'insieme dei valori che possono essere assunti da una variabile casuale Y con probabilità marginale $p_Y(y_1), p_Y(y_2), \dots, p_Y(y_M)$ e $\sum_j p_Y(y_j) = 1$. Sia $p(x_i, y_j)$ la probabilità congiunta dell'evento $X = x_i$ e $Y = y_j \forall i, j$ con $\sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) = 1$. Indichiamo con $p(x_i|y_j)$ la probabilità dell'evento $X = x_i$ condizionata alla realizzazione dell'evento $Y = y_j$; per ogni j fissato abbiamo $\sum_{i=1}^N p(x_i|y_j) = 1$.

Per l'entropia congiunta $H(X, Y)$ abbiamo

$$H(X, Y) = \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \log_2 \frac{1}{p(x_i, y_j)}$$

L'entropia $H(X)$ è modificata dalla realizzazione dell'evento $Y = y_j$ secondo la formula

$$H(X|Y = y_j) = \sum_{i=1}^N p(x_i|y_j) \log_2 \frac{1}{p(x_i|y_j)}$$

Ne segue che l'entropia di X condizionata alla realizzazione di Y si scrive come

$$H(X|Y) = \sum_{j=1}^M p_Y(y_j) H(X|Y = y_j)$$

Osservazione 3.15.1. Logaritmo del prodotto e somma dei logaritmi

La probabilità congiunta di due variabili casuali X e Y si scrive come prodotto della probabilità di Y per la probabilità di X dato Y . La dipendenza logaritmica dalle probabilità dell'informazione di Shannon trasforma il prodotto di queste probabilità nella somma delle corrispondenti entropie.

Esempio 3.15.3. Ancora sul dado equo con otto facce

Sia X la variabile casuale che assume i valori $i = 1, \dots, 8$ con $p_X(i) = 1/8$ e Y la variabile casuale che assume i valori $j = 0$ se l'esito del lancio è pari e $j = 1$ se dispari con $p_Y(j) = 1/2$. Calcoliamo le entropie $H(X)$ - quale faccia del dado, $H(Y)$ - quale parità, $H(X|Y)$ - quale faccia nota quale parità, $H(Y|X)$ - quale parità nota quale faccia, e $H(X, Y)$ - quale faccia e quale parità congiunte. Osserviamo che $p(i|j) = 1/4$ se i e j hanno la stessa parità e 0 altrimenti, mentre $p(j|i)$ è uguale a 1 se j ha la stessa parità di i e 0 altrimenti. Infine, $p(i, j) = 1/8$ se i e j hanno la stessa parità e 0 altrimenti.

$$\begin{aligned} H(X) &= \sum_{i=1}^8 p_X(i) \log_2 \frac{1}{p_X(i)} = \sum_{i=1}^8 \frac{1}{8} \log_2 8 = \log_2 8 = 3 \\ H(Y) &= \sum_{j=0}^1 p_Y(j) \log_2 \frac{1}{p_Y(j)} = \frac{1}{2} \log_2 2 + \frac{1}{2} \log_2 2 = 1 \\ H(X|Y) &= \sum_{j=0}^1 p_Y(j) H(X|j) = \sum_{j=0}^1 p_Y(j) \sum_{i=1}^8 p(i|j) \log_2 \frac{1}{p(i|j)} \\ &= \frac{1}{2} \left(4 \cdot \frac{1}{4} \log_2 4 \right) + \frac{1}{2} \left(4 \cdot \frac{1}{4} \log_2 4 \right) = \log_2 4 = 2 \\ H(Y|X) &= \sum_{i=1}^8 p_X(i) H(Y|i) = \sum_{i=1}^8 p_X(i) \sum_{j=0}^1 p(j|i) \log_2 \frac{1}{p(j|i)} = 0 \\ H(X, Y) &= \sum_{i=1}^8 \sum_{j=0}^1 p(x_i, y_j) \log_2 \frac{1}{p(x_i, y_j)} = \log_2 8 = 3 \end{aligned}$$

Notiamo che il guadagno atteso dell'informazione di Shannon legato alla conoscenza di quale faccia sia uscita, nota la parità, si riduce perché le facce possibili passano da 8 a 4. Nel caso inverso, invece, il guadagno si azzera perché nota la faccia una delle due parità è certa e l'altra impossibile.

Esempio 3.15.4. *Un dado equo a otto facce e una moneta equa*

Sia X ancora la variabile casuale che assume i valori da 1 a 8 con probabilità $p_X(i) = 1/8$, lancio di un dado equo. Ora Y è la variabile casuale che vale 0 se l'esito del lancio di una moneta equa è *testa* e 1 se *croce* (in entrambi i casi quindi con probabilità $p_Y(j) = 1/2$). Le due variabili sono chiaramente indipendenti. Pertanto $p(i|0) = p(i|1) = 1/8$ perché l'esito del lancio della moneta non modifica le probabilità del dado, $p(0|i) = p(1|i)$ perché l'esito del lancio del dado non modifica le probabilità di testa e croce. Infine, $p(i, j) = 1/16$ in virtù dell'indipendenza.

Calcoliamo le entropie $H(X)$ - quale faccia del dado sia uscita, $H(Y)$ - quale faccia della moneta, $H(X|Y)$ - quale faccia del dado nota la faccia della moneta, $H(Y|X)$ - quale faccia della moneta nota la faccia del dado, e $H(X, Y)$ - quale faccia del dado e quale della moneta congiunte.

$$\begin{aligned} H(X) &= \sum_{i=1}^8 p_X(i) \log_2 \frac{1}{p_X(i)} = 8 \cdot \left(\frac{1}{8} \log_2 8 \right) = \log_2 8 = 3 \\ H(Y) &= \sum_{j=0}^1 p_Y(j) \log_2 \frac{1}{p_Y(j)} = 2 \cdot \left(\frac{1}{2} \log_2 2 \right) = 1 \\ H(X|Y) &= \sum_{j=0}^1 p_Y(j) H(X|j) = \sum_{j=0}^1 \frac{1}{2} \sum_{i=1}^8 p(i|j) \log_2 \frac{1}{p(i|j)} = \log_2 8 = 3 \\ H(Y|X) &= \sum_{i=1}^8 p_X(i) H(Y|i) = \sum_{i=1}^8 \frac{1}{8} \sum_{j=0}^1 p(j|i) \log_2 \frac{1}{p(j|i)} = 8 \cdot \left(\frac{1}{8} \log_2 2 \right) = 1 \\ H(X, Y) &= \sum_{i=1}^8 \sum_{j=0}^1 p(x_i, j) \log_2 \frac{1}{p(x_i, j)} = \log_2 16 = 4 \end{aligned}$$

Per via dell'equiprobabilità, tutte le entropie non condizionate sono massime.

Osservazione 3.15.2. *Le apparenze non sempre ingannano*

In entrambi gli esempi risulta che

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

e che

$$\begin{aligned} H(X|Y) &\leq H(X) \\ H(Y|X) &\leq H(Y) \end{aligned}$$

Dimostriamo ora che queste relazioni, in effetti, sono di portata generale.

Uguaglianza fondamentale

Proposizione 3.15.1. *Entropia congiunta come somma di entropie*

Per ogni coppia di variabili casuali discrete X e Y si ha

$$H(X, Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$$

Dimostrazione: siccome $p(x_i, y_j) = p_Y(y_j)p(x_i|y_j)$, abbiamo

$$\begin{aligned}
 H(X, Y) &= \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \log_2 \frac{1}{p(x_i, y_j)} \\
 &= \sum_{i=1}^N \sum_{j=1}^M p_Y(y_j) p(x_i|y_j) \log_2 \frac{1}{p_Y(y_j)} + \sum_{i=1}^N \sum_{j=1}^M p_Y(y_j) p(x_i|y_j) \log_2 \frac{1}{p(x_i|y_j)} \\
 &= \sum_{j=1}^M p_Y(y_j) \log_2 \frac{1}{p_Y(y_j)} \sum_{i=1}^N p(x_i|y_j) + \sum_{j=1}^M p_Y(y_j) \sum_{i=1}^N p(x_i|y_j) \log_2 \frac{1}{p(x_i|y_j)} \\
 &= \sum_{j=1}^M p_Y(y_j) \log_2 \frac{1}{p_Y(y_j)} \cdot 1 + H(X|Y) = H(Y) + H(X|Y)
 \end{aligned}$$

Disuguaglianza fondamentale

Proposizione 3.15.2. *La realizzazione di Y non può aumentare l'entropia di X*

Se X e Y sono variabili casuali, allora

$$H(X|Y) \leq H(X)$$

Dimostrazione: ricordando che $\ln t = \log_2 t / \log_2 e$, riscriviamo $\ln t \leq (t - 1)$, vedi figura 3.1, come

$$\log_2 t \leq (t - 1) \log_2 e \quad (3.1)$$

Notiamo che il segno di uguaglianza vale solo per $t = 1$. Pertanto, usando l'equazione (3.1), otteniamo

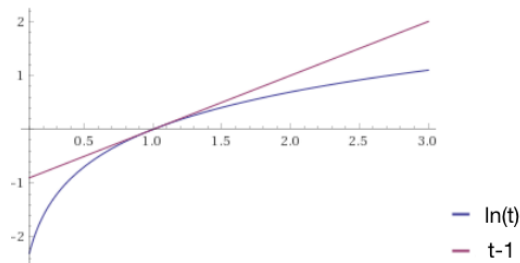


Figura 3.1: Vedi testo.

$$\begin{aligned}
H(X|Y) - H(X) &= \sum_{i=1}^N \sum_{j=1}^M p_Y(y_j) p(x_i|y_j) \log_2 \frac{1}{p(x_i|y_j)} - \sum_{i=1}^N p_X(x_i) \log_2 \frac{1}{p_X(x_i)} \\
&= \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \log_2 \frac{1}{p(x_i|y_j)} - \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \log_2 \frac{1}{p_X(x_i)} \\
&= \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \log_2 \frac{p_X(x_i)}{p(x_i|y_j)} \\
&\leq \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \left(\frac{p_X(x_i)}{p(x_i|y_j)} - 1 \right) \log_2 e \\
&= \left(\sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \frac{p_X(x_i)}{p(x_i|y_j)} - \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \right) \log_2 e \\
&= \left(\sum_{i=1}^N \sum_{j=1}^M p_X(x_i) p_Y(y_j) - 1 \right) \log_2 e \\
&= (1 - 1) \log_2 e = 0
\end{aligned}$$

Osservazione 3.15.3. *Entropia di variabili indipendenti*

Combinando le due relazioni otteniamo che se le variabili X e Y allora sono indipendenti

$$H(X, Y) = H(X) + H(Y)$$

Mutua informazione

La mutua informazione tra due variabili casuali X e Y

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

è l'informazione ottenibile da una variabile casuale dopo averne osservato un'altra. Le uguaglianze e le disuguaglianze tra le varie entropie e la mutua informazione sono desumibili dalla figura (3.2). La mutua informazione gioca un ruolo fondamentale nella comunicazione.

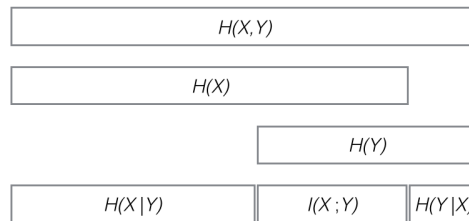


Figura 3.2: Vedi testo.

Compito 3.15.1. *Dado a 16 facce*

Considera un dado equo con 16 facce. Se X è la variabile casuale che assume i 16 possibili valori e Y la variabile casuale che vale 0 se la faccia è pari e 1 se la faccia è dispari, calcola $H(X)$, $H(Y)$, $H(X, Y)$, $H(X|Y)$ e $H(Y|X)$ e commenta i risultati che ottieni.

3.16 Lo stretto indispensabile sulla teoria dei codici

Introduciamo ora il tema delle codifiche capaci di ottenere compressione senza perdita di informazione. La nostra attenzione è circoscritta al caso particolare di codifiche binarie.

Decifrabilità univoca e istantaneità

Definizione 3.16.1. Codifica

Sia \mathcal{X} un insieme finito di simboli che possiamo pensare come i valori possibili di una variabile casuale X . Una codifica per simbolo C è una funzione dall'insieme \mathcal{X} a $\{0, 1\}^+$. \square

Indichiamo con $L_C(x)$ la lunghezza di $C(x)$ per $x \in \mathcal{X}$.

Definizione 3.16.2. Codifica estesa

La codifica estesa C^+ è una funzione dall'insieme \mathcal{X}^+ a $\{0, 1\}^+$ ottenuta concatenando le rappresentazioni senza segni di interpunzione.

Esempio 3.16.1. Simboli non equiprobabili

Sia $\mathcal{X} = \{a, b, c, d\}$ con $p(a) = 1/2$, $p(b) = 1/4$, $p(c) = 1/8$ e $p(d) = 1/8$. Introduciamo una codifica C_1 tale che

$$\begin{aligned}C_1(a) &= 1000 \\C_1(b) &= 0100 \\C_1(c) &= 0010 \\C_1(d) &= 0001\end{aligned}$$

Chiaramente $L_{C_1}(x) = 4$ per tutti gli $x \in \mathcal{X}$. Nella codifica estesa C_1^+ la stringa $acdbac$ è codificata in

$$C_1^+(acdbac) = C_1(a)C_1(c)C_1(d)C_1(b)C_1(a)C_1(c) = 1000\ 0010\ 0001\ 0100\ 1000\ 0010$$

dove aggiungiamo lo spazio tra le codifiche dei caratteri per poter leggere più facilmente la codifica estesa. Per quanto riguarda l'entropia per simbolo di C_1 abbiamo

$$H(X) = \frac{1}{2} \log_2 2 + \frac{1}{4} \log_2 4 + \frac{2}{8} \log_2 8 = \frac{1}{2} + \frac{1}{2} + \frac{3}{4} = \frac{7}{4} = 1.75$$

\square

Esaminiamo ora due proprietà importanti che consentono di caratterizzare le codifiche per simbolo senza perdita di informazione: decifrabilità univoca e istantaneità.

Definizione 3.16.3. Decifrabilità univoca

Un codifica C è *univocamente decifrabile* se per la codifica estesa C^+

$$\forall x, y \in \mathcal{X}^+ \quad x \neq y \rightarrow C^+(x) \neq C^+(y)$$

\square

È immediato verificare che la codifica C_1 è univocamente decifrabile, poiché le codifiche dei simboli sono diverse tra loro e tutte della stessa lunghezza.

Esempio 3.16.2. Codifiche con perdita di informazione

La codifica C_2 per cui

$$\begin{aligned} C_2(a) &= 1 \\ C_2(b) &= 0 \\ C_2(c) &= 10 \\ C_2(d) &= 01 \end{aligned}$$

non è univocamente decifrabile. Per esempio

$$\begin{aligned} C_2^+(acdbac) &= 1\ 10\ 01\ 0\ 1\ 10 \\ C_2^+(aabbadc) &= 1\ 1\ 0\ 0\ 1\ 01\ 10 \end{aligned}$$

Osservazione 3.16.1. Codice Morse

La decodifica del codice Morse è possibile grazie agli spazi tra punti e trattini, ovvero alle pause tra colpi secchi e accentuati. I simboli utilizzati dal codice Morse in effetti sono tre: “.”, “—” e “fine-simbolo”. Nella versione a soli due simboli, “.” e “—”, il codice Morse non è univocamente decifrabile.

Definizione 3.16.4. Istantaneità

Una codifica C è *istantanea*, o *istantaneamente decifrabile*, se per ogni $x \in \mathcal{X}$ $C(x)$ non è un prefisso di qualunque altra rappresentazione.

Osservazione 3.16.2. Istantaneità vuol dire efficienza

Una codifica istantanea consente l'identificazione di fine rappresentazione di ogni simbolo senza dover attendere l'arrivo del simbolo successivo e quindi consente implementazioni efficienti.

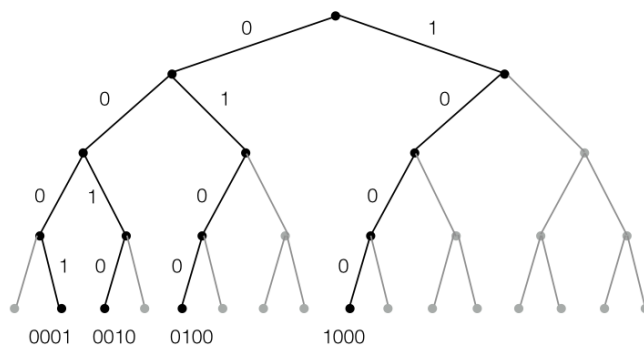


Figura 3.3: Un possibile albero per la codifica istantanea C_1 . Ogni simbolo è una foglia la cui profondità è data dalla lunghezza della rappresentazione che si ottiene concatenando i bit del cammino che parte dalla radice e raggiunge la foglia.

Osservazione 3.16.3. Istantaneità implica decifrabilità univoca

È facile verificare che una codifica istantanea è univocamente decifrabile. Il vincolo del prefisso consente di associare all'intera codifica un albero binario (vedi figura 3.3).

Consideriamo ora la codifica C_3 , sempre per lo stesso insieme di simboli a, b, c e d , con

$$\begin{aligned} C_3(a) &= 1 \\ C_3(b) &= 10 \\ C_3(c) &= 100 \\ C_3(d) &= 1000 \end{aligned}$$

La codifica C_3 è univocamente decifrabile ma, chiaramente, non istantanea.

Lunghezza attesa e disuguaglianza di Kraft-McMillian

Una codifica deve anche mirare a ottenere, in media, quanta più compressione possibile. Al riguardo definiamo la lunghezza attesa di un codice e rispondiamo a una domanda fondamentale attraverso il teorema di Kraft-McMillian.

