UNIVERSITÀ DEGLI STUDI DI TRENTO

Dipartimento Matematica

Corso di Laurea in
Matematica

Elaborato finale

Titolo

Sottotitolo (alcune volte lungo - opzionale)

Supervisore
Sonia Mazzucchi

Laureando
Claudio Meggio

Anno accademico 2017/2018

« La mia più grande preoccupazione era come chiamarla. Pensavo di chiamarla informazione, ma la parola era fin troppo usata, così decisi di chiamarla incertezza. Quando discussi della cosa con John Von Neumann, lui ebbe un'idea migliore. Mi disse che avrei dovuto chiamarla entropia, per due motivi: "Innanzitutto, la tua funzione d'incertezza è già nota nella meccanica statistica con quel nome. In secondo luogo, e più significativamente, nessuno sa cosa sia con certezza l'entropia, così in una discussione sarai sempre in vantaggio » (Claude Shannon)

Ringraziamenti

...thanks to...

Indice

Introduzione

Questa tesi si pone l'obiettivo di esplorare i concetti di Entropia ed Informazione, così come presentati nel lavoro di Claude Shannon [4] concentrandosi maggiormente su variabili casuali discrete.
Il campo dove questa teoria trova maggior applicazione è sicuramente la teoria dei codici cui verrà prestata una particolare attenzione. Ci si soffermerà soprattutto sui codici binari, i quali, attraverso i computer si sono ormai inseriti in modo indelebile nella nostra quotidianità. La portata di questi teoremi non si deve però pensare limitata a codici informatici, ma ad ogni tipo di codice che si possa definire tale, primo fra tutti il DNA il quale, grazie alla sua composizione di basi azotate, permette la codifica di amminoacidi fondamentali per la nostra esistenza. Queste applicazioni dimostrano ancora una volta che la Matematica, nonostante venga ritenuta una scienza astratta, è in grado di rapportarsi con la nostra realtà e ci indirizza nella sua comprensione.
Verrà riservato un accenno anche al *caso continuo* il quale purtroppo non ammette uno studio così elegante e proficuo a causa della mancanza di controllo che si può esercitare sulla funzione di densità di una variabile assolutamente continua.
Il "teorema fondamentale di Shannon" ci garantirà infine l'esistenza di codici che, utilizzati all'interno di canali comunicativi generici, forniscono una probabilità d'errore arbitrariamente piccola, permettendoci quindi di rivolgere le nostre attenzioni allo studio di nuovi codici piuttosto che alla costruzione di canali comunicativi.

1 Informazione ed Entropia per variabili casuali discrete

1.1 Informazione

Fondamentali in questa tesi saranno i concetti di Informazione ed entropia. Bisogna anzitutto specificare che in Probabilità il significato di Informazione ha un connotato diverso da quello della lingua parlata. Consideriamo ad esempio le seguenti frasi:

i. Quando vado in palestra mi alleno

ii. Il vincitore delle prossime elezioni sarà Claudio Baglioni

iii. QUER W LKS E W

Istintivamente diremo che la frase contente maggior informazione è (*ii*) in quanto contiene un'informazione totalmente inaspettata e nuova, seguita poi da (*i*) ed in fine (*iii*) la quale non avendo significato non conterrà nessuna informazione.
Questa scala però tiene conto sia del significato della frase sia della quantità di *sorpresa* che porta. In questo senso (*iii*) non ha significato, ma porta *sorpresa*, mentre (*ii*) contiene sia significato che sorpresa.
Nel mondo della matematica si è visto che il concetto di *significato* è difficile da esprimere e si è dunque preferito puntare sul concetto di *sorpresa* per esprimere il significato d'*informazione*.
Per definire in maniera rigorosa il concetto di **informazione** poniamoci in uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$.
Dati due eventi $E_1, E_2 \in \mathcal{F}$ vogliamo che la nostra funzione d'informazione I soddisfi alcuni criteri:

1. $I(E) \geq 0$ per ogni $E \in \mathcal{F}$
2. se $\mathbb{P}(E_1) \leq \mathbb{P}(E_2)$ allora $I(E_1) \geq I(E_2)$
3. se E_1, E_2 sono indipendenti allora $I(E_1 \cup E_2) = I(E_1) + I(E_2)$

Per soddisfare queste richieste viene naturalmente in mente la funzione log, infatti:

Definizione 1.1.1. In uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$definiamo la funzione **informazione** $I : \mathcal{F} \to \mathbb{R}^+$ come:

$$I(E) = -\log_a(\mathbb{P}(E)). \tag{1.1.1}$$

dove a è una costante positiva (in alcuni testi la funzione viene moltiplicata per K, ma tale costante è inutile dato che già scegliere la base coincide col moltiplicare per una costante, infatti: $\log_a(x) = \frac{\log_b(y)}{\log_b(a)}\Big)$.

Si verifica facilmente che la funzione I così definita rispetta le proprietà preposte. L'unico intoppo nasce per un evento E tale che $\mathbb{P}(E) = 0$ in questo caso $I(E) = \infty$, questa occorrenza può essere interpretata come l'incapacità di ottenere informazioni da un evento impossibile. La funzione *Informazione* possiede inoltre la proprietà di essere nulla qualora la probabilità di un evento sia 1 cioè se un evento è certo, la sua realizzazione non ci fornirà alcuna informazione.
Essendo questa funzione spesso associata a codici è comodo scegliere 2 come base del logaritmo, in questo modo supponendo di avere una variabile casuale X con distribuzione di Bernoulli a parametro $p = \frac{1}{2}$ (il nostro messaggio sarà definito da un codice binario ($\{0, 1\}$) abbiamo che

$$I(X = 0) = I(X = 1) = -\log_2\left(\frac{1}{2}\right) = 1 \tag{1.1.2}$$

Per questo d'ora in avanti, salvo diversa indicazione, con log si intenderà $\log_2$.

1.2 Entropia

Il secondo concetto fondamentale trattato in questa tesi è quello di *entropia*.
Data una variabile casuale discreta X a valori $\{x_1...x_n\}$ e con legge di probabilità $\{p_1...p_n\}$ ($p_i := \mathbb{P}(X = x_i)$) non possiamo conoscere a priori il valore che assumerà X e di conseguenza non possiamo sapere quanta informazione verrà inviata. Definiamo per questo l'*entropia*.

Definizione 1.2.1. Si dice **entropia** di una variabile casuale discreta X il valore

$$H(X) := \mathbb{E}(I(X)) = -\sum_{j=1}^{n} p_j \Phi(p_j) \tag{1.2.1}$$

dove

$$\Phi(p) := \begin{cases} \log_2(p) \ se \ p \neq 0 \\ 0 \ se \ p = 0 \end{cases}$$

Per capire il senso di questa definizione si immagini di voler scommettere con una moneta modificata come segue:

1. esce testa con probabilità $p_1 = 0.95$
2. esce testa con probabilità $p_2 = 0.6$
3. esce testa con probabilità $p_3 = 0.5$

usando la definizione di entropia otteniamo:

1. $H_1(p_1) = 0.286$
2. $H_2(p_2) = 0.971$
3. $H_3(p_3) = 1$

Ovviamente nel primo caso la probabilità di predire il risultato corretto è molto alta dato che la moneta è pesantemente modificata e infatti il sistema avrà una bassa entropia, nel secondo caso l'entropia aumenta, infine nel terzo l'indecisione sarà massima e l'entropia di conseguenza.
Per convincersi di quanto detto in maniera più matematica, si ha il seguente teorema:

Teorema 1.2.1. *Sia X una variabile casuale discreta, allora vale:*

1. *$H(X) \geq 0$ e $H(X) = 0$ se e solo se esiste un valore di X, x_1 t.c. $\mathbb{P}(x_1) = 1$*
2. *$H(X) \leq \log(n)$ e l'uguaglianza varrà solo quando X ha distribuzione uniforme*

Dimostrazione.

1. ovviamente $H(X) \geq 0$ perché somma di quantità positive (consideriamo gli addendi come $-\log(x)$ e ricordando che $x \in (0, 1]$). Per quanto riguarda l'uguaglianza, dato che tutti gli addendi della sommatoria sono positivi, abbiamo che $H(X) = 0$ se e solo se $p_j \log(p_j) = 0 \ \forall j$, quindi abbiamo che p_j sarà uguale ad 1 o 0, ma non può essere che tutti i p_j siano uguali a 0 e dunque deve esistere almeno un $p_j = 1$.
2. per prima cosa supponiamo che $p_j > 0$ (nel caso non lo fossero basterebbe togliere i $p_k = 0$ e dimostrare che $H(X) \leq \log(n - c) \leq \log(n)$ dove c è il numero di $p_k = 0$).