Definizione 3.16.5. *Lunghezza attesa di una codifica*

La lunghezza attesa $L(C, \mathcal{X})$ di una codifica C è

$$L(C, \mathcal{X}) = \sum_{x \in \mathcal{X}} p(x) L_C(x)$$

Per quanto riguarda gli esempi visti sopra abbiamo

$$\begin{aligned} L(C_1, \mathcal{X}) &= \frac{1}{2} \cdot 4 + \frac{1}{4} \cdot 4 + \frac{1}{8} \cdot 4 + \frac{1}{8} \cdot 4 = 4 \\ L(C_2, \mathcal{X}) &= \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 1 + \frac{1}{8} \cdot 2 + \frac{1}{8} \cdot 2 = 1.25 \\ L(C_3, \mathcal{X}) &= \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 4 = 1.875 \end{aligned}$$

Dato un insieme \mathcal{X} di simboli x_i e di interi L_i con $i = 1, \dots, |\mathcal{X}|$, esiste una codifica univocamente decifrabile C che abbia gli interi L_i come lunghezze delle rappresentazioni $C(x_i)$?

Teorema 3.16.1. *Kraft-McMillian*

Le lunghezze L_i delle rappresentazioni $C(x_i)$ di una codifica C univocamente decifrabile soddisfano la disuguaglianza

$$\sum_{i=1, \dots, |\mathcal{X}|} 2^{-L_i} \leq 1$$

Dimostrazione: definiamo $A = \sum_i 2^{-L_i}$ e, per qualche intero n , consideriamo la quantità

$$A^n = \left(\sum_i 2^{-L_i} \right)^n = \sum_{i_1} \sum_{i_2} \dots \sum_{i_n} 2^{-(L_{i_1} + L_{i_2} + \dots + L_{i_n})}$$

La quantità $L_{i_1} + L_{i_2} + \dots + L_{i_n}$ è la lunghezza della rappresentazione di $x = x_{i_1} x_{i_2} \dots x_{i_n}$. Abbiamo un termine della somma per ogni stringa x di n simboli e tutti i termini diversi per via del fatto che C è certamente univocamente decifrabile. Introduciamo un vettore v_L che conta quante stringhe x nella somma hanno una rappresentazione di lunghezza L . Siano L_m ed L_M rispettivamente il valore minimo e massimo delle lunghezze delle rappresentazioni di ogni simbolo. Possiamo scrivere

$$A^n = \sum_{L=nL_m, \dots, nL_M} 2^{-L} v_L$$

Poiché ci sono al più 2^L stringhe binarie di lunghezza L , abbiamo $v_L \leq 2^L$ e, quindi,

$$A^n = \sum_{L=nL_m, \dots, nL_M} 2^{-L} v_L \leq \sum_{L=nL_m, \dots, nL_M} 1 < nL_M$$

Ora, se A fosse maggiore di 1 questa disuguaglianza non potrebbe valere per tutti gli n , poiché il termine A^n crescerebbe più velocemente del termine lineare nL_M . Pertanto $A \leq 1$.

■

Osservazione 3.16.4. *Verifica*

Verifichiamo la disuguaglianza di Kraft-McMillian per le codifiche C_1 , C_2 e C_3 .

C_1 è univocamente decifrabile poiché $L_i = 4$ per ogni i e

$$2^{-4} + 2^{-4} + 2^{-4} + 2^{-4} = \frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{1}{4} < 1$$

C_2 non è univocamente decifrabile poiché

$$2^{-1} + 2^{-1} + 2^{-2} + 2^{-2} = \frac{1}{2} + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} = \frac{3}{2} > 1$$

C_3 è univocamente decifrabile poiché $L_i = i + 1$ per $i = 0, 1, 2$ e 3 e

$$2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{15}{16} < 1$$

Alberi binari associati a una codifica istantanea e teorema inverso

Se assumiamo che la codifica C sia anche istantanea, la disuguaglianza di Kraft-McMillian può essere apprezzata partendo dall'albero binario associato a C . Assumiamo di rilasciare un'unità di un liquido dalla radice dell'albero. In ogni nodo, ad ogni livello, il liquido si divide in due parti uguali. In una foglia alla profondità L_i troveremo, pertanto, una frazione dell'unità pari a 2^{-L_i} che associamo al valore x_i . Se alcuni rami non sono utilizzati parte del liquido andrà perso e il totale del liquido raccolto nelle foglie corrispondenti ai valori rappresentati sarà strettamente minore dell'unità. Se tutti i rami e le foglie sono utilizzati, invece, il totale del liquido raccolto sarà pari a 1 e la codifica è *completa*.

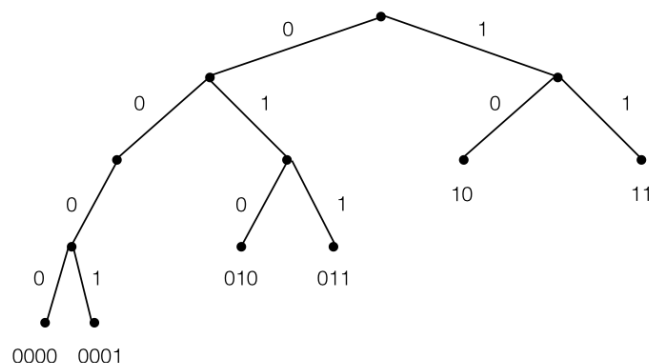


Figura 3.4: Costruzione di una codifica istantanea da un insieme di interi che soddisfano la disuguaglianza di Kraft-McMillian. Date le lunghezze $L(x_1) = L(x_2) = 4$, $L(x_3) = L(x_4) = 3$ e $L(x_5) = L(x_6) = 2$, otteniamo un albero binario di profondità 4. Partendo dall'alto si assegna il primo nodo disponibile di profondità 2 alla prima rappresentazione cancellando tutti i discendenti. Si ripete l'operazione ottenendo $C(x_1) = 0000$, $C(x_2) = 0001$, $C(x_3) = 010$, $C(x_4) = 011$, $C(x_5) = 10$ e $C(x_6) = 11$ e una codifica completa.

In effetti vale anche l'inverso del teorema **3.16.1**.

Teorema 3.16.2. *Inverso di Kraft-McMillian*

Sia x una variabile casuale che assume valori nell'insieme \mathcal{X} . Se un insieme di interi L_i , con i che varia da 1 a $|\mathcal{X}|$, soddisfa la disuguaglianza di Kraft-McMillian, allora esiste una codifica istantanea C per la quale $L_C(x_i) = L_i$.

Dimostrazione: disponiamo ogni simbolo x_i alla profondità appropriata, L_i , in un albero binario partendo dalla profondità minima ed eliminando tutti i discendenti (vedi figura 3.4). Questa operazione

è sempre possibile perché le lunghezze soddisfano la disuguaglianza di Kraft-McMillian. La codifica ottenuta è istantanea.

■

Osservazione 3.16.5. *Codifiche univocamente decifrabili e istantanee*

Per una codifica univocamente decifrabile C vale il teorema **3.16.1**. Allo stesso tempo abbiamo appena visto che, per il teorema **3.16.2**, è sempre possibile costruire una codifica C' istantanea che usa le stesse lunghezze ottenute con C . Nella pratica, pertanto, se disponiamo di una codifica univocamente decifrabile possiamo sempre pensare che sia anche istantanea (o comunque modificabile in una codifica istantanea con simboli codificati con le stesse lunghezze).

3.17 Codifiche in assenza di rumore

Prima di presentare la codifica di Huffman, codifica per simbolo istantanea ottimale, enunciamo una proprietà fondamentale della compressione in assenza di rumore e discutiamo il ruolo dell'entropia come limite inferiore di comprimibilità.

Generalità sulla compressione

Che cosa significa comprimere una data quantità di informazione? La risposta è legata al numero di bit necessari a rappresentare i valori che può assumere una variabile casuale.

Definizione 3.17.1. *Quantità di informazione grezza*

Sia X una variabile casuale e \mathcal{X} l'insieme dei possibili valori assunti da X . La quantità di informazione grezza, H_0 , è

$$H_0 = \log_2 |\mathcal{X}|$$

ed è uguale all'entropia associata a X assumendo che tutti i suoi $|\mathcal{X}|$ valori siano equiprobabili.

Esercizio 3.17.1. *Limite invalicabile*

Se $L_C(x)$ è la lunghezza della rappresentazione del valore $x \in \mathcal{X}$ in una codifica C . Esiste una codifica binaria C^* tale che $\forall x \in \mathcal{X}, L_{C^*}(x) < H_0$?

Soluzione

Conosciamo già la risposta: abbiamo visto che i bit necessari per rappresentare n numeri sono $\log_2 n$. Sappiamo quindi che per codificare $|\mathcal{X}|$ valori sono necessari *almeno* $H_0 = \log_2 |\mathcal{X}|$.

Osservazione 3.17.1. *Legge del buco della piccionaia*

Il risultato appena ottenuto può essere visto come un'applicazione del *pigeonhole principle*: se dobbiamo sistemare $n + k$ lettere in n buche con $k \geq 1$, almeno una buca conterrà due lettere.

Osservazione 3.17.2. *Non esistono compressor perfetti*

La legge del buco della piccionaia è alla base di un risultato generale sull'impossibilità di comprimere *tutti* i possibili file di dimensione M bit in file di dimensione strettamente minore di M . I possibili file di dimensione M bit sono 2^M . Tutti i file di dimensione minore di M bit sono dati dalla somma $S = 2 + 2^2 + \dots + 2^{M-1}$. Poiché $2S = 2^2 + \dots + 2^{M-1} + 2^M$, abbiamo

$$S = 2S - S = 2^2 + \dots + 2^{M-1} + 2^M - (2 + 2^2 + \dots + 2^{M-1}) = 2^M - 2 < 2^M$$

Pertanto, non abbiamo abbastanza file per comprimere *tutti* i file di dimensione M bit in file di dimensione al più $M - 1$. Conseguentemente, può capitare che un compressore reale espanda le dimensioni di un particolare file!

Ruolo dell'entropia nella compressione

Siamo ora in grado di stabilire un limite inferiore per la lunghezza attesa di una codifica univocamente decifrabile.

Teorema 3.17.1. *Entropia come limite insuperabile*

La lunghezza attesa $L(C, \mathcal{X})$ di una codifica univocamente decifrabile C non può essere minore dell'entropia $H(X)$.

Dimostrazione: siano $z = \sum_i 2^{-L_i} \leq 1$ e $q_i = 2^{-L_i}/z$ per $i = 1, \dots, |\mathcal{X}|$ con $\sum_i q_i = 1$. Applicando il logaritmo a $q_i = 2^{-L_i}/z$ si ottiene

$$L_i = \log_2 \frac{1}{q_i} - \log_2 z$$

Ora, se $p_i = p(x_i) > 0$ per $i = 1, \dots, |\mathcal{X}|$ e $\sum_i p_i = 1$ sono le probabilità dei simboli in \mathcal{X} e $t_i = q_i/p_i$, usando l'equazione (3.1) otteniamo

$$\sum_{i=1}^{|\mathcal{X}|} p_i \log_2 \frac{1}{p_i} - \sum_{i=1}^{|\mathcal{X}|} p_i \log_2 \frac{1}{q_i} = \sum_{i=1}^{|\mathcal{X}|} p_i \log_2 t_i \leq (\log_2 e) \sum_{i=1}^{|\mathcal{X}|} p_i (t_i - 1) = (\log_2 e) \sum_{i=1}^{|\mathcal{X}|} (q_i - p_i) = 0$$

con l'uguaglianza che vale se e solo se $q_i = p_i$ per tutti gli i . Pertanto,

$$L(C, \mathcal{X}) = \sum_{i=1}^{|\mathcal{X}|} p_i L_i = \sum_{i=1}^{|\mathcal{X}|} p_i \left(\log_2 \frac{1}{q_i} - \log_2 z \right) \geq \sum_{i=1}^{|\mathcal{X}|} p_i \left(\log_2 \frac{1}{p_i} - \log_2 z \right) \geq H(X)$$

dove l'uguaglianza si ha se e solo se $z = 1$ e le lunghezze delle rappresentazioni soddisfano le uguaglianze $L_i = \log_2 1/p_i$.

■

Osservazione 3.17.3. Controllo di consistenza

Coerentemente, le lunghezze attese $L(C_1, \mathcal{X})$ e $L(C_3, \mathcal{X})$ sono entrambe maggiori di $H(X)$. La lunghezza attesa di C_2 , invece, viola la disuguaglianza del teorema ma non genera contraddizione perché la codifica non è univocamente decifrabile.

Il prossimo teorema pone limiti al valore della lunghezza attesa della codifica ottimale.

Teorema 3.17.2. Codifica di Shannon

Esiste sempre una codifica istantanea C per i valori assunti da una variabile casuale X nell'insieme \mathcal{X} per la quale

$$H(X) \leq L(C, \mathcal{X}) \leq H(X) + 1$$

Dimostrazione: introduciamo l'intero $L_i = \lceil \log_2 1/p_i \rceil$ per ogni p_i , con $i = 1, 2, \dots, |\mathcal{X}|$, e osserviamo che gli interi $L_1, \dots, L_{|\mathcal{X}|}$ soddisfano la disuguaglianza di Kraft-McMillian poiché

$$\sum_{i=1}^{|\mathcal{X}|} 2^{-L_i} = \sum_{i=1}^{|\mathcal{X}|} 2^{-\lceil \log_2 1/p_i \rceil} \leq \sum_{i=1}^{|\mathcal{X}|} 2^{-\log_2 1/p_i} = \sum_{i=1}^{|\mathcal{X}|} p_i = 1$$

Per la lunghezza attesa $L(C, \mathcal{X})$ della codifica univocamente decifrabile C che utilizza le lunghezze L_i , quindi, abbiamo

$$L(C, \mathcal{X}) = \sum_{i=1}^{|\mathcal{X}|} p_i L_i = \sum_{i=1}^{|\mathcal{X}|} p_i \left\lceil \log_2 \frac{1}{p_i} \right\rceil \leq \sum_{i=1}^{|\mathcal{X}|} p_i \left(\log_2 \frac{1}{p_i} + 1 \right) = H(X) + 1$$

■

Codifica di Huffman

La codifica di Huffman è un algoritmo di compressione per simbolo molto semplice. Nel caso in cui la probabilità di ogni simbolo sia nota, la codifica di Huffman è la migliore codifica per simbolo possibile. Come vedremo in seguito, questo non significa che si avvicini al limite inferiore dato dall'entropia. In molti casi questo limite può essere avvicinato solo ricorrendo a compressori basati su flusso di dati.

La codifica di Huffman costruisce un albero da un insieme di $|\mathcal{X}|$ foglie in $|\mathcal{X}| - 1$ fusioni. Ogni carattere $x \in \mathcal{X}$ è un oggetto con un attributo dato dalla probabilità p con cui x compare nel testo da comprimere. Per identificare i due oggetti con la probabilità più piccola si utilizza una coda con priorità

con chiave basata su p . La probabilità del nuovo oggetto è data dalla somma delle probabilità degli oggetti identificati.

Consideriamo l'esempio in figura 3.5 con $\mathcal{X} = \{a, b, c, d\}$ e $p(a) = 1/2$, $p(b) = 1/4$, $p(c) = 1/8$ e $p(d) = 1/8$. L'algoritmo parte assegnando a Q , coda con priorità, i caratteri a, b, c e d . Il primo dei tre cicli rimuove dalla coda c e d , i nodi con la probabilità più bassa, e inserisce nella coda un nuovo nodo z che è il risultato dell'identificazione dei nodi c e d . Il nodo z ha come figlio sinistro c e come figlio destro d . L'arco sinistro codifica 0 e l'arco destro 1. L'algoritmo restituisce l'unico nodo che rimane nella coda, la radice dell'albero della codifica. La codifica di ogni carattere si legge concatenando i bit degli archi che uniscono la radice a ogni foglia.

Algoritmo 3.17.1. *HuffmanCoding*

Input: \mathcal{X} , insieme di $|\mathcal{X}|$ simboli con $p(x)$ probabilità del simbolo x e $\sum_{x \in \mathcal{X}} p(x) = 1$.

Output: T , albero binario in cui ognuna delle $|\mathcal{X}|$ foglie contiene una codifica binaria di uno degli x .

```

 $Q \leftarrow \mathcal{X}$ 
for  $i = 1 \rightarrow (|\mathcal{X}| - 1)$ 
     $x \leftarrow \text{left}(z) \leftarrow \text{EXTRACT-MIN}(Q)$ 
     $y \leftarrow \text{right}(z) \leftarrow \text{EXTRACT-MIN}(Q)$ 
     $p(z) \leftarrow p(x) + p(y)$ 
     $\text{INSERT}(Q, z)$ 
return  $\text{EXTRACT-MIN}(Q)$ 

```

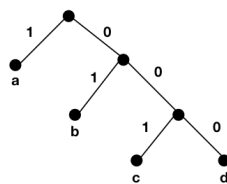


Figura 3.5: Codifica di Huffman: $C(a) = 1$, $C(b) = 01$, $C(c) = 001$ e $C(d) = 0001$.

Osservazione 3.17.4. *Cenni alla correttezza*

La dimostrazione di correttezza di *HuffmanCoding*, algoritmo di tipo *greedy*, è piuttosto elaborata e consiste di due passi. Nel primo si dimostra che una codifica ottimale può sempre essere modificata in modo tale che i due simboli con probabilità minore a e b siano figli dello stesso nodo e si trovino alla profondità massima dell'albero (e quindi differiscano nella codifica per il solo ultimo bit). Nel secondo, per induzione, si dimostra che l'albero ottenuto sostituendo i simboli a probabilità minima, a e b , con il nuovo simbolo ab di probabilità $p(ab) = p(a) + p(b)$ è ottimale.

Osservazione 3.17.5. *Ridondanza*

Nell'esempio precedente sappiamo che $H(X) = 1.75$. Come mostrato in figura 3.5 l'algoritmo produce le rappresentazioni $C(a) = 1$, $C(b) = 01$, $C(c) = 001$ e $C(d) = 0001$. La lunghezza attesa è

$$L(C, \mathcal{X}) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 4 = 1.75 = H(X)$$

Quando, come in questo caso, $L(C, \mathcal{X}) = H(X)$ la codifica è a ridondanza nulla e non ci sono margini di miglioramento per la compressione ottenibile.

Osservazione 3.17.6. Punti di forza

La codifica di Huffman, istantanea per costruzione, è ottimale. Nessun'altra codifica istantanea per simboli può fornire una migliore compressione media. Il suo grande successo è dovuto anche alla semplicità con la quale può essere implementata.

Osservazione 3.17.7. Punti di debolezza

Alcune rigidità della codifica di Huffman possono presentare un conto salato. Il teorema 3.17.2 garantisce che la lunghezza media non può mai essere oltre 1 bit dall'entropia ma la cattiva notizia è che non è difficile trovarsi nei dintorni del caso peggiore. Consideriamo il caso di una moneta truccata con $p \approx 1$. L'entropia è quasi 0 ma la codifica di Huffman è inchiodata alla scelta binaria: 0 per un risultato, 1 per l'altro e non riesce a comprimere nemmeno noiosissime sequenze quasi sempre composte da un solo simbolo. La lunghezza media raggiunta resta lontana dal valore minimo dall'entropia anche nei casi in cui le probabilità sono lontane dall'inverso di potenze di 2, come nel caso di 3 soli possibili valori ugualmente probabili. Infine, l'ipotesi che le probabilità con le quali si presentano i possibili valori siano note a priori e l'assunzione che le variabili casuali siano indipendenti e identicamente distribuite sono spesso poco realistiche.

Osservazione 3.17.8. Blocchi di simboli

Prima di concludere vediamo, per mezzo di un semplice esempio, come sia possibile migliorare l'efficienza della codifica di Huffman ricorrendo a una codifica per blocchi di simboli anziché per simbolo. Sia X una variabile casuale binaria con $p(1) = 3/4$ e $p(0) = 1/4$. Valutiamo l'entropia per simbolo.

$$H(X) = \frac{3}{4} \log_2 \frac{4}{3} + \frac{1}{4} \log_2 4 \approx 0.81$$

Ciononostante, trattandosi di un alfabeto binario, la codifica di Huffman non può portare ad alcuna compressione. Consideriamo allora una codifica di Huffman per la rappresentazione di blocchi di N . Poniamo $N = 4$. Una possibile codifica di Huffman C per i 16 possibili blocchi di 4 simboli è mostrata in tabella.

blocco	probabilità	codifica	blocco	probabilità	codifica	blocco	probabilità	codifica
1111	81/256	01	1010	9/256	00110	0010	3/256	000011
1110	27/256	111	0110	9/256	00101	0100	3/256	000010
1101	27/256	110	1001	9/256	00100	1000	3/256	0000001
1011	27/256	101	0101	9/256	00011	0000	1/256	0000000
0111	27/256	100	0011	9/256	00010			
1100	9/256	00111	0001	3/256	000001			

Per la lunghezza attesa della rappresentazione di un blocco di 4 simboli in questa codifica otteniamo

$$L(C, \mathcal{X}^4) = \frac{1 \cdot 2 \cdot 81 + 4 \cdot 3 \cdot 27 + 6 \cdot 5 \cdot 9 + 3 \cdot 6 \cdot 3 + 1 \cdot 7 \cdot 3 + 1 \cdot 7 \cdot 1}{256} \approx 3.27$$

La lunghezza attesa della codifica per simbolo, $L(C, \mathcal{X}^4)/4 \approx 0.82$, è ora più vicina al minimo valore possibile. Una codifica a blocchi di soli 4 simboli, quindi, consente di avvicinarsi considerevolmente al limite teorico.

Osservazione 3.17.9. Frequenza dei simboli e comprimibilità

È istruttivo osservare che cosa succede alla frequenza relativa degli 0 e degli 1 nella codifica di Huffman. Se tutte le probabilità sono l'inverso di potenze di 2, ovvero quando la lunghezza attesa è uguale all'entropia, 0 e 1 compaiono nelle stesse proporzioni. Nella codifica di Huffman in cui $C(a) = 1$ con $p(a) = 1/2$, $C(b) = 01$ con $p(b) = 1/4$, $C(c) = 001$ con $p(c) = 1/8$ e $C(d) = 000$ con $p(d) = 1/8$,

le occorrenze di 0 e 1 - pesate con le rispettive probabilità - sono in entrambi i casi pari a $7/8$. Che cosa possiamo dire quando la lunghezza attesa non è uguale all'entropia? Contiamo gli 1 e gli 0 nelle due colonne blocco della tabella di sopra, pesati con la corrispondente probabilità. La proporzione di partenza è 75% contro 25%. Se ripetiamo il conteggio dopo la codifica di Huffman a blocchi di 4 simboli, invece, otteniamo 51% contro 49%. La codifica a blocchi, quindi, produce sequenze difficili da comprimere ulteriormente.

Nella pratica la codifica di Huffman richiede di conoscere o di stimare le probabilità dei simboli o, come vedremo, di blocchi di simboli (e.g. byte invece di bit). Ci soffermiamo brevemente su due tra le diverse strategie possibili. Nella prima si stimano le frequenze dei simboli su un *grande numero di file* pubblicando i risultati in un luogo accessibile al compressore e al decompressore. Il vantaggio di questo modo di procedere è di non dover accludere al file compresso la lista delle probabilità utilizzate, senza le quali il decompressore non saprebbe come procedere. La seconda strategia legge una prima volta il singolo file da comprimere in modo da stimare le probabilità con le frequenze presenti in *quel file*. In questo modo si otterrà una compressione tipicamente più spinta ma a scapito di dover accludere al file compresso la lista delle probabilità utilizzate.

Compito 3.17.1. *Compressione di un testo*

Ogni carattere ASCII occupa in memoria un byte. Fissa un file in input che consiste di almeno 10^5 caratteri (spazi bianchi inclusi). Supponiamo che il file contenga M caratteri diversi. Poni la frequenza empirica del carattere x_i del file uguale alla probabilità p_i e calcola l'entropia di Shannon $H(X)$ associata con $\mathcal{X} = \{x_1, \dots, x_M\}$. Implementa una codifica di Huffman C per l'alfabeto \mathcal{X} e confronta la lunghezza attesa $L(C, \mathcal{X})$ con $H(X)$. Comprimi il testo usando la codifica e valuta la compressione assumendo che, nella codifica di Huffman, ogni 0 o 1 sia immagazzinato in un bit.

3.18 Codifica aritmetica

I metodi di compressione basati su flusso di dati aggirano molti dei problemi della codifica di Huffman (e di tutte le altre codifiche per simbolo). In particolare possono raggiungere una compressione vicina a quella ottimale. Discutiamo ora in dettaglio forse il più elegante metodo di compressione basato su flusso di dati in grado, con alta probabilità, di ottenere una lunghezza attesa uguale all'entropia per simbolo. Partiamo dalla codifica. Ci restringiamo al caso in cui i simboli sono cifre.

Algoritmo di codifica

Per ogni N fissato, la codifica aritmetica associa biunivocamente a ogni stringa $x = x_1x_2 \dots x_N$ un intervallo $\Phi_N(x) \subset [0, 1)$ di ampiezza

$$p(x|N) = p(x_1)p(x_2) \dots p(x_N)$$

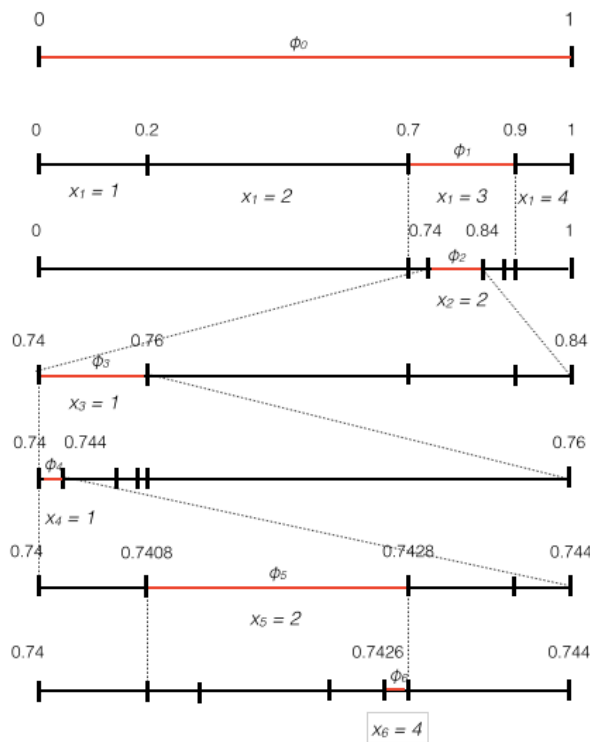


Figura 3.6: Codifica aritmetica di $x = 321124$ con $p(1) = 0.2, p(2) = 0.5, p(3) = 0.2$ e $p(4) = 0.1$.

Esempio 3.18.1. Sia $\mathcal{X} = \{1, 2, 3, 4\}$ con

$$\begin{aligned} p(1) &= 0.2 \\ p(2) &= 0.5 \\ p(3) &= 0.2 \\ p(4) &= 0.1 \end{aligned}$$

Per l'entropia otteniamo

$$H(\mathcal{X}) \approx 1.76$$

Poniamo $N = 6$ e consideriamo $x = 321124$.

Costruiamo una sequenza di intervalli innestati Φ_k chiusi a sinistra e aperti a destra

$$\Phi_k = [\alpha_k, \beta_k) \text{ per } k = 0, 1, \dots, 6$$

con α_k e β_k numeri reali tali che $0 \leq \alpha_k \leq \alpha_{k+1}$ e $\beta_{k+1} \leq \beta_k \leq 1$. Come mostrato nella figura 3.6, nel primo passo dividiamo l'intervallo Φ_0 , ovvero l'intervallo $[0, 1)$, in 4 sotto-intervalli di lunghezza proporzionale a $p(1)$, $p(2)$, $p(3)$ e $p(4)$. Il primo simbolo della stringa, $x_1 = 3$, seleziona l'intervallo $\Phi_1 = [0.7, 0.9)$ di ampiezza $p(3) = 0.2$. A questo punto la suddivisione in 4 sotto-intervalli è eseguita nuovamente, rispettando le proporzioni, sul sotto-intervallo Φ_1 . Il secondo simbolo, $x_2 = 2$, seleziona il sotto-intervallo $\Phi_2 = [0.74, 0.84)$ di ampiezza $p(3)p(2) = 0.10$. Sia l'ampiezza di Φ_2 , sia il suo estremo inferiore, sono determinati da x_1 e x_2 . Per ogni k , si ripete la suddivisione in 4 sotto-intervalli e si procede alla selezione del sotto-intervallo sulla base del k -esimo simbolo. Sia la posizione, sia l'ampiezza di ogni Φ_k , sono determinati da x_k e da tutti i simboli precedenti.

Denotiamo con $\mathcal{L}_k = \beta_k - \alpha_k$ l'ampiezza dell'intervallo $\Phi_k = [\alpha_k, \beta_k)$. Facendo uso della funzione di probabilità cumulata

$$cdf(x) = \sum_{i \leq x} p(i)$$

siamo ora in grado di descrivere un algoritmo per la codifica aritmetica.

Algoritmo 3.18.1. *ArithmeticCoding(x)*

Input: $x = x_1 \dots x_N$, stringa di N cifre x_k per $k = 1, \dots, N$

Output: Φ_N , intervallo $\in [0, 1)$ chiuso a sinistra e aperto a destra

1 $\Phi_0 = [0, 1)$

2 for $k = 1 \rightarrow N$

$$\Phi_k = [\alpha_k, \beta_k) = [\alpha_{k-1} + \mathcal{L}_{k-1}cdf(x_k - 1), \alpha_{k-1} + \mathcal{L}_{k-1}cdf(x_k))$$

Osservazione 3.18.1. *Correttezza*

Per quanto visto prima e per la ripetizione del passo 2, N volte, otteniamo che l'ampiezza \mathcal{L}_N dell'intervallo Φ_N è data da

$$\mathcal{L}_N = p(x|N) = p(x_1)p(x_2) \dots p(x_N)$$

Procedendo per bisezioni successive dell'intervallo $[0, 1)$, quindi, è possibile associare univocamente all'intervallo Φ_N , e alla stringa di lunghezza N una frazione binaria - esprimibile con L bit - dove

$$L = \left\lceil \log_2 \frac{1}{p(x|N)} \right\rceil = \left\lceil \log_2 \frac{1}{p(x_1)} + \log_2 \frac{1}{p(x_2)} + \dots + \log_2 \frac{1}{p(x_N)} \right\rceil$$

Osservazione 3.18.2. *Efficienza*

Rispetto alla codifica di Huffman, la codifica aritmetica è più efficiente perché arrotonda all'intero più vicino solo la lunghezza finale L della rappresentazione di x e non la lunghezza della rappresentazione di ognuna delle N cifre della stringa. Consideriamo il rapporto L/N al crescere di N ; se raggruppiamo la somma di sopra in $|\mathcal{X}|$ somme parziali e indichiamo con $f_j = \#j/N$ la frequenza con cui j compare nella stringa originale otteniamo

$$\frac{L}{N} = \left\lceil f_1 \log_2 \frac{1}{p(1)} + f_2 \log_2 \frac{1}{p(2)} + \dots + f_{|\mathcal{X}|} \log_2 \frac{1}{p(|\mathcal{X}|)} \right\rceil$$

Per la legge dei grandi numeri, al crescere di N , abbiamo che per tutti i j

$$f_j \rightarrow p(j) \text{ e, quindi, } \frac{L}{N} \rightarrow H$$

Algoritmo di decodifica

Analizziamo la decodifica sullo stesso esempio utilizzato per illustrare la codifica. Sia θ la rappresentazione ottenuta con la codifica aritmetica e N la lunghezza della stringa originale. La prima cifra della stringa, x_1 , è ricostruita determinando a quale dei 4 sotto-intervalli dell'insieme $[0, 1)$ appartenga θ . Quindi si procede ad aggiornare il valore di θ , ovvero nel sottrarre al valore corrente di θ la funzione di probabilità cumulata valutata in x_1 , $cdf(x_1)$, e nel dividere il risultato ottenuto per la probabilità di x_1 , $p(x_1)$. La sottrazione determina la posizione di θ all'interno del sotto-intervallo rispetto all'estremo inferiore, mentre la divisione per la probabilità aggiusta il fattore di scala. La procedura descritta è ripetuta per il nuovo valore di θ e, dopo $N - 1$ aggiornamenti, termina con la ricostruzione della stringa. La tabella qui sotto illustra i passi della codifica e della decodifica per l'esempio della stringa $x = 321124$. Il numero θ è rappresentato in base decimale. La tabella mostra anche che cosa succede all'iterazione $N + 1$: se mancasse l'informazione sulla lunghezza della stringa originale, infatti, la decodifica non saprebbe quando fermarsi.

k	carattere da codificare	estremo inferiore di Φ_k	ampiezza di Φ_k	θ_{k-1}	carattere codificato
0	-	-	0	1	-
1	3	0.7	0.2	0.74267578125	3
2	2	0.74	0.1	0.21337890625	2
3	1	0.74	0.02	0.0267578125	1
4	1	0.74	0.004	0.1337890625	1
5	2	0.7408	0.002	0.6689453125	2
6	4	0.7426	0.0002	0.937890625	4
7				0.37890625	2

Supponiamo inizialmente di conoscere $p(x)$ per $x = 1, 2, \dots, |\mathcal{X}|$ e la funzione di probabilità cumulata $cdf(x)$.

Algoritmo 3.18.2. *ArithmeticDecoding*(N, θ)

Input: N lunghezza della stringa codificata, θ codifica aritmetica in notazione decimale

Output: x , stringa decodificata

-
1. $\theta_0 = \theta$
 2. **for** $k = 1 \rightarrow N$

$$\begin{aligned}
 x_k &= \{x \in \mathcal{X} : cdf(x-1) < \theta_{k-1} < cdf(x)\} \\
 \theta_k &= \frac{\theta_{k-1} - cdf(x_k-1)}{p(x_k)}
 \end{aligned}$$

Osservazione 3.18.3. *Separazione del modello dalla codifica*

Il fatto che la decodifica proceda nella ricostruzione della stringa originale elaborando la stringa codificata un simbolo alla volta, e a partire dal primo, apre alla possibilità di concepire variazioni sul tema capaci di sfruttare la separazione del modello probabilistico sottostante dalla codifica vera e propria. Aniché trasmettere al decodificatore le probabilità *a priori* utilizzate nella codifica una volta per tutte, come per le codifiche istantanee tipo Huffman, si costruisce un modello nel quale le probabilità, inizialmente uniformi, sono aggiornate ricorsivamente sulla base delle letture dei simboli della stringa da comprimere, simbolo per simbolo. Questa strategia incorpora nella codifica il modello probabilistico

utilizzato e ne consente la ricostruzione in decodifica senza richiedere alcun passaggio esplicito di informazioni. **La codifica aritmetica, pertanto, è in grado di ottenere compressioni efficienti anche quando le probabilità *a priori* dei simboli non siano disponibili o in presenza di dati caratterizzati da frequenze e occorrenze empiriche particolari.**

Osservazione 3.18.4. *Compressioni a confronto*

Nel nostro esempio, la codifica aritmetica ha determinato un intervallo di ampiezza $A = 0.0002$ per cui la stringa $x = 321124$ è rappresentata da una frazione dell'unità che in base 2 ha lunghezza

$$L = \left\lceil \log_2 \frac{1}{A} \right\rceil = \lceil \log_2 5000 \rceil = \lceil 12.28 \rceil = 13$$

Poiché

$$H(\mathcal{X}) \approx 6 \cdot 1.76 \approx 10.56$$

la codificata ottenuta non è ottimale. Questa inefficienza dipende da due motivi distinti. In primo luogo, il valore calcolato non tiene conto del fatto che gli 0 che seguono l'ultima cifra significativa dopo la virgola sono eliminabili. Nel caso specifico, l'unica frazione dell'unità che appartiene all'intervallo $[0.7426, 0.7428)$ costituita da 13 bit è

$$C(x) = 0.74267578125 = 0.1011111000100_2$$

Eliminando gli ultimi due 0, che non sono significativi, la lunghezza della rappresentazione si riduce a 11 bit. La lunghezza media della rappresentazione di tutte le possibili stringhe di 6 simboli si avvicina asintoticamente a $L - 1$ al crescere di N , ma manca ancora qualcosa. Il secondo motivo è dovuto alla discrepanza tra le probabilità *a priori* dei simboli e la loro frequenza empirica nella stringa. In altre parole, N non è abbastanza grande per far scattare le conseguenze della legge dei grandi numeri. Per valori piccoli di N i vantaggi della codifica aritmetica non sono sempre apprezzabili.