Dalla definizione abbiamo:

$$\begin{aligned} H(x) - \log(n) &= -\frac{1}{ln(2)}\left(\sum_{j=1}^{n} p_j ln(p_j) + ln(n)\right) \\ &= -\frac{1}{ln(2)}\left(\sum_{j=1}^{n} p_j (ln(p_j) + ln(n))\right) \\ &= -\frac{1}{ln(2)}\left(\sum_{j=1}^{n} p_j ln(p_j n)\right) \\ &= \frac{1}{ln(2)}\left(\sum_{j=1}^{n} p_j ln\left(\frac{1}{p_j n}\right)\right) \\ &\leq \frac{1}{ln(2)}\left(\sum_{j=1}^{n} p_j \left(\frac{1}{p_j n} - 1\right)\right) \\ &= \frac{1}{ln(2)}\left(\sum_{j=1}^{n} \left(\frac{1}{n} - p_j\right)\right) \leq 0 \end{aligned}$$

dove nel per passare dalla quarta alla quinta riga abbiamo usato il fatto che $ln(x) \leq x - 1$ con l'uguaglianza solo se $x = 1$. Quindi abbiamo che le disuguaglianze si trasformano in uguaglianze solo se $\frac{1}{p_j n} = 1$ cioè se $p_j = \frac{1}{n}$ cioè se si ha distribuzione uniforme.

□

1.3 Unicità dell'Entropia

Si può dimostrare che la scelta della funzione di entropia come *misura di incertezza* è unica a meno di una costante moltiplicativa.
Ptrima di definire la *misura di incertezza* premettiamo una precisazione sulla notazione.
Indicheremo la probabilità condizionata ($\mathbb{P}(Y = k|X = j)$) con la notazione $p_j(k)$ oppure, in modo totalmente equivalente, $p(k|j)$.

Definizione 1.3.1. sia $(\Omega, \mathcal{F}, \mathbb{P})$un spazio di probabilità e X variabile casuale discreta di legge $\{p_1....p_n\}$, una funzione U viene detta **misura di incertezza** se soddisfa le seguenti condizioni:

1. $U(X)$ è un massimo quando ha distribuzione uniforme
2. $U(p_1...p_n, 0) = U(p_1...p_n)$
3. $U(p_1....p_n)$ è continua per tutti i suoi argomenti.
4. presa Y variabile casuale allora $U(X, Y) = U_X(Y) + U(X)$ dove $U_X(Y) = \sum_{j=1}^{n} p_j U(Y|X = j)$ dove $(Y|X = j)$ può essere vista come una variabile casuale di legge di probabilità $\{p_j(1)...p_j(m)\}$

Teorema 1.3.1. *In uno spazio di probabilità* $(\Omega, \mathcal{F}, \mathbb{P})$*consideriamo una variabile casuale* X *con legge di probabilità* $\{p_1...p_n\}$ *allora*
$U(X)$ *è una misura di incertezza se e solo se*

$$U(X) = KH(X)$$

dove K *è una costante* $K \geq 0$

Per la dimostrazione si veda [2] pag.10.

1.4 Proprietà dell'entropia

In questa sezione indagheremo le prime proprietà dell'entropia e dimostreremo i primi risultati che getteranno le basi per le costruzioni successive. Può essere interessante capire come si comporta l'entropia nel caso in cui le variabili in considerazione siano dipendenti, per fare ciò definiamo l'*entropia condizionata.*

Definizione 1.4.1. Si dirà **entropia condizionale di Y data $X = j$** la funzione:

$$H_j(Y) := -\sum_{k=1}^{m} p_j(k) \log(p_j(k)) \tag{1.4.1}$$

Prendiamo ora una variabile casuale X, possiamo considerare la variabile casuale $H.(Y)$ che avrà immagine $\{H_1(Y)...H_n(Y)\}$ e legge di probabilità $\{p_1...p_n\}$. Avremo quindi che $H.(Y)$ sarà funzione di X.

Definizione 1.4.2. Si dirà **entropia condizionale di Y data X**,la funzione:

$$H_X(Y) := \mathbb{E}[H.(Y)] = \sum_{j=1}^{n} p_j H_j(Y) \tag{1.4.2}$$

Osservazione 1. Più avanti, analogamente a quanto detto per la probabilità condizionata, ci sarà più comodo scrivere $H_X(Y)$ come $H(Y|X)$.

Lemma 1.4.1.

$$H_X(Y) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} \log(p_j(k)) \tag{1.4.3}$$

Dimostrazione. Sostituendo 1.4.1 in 1.4.2 otteniamo

$$H_X(Y) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_j p_j(k) \log(p_j(k)) \tag{1.4.4}$$

Ricordando che

$$p_j(k) = \mathbb{P}(Y = k|X = j) \text{ } e \text{ } p_j = \mathbb{P}(X = j)$$

otteniamo che

$$p_j p_j(k) = \mathbb{P}(X = j)\mathbb{P}(Y = k|X = j) = \mathbb{P}(X = j, Y = k) = p_{jk}$$

sostituendo questo risultato in 1.4.4 possiamo concludere. □

Lemma 1.4.2. *se X e Y sono indipendenti allora vale:*

$$H_X(Y) = H(Y) \tag{1.4.5}$$

Dimostrazione. Sia $\{q_1...q_m\}$ la legge di probabilità di Y allora ci basterà notare che nel caso in cui X e Y siano indipendenti $p_j(k) = \mathbb{P}(Y = k|X = j) = \mathbb{P}(Y = k) = q_k$ e dunque 1.4.4 diventerà

$$H_X(Y) = -\sum_{k=1}^{m} q_k \log(q_k) \sum_{j=1}^{n} p_j = -\sum_{k=1}^{m} q_k \log(q_k) 1 = H(Y)$$

□

Definizione 1.4.3. Siano X e Y due variabili casuali definite sullo stesso spazio di probabilità, definiamo la loro **entropia congiunta** $H(X,Y)$ come:

$$H(X,Y) := -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\log(p_{jk}) \tag{1.4.6}$$

dove con p_{jk} intendiamo $\mathbb{P}(X=j, Y=k)$

Osservazione 2. Dalla definizione si ha immediatamente che $H(X,Y) = H(Y,X)$.

Teorema 1.4.1. *Date due variabili casuali X, Y vale:*

$$H(X,Y) = H(X) + H_X(y). \tag{1.4.7}$$

Dimostrazione. sapendo che $\mathbb{P}(A\cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$ e quindi che $p_{jk} = p_j p_j(k)$ possiamo sostituire direttamente nella definizione di entropia congiunta 1.4.3 ottenendo:

$$H(X,Y) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\log(p_j p_j(k)) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\log(p_j(k)) - \sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\log(p_j)$$

possiamo concludere ricordando che $\sum_{k=1}^{m} p_{jk} = p_j$ □

Corollario 1.4.1. *se X e Y sono indipendenti allora vale:*

$$H(X,Y) = H(X) + H(Y) \tag{1.4.8}$$

Dimostrazione. basta applicare il lemma 1.4.2 al teorema precedente □

Teorema 1.4.2. *(Disuguaglianza fondamentale di Shannon)*

$$H_X(Y) \leq H(Y) \tag{1.4.9}$$

Dimostrazione. Per la dimostrazione utilizziamo la disuguaglianza di Jensen:
data f funzione convessa vale

$$\sum_{j=1}^{n}\lambda_j f(x_j) \geq f\left(\sum_{j=1}^{n}\lambda_j x_j\right) \tag{1.4.10}$$

con $\lambda_j > 0$ e $\sum_{j=1}^{n}\lambda_j = 1$ per la dimostrazione si veda [6].
Ora applicando la disuguaglianza con:

$$\lambda_j = p_j,\ f(x) = x\log(x),\ x_j = p_j(k)$$

per k fissato, otteniamo quindi:

$$\sum_{j=1}^{n} p_j p_j(k)\log(p_j(k)) \geq \sum_{j=1}^{n}\left(p_j p_j(k)\right)\log\left(\sum_{j=1}^{n} p_j p_j(k)\right) = q_k\log(q_k)$$

dove l'uguaglianza la ricaviamo da: $\sum_{j=1}^{n} p_j p_j(k) = \sum_{j=1}^{n}\left(\mathbb{P}(X=j)\mathbb{P}(Y=k|X=j)\right) = \mathbb{P}(Y=k) = q_k$. Sommando su k abbiamo che la parte sinistra della disuguaglianza diventa:

$$\sum_{j=1}^{n} p_j \sum_{k=1}^{m} p_j(k)\log(p_j(k)) = -\sum_{j=1}^{n} p_j H_k(Y) = -H_X(Y)$$

mentre a destra otteniamo

$$\sum_{k=1}^{m} q_k\log(q_k) = -H(Y)$$

e quindi:

$$-H_X(Y) \geq -H(Y) \tag{1.4.11}$$

Da cui possiamo concludere direttamente. □

Questo risultato può essere pensato come: aggiungendo informazione (il valore di X) l'entropia del sistema diminuisce.