Osservazione 3.18.5. *Dati come flusso*

Diversamente da Huffman, la codifica aritmetica affronta il problema di come comprimere l'informazione contenuta in un flusso di dati in input. Sia la codifica, sia la decodifica, infatti, elaborano esclusivamente l'informazione acquisita fino a quel momento.

Compito 3.18.1. *Codifica e decodifica adattive*

Implementa una versione di *ArithmeticCoding* che codifica il k -esimo bit b_k per $k = 1, \dots, N$ con la probabilità $p_{k-1}(0)$ per $b_k = 0$ e $p_{k-1}(1) = 1 - p_{k-1}(0)$ per $b_k = 1$. Ipotizza che inizialmente 0 e 1 siano equiprobabili, per cui per il primo bit, b_1 , poni $p_0(0) = p_0(1) = 1/2$. Per i successivi utilizza le ricorsioni

$$\begin{aligned} p_k(0) &= p_{k-1}(0) + \frac{(1 - b_k) - p_{k-1}(0)}{k + 2} \\ p_k(1) &= p_{k-1}(1) + \frac{b_k - p_{k-1}(1)}{k + 2} \end{aligned}$$

Il “+2” al denominatore tiene conto del fatto che tutte le sequenze, inizialmente, sono costituite da una coppia 01 non codificata.

Implementa *ArithmeticDecoding* usando come funzione di probabilità cumulata al passo k

$$cdf_k(x) = \sum_{i \leq x} p_k(i)$$

sempre con $p_0(0) = p_0(1) = 1/2$.

3.19 Codifiche in presenza di rumore

Un segnale digitale trasmesso attraverso un canale può essere affetto da rumore e subire modifiche. Nel caso dell'archiviazione di dati in memoria, per esempio, anche nella ragionevole assunzione che la probabilità di errore per singolo bit sia molto piccola, la grande quantità di bit da archiviare rende molto alta la probabilità che qualche bit sia invertito. Nel caso delle telecomunicazioni spaziali, invece, il rumore è particolarmente intenso e pressoché ineliminabile. Invariabilmente il problema è lo sviluppo di una codifica in grado di proteggere il segnale dal rumore. Le soluzioni individuate si basano sempre su codifiche ridondanti.

Distanza di Hamming

I valori 0 e 1 assunti da ogni bit sono il risultato di un confronto tra una misura analogica e un valore di riferimento. Piccole variazioni del segnale sottostante o interferenze con circuiti o dispositivi vicini introducono un disturbo, o rumore, che può provocare l'inversione del bit ricevuto rispetto al bit inviato attraverso un certo canale. Il modello più semplice di rumore assume che la probabilità che un bit sia invertito è indipendente dalla probabilità che un bit vicino nel tempo sia pure invertito. Modelli più realistici tengono conto del fatto che la probabilità che un bit sia invertito è maggiore se si sono verificate inversioni negli istanti immediatamente precedenti. Nella pratica non è raro imbattersi in sequenze, anche piuttosto lunghe, di bit invertiti. Per mitigare il problema si invia attraverso il canale una permutazione pseudo-casuale dei bit che compongono il segnale per poi sottoporre i bit ricevuti alla permutazione inversa.

Prima di discutere un esempio di codifica in presenza di rumore, introduciamo una nozione di distanza utile per confrontare sequenze di bit.

Definizione 3.19.1. *Distanza di Hamming*

Per due sequenze di bit $u = u_1 u_2 \dots u_N$ e $v = v_1 v_2 \dots v_N$ di lunghezza N la distanza di Hamming $d_H(u, v)$ è definita come il numero di posizioni nelle quali i bit delle due sequenze sono diversi.

Proposizione 3.19.1. *La distanza di Hamming è una metrica.*

Dimostrazione

Simmetria Per tutte le sequenze di bit u e v di lunghezza N il fatto che $d_H(u, v) = d_H(v, u)$ segue immediatamente dal fatto che $d_H(u, v)$ e $d_H(v, u)$ contano lo stesso numero di posizioni

Non negatività Per tutte le sequenze di bit u e v di lunghezza N il fatto che $d_H(u, v) \geq 0$ e $d_H(u, v) = 0$ se e solo se $u = v$ segue immediatamente dalla definizione

Disuguaglianza triangolare Per tutte le sequenze di bit u , v e w di lunghezza N deve valere

$$d_H(u, w) + d_H(w, v) \geq d_H(u, v)$$

Partiamo osservando che per ogni posizione i in cui $u_i = v_i$ la disuguaglianza è soddisfatta. Possiamo pertanto limitarci alle sole posizioni nelle quali u e v differiscono. Per ogni posizione i in cui $u_i \neq v_i$ non può essere che $u_i = w_i$ e $v_i \neq w_i$. Pertanto, quando $d_H(u, v)$ aumenta di un'unità, anche $d_H(u, w) + d_H(w, v)$ aumenta di un'unità. ■

Fissiamo un intero $K > 0$ e supponiamo di voler trasmettere tutte le 2^K sequenze possibili di lunghezza K attraverso un canale affetto da rumore. Per ogni sequenza x di lunghezza K ci sono K sequenze x^i a distanza di Hamming uguale a 1 da x , con x^i la sequenza che differisce da x nel bit i -esimo con $i = 1, \dots, K$. Se nella trasmissione di x si invertisse un solo bit, pertanto, la sequenza ricevuta sarebbe una di queste K sequenze e non saremmo in grado di rilevare la presenza di un errore. L'idea della codifica in presenza di rumore è allora aggiungere M bit a ogni blocco in modo tale da ottenere 2^K sequenze di $K + M$ bit che possano mantenersi distinguibili in presenza di errori nella trasmissione, in quanto a distanza di Hamming più grande fra loro.

Codifica convoluzionale

La codifica convoluzionale fornisce un'elegante soluzione al problema di trovare codifiche in presenza di rumore.

L'idea alla base della codifica convoluzionale è molto semplice. Nella fase di codifica una finestra di lunghezza K scorre su una sequenza in input di N bit avanzando di una posizione alla volta. In ogni posizione i K bit all'interno della finestra sono combinati per calcolare un blocco di P bit di parità che è poi trasmesso attraverso il canale. La sequenza originale di N bit è pertanto codificata in una sequenza di N blocchi di P bit, $N \times P$ bit in tutto. È immediato rendersi conto che la velocità di trasmissione della codifica convoluzionale è $V = 1/P$.

Indichiamo con $x[n]$ l' n -esimo bit della sequenza x in input con $n = 1, 2, \dots, N$. Sia K la lunghezza dei vincoli, ovvero la dimensione della finestra W che scorre su x avanzando di una posizione a ogni passo e P il numero di sottoinsiemi dei K bit selezionati per il calcolo dei valori dei P bit di parità $C_n = (y_1[n], y_2[n], \dots, y_P[n])$ per $n = 1, \dots, N$. Posto $x[n] = 0$ per i $K - 1$ valori di n compresi tra $-K + 2$ e 0 , il blocco di P bit di parità $C_n = (y_1[n]y_2[n] \dots y_P[n])$ per ogni $n = 1, \dots, N$ è ottenuto dalle equazioni

$$\begin{aligned} y_1[n] &\equiv w_1[0]x[n] + w_1[1]x[n-1] + \dots + w_1[K-1]x[n-K+1] \pmod{2} \\ y_2[n] &\equiv w_2[0]x[n] + w_2[1]x[n-1] + \dots + w_2[K-1]x[n-K+1] \pmod{2} \\ &\dots \equiv \dots \\ y_P[n] &\equiv w_P[0]x[n] + w_P[1]x[n-1] + \dots + w_P[K-1]x[n-K+1] \pmod{2} \end{aligned}$$

dove i polinomi generatori w_1, w_2, \dots, w_P agiscono da selezionatori dei bit della sequenza di input all'interno della finestra W .

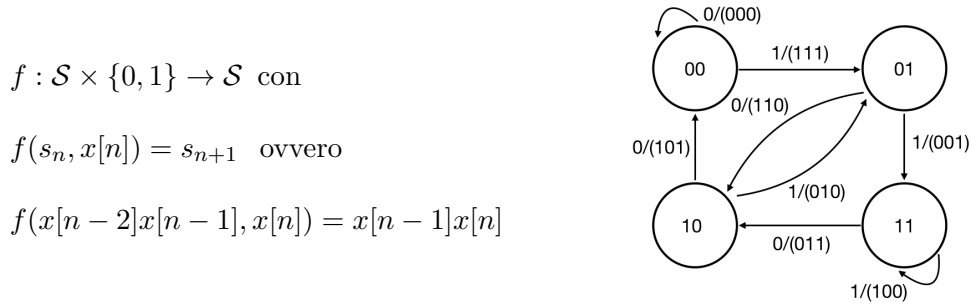


Figura 3.7: Funzione di transizione e FSM per l'esempio 3.19.1.

Osservazione 3.19.1. Macchina a stati finiti per la codifica

Nel linguaggio delle macchine a stati finiti (FSM), l'input è il bit $x[n]$, lo stato corrente s_n i $K - 1$ bit $x[n-1]x[n-2] \dots x[n-K+1]$ e l'output della FSM il blocco di P bit $C_n = (y_1[n]y_2[n] \dots y_P[n])$. Gli stati possibili sono 2^{K-1} . Il comportamento della FSM è completamente determinato dalle equazioni di parità. Il diagramma nella figura 3.7 si riferisce al caso del prossimo esempio. Lo stato corrente è scritto all'interno di ogni nodo, mentre ogni arco punta al nuovo stato riportando il valore dell'input e il blocco dei $P = 3$ bit codificati. La funzione di transizione f che sulla base dello stato s_n e dell'input $x[n]$ restituisce lo stato s_{n+1} è esplicitata nella figura 3.7.

Esempio 3.19.1. Per $K = 3$ e $P = 3$, se $w_1 = [1 \ 1 \ 1]$, $w_2 = [1 \ 1 \ 0]$ e $w_3 = [1 \ 0 \ 1]$ per le equazioni di parità abbiamo

$$\begin{aligned} y_1[n] &\equiv x[n] + x[n-1] + x[n-2] \pmod{2} \\ y_2[n] &\equiv x[n] + x[n-1] \pmod{2} \\ y_3[n] &\equiv x[n] + x[n-2] \pmod{2} \end{aligned}$$

Per ogni n da 1 a $N + 1$ lo stato s_n appartiene all'insieme $\mathcal{S} = \{00, 01, 10, 11\}$. Utilizzando la macchina a stati finiti in figura 3.7 per codificare la sequenza $x = 1101$ otteniamo

$$C(x) = (C_1(x))(C_2(x))(C_3(x))(C_4(x)) = (111)(001)(011)(010)$$

Algoritmo di Viterbi

Presentiamo ora un algoritmo di decodifica. Descriviamo il suo funzionamento sullo stesso esempio utilizzato per la codifica. Sappiamo che la sequenza $x = 1101$ è codificata in output nella sequenza di $N = 4$ blocchi di $P = 3$ bit, C_1, \dots, C_4 . Se per effetto del rumore il terzo bit di C_1^{ric} e il primo bit di C_3^{ric} sono invertiti all'uscita del canale riceveremo

$$C^{ric} = (C_1^{ric})(C_2^{ric})(C_3^{ric})(C_4^{ric}) = (110)(001)(111)(010)$$

Misuriamo l'accordo tra $C(x)$ e C^{ric} contando il numero di posizioni $L(x)$ in cui le due sequenze di 4 blocchi di 3 bit non coincidono. Agendo sulle 4 triplette separatamente, otteniamo

$$L(x) = d_H(C(x), C^{ric}) = \sum_{n=1}^4 d_H(C_n(x), C_n^{ric})$$

Chiaramente, $L(x) = 0$ se e solo se non ci sono errori di trasmissione, ovvero se e solo se $C^{ric} = C(x)$. Pertanto possiamo enunciare un principio generale.

Principio 3.19.1. Piccolo è bello

La sequenza in input è quella che codifica in output la sequenza di bit di parità a distanza di Hamming minima dalla sequenza ricevuta in uscita dal canale. \square

Poichè la tripletta $C_n(x)$ è funzione del bit $x[n]$ da codificare e dello stato $s_n = x[n-2]x[n-1]$, possiamo scrivere

$$C_n(x) = C_n(s_n, x[n])$$

e, quindi,

$$L(s_1, x[1], \dots, x[N]) = \sum_{n=1}^N d_H(C_n(s_n, x[n]), C_n^{ric})$$

L'algoritmo di Viterbi per la decodifica di C^{ric} si basa sulla ricerca della sequenza $x^* = x^*[1] \dots x^*[N]$ tale che

$$\begin{aligned} x^*[1] \dots x^*[N] &= \arg \min_{x[1] \dots x[N]} L(s_1, x[1], \dots, x[N]) \\ \text{con } x[n] &= \{0, 1\} \text{ per } n = 1, \dots, N \text{ e} \\ s_1 &= 00 \end{aligned} \tag{3.2}$$

Nel nostro caso, ovviamente, abbiamo $N = 4$. Il valore del minimo, $L^*(s_1)$, è pari al numero di bit di parità invertiti nella trasmissione.

Osservazione 3.19.2. Una forma distinguibile

Il problema (3.2) ha una struttura particolare: **la parte finale della decodifica ottimale per l'intera sequenza ricevuta è la decodifica ottimale per la corrispondente parte finale della sequenza ricevuta.** Il problema di ottimizzazione sottostante è quindi affrontabile con metodologie di programmazione dinamica (*DP*). In *DP* la sequenza di bit di input $x[1] \dots x[N]$ è la sequenza di controllo e $x^*[1] \dots x^*[N]$ la sequenza ottimale. \square

Prima di procedere con la descrizione dell'algoritmo di decodifica, precalcoliamo le distanze di Hamming tra ogni possibile blocco ricevuto al passo n con il blocco dalla FSM di figura 3.7 per i due possibili diversi input, $x[n] = 0$ e $x[n] = 1$ e per tutti i possibili stati. Otteniamo la seguente tabella

$x[n-2]x[n-1]$ stato: 00	$x[n] \rightarrow (y_1[n]y_2[n]y_3[n])$ $0 \rightarrow (000) \quad 1 \rightarrow (111)$	$x[n-2]x[n-1]$ stato: 10	$x[n] \rightarrow (y_1[n]y_2[n]y_3[n])$ $0 \rightarrow (101) \quad 1 \rightarrow (010)$
blocco ricevuto	distanza di Hamming	blocco ricevuto	distanza di Hamming
(000)	0 3	(000) (011) (110)	2 1
(001) (010) (100)	1 2	(001) (100) (111)	1 2
(011) (101) (110)	2 1	(010)	3 0
(111)	3 0	(101)	0 3
$x[n-2]x[n-1]$ stato: 01	$x[n] \rightarrow (y_1[n]y_2[n]y_3[n])$ $0 \rightarrow (110) \quad 1 \rightarrow (001)$	$x[n-2]x[n-1]$ stato: 11	$x[n] \rightarrow (y_1[n]y_2[n]y_3[n])$ $0 \rightarrow (011) \quad 1 \rightarrow (100)$
blocco ricevuto	distanza di Hamming	blocco ricevuto	distanza di Hamming
(000) (011) (101)	2 1	(000) (101) (110)	2 1
(001)	3 0	(001) (010) (111)	1 2
(010) (100) (111)	1 2	(011)	0 3
(110)	0 3	(100)	3 0

Per la decodifica ricorriamo al traliccio di figura 3.8. Ogni riga si riferisce a uno dei quattro stati possibili, indicati nella colonna di sinistra. Le quattro triplette ricevute sono indicate nella riga superiore nell'ordine nel quale compaiono nella sequenza ricevuta. Ogni colonna di rettangoli si riferisce allo stato corrente $n = 1, \dots, 5$ includendo lo stato finale nell'ultima colonna.

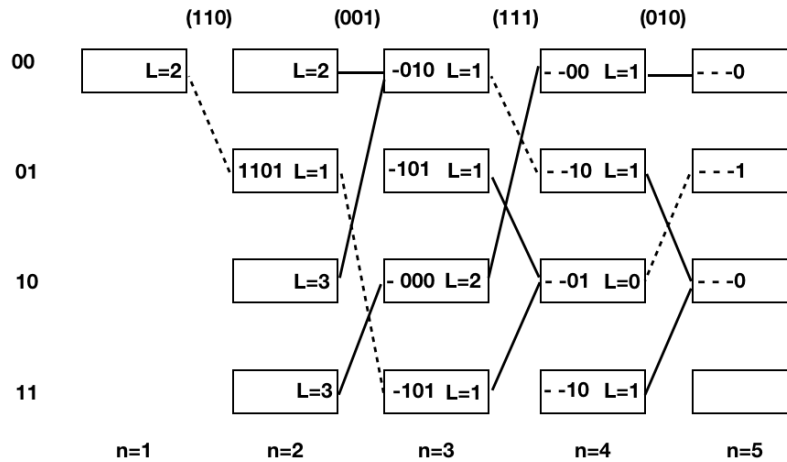


Figura 3.8: Vedi testo.

Sfruttando la programmazione dinamica possiamo riempire il traliccio da destra a sinistra in quattro passi. Partiamo dall'arrivo dell'ultima tripletta, C_4^{ric} . Dobbiamo determinare, per ogni possibile stato di partenza della colonna $n = 4$, quale dei due possibili valori di $x[4]$ produce la tripletta $C_4(s_4, x[4])$ più vicina a C_4^{ric} . In figura 3.8 delle due possibili linee (piena se $x[4] = 1$ e tratteggiata se $x[4] = 0$) che partendo da ognuno dei quattro possibili stati della colonna $n = 4$ terminano in uno stato della colonna $n = 5$ è indicata sola quella corrispondente alla codifica della tripletta a distanza più piccola da C_4^{ric} . Il bit decodificato, il quarto, è rappresentato all'interno del rettangolo corrispondente allo stato finale cui punta la linea. Nel rettangolo da cui parte la linea è indicata la distanza di Hamming tra la tripletta ricevuta e la tripletta codificata.