Osservazione 3. Nel caso di *processi stocastici* (si veda 1.6 per la definizione) è comodo osservare che considerando $Y = (X_{n+1})$ e $X = X_0$ nel teorema precedente si ottiene:

$$H(X_{n+1}|X_0, X_1...X_n) \leq H(X_{n+1}|X_1...X_n)$$

Definizione 1.4.4. date due variabili casuali X, Ydefiniamo **mutua informazione di X e Y**

$$I(X,Y) := H(Y) - H_X(Y) \tag{1.4.12}$$

Notiamo che $H_X(Y)$ è l'informazione contenuta in Y che non è contenuta in X e quindi l'informazione di Y contenuta in X sarà $H(Y) - H_X(Y) = I(X,Y)$

Teorema 1.4.3. *Siano X e Y due variabili casuali rispettivamente legge di probabilità $\{p_1...p_n\}$ e $\{q_1...q_n\}$*

1. $I(X,Y) = \sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\log\left(\frac{p_{jk}}{p_j q_k}\right)$
2. $I(X,Y) = I(Y,X)$
3. *se X e Y sono indipendenti allora $I(X,Y) = 0$*

Dimostrazione. si proceda come segue:

1. sempre ricordando che $\sum_{k=1}^{m} p_{jk} = p_j$ possiamo scrivere

$$H(Y) = -\sum_{k=1}^{m} q_k \log(q_k) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\log(q_k)$$

   e dunque per 1.4.3 otteniamo

$$I(X,Y) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\log(q_k) + \sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\log\left(p_j(k)\right)$$

2. immediato da 1.
3. semplicemente ricordando che se X e Y sono indipendenti $H_X(Y) = H(Y)$

□

1.5 Principio dell'Entropia Massima

Spesso ci si trova in condizioni in cui è data una variabile casuale X a valori $\{x_1...x_n\}$ di cui non si conosce la legge di probabilità $\{p_1...p_n\}$ in questi casi si può applicare il principio di entropia massima:

Definizione 1.5.1. Data una una variabile casuale X con legge di probabilità $\{p_1...p_n\}$ incognita il **principio dell'entropia massima** ci impone di scegliere i p_j in modo tale che $H(X)$ sia massima

Esempio. Sia X una variabile casuale a valori $\{x_1...x_n\}$ di cui non si conosce la legge di probabilità $\{p_1...p_n\}$. Sappiamo già che, se non ci sono altre condizioni, l'entropia sarà massima se X sarà uniformemente distribuita. Prendiamo ora il caso in cui ci venga fornita la media di $\mathbb{E}[X] = E$. Per trovare il massimo dell'entropia $H(X)$ utilizziamo il metodo dei moltiplicatori di Lagrange:
come costrizioni abbiamo:

1. $\sum_{j=1}^{n} p_j = 1$
2. $\sum_{j=1}^{n} x_j p_j = E$

Dunque dobbiamo trovare il massimo valore di:

$$L(p_1...p_n; \lambda, \mu) := -\sum_{j=1}^{n} p_j \log(p_j) + \lambda\left(\sum_{j=1}^{n} p_j - 1\right) + \mu\left(\sum_{j=1}^{n} x_j p_j - E\right) \tag{1.5.1}$$

dove λ, μ sono i moltiplicatori di Lagrange.
Imponendo le derivate parziali uguali a 0 otteniamo:

$$\frac{\partial L}{\partial p_j} = -\frac{1}{ln(2)}(ln(p_j) + 1) + \lambda + \mu x_j = 0 \quad (1 \leq j \leq n)$$

$$quindi$$

$$p_j = e^{\lambda' + \mu' x_j} \quad (1 \leq j \leq n)$$

dove $\lambda' = ln(2)\lambda - 1$ e $\mu' = ln(2)\mu$
da 1. possiamo ricavare l'equazione $0 = \sum_{j=1}^{n} p_j - 1 = \sum_{j=1}^{n} e^{\lambda' + \mu' x_j} - 1$ che risolta ci restituisce:

$$\lambda' = -ln(Z(\mu'))$$

dove $Z(\mu') := \sum_{j=1}^{n} e^{\mu' x_j}$.
questo ci permette di riscrivere p_j come:

$$\begin{aligned} p_j &= e^{\lambda' + \mu' x_j} \\ &= e^{-ln(Z(\mu')) + \mu' x_j} \\ &= \frac{e^{\mu' x_j}}{Z(\mu')} \end{aligned}$$

Per μ' invece possiamo usare 2. ottenendo

$$\begin{aligned} 0 &= \sum_{j=1}^{n} x_j p_j \\ &= \sum_{j=1}^{n} x_j \frac{e^{\mu' x_j}}{Z(\mu')} \end{aligned}$$

riassumendo quindi abbiamo:

$$p_j = \frac{e^{\mu' x_j}}{Z(\mu')} \quad (1 \leq j \leq n) \tag{1.5.2}$$

1.6 Entropia nelle catene di Markov

Definizione 1.6.1. Si consideri una famiglia di variabili casuali tutte definite sullo stesso spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$, $(X(t), t \geq 0)$, tale famiglia è detta **processo stocastico**.

Nella nostra trattazione ci limiteremo a considerare una piccola classe di processi stocastici chiamati catene di Markov.

Definizione 1.6.2. Un processo stocastico è detto **catena di Markov** se

1. l'insieme S che comprende i valori ammissibili delle variabili X_n è discreto (se S è denso si dirà processo di Markov)
2. possiede la 'proprietà di Markov' cioè

$$\mathbb{P}(X_{n+1} = k_{n+1} | X_n = k_n, ., X_0 = k_0) = \mathbb{P}(X_{n+1} = k_{n+1} | X_n = k_n)$$

Per i nostri scopi inoltre considereremo l'insieme del tempo come come un insieme discreto contenuto in $\mathbb{N}$.
Infine, definita $p_{ij}^n := \mathbb{P}(X_{n+1} = j | X_n = i)$ vogliamo che la nostra matrice di transizione P, formata dai vari p_{ij}, sia stazionaria cioè:

$$p_{ij}^n = p_{ij}^0 =: p_{ij} \ \forall n$$

Ci domandiamo ora se è sempre possibile definire una catena di Markov $X = (X_n, n \in \mathbb{Z})$ per la quale ogni X_n ammette entropia massima cioè per il teorema 1.2.1 X_n ha distribuzione uniforme (se non vi sono altre restrizioni). Premettiamo alcune definizioni

Definizione 1.6.3. Sia $X = (X_n, n \in \mathbb{N})$ una catena di Merkov con matrice di transizione P. Un vettore di probabilità ρ è detto **distribuzione invariante o stazionaria per X** se:

$$\rho = \rho P$$

Definizione 1.6.4. Una matrice A ad elementi positivi è detta **bistocastica** se per ogni riga e per ogni colonna la somma dei suoi elementi è pari ad 1.

Teorema 1.6.1. *Una catena di Markov con matrice di transizione P ammette la distribuzione uniforme come distribuzione invariante se e solo se P è bistocastica*

Dimostrazione. Se P è bistocastica allora $\sum_{i=1}^N P_{ij} = 1 \forall 1 \leq j \leq N$ e quindi:

$$\sum_{i=1}^N \frac{1}{N} P_{ij} = \frac{1}{N} \sum_{i=1}^N P_{ij} = \frac{1}{N}$$

e quindi $\frac{1}{N}$ è una distribuzione invariante.
Supponiamo ora che la distribuzione uniforme sia invariante e dimostriamo che P è bistocastica. Essendo P una matrice di transizione abbiamo già che la somma degli elementi di una riga sarà 1. Dimostriamo che anche la somma degli elementi di una colonna è pari a 1. Procedendo al contrario di prima abbiamo:

$$\frac{1}{N} = \frac{1}{N} \sum_{i=1}^N P_{ij} \Rightarrow \sum_{i=1}^N P_{ij} = 1$$

Abbiamo che per ogni colonna la somma dei suoi elementi è 1, quindi possiamo concludere □

Definizione 1.6.5. Siano X e Y due variabili casuali della stessa dimensione con legge di probabilità rispettivamente $\{p_1 ... p_n\}$ e $\{q_1 ... q_n\}$. Definiamo **entropia relativa** il valore:

$$D(X, Y) := \sum_{j=1}^n p_j \log\left(\frac{p_j}{q_j}\right) \tag{1.6.1}$$

Teorema 1.6.2. *Per l'entropia relativa vale:*

1. $D(X, Y) \geq 0, =$ *se e solo se X e Y sono identicamente distribuite*
2. *se Y è uniformemente distribuita allora vale*

$$D(X, Y) = \log(n) - H(X) \tag{1.6.2}$$

Dimostrazione.