Consideriamo ora la tripletta C_3^{ric} . Nuovamente, da ognuno dei quattro possibili stati nella colonna $n = 3$ delle due possibili linee (piena se $x[3] = 1$ e tratteggiata se $x[3] = 0$) tracciamo solo quella che corrisponde al percorso che *minimizza la somma* delle distanze di Hamming delle triplette decodificate con le triplette ricevute per $n = 3$ e $n = 4$. Se le due somme forniscono lo stesso valore, scegliamo una delle due linee a caso. I bit decodificati, il terzo e il quarto, sono mostrati sulla sinistra del rettangolo corrispondente allo stato di arrivo (colonna $n = 4$), mentre il valore corrente della somma delle distanze di Hamming per le ultime due triplette decodificate ottimali è indicato nel rettangolo corrispondente allo stato di partenza (colonna $n = 3$). Ripetendo per altre due volte la stessa procedura otteniamo una decodifica ottimale, nonché il numero e le posizioni dei bit ricevuti invertiti dal rumore.

Osservazione 3.19.3. Robustezza

Sequenze di N bit codificate per mezzo di $3N$ bit di parità individuano 2^N vertici dei 2^{3N} vertici di un cubo in $3N$ dimensioni. La tabella qui sotto mette a confronto il numero totale di vertici a distanza di Hamming 1 e 2 da ciascuno dei 2^N vertici nei casi in cui i bit di parità siano $2N$ e $3N$. All'aumentare di N il numero totale di vertici a distanza di Hamming 2 dai 2^N vertici è una frazione trascurabile di 2^{2N} o 2^{3N} , numero totale di vertici nei due casi.

N	2^{2N}	$d_H = 1$ $2^N 2N$	$d_H = 2$ $2^N 2N(2N - 1)/2$	2^{3N}	$d_H = 1$ $2^N 3N$	$d_H = 2$ $2^N 3N(3N - 1)/2$
4	256	128	448	4096	192	1056
8	$\approx 66K$	4096	$\approx 31K$	$\approx 17M$	6144	$\approx 71K$
12	$\approx 17M$	$\approx 98K$	$\approx 1M$	$\approx 69G$	$\approx 147K$	$\approx 3M$

Capitolo 4

Elementi di Inferenza

4.20 Inferenza frequentista

Il primo principio che discutiamo per inferire dai dati è il *principio di massima verosimiglianza*.

Principio di massima verosimiglianza

Disponiamo di n dati (x_1, x_2, \dots, x_n) realizzazioni di n variabili casuali X_i *indipendenti e identicamente distribuite, iid*, con $i = 1, 2, \dots, n$. Supponiamo di sapere che la funzione di distribuzione sottostante f , funzione di *probabilità di massa* nel caso discreto e di *densità* nel caso continuo, dipende da un parametro θ **che non conosciamo**. Scriviamo pertanto f come f_θ . Vogliamo *inferire* dagli n dati il valore di θ che determina la distribuzione che ha effettivamente generato i dati. Indichiamo questo valore con θ_0 e con f_{θ_0} la distribuzione corrispondente. Tipicamente il parametro θ è reale e appartiene all'insieme $\Theta \subseteq \mathbb{R}$, ma più in generale θ potrebbe essere un vettore di parametri.

Una possibile strategia è calcolare la verosimiglianza

$$L(\mathbf{x}|\theta) = \prod_{i=1}^n f_\theta(x_i)$$

ovvero la probabilità con la quale ci aspettiamo di ottenere i dati che abbiamo effettivamente osservato in funzione del parametro θ .

Osservazione 4.20.1. Verosimiglianza o probabilità

La verosimiglianza è una *probabilità* per ogni valore di θ ma *non è una probabilità* come funzione di θ . Quando scriviamo $L(\mathbf{x}|\theta)$, pertanto, non stiamo studiando probabilità di eventi diversi ma quanto sia verosimile il verificarsi di un evento fissato al variare della distribuzione sottostante. \square

Ci aspettiamo che per valori di θ vicini θ_0 la verosimiglianza assuma valori più grandi che per valori lontani da θ_0 . Stimiamo θ_0 , pertanto, come il **massimo della verosimiglianza**, ovvero

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \prod_{i=1}^n f_\theta(x_i)$$

Vediamo con qualche esempio in che senso $\hat{\theta}$ è una *buona* stima di θ_0 .

Distribuzione di Bernoulli Supponiamo di voler inferire il parametro p_0 di una distribuzione di Bernoulli. Indicando il parametro θ con p , abbiamo $\Theta = (0, 1)$ e $\theta_0 = p_0$. I dati $x_i \in \{0, 1\}$ per $i = 1, \dots, n$

corrispondono, per esempio, all'esito di n lanci di una stessa moneta (1 e 0 i valori assunti dalla variabile casuale X con probabilità p_0 e $1 - p_0$). Per p fissato, la funzione di probabilità di massa può essere scritta come

$$f_p(X) = p^X (1 - p)^{(1-X)} \quad (4.1)$$

L'equazione (4.1) può essere studiata analiticamente al variare del parametro p (notiamo che, correttamente, $f_p(1) = p$ e $f_p(0) = 1 - p$). Per la verosimiglianza abbiamo

$$L(x_1, \dots, x_n | p) = \prod_{i=1}^n p^{x_i} (1 - p)^{(1-x_i)} \quad (4.2)$$

Applichiamo il logaritmo (che in quanto funzione monotona non modifica il massimo)

$$\ln L(\mathbf{x} | p) = \ln p \sum_{i=1}^n x_i + \ln(1 - p) \sum_{i=1}^n (1 - x_i)$$

e annulliamo la derivata rispetto a p ottenendo

$$\frac{d \ln L(\mathbf{x} | p)}{dp} = 0 = \frac{1}{p} \sum_{i=1}^n x_i - \frac{n}{1 - p} + \frac{1}{1 - p} \sum_{i=1}^n x_i$$

la cui soluzione è

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$$

Quindi \hat{p} , la stima di massima verosimiglianza di p_0 , è la media empirica dei valori ottenuti. Non sorprendentemente, \hat{p} stima p_0 come frazione delle *teste* ottenute su n lanci.

Distribuzione uniforme Indicando con S il parametro θ , consideriamo ora il caso di una distribuzione uniforme f_S

$$f_S(x) = \begin{cases} 1/S & x \in [0, S] \\ 0 & \text{altrove} \end{cases}$$

dove $S > 0$ è il supporto di f_S . In questo caso abbiamo $\Theta = (0, +\infty)$ e $\theta_0 = S_0$. I dati x_i per $i = 1, \dots, n$ sono valori campionati uniformemente nell'intervallo $[0, S_0]$ con S_0 fissato ma incognito. Un esempio nel caso discreto è dato da una città in cui tutti i taxi sono registrati con un numero da 0 a S_0 . Vogliamo stimare S_0 osservando n numeri x_1, \dots, x_n . Sia x_{max} il massimo valore tra gli x_i . La verosimiglianza vale il prodotto di n fattori uguali a $1/S$ se $S \geq x_{max}$ e 0 altrimenti, ovvero

$$L(\mathbf{x} | S) = \begin{cases} 1/S^n & \text{se } S \geq x_{max} \\ 0 & \text{se } S < x_{max} \end{cases} \quad (4.3)$$

In questo caso non ci serve calcolare la derivata per determinare il massimo. La verosimiglianza è massima per il più piccolo valore di S , e quindi per

$$\hat{S} = x_{max}$$

Pertanto \hat{S} , la stima di massima verosimiglianza di S_0 , è il massimo dei valori campionati (ovvero, nell'esempio dei taxi, il numero di registrazione più grande tra quelli osservati).

Discutiamo ora un caso in cui il parametro θ è bi-dimensionale.

Distribuzione normale Supponiamo ora che i dati x_1, \dots, x_n provengano dalla distribuzione normale $\mathcal{N}(\mu_0, \sigma_0^2)$ con μ_0 e σ_0^2 non noti. In questo caso θ corrisponde alla coppia $\theta = (\mu, \sigma^2)$. Per la verosimiglianza abbiamo

$$L(\mathbf{x}|\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - \mu)^2 / 2\sigma^2} \quad (4.4)$$

Applicando il logaritmo otteniamo

$$\ln L(\mathbf{x}|\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

Uguagliando a zero la derivata rispetto a μ otteniamo

$$\begin{aligned} \frac{\partial}{\partial \mu} \ln L(\mathbf{x}|\mu, \sigma^2) &= \frac{\partial}{\partial \mu} \left(-\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \end{aligned}$$

che è risolta da

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

Anche in questo caso $\hat{\mu}$, la stima di massima verosimiglianza di μ_0 , coincide con la media empirica.

Ponendo $\mu = \hat{\mu}$ e uguagliando a zero la derivata rispetto a σ^2 otteniamo

$$\begin{aligned} \frac{\partial}{\partial \sigma^2} \ln L(\mathbf{x}|\mu, \sigma^2) &= \frac{\partial}{\partial \sigma^2} \left(-\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \hat{\mu})^2 \right) \\ &= -\frac{n}{2\sigma^2} + \left(\frac{1}{2} \sum_{i=1}^n (x_i - \hat{\mu})^2 \right) \frac{1}{\sigma^4} = 0 \end{aligned}$$

che è risolta da

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 \quad (4.5)$$

Proprietà asintotiche

Poniamoci ora il problema di confrontare i valori ottenuti applicando il principio della massima verosimiglianza con il valore atteso dei parametri.

Correttezza

Distribuzione binomiale Sia p_0 il valore del parametro della distribuzione di Bernoulli che ha effettivamente generato i dati osservati. Per \hat{p} abbiamo

$$\mathbb{E}[\hat{p}] = \mathbb{E} \left[\frac{1}{n} \sum_i x_i \right] = \frac{1}{n} \mathbb{E} \left[\sum_i x_i \right] = \frac{1}{n} \sum_i \mathbb{E}[x_i] = \frac{np_0}{n} = p_0$$

Poiché $\mathbb{E}[\hat{p}] = p_0$, lo stimatore di massima verosimiglianza \hat{p} è *corretto*.

Distribuzione uniforme Poiché $x_{max} \leq S_0$, lo stimatore di massima verosimiglianza $\hat{S} = x_{max}$ è intrinsecamente *distorto* per difetto. Ripetendo la stima tante volte, in questo caso, i risultati fluttuano sempre dalla stessa parte non essendo possibile osservare un numero di registrazione più grande di S_0 .

Distribuzione normale Sia $\mathcal{N}(\mu_0, \sigma_0^2)$ la distribuzione normale che ha effettivamente generato i dati. Per $\hat{\mu}$ abbiamo

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{n} \sum_i x_i\right] = \frac{1}{n} \mathbb{E}\left[\sum_i x_i\right] = \frac{1}{n} \sum_i \mathbb{E}[x_i] = \frac{1}{n} \sum_i \mu_0 = \frac{n\mu_0}{n} = \mu_0$$

Lo stimatore $\hat{\mu}$, quindi, è *corretto*. Tenuto conto che per la varianza di $\hat{\mu}$ abbiamo

$$Var(\hat{\mu}) = \mathbb{E}[(\hat{\mu} - \mu_0)^2] = Var\left(\frac{1}{n} \sum_i x_i\right) = \frac{1}{n^2} \sum_i Var(x_i) = \frac{n\sigma_0^2}{n^2} = \frac{\sigma_0^2}{n}$$

per il valore atteso di $\hat{\sigma}^2$ otteniamo allora

$$\begin{aligned} \mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{n} \sum_i (x_i - \hat{\mu})^2\right] = \frac{1}{n} \sum_i \mathbb{E}[(x_i - \hat{\mu})^2] = \frac{1}{n} \sum_i \mathbb{E}[(x_i - \mu_0 + \mu_0 - \hat{\mu})^2] \\ &= \frac{1}{n} \sum_i \mathbb{E}[(x_i - \mu_0)^2] + \frac{1}{n} \sum_i \mathbb{E}[(\mu_0 - \hat{\mu})^2] - 2\mathbb{E}\left[(\mu_0 - \hat{\mu}) \frac{1}{n} \sum_i (\mu_0 - x_i)\right] \\ &= \frac{1}{n} n\sigma_0^2 + \frac{1}{n} \frac{n\sigma_0^2}{n} - 2\mathbb{E}[(\mu_0 - \hat{\mu})^2] = \sigma_0^2 + \frac{\sigma_0^2}{n} - 2\frac{\sigma_0^2}{n} = \frac{n-1}{n} \sigma_0^2 \end{aligned}$$

Lo stimatore $\hat{\sigma}^2$, pertanto, è *distorto*. Tuttavia, poiché $(n-1)/n \rightarrow 1$ per $n \rightarrow \infty$, è *asintoticamente corretto*. Si può *correggere* lo stimatore dividendo per $n-1$ anziché per n nella formula (4.5).

Consistenza

Una seconda importante proprietà utile per valutare la qualità di uno stimatore è la *consistenza*, ovvero la convergenza in probabilità della stima ottenuta rispetto al valore vero nel senso della legge dei grandi numeri. Per le distribuzioni binomiale e normale la consistenza è garantita dalla convergenza della media empirica al valore atteso. Per la distribuzione uniforme la convergenza, sempre per la legge dei grandi numeri, è garantita dal fatto che

$$\text{per } n \rightarrow \infty, \forall \epsilon \Pr(S - x_{max} \geq \epsilon) \rightarrow 0$$

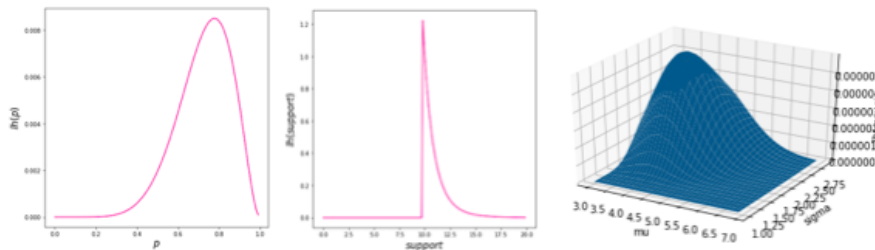


Figura 4.1: Vedi testo

Compito 4.20.1. *La forma della verosimiglianza*

1. Fissa $0 < p_0 < 1$ e campiona n punti x_i con $i = 1, \dots, n$ dalla distribuzione uniforme in $[0, 1]$ e poni $x_i = 1$ se $x_i < p_0$ e 0 altrimenti. Produci il grafico della verosimiglianza dell'equazione (4.2), vedi figura 4.1 a sinistra, per $0 < p < 1$. Confronta \hat{p} con p_0 ripetendo l'esperimento 10 volte con $n = 10$ ed $n = 100$ e per diversi valori di p_0 .
2. Fissa $S_0 > 0$ e campiona n punti da una distribuzione uniforme in $[0, S_0]$. Produci il grafico della verosimiglianza dell'equazione (4.3), vedi figura 4.1 al centro, per $0 < S < 2S_0$. Confronta \hat{S} con S_0 ripetendo l'esperimento 10 volte con $n = 10$ ed $n = 100$.
3. Fissa μ_0 e σ_0^2 e campiona n punti da una distribuzione normale $\mathcal{N}(\mu_0, \sigma_0^2)$. Produci il grafico della verosimiglianza dell'equazione (4.4), vedi figura 4.1 a destra, per $0 < \mu < 2\mu_0$ e $0 < \sigma^2 < 2\sigma_0^2$. Confronta $\hat{\mu}$ con μ_0 e $\hat{\sigma}^2$ con σ_0^2 ripetendo l'esperimento 10 volte con $n = 10$ ed $n = 100$.

4.21 Inferenza Bayesiana

Principio di Massima Probabilità a Posteriori

Un altro metodo di inferenza consiste nell'applicazione diretta del teorema di Bayes. L'inferenza in questo caso si appoggia al principio di *Massima A Posteriori* (MAP).

Le apparenze possono ingannare

Un tizio si reca dal medico perché sospetta di essere affetto da una certa patologia. Effettua un test con *sensibilità* del 95%, probabilità che una persona affetta dalla patologia risulti positiva, e con *specificità* del 99%, probabilità che una persona sana risulti negativa. Sapendo che la patologia colpisce lo 0.2% della popolazione, con quale probabilità soffre della patologia il tizio se risulta positivo al test?

Rifuggiamo l'idea di affidarci all'intuizione e usiamo il teorema di Bayes. Se Pos è l'evento di risultare positivo al test e Neg l'evento di risultare negativo abbiamo

$$sensibilità = \Pr(Pos | malato) = 0.95 \quad e \quad specificità = \Pr(Neg | sano) = 0.99$$

Dalla seconda possiamo evincere che

$$\Pr(Pos | sano) = 1 - \Pr(Neg | sano) = 0.01$$

Inoltre, sappiamo che per la probabilità *a priori* abbiamo

$$\Pr(malato) = 0.002$$

Dal teorema di Bayes, pertanto, otteniamo che per la probabilità *a posteriori* si ha

$$\begin{aligned} \Pr(malato | Pos) &= \frac{\Pr(malato)\Pr(Pos | malato)}{\Pr(malato)\Pr(Pos | malato) + \Pr(sano)\Pr(Pos | sano)} \\ &= \frac{0.002 \times 0.95}{0.002 \times 0.95 + 0.998 \times 0.01} \\ &= \frac{0.00190}{0.00190 + 0.00998} \approx 0.016 \end{aligned} \quad (4.6)$$

La probabilità *a posteriori* ottenuta è circa 8 volte più grande di quella *a priori*, ma per la rarità della patologia la probabilità che il paziente sia effettivamente affetto dalla patologia continua a essere piuttosto piccola. Più del 98% dei soggetti positivi al test, inoltre, sono sani!