1. riscriviamo il primo punto come:

$$\begin{aligned} D(X,Y) &= \sum_{j=1}^{n} p_j \log\left(\frac{p_j}{q_j}\right) \\ &= \sum_{j=1}^{n} p_j \log(p_j) - \sum_{j=1}^{n} p_j \log(q_j) \end{aligned}$$

   E da qui si può concludere applicando la disuguaglianza di Gibbs: $-\sum_{i=1}^{n} p_i \log(p_i) \leq -\sum_{i=1}^{n} p_i \log(q_i)$ per $\{p_1...p_n\}$ e $\{q_1...q_n\}$ distribuzioni di probabilità (deriva immediatamente dal caso continuo 2.1.1)

2. essendo Y uniformemente distribuita avremo che $q_j = 1\frac{1}{n}$ e quindi dalla definizione di entropia relativa abbiamo:

$$\begin{aligned} D(X,Y) &= \sum_{j=1}^{n} p_j \log\left(\frac{p_j}{q_j}\right) \\ &= \sum_{j=1}^{n} p_j \log(p_j n) \\ &= \sum_{j=1}^{n} p_j \log(p_j) + \sum_{j=1}^{n} p_j \log(n) \\ &= -H(X) + \log(n) \sum_{j=1}^{n} p_j = \log(n) - H(X) \end{aligned}$$

□

Definizione 1.6.6. Un processo stocastico è detto **stazionario** se presi $m, k \in \mathbb{N}$ vale:

$$\mathbb{P}(X_{n_1} = i_1 ... X_{n_k} = i_k) = \mathbb{P}(X_{n_1+m} = i_1 ... X_{n_k+m} = i_k) \ \forall i_s \in S$$

Consideriamo una catena di Markov stazionaria $(X_n, n \in \mathbb{Z}_+)$ con una matrice di transizione bistocastica P e sia X_∞ una variabile casuale di dimensione n uniformemente distribuita, abbiamo quindi che $D(X_n, X_\infty) = \log(N) - H(X_n)$. Si dimostra che $D(X_n, X_\infty)$ è una funzione decrescente e che se la distribuzione uniforme è l'unica distribuzione invariante, allora $\lim_{n\to\infty} D(X_n, X_\infty) = 0$. Segue quindi che $H(X_n)$ è crescente. [5]

1.7 La Regola della Catena

Vediamo ora come cambia l'informazione in un processo stocastico. L'approccio più naturale può sembrare quello di considerare l'entropia come funzione del tempo,come si è cominciato a fare sopra, in questo modo però ci si dimentica della relazione che esiste tra due passaggi successivi, dal tempo t_n al tempo t_{n+1}. Procederemo quindi in modo differente cominciando dal generalizzare i risultati visti nel caso di due sole variabili.
Estendiamo la definizione di entropia congiunta (1.4.3) in questo modo:

$$H(X_0...X_n) := -\sum_{i_0...i_n=1}^{N} p(i_0...i_n) \log(p(i_o...i_n)). \tag{1.7.1}$$

Mentre la definizione di entropia condizionata (1.4.1) nel caso multivariato sarà:

$$H(Y|X_1...X_n) = -\sum_{j,i_1...i_n=1}^{N} \mathbb{P}(Y = j, X_1 = i_1 ... X_n = i_n) \log(\mathbb{P}(Y = j | X_1 = i_1 ... X_n = i_n)) \tag{1.7.2}$$

Non ci rimane che generalizzare il teorema 1.4.1.

Teorema 1.7.1. ***Regola della catena***

$$H(X_0...X_n) = H(X_0) + \sum_{i=1}^{n} H(X_i|X_0,...X_{i-1}) = H(X_0) + H(X_1|X_0) + ... + H(X_n|X_0...X_{n-1}) \quad (1.7.3)$$

Inoltre l'entropia congiunta cresce al crescere di n.

Dimostrazione. Procediamo per induzione:
Il caso base con $n = 1$ è esattamente il teorema 1.4.1, procediamo con il passo induttivo. Quindi assumiamo che valga per n, dimostriamo che vale per $n + 1$.

$$H(X_0...X_n, X_{n+1}) = -\sum_{i_0...i_n,i_{n+1}=1}^{N} p(i_0...i_n, i_{n+1}) \log(p(i_o...i_n, i_{n+1}))$$

$$= -\sum_{i_0...i_n,i_{n+1}=1}^{N} p(i_0...i_n, i_{n+1}) \log(\mathbb{P}(X_{n+1} = i_{n+1}|X_0 = i_0, ...X_n = i_n)) - \sum_{i_0...i_n,i_{n+1}=1}^{N} p(i_0...i_n, i_{n+1}) \log(p(i_0...i_n)$$

dato che

$$\sum_{i_0...i_n,i_{n+1}=1}^{N} p(i_0...i_{n+1}) \log(p(i_0...i_n)) = \sum_{i_0...i_n=1}^{N} p(i_0...i_n) \log(p(i_0..i_n))$$

abbiamo che

$$H(X_0...X_{n+1}) = H(X_0...X_n) + H(X_{n+1}|X_0...X_n) \quad (1.7.4)$$

Applicando l'ipotesi induttiva otteniamo il risultato. Inoltre da 1.7.4 e dal fatto che l'entropia condizionata è sempre maggiore di zero otteniamo che l'entropia congiunta cresce nel tempo. □

1.8 Velocità dell'Entropia

Definizione 1.8.1. Quando il limite esiste, $h(X)$ si dice **velocità dell'entropia** dove

$$h(X) := \lim_{n\to\infty} \frac{1}{n} H(X_0...X_{n-1})$$

Teorema 1.8.1. *se $X = (X_i, i \in \mathbb{N})$ è un processo stocastico stazionario, allora $h(X)$ esiste e:*

$$h(X) = \lim_{n\to\infty} H(X_{n-1}|X_0...X_{n-2}) \quad (1.8.1)$$

Dimostrazione. Applicando la regola della catena 1.7.1 otteniamo subito che

$$h(X) = \lim_{n\to\infty} \frac{1}{n} H(X_0...X_{n-1}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} H(X_i|X_0...X_{i-1}) \quad (1.8.2)$$

Passiamo ora a dimostrare l'esistenza del secondo membro di 1.8.1:
dall'osservazione 3 otteniamo:

$$H(X_{n+1}|X_0, X_1...X_n) \leq H(X_{n+1}|X_1...X_n) \quad (1.8.3)$$

Grazie al Teorema 1.4.1 possiamo scrivere:

$$H(X_{n+1}|X_1...X_n) = H(X_{n+1}, X_1...X_n) - H(X_1....X_n)$$

e ricordandoci che il processo è stazionario abbiamo:

$$H(X_{n+1}, X_1...X_n) - H(X_1....X_n) = H(X_n, X_0...X_{n-1}) - H(X_0....X_{n-1})$$

infine applicando il Teorema 1.4.1 in modo inverso rispetto a prima

$$H(X_n, X_0...X_{n-1}) - H(X_0....X_{n-1}) = H(X_n|X_0...X_{n-1})$$

riassumendo quindi

$$H(X_{n+1}|X_1...X_n) = H(X_n|X_0...X_{n-1}) \tag{1.8.4}$$

Sostituendo 1.8.4 in 1.8.3 otteniamo:

$$H(X_{n+1}|X_0, X_1...X_n) \leq H(X_n|X_0...X_{n-1}) \tag{1.8.5}$$

Quindi definendo $a_n := H(X_n|X_0...X_{n-1})$ otteniamo una successione $\{a_n\}_{n\in\mathbb{N}}$ monotona non crescente limitata dal basso visto che $a_k = H(X_k|X_0...X_{k-1}) \geq 0$ e dunque $\lim_{n\to\infty} a_n$ esiste ed è finito dato che $H(Y) < \infty$. Proviamo adesso che la serie $\frac{1}{n}\sum_{i=1}^{n} a_i = a$ dove $a := \lim_{n\to\infty} a_n$:

$$\lim_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} a_i - a\right| \leq \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}|a_i - a| = \lim_{n\to\infty}\frac{1}{n}\left(\sum_{i=1}^{N_0}|a_i - a| + \sum_{i=N_0+1}^{n}|a_i - a|\right) = \lim_{n\to\infty}\sum_{i=N_0+1}^{n}\frac{|a_i - a|}{n}$$

E da qui possiamo concludere scegliendo N_0 tale che $\frac{1}{n}|a_i - a|$ sia piccolo a piacere, cosa sempre possibile dato che $\lim_{n\to\infty} a_n = a$.
Ricordando come abbiamo definito a_n otteniamo quindi:

$$\lim_{n\to\infty} H(X_{n-1}|X_0...X_{n-2}) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} H(X_i|X_0...X_{i-1}) \tag{1.8.6}$$

ricordando infine 1.8.2 possiamo concludere:

$$h(X) = \lim_{n\to\infty}\frac{1}{n}H(X_0...X_{n-1}) = \lim_{n\to\infty} H(X_{n-1}|X_0...X_{n-2}).$$