Ipotesi a confronto

Consideriamo un secondo semplice esercizio che chiarisce come l'approccio bayesiano si basi sull'accumulo di evidenze basate sui dati.

Esercizio 4.21.1. Quale tipo di moneta?

Sono dati tre tipi di monete. Le monete di tipo A per le quali la probabilità di ottenere *testa*, T , con un lancio è $1/2$, quelle di tipo B per le quali è $3/5$ e quelle di tipo C per le quali è $9/10$. Un cassetto contiene due monete di tipo A , una di tipo B e una di tipo C . Pesco una moneta a caso. Qual è la probabilità che la moneta sia di tipo A , B o C ? Se lanciando la moneta una prima volta ottengo *testa*, come cambiano le probabilità?

Per ottenere la risposta alla prima domanda basta applicare le probabilità *a priori*. Per la seconda, invece, il teorema di Bayes. Se indichiamo con T_1 il risultato *testa* al primo lancio il teorema di Bayes ci dice che la probabilità *a posteriori* che la moneta sia di tipo A si ottiene come

$$\Pr(A|T_1) = \frac{\Pr(A)\Pr(T_1|A)}{\Pr(A)\Pr(T_1|A) + \Pr(B)\Pr(T_1|B) + \Pr(C)\Pr(T_1|C)}$$

e, similmente, se di tipo B o C . La tabella qui sotto riporta le probabilità *a priori* e le probabilità *a posteriori* ottenute con il teorema di Bayes.

ipotesi	<i>a priori</i>	verosimiglianza	<i>a posteriori</i> non normalizzata 1	<i>a posteriori</i> 1
H	$\Pr(H)$	$\Pr(T_1 H)$	$\Pr(T_1 H)\Pr(H)$	$\Pr(H T_1)$
A	0.5	0.5	0.25	0.4
B	0.25	0.6	0.15	0.24
C	0.25	0.9	0.225	0.36
totale	1		0.625	1

Osservazione 4.21.1. Metodi a confronto

La moneta di tipo A è ancora la più probabile anche se il margine di vantaggio sulla moneta di tipo C si è ridotto notevolmente. Se applicassimo il principio di massima verosimiglianza, invece, otterremmo che la moneta di tipo C è la più probabile.

La forza dell'inferenza bayesiana risiede nel fatto che lo schema appena visto può essere iterato nel caso si accumuli ulteriore evidenza. Per esempio: come si modificano le probabilità se lanciando una seconda volta la moneta ottenessimo nuovamente *testa*? Per rispondere a questa domanda ripetiamo l'aggiornamento delle probabilità *a posteriori* non normalizzate moltiplicandole per le verosimiglianze (che non sono cambiate) e imponiamo la normalizzazione all'ultimo passo. I risultati sono mostrati nella tabella qui sotto.

ipotesi	<i>a posteriori</i> non normalizzata 1	<i>a posteriori</i> non normalizzata 2	<i>a posteriori</i> 2
H	$\Pr(T_1 H)\Pr(H)$	$\Pr(T_1 H)\Pr(T_2 H)\Pr(H)$	$\Pr(H T_1T_2)$
A	0.25	0.125	0.299
B	0.15	0.09	0.216
C	0.225	0.2025	0.485
totale	0.625	0.41750	1

Osservazione 4.21.2. Più osservazioni

Le differenze tra i metodi frequentisti e i metodi bayesiani tendono a ridursi all'aumentare del numero di osservazioni.

Aggiornamento delle probabilità predittive

Riconsideriamo lo schema descritto nell'esercizio 4.21.1 e poniamoci il problema di stimare le probabilità di ottenere *testa* nelle tre condizioni, ovvero prima di effettuare lanci, dopo aver ottenuto *testa* con un primo lancio, T_1 , e dopo aver ottenuto *testa* anche con un secondo lancio, T_2 .

Prima di effettuare lanci, la probabilità di ottenere *testa* con un primo lancio è

$$\begin{aligned}\Pr(T_1) &= \Pr(T_1|A)\Pr(A) + \Pr(T_1|B)\Pr(B) + \Pr(T_1|C)\Pr(C) \\ &= 0.5 \cdot 0.5 + 0.25 \cdot 0.6 + 0.25 \cdot 0.9 = 0.625\end{aligned}$$

Se l'esito del primo lancio è *testa*, invece, la probabilità di ottenere *testa* con un secondo lancio è

$$\begin{aligned}\Pr(T_2|T_1) &= \Pr(T_2|A)\Pr(A|T_1) + \Pr(T_2|B)\Pr(B|T_1) + \Pr(T_2|C)\Pr(C|T_1) \\ &= 0.5 \cdot 0.4 + 0.6 \cdot 0.24 + 0.9 \cdot 0.36 = 0.668\end{aligned}$$

Osservazione 4.21.3. *Probabilità predittive*

Sia $\Pr(T_1)$ sia $\Pr(T_2|T_1)$ esprimono una predizione. Nel primo caso, $\Pr(T_1)$ è una probabilità *predittiva a priori*. Nel secondo, $\Pr(T_2|T_1)$ è una probabilità *predittiva a posteriori*. Notiamo che il risultato del primo lancio aumenta la probabilità di ottenere *testa* in un secondo lancio. In entrambi i casi riscriviamo $\Pr(T_1)$ e $\Pr(T_2|T_1)$ come valore atteso delle probabilità condizionate alla scelta del tipo di moneta di ottenere *testa* al primo e al secondo lancio, $\Pr(T_1|\cdot)$ e $\Pr(T_2|\cdot)$, rispetto alla probabilità di aver scelto quel tipo di moneta, $\Pr(\cdot)$ e $\Pr(\cdot|T_1)$. \square

Nell'approccio bayesiano l'acquisizione di osservazioni modifica la distribuzione delle probabilità *a priori* sia delle ipotesi sia dei risultati, trasformandole da probabilità *a priori* a probabilità *a posteriori*. Dopo ogni acquisizione le probabilità *a posteriori* diventano le nuove probabilità *a priori*.

Parametri di una distribuzione di probabilità come variabili casuali

L'inferenza ottenuta mediante il principio di massima verosimiglianza segue il punto di vista *frequentista*: ipotizza l'esistenza di una distribuzione di probabilità che dipende da un parametro θ e cerca di determinare il valore vero del parametro θ_0 , fissato ma ignoto, a partire dai dati disponibili. Il modello sottostante è quindi uno solo, $\theta = \theta_0$, e l'inferenza si basa sulla manipolazione delle probabilità con le quali i dati potrebbero essere stati generati al variare di θ .

Nel punto di vista *bayesiano*, invece, l'inferenza si basa sulla convinzione che anche i parametri che definiscono una distribuzione di probabilità, come i dati, sono variabili casuali: ipotizza una distribuzione di probabilità *a priori* per il parametro θ , nota **prima** di aver acquisito i dati. **Dopo** aver acquisito i dati, l'inferenza si ottiene modificando la distribuzione di probabilità *a priori* in quella *a posteriori*. Per il teorema di Bayes la distribuzione di probabilità *a posteriori* è proporzionale al prodotto della probabilità *a priori* con la verosimiglianza.

Abbiamo nuovamente a disposizione n realizzazioni delle variabili X_i con $i = 1, \dots, n$ identicamente e indipendentemente distribuite secondo una distribuzione $\mathcal{N}(\mu_0, \sigma_0^2)$ con σ_0^2 nota.

Sia μ una **variabile casuale** distribuita secondo una distribuzione normale $\mathcal{N}(\mu_{pr}, \sigma_{pr}^2)$, ovvero

$$\Pr(\mu) = \frac{1}{\sqrt{2\pi\sigma_{pr}^2}} e^{-(\mu_{pr}-\mu)^2/2\sigma_{pr}^2}$$

Questa ipotesi è basata sulle nostre conoscenze *a priori* sul problema.

Dimostriamo ora che, sotto queste condizioni, anche la distribuzione di probabilità *a posteriori* del parametro μ è normale. Per la verosimiglianza abbiamo

$$L(\mathbf{x}|\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(x_i-\mu)^2/2\sigma_0^2}$$

Da Bayes sappiamo che la probabilità *a posteriori* è proporzionale alla verosimiglianza moltiplicata per la probabilità *a priori*.

Applicando il logaritmo alla distribuzione *a posteriori*

$$\Pr(\mu|\mathbf{x}) = \frac{\Pr(\mu)L(\mathbf{x}|\mu)}{\Pr(\mathbf{x})}$$

si ottiene

$$\ln \Pr(\mu|\mathbf{x}) = -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{1}{2\sigma_{pr}^2} (\mu_{pr} - \mu)^2 + a$$

dove la costante a raggruppa tutti i termini che non dipendono da μ , ovvero

$$a = -n \ln \sqrt{2\pi} - \ln \sqrt{2\pi} - \ln \Pr(\mathbf{x})$$

Sviluppando i quadrati e ponendo nuovamente $\hat{\mu} = \sum_{i=1}^n x_i/n$, si ha

$$\begin{aligned} \ln \Pr(\mu|\mathbf{x}) &= -\frac{1}{2} \left(\frac{\sum_{i=1}^n x_i^2 + n\mu^2 - 2n\hat{\mu}\mu}{\sigma_0^2} + \frac{\mu_{pr}^2 + \mu^2 - 2\mu_{pr}\mu}{\sigma_{pr}^2} \right) + a \\ &= -\frac{1}{2} \left(\frac{(\sigma_0^2 + n\sigma_{pr}^2)\mu^2 - 2(\sigma_0^2\mu_{pr} + \sigma_{pr}^2 n\hat{\mu})\mu}{\sigma_0^2\sigma_{pr}^2} \right) + b \end{aligned} \quad (4.7)$$

dove b riassume tutti i termini che non dipendono da μ , ovvero

$$b = a - \frac{1}{2} \frac{\sum_{i=1}^n x_i^2}{\sigma_0^2} - \frac{1}{2} \frac{\mu_{pr}^2}{\sigma_{pr}^2}$$

L'equazione (4.7) può essere riscritta come

$$\ln \Pr(\mu|\mathbf{x}) = -\frac{\sigma_0^2 + n\sigma_{pr}^2}{2\sigma_0^2\sigma_{pr}^2} \left(\mu^2 - 2\frac{\sigma_0^2\mu_{pr} + \sigma_{pr}^2 n\hat{\mu}}{\sigma_0^2 + n\sigma_{pr}^2} \mu \right) + b$$

Ponendo

$$\mu_{post} = \frac{\sigma_0^2\mu_{pr} + \sigma_{pr}^2 n\hat{\mu}}{\sigma_0^2 + n\sigma_{pr}^2} \quad \text{e} \quad \sigma_{post}^2 = \frac{\sigma_0^2\sigma_{pr}^2}{\sigma_0^2 + n\sigma_{pr}^2}$$

e completando i quadrati otteniamo allora

$$\ln \Pr(\mu|\mathbf{x}) = -\frac{1}{2\sigma_{post}^2} (\mu^2 - 2\mu_{post}\mu + \mu_{post}^2) + \frac{1}{2} \frac{\mu_{post}^2}{\sigma_{post}^2} + b$$

Raggruppando nuovamente tutti i termini che non dipendono da μ in una sola costante

$$c = \frac{1}{2} \frac{\mu_{post}^2}{\sigma_{post}^2} + b$$

e ripristinando l'esponenziale, otteniamo

$$\Pr(\mu|\mathbf{x}) = C e^{-\frac{(\mu - \mu_{post})^2}{2\sigma_{post}^2}} = \mathcal{N}(\mu_{post}, \sigma_{post}^2)$$

con $C = e^c$ che non dipende da μ .

Osservazione 4.21.4. *Consistenza e correttezza*

Per $n = 0$ abbiamo $\mu_{post} = \mu_{pr}$ e $\sigma_{post}^2 = \sigma_{pr}^2$. Per $n \rightarrow \infty$, invece, abbiamo che $\mu_{post} \rightarrow \hat{\mu} \rightarrow \mu_0$ e $\sigma_{post}^2 \rightarrow \sigma_0^2/n$, coerentemente col fatto che all'aumentare del numero dei dati la distribuzione a priori perde progressivamente peso ed entra in gioco la legge dei grandi numeri. Notiamo che μ_{post} è una media pesata di μ_{pr} e $\hat{\mu}$ con le varianze rispettive scambiate e il peso di $\hat{\mu}$ che cresce al crescere di n . Anche se *asintoticamente* corretto, quindi, μ_{post} è distorto. Nel caso di pochissimi campioni, tuttavia, μ_{post} potrebbe risultare più efficace di $\hat{\mu}$, ovviamente nel caso in cui l'assunzione *a priori* sul modello fosse appropriata.

Notiamo infine che applicando due diversi metodi di inferenza, il principio di massima verosimiglianza e la stima della distribuzione di probabilità *a posteriori* via il teorema di Bayes, si ottengono risultati diversi. In generale, non possiamo concludere che uno sia meglio dell'altro, ma solo osservare che **l'inferenza dipende dal principio che decidiamo di adottare**.

Compito 4.21.1. *Confronto*

Fissa $\sigma_0^2 = 1$ e campiona μ_0 dalla distribuzione normale $\mathcal{N}(\mu_{pr}, \sigma_{pr}^2)$. Campiona n punti x_i (con $n = 2, 10$ e 20) dalla distribuzione normale $\mathcal{N}(\mu_0, 1)$ e confronta $\hat{\mu}$ e μ_{post} con μ_0 al variare di n ripetendo l'esperimento 100 volte.

4.22 Metodi Monte Carlo

Integrale definito

Riprendiamo brevemente, dapprima, un risultato fondamentale.

Legge dei grandi numeri

Siano X_m con $m = 1, 2, \dots, M$ variabili casuali indipendenti e identicamente distribuite. Indichiamo con $\langle X \rangle_M$ la media empirica di una realizzazione delle M variabili casuali

$$\langle X \rangle_M = \frac{1}{M} \sum_{m=1}^M X_m$$

Se $\mathbb{E}[X_m] = \mu$ e $\text{Var}(X_m) = \sigma^2$, per il valore atteso e la varianza di $\langle X \rangle_M$ sappiamo che

$$\mathbb{E}[\langle X \rangle_M] = \mu \text{ e } \text{Var}(\langle X \rangle_M) = \frac{\sigma^2}{M}$$

e, dalla legge dei grandi numeri, otteniamo che con alta probabilità per M grande $\langle X \rangle_M \rightarrow \mu$, ovvero

$$\forall \epsilon > 0 \quad \lim_{M \rightarrow \infty} \Pr \{ |\langle X \rangle_M - \mu| \geq \epsilon \} = 0$$

Osservazione 4.22.1. Frequenza e probabilità, finalmente!

Al crescere di M la probabilità che $\langle X \rangle_M$ sia apprezzabilmente diverso da μ tende a 0, ma non possiamo dire che $\langle f \rangle_M \rightarrow \mu$. Differenze significative per M grande possono essere rilevate anche se non frequentemente.

Area del cerchio

Dobbiamo calcolare l'area A di un cerchio di raggio R ma non ricordiamo la formula $A = \pi R^2$. Abbiamo trovato da qualche parte che

$$A = 2 \int_{-R}^R \sqrt{R^2 - x^2} dx$$

ma non sappiamo come calcolare l'integrale. Come possiamo uscirne?

Se sappiamo campionare da una distribuzione uniforme tra $[-R, R]$ possiamo stimare A valutando $\langle f \rangle_M$, ovvero la frazione di M punti (x, y) campionati nel quadrato azzurro di lato $2R$ (vedi figura 4.2, a sinistra) che cadono all'interno del cerchio giallo inscritto, ovvero per i quali $x^2 + y^2 \leq R^2$. L'area

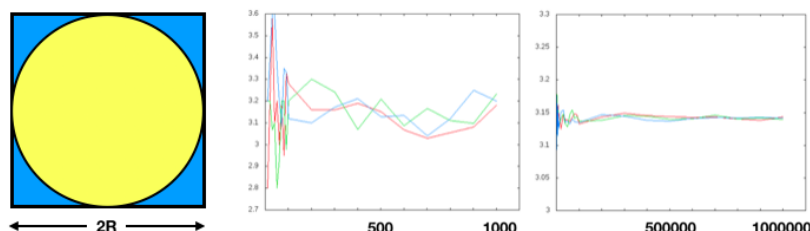


Figura 4.2: Vedi testo.

del quadrato è $4R^2$ per cui se f è il rapporto tra le aree del cerchio e del quadrato abbiamo $A = 4fR^2$.

I due grafici in figura 4.2, al centro e a destra rispettivamente, mostrano stime di $\pi = 4f$ ottenute come $4\langle f \rangle_M$, con $\langle f \rangle_M$ la stima *Monte Carlo* di f ottenuta per M da 1 a 10^3 e da 10^3 a 10^6 . Al crescere di M le oscillazioni di $\langle f \rangle_M$ diminuiscono in ampiezza: questo risultato non ci sorprende perché stiamo di fatto stimando un valore atteso, f , con una media empirica, $\langle f \rangle_M$, e la varianza sottostante decresce come $1/M$. Stimato π come $4\langle f \rangle_M$, infine, otteniamo $A = 4\langle f \rangle_M R^2$.

Caso generale

Supponiamo di dover calcolare l'integrale

$$I = \int_a^b f(x) dx$$

per una qualche f continua nell'intervallo $[a, b]$. Per il *Teorema del valor medio* sappiamo che

$$\int_a^b f(x) dx = (b - a) \cdot \bar{f}$$

con \bar{f} un qualche valore assunto dalla funzione f nell'intervallo. Per stimare \bar{f} basta campionare M valori nell'intervallo $[a, b]$ e valutare la media

$$\langle f \rangle_M = \frac{1}{M} \sum_{m=1}^M f(x_m)$$

Per la legge dei grandi numeri sappiamo che per ogni ϵ fissato

$$\Pr \{ |\langle f \rangle_M - \bar{f}| \geq \epsilon \} \rightarrow 0 \quad \text{per } M \rightarrow \infty$$

Mettendo insieme i pezzi abbiamo infine che

$$\int_a^b f(x) dx = \frac{b - a}{M} \sum_{m=1}^M f(x_m)$$

Osservazione 4.22.2. Implementazione efficiente

Per grandi valori di M , e soprattutto nel caso dovessimo ripetere il calcolo della media al variare di M , è preferibile calcolare $\langle f \rangle_M$ ricorsivamente. Poniamo $\langle f \rangle_0 = 0$ e per $m = 1, \dots, M$ campioniamo x_m uniformemente nell'intervallo $[a, b]$ e aggiorniamo il valor medio con

$$\langle f \rangle_m = \langle f \rangle_{m-1} + \frac{f(x_m) - \langle f \rangle_{m-1}}{m}$$

Campionare dove serve per ridurre la varianza

Se l'integrale di una funzione che assume valori diversi da 0 su parti piccole del dominio di integrazione, il campionamento da una distribuzione uniforme presenta due inconvenienti seri: la convergenza si otterrebbe per valori di n spropositati e la varianza del risultato sarebbe inoltre molto grande. La figura (4.3) a sinistra mostra che nel calcolo dell'integrale

$$\int_0^{10} e^{-2|x-5|} dx$$

il campionamento da una densità uniforme finirebbe con lo scegliere punti nei quali il valore della funzione integranda è molto piccolo. È immediato rendersi conto che in questo caso la convergenza rallenta e, conseguentemente, la varianza della stima aumenta.

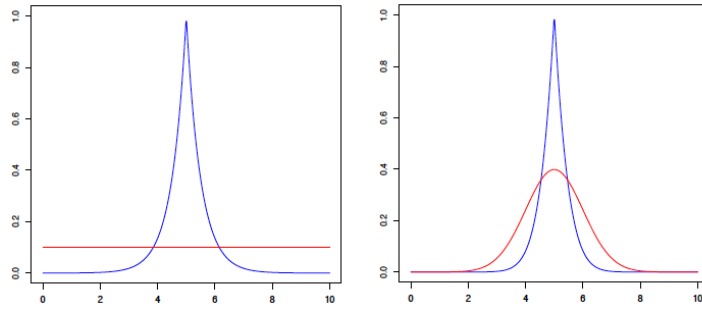


Figura 4.3: Vedi testo.

Una possibilità per ovviare a questi difetti è di fare uso di una densità f dalla quale si sappia campionare concentrata nelle regioni importanti del dominio di integrazione (regioni nelle quali la funzione integranda assume valori diversi da 0) e moltiplicare la funzione integranda per il reciproco di f . Nel caso di prima, vedi figura (4.3) a destra, si potrebbe per esempio utilizzare la densità

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-5)^2/2}$$

e riscrivere l'integrale come

$$\int_0^{10} \frac{e^{-2|x-5|}}{f(x)} f(x) dx = \int_0^{10} e^{-2|x-5|} \frac{1}{\sqrt{2\pi}} e^{-(x-5)^2/2} dx.$$

Il calcolo dell'integrale è del tutto simile al caso precedente con la differenza che i punti sono campionati da $f(x)$ e la funzione integranda è modificata dividendo la funzione originale per $f(x)$.

Risulta pertanto cruciale disporre di un metodo in grado di campionare da una distribuzione arbitraria.

Campionamento da una densità arbitraria

Supponiamo di dover campionare una variabile casuale X da una funzione di densità $f(x)$. Come spesso accade in pratica conosciamo $f(x)$ a meno di un fattore di proporzionalità. Inoltre, sappiamo campionare da una densità $g(x)$ tale che, per qualche costante M , $f(x) \leq Mg(x)$.

Algoritmo 4.22.1. *AcceptReject*

[topsep=0pt, partopsep=0pt, itemsep=2pt, parsep=2pt]

1. campiona x dalla densità g e u uniformemente nell'intervallo $(0, 1)$
 2. if $u \leq f(x)/(Mg(x))$
 - return x
 - else
 - goto 1.
-

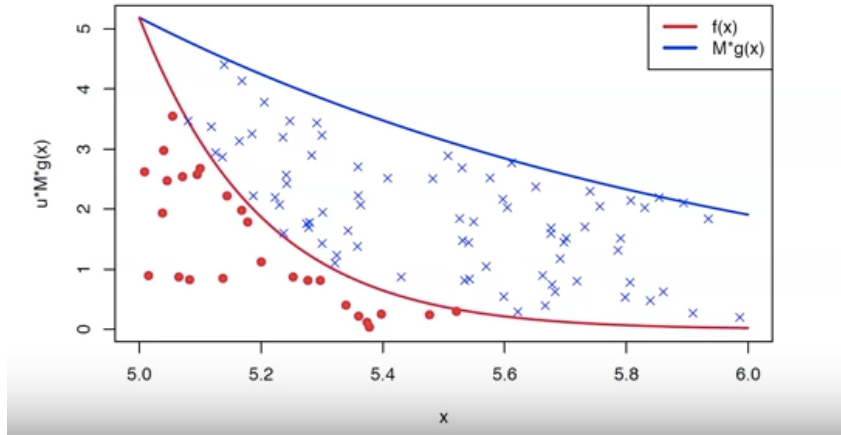


Figura 4.4: I campioni estratti con l'algoritmo *AcceptReject* sono accettati solo se cadono sotto il grafico della densità $f(x)$. Quelli che cadono sopra, ovviamente sotto il grafico della funzione $Mg(x)$, sono invece rifiutati.

Osservazione 4.22.3. Correttezza

Verifichiamo che la funzione di densità da cui campioniamo con l'algoritmo *AcceptReject* è proprio $f(x)$. Dobbiamo dimostrare che

$$\Pr\{X \leq a\} = \int_{-\infty}^a f(x)dx = F(a)$$

dove $F(a)$ è la funzione cumulativa di probabilità. Supponiamo che le iterazioni necessarie a ottenere un campionamento accettato siano state N . Abbiamo

$$\Pr\{X \leq a\} = \Pr\left\{X \leq a \mid U \leq \frac{f(X)}{Mg(X)}\right\} = \frac{\Pr\left\{X \leq a, U \leq \frac{f(X)}{Mg(X)}\right\}}{\Pr\left\{U \leq \frac{f(X)}{Mg(X)}\right\}}.$$

Per quanto riguarda il numeratore l'indipendenza dei due eventi consente di scrivere la funzione di densità congiunta come

$$g(x)\mathbf{1}_{[0,1]}(u)$$

per cui abbiamo

$$\begin{aligned} \Pr\left\{X \leq a, U \leq \frac{f(X)}{Mg(X)}\right\} &= \int_{-\infty}^a \left(\int_0^{f(x)/(Mg(x))} du \right) g(x)dx \\ &= \int_{-\infty}^a \frac{f(x)}{Mg(x)} g(x)dx = \frac{1}{M} \int_{-\infty}^a f(x)dx = \frac{1}{M} F(a). \end{aligned}$$

Pertanto, poiché $F(a) \rightarrow 1$ per $a \rightarrow +\infty$, abbiamo

$$\Pr\left\{U \leq \frac{f(X)}{Mg(X)}\right\} = \Pr\left\{X \leq +\infty, U \leq \frac{f(X)}{Mg(X)}\right\} = \frac{1}{M}$$

e infine

$$\Pr\{X \leq a\} = F(a).$$

■

Osservazione 4.22.4. *Importanza della costante*

Il numero di iterazioni N segue una distribuzione geometrica con valore atteso M . Conseguentemente, se la costante M è troppo grande, l'algoritmo *AcceptReject* tende a scartare la maggior parte dei campionamenti e perde in efficienza.

4.23 Catene di Markov

Introduciamo le catene di Markov discrete nel tempo e con spazio degli stati finito. Ci interessa soprattutto il comportamento delle catene di Markov per tempi grandi. Per semplicità ci restringiamo alle catene omogenee.

Processi senza memoria

Concetti fondamentali

Data una sequenza di variabili casuali discrete S_t con $t = 0, 1, \dots$, un *processo di Markov* (MP) è costituito da una coppia (S, \mathbf{P}) dove

$S = \{s_1, s_2, \dots, s_N\}$ è lo *spazio degli stati*, insieme dei possibili N valori che possono essere assunti dalle variabili casuali S_t

$\mathbf{P} \in \mathbb{R}^{N \times N}$ è la *matrice di transizione* (TM) che esprime la probabilità di transizione allo stato s_j **al tempo** $t + 1$ condizionata al fatto che il processo si trova **al tempo** t nello stato s_i con $i, j = 1, \dots, N$. Pertanto per l'elemento di riga i e colonna j scriviamo

$$(\mathbf{P})_{ij} = P_{ij} = \Pr(S_{t+1} = s_j | S_t = s_i) \quad (4.8)$$

con $0 \leq P_{ij} \leq 1$ per ogni i e j e $\sum_j P_{ij} = 1$ per ogni riga i .

□

Restringiamoci al caso in cui un MP è **omogeneo nel tempo**. In un MP omogeneo gli elementi P_{ij} descrivono probabilità che **non dipendono dal tempo in cui avviene la transizione**. Possiamo quindi riscrivere la (4.8) semplicemente come

$$P_{ij} = \Pr(s_j | s_i)$$

Definizione 4.23.1. Catena di Markov (MC)

Sia \mathbf{P} una TM. Partendo da uno stato iniziale $\bar{s} \in S$ scelto casualmente al tempo $t = 0$, una **catena di Markov** (MC) è una realizzazione del MP sottostante, ovvero una sequenza di stati per la quale se lo stato al tempo t è s_i , lo stato della catena al tempo $t + 1$ è campionato dalla distribuzione di probabilità data dalla i -esima riga di \mathbf{P} .

Osservazione 4.23.1. Una catena senza memoria

Omogenea o no nel tempo, una MC è senza memoria: P_{ij} , la probabilità di transizione allo stato s_j da s_i , non dipende dagli stati eventualmente occupati dalla MC nei tempi precedenti.

Esempio 4.23.1. Sole o pioggia?

Assumiamo che lo stato del tempo in una città possa essere descritto come *sole*, s_1 , o *pioggia*, s_2 , e che la TM sia

$$\mathbf{P} = \frac{1}{5} \begin{pmatrix} 3 & 2 \\ 1 & 4 \end{pmatrix}$$

Osservato che la matrice non dipende da t (fatto strano in meteorologia), qual è la frazione di giorni di sole che possiamo aspettarci per t grande?

Esempio 4.23.2. Gioco d'azzardo

A ogni mano di un gioco G Alice può trovarsi in uno di cinque stati s_1, \dots, s_5 . Nello stato s_i Alice ha un patrimonio di $(i - 1)\$$. Dallo stato s_i con $i = 2, 3$ e 4 Alice *o* passa allo stato s_{i+1} con probabilità p e vince $1\$$, *o* passa allo stato s_{i-1} e perde $1\$$. Una volta che Alice si trova in s_1 o s_5 , invece, non succede

più nulla: nello stato s_1 Alice è rovinata (non ha più dollari), nello stato s_5 ha vinto definitivamente e ha un patrimonio di 4\$. La TM \mathbf{P} è data da

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1-p & 0 & p & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Che cosa possiamo dire sul patrimonio di Alice per t grande?

Esempio 4.23.3. *Visita a caso in un grafo diretto e pesato*

In un cammino casuale in un grafo diretto pesato il vertice da visitare tra i vertici raggiunti dagli archi uscenti da un vertice origine è campionato con probabilità proporzionale al peso degli archi uscenti (come nel caso del grafo e della TM di figura 4.5). *Google's Page Rank* è basato su un modello simile: internet è il grafo, i nodi le pagine e un arco connette A a B se A contiene un collegamento a B . Se il navigatore segue collegamenti a caso quale frazione di tempo trascorre su ogni pagina per t grande?

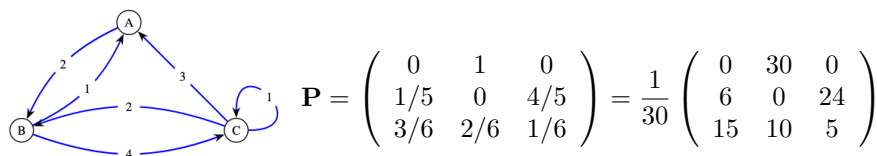


Figura 4.5: Vedi testo.

Esempio 4.23.4. *Saltatore ostinato*

Consideriamo infine un esempio utile per capire una proprietà importante di alcune MC: se la catena è nello stato s_1 al tempo t sarà sempre nello stato s_2 al tempo $t + 1$ e viceversa, o

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Evoluzione di una catena di Markov

La distribuzione di probabilità dell'evoluzione nel tempo di una MC che si trova in uno stato iniziale s_i è calcolabile a partire dalla TM \mathbf{P} .

Due passi Per calcolare la probabilità che l'elemento $t + 2$ di una MC sia s_j se l'elemento t è s_i dobbiamo sommare su tutti i possibili stati s_k in cui la MC può trovarsi al tempo $t + 1$, ovvero

$$\begin{aligned} \Pr(s_j | s_i, \text{in due passi}) &= \sum_k \Pr(s_j | s_k, s_i) \Pr(s_k | s_i) \\ &= \sum_k \Pr(s_j | s_k) \Pr(s_k | s_i) = \sum_k P_{kj} P_{ik} = (\mathbf{P}^2)_{ij} \end{aligned}$$

dove $\Pr(s_j | s_k, s_i) = \Pr(s_j | s_k)$ perché si tratta di una MC.

Caso generale Applicando ripetutamente i passaggi del punto precedente per t arbitrario, otteniamo l'equazione di Chapman-Kolmogorov, ovvero

$$\Pr(s_j | s_i, \text{in } t \text{ passi}) = (\mathbf{P}^t)_{ij}$$

Esercizio 4.23.1. *Sempre una matrice di transizione*

Dimostra che \mathbf{P}^t è una TM per tutti i $t \in \mathbb{N}$.

Dimostrazione: se $\mathbf{1} = (1 \ 1 \dots 1)^\top$, abbiamo $\mathbf{P}^t \mathbf{1} = \mathbf{P}^{t-1} \mathbf{P} \mathbf{1} = \mathbf{P}^{t-1} \mathbf{1} = \dots = \mathbf{P} \mathbf{1} = \mathbf{1}$. \square

Introduciamo ora due concetti utili per descrivere l'importante proprietà di raggiungibilità.

Definizione 4.23.2. *Irriducibilità*

Una TM \mathbf{P} è *irriducibile* se per ogni s_i ed s_j con $i, j = 1, \dots, N$ esiste $t(i, j)$ tale che

$$(\mathbf{P}^{t(i,j)})_{ij} > 0$$

Definizione 4.23.3. *Regolarità*

Una TM \mathbf{P} è *regolare* se per ogni s_i ed s_j con $i, j = 1, \dots, N$ esiste \bar{t} tale che per $\forall t > \bar{t}$ si ha

$$(\mathbf{P}^t)_{ij} > 0$$

\square

Intuitivamente, in una MC irriducibile per ogni coppia s_i ed s_j , se la MC si trova in s_i , prima o poi si troverà in s_j . Il tempo minimo di attesa potrebbe essere più lungo per alcune coppie ma sempre finito. Se la MC è regolare, il tempo minimo di attesa è lo stesso per tutte le coppie. Chiaramente una MC regolare è irriducibile, poiché è sufficiente porre $\tau(i, j) = \bar{\tau}$ per tutti le coppie i e j .

Esercizio 4.23.2. *Una MC regolare è irriducibile ma non viceversa*

Determina se le MC degli esempi 4.23.1, 4.23.2, 4.23.3 e 4.23.4 sono regolari o irriducibili.

Soluzione

4.23.1 : banalmente regolare, e quindi irriducibile, poiché $\bar{\tau} = \tau(i, j) = 1$

4.23.2 : non irriducibile poiché gli ultimi quattro elementi della prima riga di \mathbf{P}^t sono sempre nulli

4.23.3 : regolare con $\bar{\tau} = 3$ poiché

$$\mathbf{P}^2 = \frac{1}{180} \begin{pmatrix} 36 & 0 & 144 \\ 72 & 84 & 24 \\ 27 & 100 & 53 \end{pmatrix} \quad \text{e} \quad \mathbf{P}^3 = \frac{1}{5400} \begin{pmatrix} 2160 & 2520 & 720 \\ 864 & 2400 & 2136 \\ 1395 & 1340 & 2665 \end{pmatrix}$$

4.23.4 : irriducibile poiché

$$\tau(i, j) = \begin{cases} 2 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases}$$

ma non regolare poiché per le potenze dispari e pari di \mathbf{P} si ha

$$\mathbf{P}^{2t-1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{e} \quad \mathbf{P}^{2t} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Distribuzioni limite e stazionarie**Esercizio 4.23.3.** *Da una distribuzione di probabilità a un'altra*

Un vettore \mathbf{p} in N dimensioni con componenti non negativi che sommano a 1 è una distribuzione di probabilità per gli stati s_1, \dots, s_N ed è noto come vettore *stocastico*. Dimostra che se \mathbf{p} è un vettore stocastico e \mathbf{P} una TM il vettore \mathbf{q} ottenuto come

$$\mathbf{q}^\top = \mathbf{p}^\top \mathbf{P}$$

è stocastico.

Soluzione

Per i vincoli di non negatività degli elementi di \mathbf{p} , per $j = 1, \dots, N$, abbiamo $q_j = \sum_i p_i P_{ij} \geq 0$. Inoltre, poiché $\sum_j p_j = 1$ e $\sum_j P_{ij} = 1$ per $i = 1, \dots, N$, otteniamo

$$\sum_j q_j = \sum_j \sum_i p_i P_{ij} = \sum_i p_i \left(\sum_j P_{ij} \right) = \sum_i p_i = 1$$

Definizione 4.23.4. *Distribuzione stazionaria*

Una distribuzione di probabilità π con $\pi_i \geq 0$ per ogni s_i con $i = 1, \dots, N$ e $\sum_i \pi_i = 1$ è *stazionaria* per una MC con TM \mathbf{P} se

$$\pi^\top = \pi^\top \mathbf{P}, \quad \text{ovvero} \quad \pi_j = \sum_i \pi_i P_{ij} \quad j = 1, \dots, N$$

Osservazione 4.23.2. *Autovalore massimale per una TM \mathbf{P}*

Poiché $\mathbf{P}\mathbf{1} = \mathbf{1}$, 1 è sempre un autovalore di \mathbf{P} . È facile dimostrare che è anche il più grande. Se $\lambda > 1$ fosse un autovalore e \mathbf{x} il corrispondente autovettore, infatti, avremmo $\mathbf{P}^n \mathbf{x} = \lambda^n \mathbf{x}$. Ma per valori crescenti di n questo porterebbe a una TM \mathbf{P}^n con elementi maggiori di 1! \square

Enunciamo un importante teorema basato sul teorema del punto fisso di Brouwer.