□

Teorema 1.8.2. *Se $(X_i \in \mathbb{N})$ è una catena di Markov stazionaria con distribuzione iniziale $\pi^{(0)}$ e matrice di transizione P allora vale*

$$h(X) = -\sum_{i,j=1}^{n} \pi^{(0)} P_{ij} \log(P_{ij}) \tag{1.8.7}$$

Dimostrazione. dal teorema precedente 1.8.1 abbiamo:

$$\begin{aligned}
h(X) &= \lim_{n\to\infty} H(X_{n-1}|X_0...X_{n-2}) \\
&= \lim_{n\to\infty} H(X_{n-1}|X_{n-2}) \\
&= H(X_1|X_2) \\
&= -\sum_{i,j=1}^{n} \mathbb{P}(X_0 = i, X_1 = j)\log(P_{ij}) \\
&= -\sum_{i,j=1}^{n} \mathbb{P}(X_0 = i) P_{ij}\log(P_{ij}) \\
&= -\sum_{i,j=1}^{n} \pi^{(0)} P_{ij}\log(P_{ij})
\end{aligned}$$

Dove per passare dalla prima alla seconda riga abbiamo usato la 'proprietà di Markov' 1.6.2, per passare dalla seconda alla terza abbiamo usato il fatto che il processo è stazionario, dalla terza alla quarta il lemma 1.4.1 □

2 Entropia per Variabili Casuali Assolutamente Continue

In questo capitolo estenderemo la definizione di entropia data per il caso di una variabile aleatoria discreta al caso in cui la nostra variabile casuale X sia assolutamente continua.

Definizione 2.0.1. Data una variabile casuale X, chiamiamo **funzione di distribuzione di X** l'applicazione $F_X : \mathbb{R} \to \mathbb{R}$ data da:

$$F_X(t) := \mathbb{P}(X \in (-\infty, t])$$

Definizione 2.0.2. Una funzione di distribuzione F è detta **assolutamente continua** se esiste una funzione $f \in L^1(\mathbb{R})$, $f \geq 0$ e $\int_{\mathbb{R}} f(u)du = 1$ tale che:

$$F(t) = \int_{-\infty}^{t} f(u)du, \ t \in \mathbb{R} \tag{2.0.1}$$

dove l'integrale è definito nel senso di Lebesgue. Tale f verrà detta **funzione di densità**
Una variabile casuale che ha funzione di distribuzione della forma 2.0.1 è detta **variabile casuale assolutamente continua**

Per le proprietà degli elementi appena definiti si veda [3]

2.1 Entropia nel caso Continuo

Definizione 2.1.1. Sia X una variabile casuale con immagine (a, b) e funzione di densità f. $H(X)$ detta **entropia di X** dove:

$$H(X) = -\int_a^b \log(f(x))f(x)dx = \mathbb{E}\left[\log\left(\frac{1}{f(X)}\right)\right]$$

Anche qui per convenzione log sarà il logaritmo in base 2.
Purtroppo la proprietà di essere misura di incertezza, valida nel caso discreto 1.3.1, non è più valida con questa definizione.
Questo deriva dal fatto che, mentre nel caso discreto l'argomento del logaritmo è sempre compreso tra 0 e 1, nel caso continuo la funzione di densità, argomento del logaritmo, può assumere valori su tutto $\mathbb{R}$.
Per un esempio si calcoli l'entropia associata alla variabile casuale uniforme, ricordando che la su funzione di densità è $f(x) = \frac{1}{b-a}$ si ottiene:

$$\begin{aligned} H(X) &= \int_a^b \log\left(\frac{1}{b-a}\right)\frac{1}{b-a}dx \\ &= \log(b-a) \end{aligned}$$

che sarà negativa se $0 < b - a < 1$.
L'entropia per le variabili casuali non potrà quindi giocare un ruolo così importante come quello giocato per variabili casuali discrete. Esistono tuttavia alcuni teoremi degni di nota.

Prima di introdurli però calcoliamo l'entropia di $X \backsim N(\mu, \sigma^2)$.

$$\begin{aligned} H(X) &= -\frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{\infty} \exp\Big(-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big) \log\Big(\frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big)\Big) dx \\ &= \log(\sigma(2\pi)^{\frac{1}{2}}) + \frac{\log(e)}{\pi^{\frac{1}{2}}} \int_{-\infty}^{\infty} e^{-y^2} - y^2 dy \end{aligned}$$

dove abbiamo usato la sostituzione $y = \frac{x-\mu}{2^{\frac{1}{2}}\sigma}$.
Il calcolo diretto dell'integrale al secondo membro risulta:

$$\int_{-\infty}^{\infty} e^{-y^2} - y^2 dy = \frac{\pi^{\frac{1}{2}}}{2}$$

e quindi sostituendo nell'equazione sopra:

$$H(X) = \log(\sigma(2\pi e)^{\frac{1}{2}}) + \frac{\log(e)}{\pi^{\frac{1}{2}}} \frac{\pi^{\frac{1}{2}}}{2} = \log(\sigma(2\pi e)^{\frac{1}{2}}) \tag{2.1.1}$$

D'ora in avanti indicheremo $\log(\sigma(2\pi e)^{\frac{1}{2}})$ con $H_N(\sigma)$. Questo evidenzia il fatto che varianza ed entropia sono due concetti molto legati.

Teorema 2.1.1. *(Disuguaglianza di Gibbs nel caso continuo) Siano f, g due funzioni di densità allora vale*

$$-\int_{-\infty}^{\infty} \log(f(x)) f(x) dx \leq -\int_{-\infty}^{\infty} \log(g(x)) f(x) dx \tag{2.1.2}$$

dove l'uguaglianza è valida solo se $g(x) = f(x)$.

Dimostrazione. Dato che $\log_b(a) = \frac{\ln(a)}{\ln(b)}$ possiamo limitarci al caso in cui abbiamo $\ln(x)$, il quale ha la proprietà di essere sempre maggiore di $x - 1$ e uguale solo nel caso $x = 1$. Quindi

$$\begin{aligned} -\int_{-\infty}^{\infty} \log(g(x)) f(x) + \int_{-\infty}^{\infty} \log(f(x)) f(x) dx &= -\int_{-\infty}^{\infty} \Big[\log(g(x)) - \log(f(x))\Big] f(x) dx \\ &= -\int_{-\infty}^{\infty} \log\Big(\frac{g(x)}{f(x)}\Big) f(x) dx \\ &\geq -\int_{-\infty}^{\infty} \Big(\frac{g(x)}{f(x)} - 1\Big) f(x) dx \\ &= -\int_{-\infty}^{\infty} g(x) dx + \int_{-\infty}^{\infty} f(x) dx = 0 \end{aligned}$$

Dove l'ultima uguaglianza si ottiene ricordando che f, g sono funzioni di densità e quindi $\int_{-\infty}^{\infty} f(x) dx = 1$, $\int_{-\infty}^{\infty} g(x) dx = 1$.
La disuguaglianza nel caso precedente diventa un uguaglianza solo se $\frac{g(x)}{f(x)} = 1$ cioè solo se $g(x) = f(x)$. □

Teorema 2.1.2. *Sia X una variabile casuale assolutamente continua con immagine $\mathbb{R}$ di media μ, varianza σ^2 e funzione di densità f allora*

$$H(X) \leq H_N(\sigma)$$

con l'uguaglianza se e solo se $X \backsim N(\mu, \sigma^2)$

Dimostrazione. Dalla disuguaglianza di Gibbs 2.1.1 appena dimostrata otteniamo che, per ogni funzione di densità g:

$$H(X) \leq -\int_{-\infty}^{\infty} \log(g(x)) f(x) dx$$

con l'uguaglianza solo se $f(x) = g(x)$. Come g prendiamo $\frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$ cioè la funzione di densità di una Normale $N \backsim N(\mu, \sigma^2)$

$$
\begin{aligned}
-\int_{-\infty}^{\infty} \log(g(x)) f(x) dx &= -\int_{-\infty}^{\infty} \log\left(\frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\right) f(x) dx \\
&= \log(\sigma(2\pi)^{1/2}) - \frac{\log(e)}{2\sigma^2} \int_{-infty}^{infty} (y-\mu)^2 f(x) dx \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{\log(e)}{2\sigma^2} Var(X) \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{\log(e)}{2} \\
&= \frac{1}{2}\log(2\pi e\sigma^2) = H_N(\sigma)
\end{aligned}
$$

Dove c'è la disuguaglianza avere l'uguaglianza dobbiamo avere $f = g$ e quindi $X \backsim N(\mu, \sigma^2)$ □

3 Comunicazione

In questo capitolo sarà proposta una modellizzazione della trasmissione di informazione attraverso canali di comunicazione.

3.1 Trasmissione di informazione

Il modello più semplice è costituito da una sorgente, un canale di comunicazione, ed un ricevente.
La sorgente sarà modellata da una variabile aleatoria S a valori $\{a_1...a_n\}$ detti alfabeto sorgente e con legge di probabilità $\{p_1...p_n\}$. Il fatto che la sorgente S sia una variabile casuale va interpretato come l'incertezza su quale messaggio verrà inviato. In questo contesto un messaggio sarà formato da una serie di simboli presi da $\{a_1...a_n\}$ e posti uno di seguito all'altro.
Il ricevente sarà un'altra variabile casuale R a valori $\{b_1...b_m\}$ detti alfabeto ricevente e con legge di probabilità $\{q_1...q_m\}$, solitamente si avrà $m \geq n$.
Infine l'effetto di distorsione del canale sarà modellato dalla famiglia di probabilità condizionate $\{p(j|i); 1 \leq i \leq n,\ 1 \leq j \leq m\}$ dove $p(j|i) := \mathbb{P}(R = b_j|S = a_i)$ (si noti che $p(j|i)$ corrisponde a $p_i(j)$ definito in 1.4).
Un sistema di trasmissione ottimale avrà i due alfabeti di trasmissione e ricezione identici e nella distorsione avremo $p(i|i)$ il più vicino possibile ad 1, in questo modo quindi i valori ricevuti saranno quasi sicuramente gli stessi che sono stati inviati.

Definizione 3.1.1. viene detta **mutua informazione** tra E ed F il valore:

$$I(a_j, b_k) = -\log(q_k) + \log(p(k|j)) \tag{3.1.1}$$

se $p_j = 0$ allora diremo $I(a_j, b_k) = 0$.
Dove E è l'evento $(S = a_j)$ che ha probabilità p_j, mentre F è l'evento $(R = b_k)$ che avverrà con probabilità q_k.

È importante notare che questa definizione di mutua informazione è diversa da quella data in 1.4.12 la quale si riferisce a due variabili casuali e non a due eventi come in questo caso.
Dato che $-\log(q_k)$ è l'informazione dell'evento $R = b_k$, mentre $-\log(p(k|j))$ è l'informazione aggiuntiva che ci darebbe la ricezione di b_k sapendo già per certo che è stato spedito a_j, possiamo interpretare $I(a_j, b_k)$ come la quantità di informazione su $R = b_k$ che ci è data dall'evento $S = a_j$. In altre parole è la quantità di informazione che è spedita attraverso il canale.

Teorema 3.1.1. *Per ogni* $1 \leq j \leq n, 1 \leq k, \leq m$ *si ha:*

1. $I(a_j, b_k) = -\log(\frac{p_{jk}}{p_j q_k})$
2. $I(a_j, b_k) = -\log(p_j) + \log(q(j|k))$
3. $I(a_j, b_k) = I(b_k, a_j)$
4. *se gli eventi* $S = a_j$ *e* $R = b_k$ *sono indipendenti allora* $I(a_j, b_k) = 0$
5. $I(S, R) = \sum_{j=1}^{n} \sum_{k=1}^{m} p_{jk} I(a_j, b_k)$.

Dimostrazione. 1. deriva banalmente da $p(k|j) = \frac{p_{jk}}{q_k}$

2. si ricava sostituendo in 1. $q(j|k) = \frac{p_{ik}}{q_k}$

3. deriva da 2.

4. ricordando che nel caso siano indipendenti $p_{jk} = p_j q_k$ si ricava immediatamente da da 1.

5. si ricava da 1. e dal primo punto del teorema 1.4.3

□

Il punto 3. del sistema ci mostra la curiosa caratteristica per cui se in un sistema si invertono sorgente e ricevente abbiamo che l'informazione su a_j contenuta in b_k è la stessa di quella contenuta in a_j su b_k quando il canale funziona normalmente. Si può dimostrare che $I(S,R) \geq 0$ sempre.

Supponiamo ora di scegliere un canale e di fissare $\{p(j|i); 1 \leq i \leq n,\ 1 \leq j \leq m\}$. Vogliamo ora fare in modo che il canale trasmetta più informazione possibile, per fare ciò le uniche variabili del sistema rimaste ancora libere con cuoi possiamo lavorare sono $\{p_1 ... p_n\}$.

Definizione 3.1.2. viene definita **capacità del canale C** la quantità:

$$C := \max_{\{p_1 \cdots p_n\}} I(S,R) \tag{3.1.2}$$

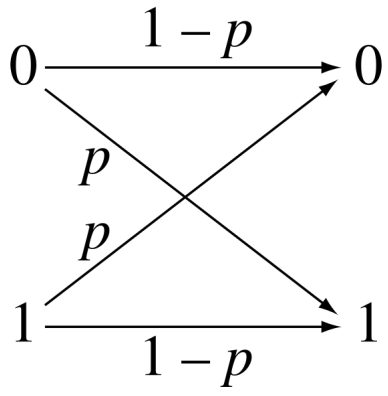
dove il massimo è scelto tra tutte le possibili leggi di probabilità della variabile S

Operativamente spesso è preferibile vedere la capacità del canale C come:

$$C = max(H(R) - H_S(R)) \tag{3.1.3}$$

ottenuta utilizzando la definizione di mutua informazione tra variabili casuali 1.4.12.

Osservazione 4. Il più semplice esempio di canale di comunicazione che possiamo trovare è un canale binario simmetrico, esso avrà grande rilevanza in seguito. È formato da una sorgente con alfabeto $\{0,1\}$ e come specificato nella figura il *rumore* del canale è definito attraverso un parametro p



3.2 Codici

In questo paragrafo daremo un idea di ciò che si intende con *codice* nella matematica per poi applicarci la nostra conoscenza sulla trasmissione di informazione.

Definizione 3.2.1. L'**alfabeto di un codice, C** è un insieme $\{c_1 ... c_r\}$i cui elementi c_i sono chiamati **simboli**.
Una **parola-codice** o **parola del codice** è una serie di simboli $c_{i_1} ... c_{i_n}$.
Il numero n sarà la **lunghezza** della parola-codice.
Un **messaggio** sarà una successione di parole-codice.

Il processo di codifica di un messaggio è quello di mappare ogni singolo simbolo dell'alfabeto di un linguaggio con una parola-codice.
Un esempio di codice che poi utilizzeremo lungo tutto il capitolo è dato dal codice binario:

$$C = \{0,1\}$$

Il nostro obiettivo sarà capire cosa succede all'informazione trasmessa modificando il percorso del messaggio nel modo seguente:

$$SORGENTE \to codificatore \to CANALE \to decodificatore \to RICEVENTE$$

Per fare cioè ci serviremo di un'importantissima classe di codici: quella dei *codici istantanei o codici prefisso.*

Definizione 3.2.2. Sia $c_{i_1}...c_{i_n}$ una parola del codice. Preso $k < n$ se $c_{i_1}...c_{i_k}$ è anch'essa una parola tale parola si dirà **prefisso**.
Un codice in cui non esistono parole che sono prefisso di altre è detto **codice istantaneo** o **codice prefisso**.

.

Lemma 3.2.1. *Ogni codice istantaneo è decodificabile in modo univoco, inoltre per avere una codifica univoca non è necessario aspettare di ricevere tutto il messaggio.*

Viste le forti proprietà dei codici istantanei è naturale chiedersi quando sia possibile creare codici con queste caratteristiche.

Teorema 3.2.1. *(Disugiaglianza di Kraft-McMillan)*
Dato un alfabeto sorgente composto da n simboli che deve essere codificato allora esiste un codice istantaneo con alfabeto di r simboli e parole di lunghezza l_i $(1 \leq i \leq n)$ se e solo se

$$\sum_{i=1}^{n} r^{-l_i} \leq 1. \tag{3.2.1}$$

La disuguaglianza di *Disugiaglianza di Kraft-McMillan* ci garantisce l'esistenza di codici che soddisfano le nostre richieste e, attraverso 3.2.1 , ci aiuta a trovare tali codici. Il passo successivo sarà chiederci come si può scegliere il migliore tra tutti i codici istantanei. Cominciamo con una definizione

Definizione 3.2.3. Dato un alfabeto sorgente S $\{a_1...a_n\}$ con legge di probabilità $\{p_1...p_n\}$ a cui viene associato un codice istantaneo posiamo considerare una variabile casuale L con immagine $\{l_1...l_n\}$ (dove l_i corrisponderà al numero di simboli necessari per scrivere in codice il simbolo a_i) e legge di probabilità $\{p_1...p_n\}$ la stessa di S. Preso il valore di aspettazione di L:

$$\mathbb{E}(L) = \sum_{i=1}^{n} p_j l_j$$

diremo che **il codice è ottimale** se minimizza $\mathbb{E}(L)$.