Teorema 4.23.1. *Perron Frobenius*

Se una TM \mathbf{P} è regolare, l'autovettore sinistro π corrispondente all'autovalore 1 è unico e con tutti gli elementi strettamente positivi.

Osservazione 4.23.3. *Facile da trovare*

Una distribuzione stazionaria π può quindi essere facilmente determinata come soluzione del sistema lineare

$$\pi^\top \mathbf{P} = \pi^\top$$

Esercizio 4.23.4. *Ancora i nostri quattro esempi*

Trova le distribuzioni stazionarie per gli esempi 4.23.1, 4.23.2, 4.23.3 e 4.23.4.

Soluzione

$$4.23.1 : \quad \pi_{sr} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \end{pmatrix}^\top$$

$$4.23.2 : \quad \pi_l = (1 \ 0 \ 0 \ 0 \ 0)^\top \text{ e } \pi_w = (0 \ 0 \ 0 \ 0 \ 1)^\top$$

$$4.23.3 : \quad \pi_{rw} = \begin{pmatrix} \frac{17}{66} & \frac{25}{66} & \frac{24}{66} \end{pmatrix}^\top$$

$$4.23.4 : \quad \pi_{sj} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}^\top$$

Definizione 4.23.5. *Distribuzione limite*

Un vettore λ con $\lambda_i \geq 0$ per ogni i e $\sum_i \lambda_i = 1$ è una *distribuzione limite* per \mathbf{P} se per i con $i = 1, \dots, N$

$$\lim_{t \rightarrow \infty} (\mathbf{P}^t)_{ij} = \lambda_j \quad \text{per ogni stato } s_j \text{ con } j = 1, \dots, N$$

Se una MC ammette una distribuzione limite ed è inizializzata da λ , a ogni passo l'evoluzione della MC sarà sempre determinata da λ . \square

Quando una distribuzione limite esiste, \mathbf{P}^t converge alla matrice con tutte le righe uguali a λ , ovvero

$$\mathbf{P}^t \xrightarrow{t \rightarrow \infty} \begin{pmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_N \\ \lambda_1 & \lambda_2 & \dots & \lambda_N \\ \lambda_1 & \lambda_2 & \dots & \lambda_N \end{pmatrix}$$

In questo caso, la probabilità di transizione allo stato s_j for $j = 1, \dots, N$ al tempo $t + 1$, per t sufficientemente grande, non dipende dallo stato della MC al tempo $t = 0$.

Per una MC regolare l'esistenza della distribuzione limite è assicurata da un teorema - basato sulla legge dei grandi numeri - che enunciamo ma non dimostriamo.

Teorema 4.23.2. *Senza via di fuga*

Una MC regolare ha un'unica distribuzione limite λ with $\lambda_j > 0$ for $j = 1, \dots, N$. Inoltre, poiché per s_j con $j = 1, \dots, N$

$$\sum_i \lambda_i P_{ij} = \sum_i \lim_{t \rightarrow \infty} (\mathbf{P}^t)_{ki} P_{ij} = \lim_{t \rightarrow \infty} \sum_i (\mathbf{P}^t)_{ki} P_{ij} = \lim_{t \rightarrow \infty} (\mathbf{P}^{t+1})_{kj} = \lambda_j$$

la distribuzione limite è l'unica distribuzione stazionaria per \mathbf{P} .

Esercizio 4.23.5. *Quale esempio ammette distribuzione limite?*

Discuti l'esistenza della distribuzione limite per i nostri quattro esempi.

Soluzione

4.23.1 : la MC è regolare e π_{sr} , quindi, è anche la distribuzione limite. Per t grandi, e indipendentemente dallo stato iniziale, possiamo attenderci 1/3 di giorni di sole e 2/3 di pioggia.

4.23.2 : non può avere una distribuzione limite poiché π_l e π_w , che corrispondono a due stati *assorbenti*, sono stazionari. Per t grandi, la MC si troverà in uno dei due stati assorbenti: Alice, quindi, o sarà in rovina o avrà sbancato. Se il banco fosse infinitamente ricco, Alice finirebbe sempre in rovina e otteniamo un caso particolare del *teorema della rovina del giocatore*.

4.23.3 : la MC è regolare e π_{rw} è anche la distribuzione limite che fornisce una misura del tempo atteso trascorso su ogni pagina.

4.23.4 : nessuna distribuzione limite. La MC oscilla indefinitamente tra i due stati. \square

Consideriamo infine un'ulteriore proprietà di cui può godere una MC.

Definizione 4.23.6. *Bilancio dettagliato*

Una MC con TM \mathbf{P} è reversibile se esiste una distribuzione di probabilità π tale che

$$\pi_j P_{ji} = \pi_i P_{ij} \quad \forall i, j \quad (4.9)$$

Osservazione 4.23.4. *Stazionarietà garantita*

La distribuzione π dell'equazione (4.9) è stazionaria per \mathbf{P} . Infatti sommando su j entrambi i membri e tenendo conto che $\sum_j P_{ij} = 1$, si ottiene

$$\sum_j \pi_j P_{ji} = \sum_j \pi_i P_{ij} = \pi_i \sum_j P_{ij} = \pi_i$$

4.24 Catene di Markov *Monte Carlo*

Le catene di Markov *Monte Carlo* sono metodi utilizzati per simulare l'estrazione di una sequenza di campioni da una distribuzione di probabilità arbitraria. La sequenza ottenuta costituisce uno strumento computazionale indispensabile sia per approssimare una distribuzione di probabilità sia per stimare valori attesi in alte dimensioni.

Metropolis-Hastings

La reversibilità espressa dall'equazione (4.9), ovvero il fatto che la probabilità di passare allo stato i dallo stato j è la stessa di passare dallo stato j allo stato i , è alla base dell'algoritmo di *Metropolis-Hastings*.

Algoritmo 4.24.1. *Metropolis-Hastings*

Input: π distribuzione di probabilità discreta per N stati, \mathbf{H} matrice di transizione regolare di dimensione $N \times N$ utilizzabile per proporre un nuovo stato, \bar{s} stato iniziale della catena arbitrario e $T \in \mathbb{N}$ grande
Output: Catena di Markov X_t con matrice di transizione \mathbf{P} e distribuzione limite π

for $t = 0$ to T

1. proponi la transizione allo stato s_j campionato con probabilità $\Pr(X_{t+1} = s_j | X_t = s_i) = H_{ij}$

2. calcola

$$a_{ij} = \min \left\{ 1, \frac{\pi_j H_{ji}}{\pi_i H_{ij}} \right\} \quad (4.10)$$

3. campiona un numero U in $[0, 1]$ con probabilità uniforme

4. if $U \leq a_{ij}$

 then $X_{t+1} = s_j$

 else $X_{t+1} = s_i$

La matrice \mathbf{P}

$$P_{ij} = \begin{cases} H_{ij} a_{ij} & (i \neq j) \\ 1 - \sum_{i \neq j} H_{ij} a_{ij} & (i = j) \end{cases} \quad (4.11)$$

è la matrice di transizione della catena X_t e ha per distribuzione limite π .

Correttezza Poiché \mathbf{H} è regolare anche la matrice \mathbf{P} è regolare. Sostituendo inoltre l'equazione (4.10) nell'equazione (4.11), se $\pi_j H_{ji} < \pi_i H_{ij}$ otteniamo

$$P_{ij} = H_{ij} \min \left\{ 1, \frac{\pi_j H_{ji}}{\pi_i H_{ij}} \right\} = \frac{\pi_j H_{ji}}{\pi_i} \quad \text{e} \quad P_{ji} = H_{ji} \min \left\{ 1, \frac{\pi_i H_{ij}}{\pi_j H_{ji}} \right\} = H_{ji}$$

da cui ricaviamo

$$\pi_i P_{ij} = \pi_i \frac{\pi_j}{\pi_i} H_{ji} = \pi_j H_{ji} = \pi_j P_{ji}$$

È facile verificare che per $\pi_j H_{ji} > \pi_i H_{ij}$ si ottiene nuovamente l'equazione (4.9). In entrambi i casi, quindi, la matrice regolare \mathbf{P} soddisfa l'equazione del bilancio dettagliato e la distribuzione di probabilità π è la distribuzione limite e stazionaria di \mathbf{P} .

Osservazione 4.24.1. Matrice H e probabilità di accettazione della proposta

La scelta della matrice H deve ricadere su una distribuzione di probabilità dalla quale sia possibile estrarre campioni a caso. Se H è simmetrica, a_{ij} - nota come *probabilità di accettazione della proposta del nuovo stato* - dipende solo dal rapporto π_j/π_i .

Osservazione 4.24.2. Conoscenza parziale della distribuzione π

La dipendenza di a_{ij} dal rapporto π_j/π_i rende possibile utilizzare l'algoritmo di *Metropolis-Hastings* anche nel caso in cui la distribuzione di probabilità π sia nota a meno di una costante di proporzionalità. Questa proprietà ha importanti ripercussioni nell'ambito della statistica bayesiana.

Osservazione 4.24.3. Convergenza

Per raggiungere il regime asintotico è necessario scartare i primi elementi della catena, operazione nota come *burn-in*, in modo tale da ottenere una sequenza che non dipende dalla scelta dello stato iniziale.

Osservazione 4.24.4. Dipendenza

Un limite particolarmente pronunciato dell'algoritmo di Metropolis-Hastings, e che affligge le catene di Markov *Monte Carlo* ottenute anche con altri algoritmi, è quello di produrre sequenze di campioni correlati. Al fine di ottenere campioni indipendenti spesso si sottocampiona la sequenza.

Modello di Ising in 2D

Consideriamo un esempio dal mondo fisico. Un modello di Ising in 2D è un sistema di A atomi disposti su un reticolo quadrato regolare che useremo come semplice modello di un materiale ferromagnetico.

Descrizione del modello

Ogni atomo a con $a = 1, \dots, A$ può trovarsi con il proprio spin σ_a *up* ($\sigma_a = 1$) o *down* ($\sigma_a = -1$). Per ogni configurazione degli A spin $\sigma = (\sigma_1, \dots, \sigma_A)$ o *stato* del sistema, l'energia è

$$E(\sigma) = E(\sigma_1, \dots, \sigma_A) = -\frac{1}{2} \sum_{a=1}^A \sum_{t \in I(a)} \sigma_a \sigma_t \quad (4.12)$$

dove $I(a)$ è l'insieme dei 4 atomi primi vicini di a sul reticolo. Poiché se $t \in I(a)$ allora $a \in I(t)$, il fattore $1/2$ evita che il contributo di una coppia di spin vicini sia contata due volte. Fissata la temperatura, **all'equilibrio il sistema si troverà in uno stato a energia minima**. Dalla fisica sappiamo che se $\beta = 1/T$ è un parametro inversamente proporzionale alla temperatura ed \mathcal{S} lo spazio dei possibili 2^A stati σ , la distribuzione stazionaria di probabilità degli stati è

$$\pi(\sigma) = \frac{e^{-\beta E(\sigma)}}{Z} \quad \text{con } Z = \sum_{\sigma \in \mathcal{S}} e^{-\beta E(\sigma)} \quad (4.13)$$

A temperatura fissata, pertanto, stati con più spin adiacenti allineati hanno energia minore e sono più probabili. Al crescere della temperatura le differenze in energia si assottigliano e spin adiacenti possono più facilmente assumere valori opposti.

La magnetizzazione del reticolo,

$$M(\sigma) = \frac{1}{A} \sum_{a=1}^A \sigma_a$$

misura la differenza tra la frazione di spin *up* e spin *down*. Dall'esperienza sappiamo che a basse temperature un materiale ferromagnetico può mostrare una magnetizzazione permanente importante. A temperature più alte, tuttavia, lo stesso materiale perde questa proprietà: una calamita dimenticata

vicino a un calorifero si smagnetizza. La fisica ci dice che questa transizione avviene in modo brusco a una temperatura ben precisa, nota come *temperatura critica* T^* . Nel caso di reticolo bidimensionale per $A \rightarrow \infty$ è possibile ottenere che la transizione avviene per un valore della temperatura

$$T^* = \frac{2}{\ln(1 + \sqrt{2})} \approx 2.26$$

Per verificare in che misura il modello di Ising è in accordo con l'esperienza vogliamo stimare

$$\mathbb{E}[|M|] = \frac{1}{Z} \sum_{\sigma \in S} e^{-\beta E(\sigma)} |M(\sigma)| = \frac{1}{AZ} \sum_{\sigma \in S} e^{-\beta E(\sigma)} \left| \sum_{a=1}^A \sigma_a \right|$$

al variare del parametro β **all'equilibrio**.

Osservazione 4.24.5. *Dimensioni spropositate*

Anche limitandoci a un sistema costituito da soli $32 \times 32 = 1024$ atomi lo spazio degli stati conta $N = 2^{1024}$ elementi. La matrice di transizione $\mathbf{P} \in \mathbb{R}^{2^{1024}} \times \mathbb{R}^{2^{1024}}$ e la distribuzione $\pi \in \mathbb{R}^{2^{1024}}$ appartengono entrambe a spazi dimensionalmente e computazionalmente intrattabili.

Stima del valore atteso della magnetizzazione

Supponiamo che il sistema si trovi in un stato σ e riscriviamo la funzione energia dell'equazione (4.12) isolando i termini che dipendono dallo spin di un particolare atomo s . Avremo

$$E(\sigma_1, \dots, \sigma_s, \dots, \sigma_A) = -\sigma_s \sum_{t \in I(s)} \sigma_t - \frac{1}{2} \sum_{a \neq s} \sum_{t \in I(a) \setminus s} \sigma_a \sigma_t \quad (4.14)$$

dove nella prima somma il fattore $1/2$ scompare per via del fatto che ogni coppia di primi vicini di s compare due volte. Per la differenza in energia ΔE tra due stati che differiscono tra loro solo per la variazione dello spin di s da σ_s a $-\sigma_s$ avremo, utilizzando l'equazione (4.14) due volte,

$$\begin{aligned} \Delta E &= E(\sigma_1, \dots, -\sigma_s, \dots, \sigma_A) - E(\sigma_1, \dots, \sigma_s, \dots, \sigma_A) \\ &= -(-\sigma_s) \sum_{t \in I(s)} \sigma_t - \frac{1}{2} \sum_{a \neq s} \sum_{t \in I(a) \setminus s} \sigma_a \sigma_t + \sigma_s \sum_{t \in I(s)} \sigma_t + \frac{1}{2} \sum_{a \neq s} \sum_{t \in I(a) \setminus s} \sigma_a \sigma_t \\ &= 2\sigma_s \sum_{t \in I(s)} \sigma_t \end{aligned}$$

Per l'additività della funzione energia, quindi, ΔE dipende solo dall'orientamento dello spin dell'atomo s e da quello dei suoi primi vicini. Per la forma esponenziale della distribuzione di probabilità π , inoltre, anche il rapporto delle probabilità dipende solo da ΔE ovvero

$$\frac{\pi(\sigma_1, \dots, -\sigma_s, \dots, \sigma_A)}{\pi(\sigma_1, \dots, \sigma_s, \dots, \sigma_A)} = e^{-\beta \Delta E} \quad (4.15)$$

Abbiamo ora tutti gli elementi per stimare $\mathbb{E}[|M|]$ in funzione di β . La distribuzione di probabilità π per gli $N = 2^A$ stati è data dall'equazione (4.13).

Osservazione 4.24.6. *Matrice regolare per la proposta di un nuovo stato*

Per ogni stato i , \mathbf{H} propone la transizione a uno degli A stati che differiscono da i per la variazione di uno solo degli A spin con probabilità uniforme e uguale a $1/A$.

h

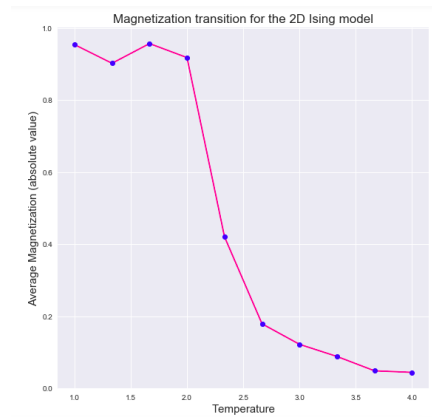


Figura 4.6: Stima di $\mathbb{E}[|M|]$ ottenuta su un reticolo 32×32 con *burn-in* di 10^7 e media empirica calcolata su 10 catene di 10^5 passi. Nello stato iniziale ogni spin è *up* o *down* con probabilità 1/2.

Osservazione 4.24.7. Regolarità di \mathbf{H}

La drastica riduzione del numero degli stati verso i quali si può muovere la catena, da 2^A ad A , è sufficiente a rendere la simulazione computazionalmente possibile. Pur consentendo transizioni verso una frazione molto piccola degli stati possibili la regolarità di \mathbf{H} non è in discussione: modificando uno spin alla volta, infatti, è evidente che è possibile raggiungere qualunque stato j indipendentemente dallo stato iniziale i con un numero di passi uguale al numero di spin il cui orientamento nello stato j è diverso rispetto allo stato i . Il calcolo di a_{ij} nell'equazione (4.10), infine, è basato sull'equazione (4.15).

Esercizio 4.24.1. Verifica dell'esistenza di un punto critico

Prova a verificare la capacità del modello di Ising di catturare la brusca variazione nel valore atteso della magnetizzazione al variare di β . Dovresti ottenere un grafico simile a quello di figura 4.6.