È chiaro che in generale un codice ottimale non è unico infatti dato un qualsiasi codice ottimale che utilizzi un alfabeto di almeno due lettere ci basterà considerare un codice in cui le lettere vengono permutate per ottenere un nuovo codice ottimale.
Enunciamo ora un sorprendente teorema che ci permette di mettere in relazione il valore di aspettazione di L con l'entropia dell'alfabeto sorgente, dandoci quindi utili informazioni sul valore di aspettazione di un codice ottimale.

Teorema 3.2.2. *(Teorema della codifica di sorgente per simboli di codice)*
Dato un alfabeto sorgente S con legge di probabilità $\{p_1...p_n\}$ vale:

1. *Per ogni codice istantaneo con una alfabeto di r simboli abbiamo che*

$$\frac{H(S)}{\log(r)} \leq \mathbb{E}(L) \tag{3.2.2}$$

con l'uguaglianza se e solo se $p_j = r^{l_j}$ $(1 \leq j \leq n)$

2. esiste un codice istantaneo formato da r simboli per cui

$$\frac{H(S)}{\log(r)} \leq \mathbb{E}(L) < \frac{H(S)}{\log(r)} + 1 \tag{3.2.3}$$

Dimostrazione.

Definiamo $\{q_1...q_n\}$ con

$$q_j = \frac{r^{-l_j}}{\sum_{i=1}^{N} r^{-l_i}} \tag{3.2.4}$$

abbiamo che l'insieme dei q_i forma una distribuzione di probabilità infatti:

$$\begin{aligned}\sum_{j=1}^{n} q_j &= \sum_{j=1}^{n} \frac{r^{-l_j}}{\sum_{i=1}^{N} r^{-l_i}} \\ &= \frac{\sum_{j=1}^{n} r^{-l_j}}{\sum_{i=1}^{N} r^{-l_i}} = 1\end{aligned}$$

e ovviamente $q_j \geq 0$.

Possiamo quindi utilizzare la disuguaglianza di Gibbs nel caso discreto: $-\sum_{i=1}^{n} p_i \log(p_i) \leq -\sum_{i=1}^{n} p_i \log(q_i)$ per $\{p_1...p_n\}$ e $\{q_1...q_n\}$ distribuzioni di probabilità (deriva immediatamente dal caso continuo 2.1.1) ottenendo:

$$\begin{aligned}H(S) &= -\sum_{j=1}^{n} p_j \log(p_j) \\ &\leq -\sum_{j=1}^{n} p_j \log(q_j) \\ &= -\sum_{j=1}^{n} p_j \log\left(\frac{r^{-l_j}}{\sum_{i=1}^{N} r^{-l_i}}\right) \\ &= -\sum_{j=1}^{n} p_j \log(r^{-l_j}) + \sum_{j=1}^{n} p_j \log\left(\sum_{i=1}^{N} r^{-l_i}\right) \\ &\leq -\sum_{j=1}^{n} p_j \log(r^{-l_j}) + \sum_{j=1}^{n} p_j \log(1) \\ &= \sum_{j=1}^{n} p_j l_j \log(r) \\ &= \mathbb{E}(L)\log(r)\end{aligned}$$

dove per ottenere la quinta riga abbiamo utilizzato la disugiaglianza di Kraft-McMillan 3.2.1.

Dalle condizioni delle disuguaglianze di Gibbs e Kraft abbiamo che si ha l'uguaglianza se e solo se $p_j = r^{-l_j}$

Il primo punto dimostra anche il primo membro della disuguaglianza, mostriamo ora la validità della seconda parte.

Imponiamo $l_j = -\lceil \log_r(p_j) \rceil$ cioè $-\log_r(p_j) \leq l_j < -\log_r(p_j) + 1$ e quindi

$$r^{-l_j} \leq p_j \implies$$

$$\sum_{j=1}^{n} r^{-l_j} \leq \sum_{j=1}^{n} p_j = 1.$$

Quindi per la disuguaglianza di Kraft esiste un codice di tale lunghezza con variabile casuale associata

L. Abbiamo:

$$\begin{aligned} H(L) &= \sum_{j=1}^{n} p_j l_j \\ &< -\sum_{j=1}^{n} p_j \log_r(p_j) + 1 \\ &= -\sum_{j=1}^{n} p_j \frac{\log(p_j)}{\log(r)} + 1 \\ &= \frac{H(S)}{\log(r)} + 1 \end{aligned}$$

□

Il teorema appena dimostrato è detto primo teorema di Shannon e fu dimostrato proprio dal matematico Americano nel 1948.

3.3 Regole di decisione

Mettiamoci nella situazione

$$SORGENTE \to codificatore \to CANALE \to decodificatore \to RICEVENTE$$

e concentriamoci sul segmento $codificatore \to CANALE \to decodificatore$.
Supponiamo che $C := \{x_1..x_n\}$ sia l'insieme di tutte le possibili parole del codice che possono essere trasmesse dal e che y sia la parola ricevuta. Per decidere quale parola x_i è stata trasmessa possiamo utilizzare il *principio di massima verosimiglianza*:
Date le probabilità condizionate $p(y|x_i) := (R = y|S = x_i)$ decideremo che la parola inviata è x_k se

$$p(y|x_k) \geq p(y = x_i) \ \forall i \tag{3.3.1}$$

Nel caso in cui più x_s soddisfino 3.3.1 x_k verrà scelta in modo casuale tra le varie x_s.
Ovviamente la nostra scelta di x_k non ci garantisce che sia stata effettivamente inviata x_k.
Esistono altri principi sui quali basarsi per la scelta di x_k nel caso di codici binari ad esempio si può definire *distanza di Hammning* per aiutarsi nella decisione:

Definizione 3.3.1. date due parole di un codice binario a, b si definisce **distanza di Hamming** il numero di simboli per cui a è differente da b

Utilizzando questa distanza è naturale scegliere come parola x_k inviata quella che dista meno dalla parola ricevuta y.

Teorema 3.3.1. *Per un canale binario simmetrico come quello visto in **Osservazione** 4 dove $0 \leq p < \frac{1}{2}$, fissata una parola y l'insieme $\{x_s\}$ delle parole con distanza di Hamming minima da y coincide con quello delle parole a verosimiglianza massima rispetto a y*

Dimostrazione. Sia m la lunghezza di y, la probabilità che sia stata inviata una parola x tale che $d(x, y) = \epsilon \leq m$ è:

$$\mathbb{P}(Y = y|X = x) = p^{\epsilon}(1-p)^{m-\epsilon} = (1-p)^m \left(\frac{p}{1-p}\right)^{\epsilon}$$

Dato che $0 \leq p < \frac{1}{2} \implies \frac{p}{1-p} < 1$ e quindi $\mathbb{P}(Y = y|X = x)$ ha massimo quando ϵ è minimo. □

Come è già stato accennato in precedenza la scelta della parola inviata, x, non è mai certa e si possono commettere errori, in particolare detto E l'evento *"viene commesso un errore"* chiamiamo $\mathbb{P}(E|S = x_j)$ la probabilità che venga commesso un errore sapendo che è stato inviato x_j.

Definizione 3.3.2. La **probabilità media di errore** è naturalmente definita come:

$$P(E) = \sum_{j=1}^{N} \mathbb{P}(E|x_j)\mathbb{P}(S = x_j)$$

Osserviamo che $\mathbb{P}(E|x_j)$ e di conseguenza anche $P(E)$ dipenderanno dal tipo di regola che adotteremo per ipotizzare chi la parola inviata.
Per semplicità d'ora in avanti utilizzeremo una distribuzione uniforme sull'insieme C delle parole del codice cioè $\mathbb{P}(S = x_j) = \frac{1}{n}$.
Un'importante regola di decisione utilizzata nei canali binari simmetrici si basa sul fatto che avendo ricevuto una parola codice di lunghezza d, si ottiene che, sotto determinate ipotesi, la variabile casuale i cui valori sono il numero di errori commessi è una variabile binomiale. Più rigorosamente presa y la probabilità di aver commesso un errore al $j - esimo$ posto è una variabile casuale X_j di Bernoulli con parametro p e quindi il numero di errori totali commessi in y sarà la somma di d variabili di Bernoulli cioè una variabile casuale $S(d, p)$ con distribuzione binomiale e parametri d, p la cui media è

$$\mathbb{E}[S(d, p)] = dp$$

Per enunciare la nostra regola di decisione pensiamo le parole del codice inviate e ricevute come vettori di dimensione d, si consideri poi una palla $d - dimensionale$ $\mathcal{B}_d(y, d(p + v)$ con v numero arbitrario piccolo a piacere.

Definizione 3.3.3. Ricevuta y diremo che la parola che è stata inviata è x solo se x è l'unica parola all'interno della palla, diremo che è stato commesso un errore se nella palla non sono presenti parole oppure ce ne sono due o più.

In particolare questa regola nel caso in cui $S(d, p) > d(p + v)$ dichiarerà che è stato commesso un errore

3.4 Teorema di Shannon

È chiaro che per noi l'aspetto più importante di un canale comunicativo è la quantità di informazione media che viene effettivamente trasmessa dal canale. Per rendere in modo matematico questo concetto ricordiamo che il nostro canale di comunicazione altro non è che due variabili casuali S, R legate da una legge di probabilità condizionata e come si è visto nella sezione 1.4 l'informazione scambiata tra queste due variabili è data dalle definizione 1.4.4. Riscriviamo quanto detto.

Definizione 3.4.1. Si dice **velocità di trasmissione**, V la quantità media di informazione contenuta in un simbolo dell'alfabeto sorgente che riesce ad essere trasmessa da un canale

$$V := H(R) - H_S(R)$$

Prendiamo il nostro solito canale binario abbiamo quindi che se sono stati emessi n simboli, allora i *bit* di informazione trasmessi saranno $[2^{nV}]$ dove con le parentesi quadre intendiamo approssimare per eccesso all'intero più vicino.
Notiamo che indicata con $H_b(p)$ l'entropia di una variabile casuale di Bernouli di parametro p abbiamo che $H_b(p) = -p\log(p) - (1 - p)\log(1 - p) = H(R|S)$ dove S è una variabile casuale di Bernoulli con parametro ϵ e con un errore del canale pari a p.
Prima di vedere il teorema di Shannon prepariamo il terreno dimostrando un lemma tecnico utile poi per la dimostrazione del teorema.

Lemma 3.4.1. *Sia* $0 \leq p < \frac{1}{2}$ *e* $m \in \mathbb{N}$ *allora vale*

$$\sum_{k=0}^{[mp]} \binom{m}{k} \leq 2^{mH_b(p)} \tag{3.4.1}$$

Dimostrazione.

$$\begin{aligned}1 &= (p+(1-p))^m \\ &= \sum_{k=0}^{[m]} \binom{m}{k} p^k (1-p)^{m-k} \\ &\geq \sum_{k=0}^{[mp]} \binom{m}{k} p^k (1-p)^{m-k} \\ &= (1-p)^m \sum_{k=0}^{[mp]} \binom{m}{k} \left(\frac{p}{1-p}\right)^k\end{aligned}$$

Dato che $0 \leq p < \frac{1}{2}$ allora anche $\left(\frac{p}{1-p}\right) < 1$ e quindi

$$\left(\frac{p}{1-p}\right)^k \geq \left(\frac{p}{1-p}\right)^{mp} \quad \forall 0 \leq k \leq [mp]$$

riprendendo da sopra abbiamo:

$$1 \geq (1-p)^m \left(\frac{p}{1-p}\right)^{mp} \sum_{k=0}^{[mp]} \binom{m}{k}$$

da cui riordinando la disequazione:

$$\sum_{k=0}^{[mp]} \binom{m}{k} \leq [p^{-p}(1-p)^{-(1-p)}]^m = 2^{mH_b(p)}$$

dove l'ultima uguaglianza si ha ricordando che in generale vale $2^{-H(X)} = p_1^{p_1} ... p_n^{p_n}$ □

Rifacendoci a 3.3.3 definiamo A come l'evento in cui non ci sono parole del codice all'interno della palla, B l'evento in cui ve ne sono più di una infine E quello in cui è stato commesso un errore. Chiaramente $E = A \cup B$ ed inoltre

$$\mathbb{P}(E) = \mathbb{P}(A) + \mathbb{P}(B) \tag{3.4.2}$$

essendo $\mathbb{P}(A \cap B) = 0$. Premettiamo al teorema due lemmi che ne renderanno immediata la dimostrazione.

Lemma 3.4.2. *Per ogni fissato $\delta_1 > 0$, scelto d sufficientemente grande vale:*

$$\mathbb{P}(A) \leq \delta_1$$

Dimostrazione. Ricordando cos'è l'evento A troviamo che

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(S(d) > d(p+v)) \\ &= \mathbb{P}(S(d) - dp > dv) \\ &\leq \mathbb{P}(|S(d) - dp| > dv)\end{aligned}$$

Ora applicando la disuguaglianza di Chebyshev otteniamo:

$$\mathbb{P}(A) \leq \frac{p(1-p)}{dv}$$

che conclude la nostra dimostrazione □

Lemma 3.4.3. *Siano ρ e δ_2 due numeri reali non negativi e supponiamo che le parole del codice siano $M = 2^{d(C-\rho)}$ dove $C = 1 - H_b(p)$ è la capacità del canale allora, per d sufficientemente grande vale:*

$$\mathbb{P}(B) \leq \delta_2$$

Dimostrazione. Supponiamo che nella palla $\mathcal{B}(y, r)$ (dove $r = d(p + v)$) ci siano due o più parole. Sia x_i quella con distanza di Hamming da y minore. Abbiamo che

$$\begin{aligned}
\mathbb{P}(B) &= \mathbb{P}\Bigg((x_i \in \mathcal{B}(y,r)) \cup \Bigg(\bigcup_{j=1, j\neq i}^{M} (x_j \in \mathcal{B}(y,r))\Bigg)\Bigg) \\
&\leq \mathbb{P}\Bigg(\bigcup_{j=1, j\neq i}^{M} (x_j \in \mathcal{B}(y,r))\Bigg) \\
&\leq \sum_{j=1, j\neq i}^{M} \mathbb{P}(x_j \in \mathcal{B}(y,r)) \\
&\leq (M-1)\mathbf{P}(x_j \in \mathcal{B}(y,r)) \textit{per alcuni } 1 \leq j \leq M
\end{aligned}$$

Troviamo ora $\mathbb{P}(x_j \in \mathcal{B}(y, r))$. Abbiamo che x_j appartiene a $\mathcal{B}(y, r)$ solo se ha almeno $[r]$ errori e ricordando che la probabilità di avere esattamente k errori è: $\frac{1}{2^d}\binom{d}{k}$ abbiamo che:

$$\mathbb{P}(x_j \in \mathcal{B}(y,r)) = \frac{1}{2^d}\sum_{k=0}^{[r]}\binom{d}{k} \leq \frac{2^{dH_b(p+v)}}{2^d} = 2^{-d(1-H_b(p+v))}$$

quindi unendo questi due ultimi risultati otteniamo:

$$\begin{aligned}
\mathbb{P}(B) &\leq (M-1)2^{-d(1-H_b(p+v))} \\
&\leq M2^{-d(1-H_b(p+v))} \\
&= 2^{d(C-\rho)}2^{-d(1-H_b(p+v))} \\
&= 2^{d(-H_b(p)-\rho}2^{-d(1-H_b(p+v))} \\
&= 2^{d(H_b(p+v)-H_b(p)-\rho)}
\end{aligned}$$

Dato che $H_b(x)$ è una funzione continua crescente per $x < \frac{1}{2}$ possiamo trovare v abbastanza piccolo tale che $H_b(p + v) - H_b(p) < \rho$ in modo che

$$(H_b(p+v) - H_b(p)) - \rho < 0$$

e quindi prendendo d sufficientemente grande possiamo fare in modo che $2^{d((H_b(p+v)-H_b(p))-\rho)} < \delta_2$ permettendoci di concludere. □

Teorema 3.4.1. *(Teorema Fondamentale di Shannon) Dati $\delta, \rho > 0$ possiamo trovare un codice tale per cui se la velocità di trasmissione in un canale binario simmetrico è $R = C - \rho$ allora*

$$\mathbb{P}(E) < \delta$$

Dimostrazione. Ricordando che $E = A \cup B$, il risultato discende direttamente dai due lemmi precedenti □

Seguendo una dimostrazione analoga il teorema è dimostrabile per ogni canale, inoltre è possibile rafforzare il risultato mostrando che è possibile controllare non solo la probabilità media di errore, ma anche quella massima ($\max_{1\leq i\leq M} \mathbb{P}(E|x_i)$).
È stato dimostrato che l'inverso non è possibile cioè non si può avere una probabilità d'errore arbitrariamente piccola se si trasmette ad una capacità superiore a quella del canale.

La grandissima forza del teorema Di Shannon viene però resa effimera dalla mancanza di una dimostrazione costruttiva di tale codice.

4 Conclusioni

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

Bibliografia

[1] David Applebaum. *Probability and: An Integrated Approach*. Cambridge, University Press, second edition edition, 2008.

[2] A. I. Khinchin. *Mathematical Foundations of Information Theory*. Dover, second edition edition, 1957.

[3] S. Mazzucchi. Note di calcolo delle probabilità, 2017.

[4] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 1948.

[5] Joy A. Thomas Thomas M. Cover. *Elements of Information Theory*. Wiley, second edition edition, 2006.

[6] Wikipedia. Jensen's inequality.