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

Ringraziamenti

...thanks to...

Indice

Sommario

« 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)

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 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 di *informazione*.
Per definire in maniera rigorosa il nostro concetto di **informazione** poniamoci in uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$.
Dati due eventi E_1, E_2 vogliamo che la nostra funzione d'informazione I soddifi 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 $I(E_1 \cup E_2) = I(E_1) + I(E_2)$

Con queste richieste ci viene naturalemte in mente una funzione che le soddisfa.

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

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

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 problema 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 desiderabile proprietà di essere nulla qualora la probabilità di un evento sia 1.
Essendo questa funzione spesso associata a codici si preferisce scegliere 2 come base del logaritmo in questo modo supponendo di avere una variabile casuale X con distribuzione di Bernoulli a parametro $p = \frac{1}{2}$ 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 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 possiam conoscere a priori il valore che assumerà X e di conseguenza conoscere 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

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

Per convincersi della sensatezza 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 predirre 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 nel terzo l'indecisione è massima e l'entropia di coneguenza ha anch'essa massimo.
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 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 $H(X) \geq 0$ perchè somma di quantità positive (consideriamo gli addendi come $-log(x)$). Supponiamo ora che $H(X) = 0$ allora $p_j log(p_j) = 0 \forall j$, 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 $p_j \geq 0$(nel caso non lo fossero basterebbe togliere i c $p_k = 0$ e dimostrare che $H(X) \leq log(n-c) \leq log(n)$).
   Dalla definizione abbiamo:
$$H(x) - log(n) =$$
$$= -\frac{1}{ln(2)}\left(\sum_{j=1}^{n} p_j ln(p_j) + ln(n)\right)$$
$$= -\frac{1}{ln(2)}\left(\sum_{j=1}^{n} p_j (ln(p_j) + ln(n))\right)$$
$$= -\frac{1}{ln(2)}\left(\sum_{j=1}^{n} p_j ln(p_j n)\right)$$
$$= \frac{1}{ln(2)}\left(\sum_{j=1}^{n} p_j ln\left(\frac{1}{p_j n}\right)\right)$$
$$\leq \frac{1}{ln(2)}\left(\sum_{j=1}^{n} p_j \left(\frac{1}{p_j n} - 1\right)\right)$$
$$= \frac{1}{ln(2)}\left(\sum_{j=1}^{n} \left(\frac{1}{n} - p_j\right)\right)$$
$$\leq 0$$
   dove nel terzultimo passaggio abbiamo usato il fatto che $ln(x) \leq x - 1$ con l'uguaglianza solo se $X = 1$. Quindi abbiamo che le disuguaglianza si trasformano in uguaglianze solo se $\frac{1}{p_j n} = 1$ cioè se $p_j = \frac{1}{n}$ cioè la 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. Inanzitutto definiamo la *misura di incertezza*:

Definizione 1.3.1. sia $(\Omega, \mathcal{F}, \mathbb{P})$un spazio di probabilità e X variabile casuale 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. presa Y variabile casuale allora $U(X,Y) = U_x(Y) + U(X)$
3. $U(p_1...p_n, 0) = U(p_1...p_n)$
4. $U(p_1....p_n)$ è continua per tutti i suoi argomenti.

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 $K \geq 0$

1.4 Proprietà dell'entropia

Definizione 1.4.1. Siano X e Y due variabili casuali definite sullo stesso spazio di probabilità, definiamo la loro **entropia congiunta** $H(X,Y)$ come:

$$H(X,Y) := -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}log(p_{jk}) \tag{1.4.1}$$

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

Osserviamo subito che $H(X,Y) = H(Y,X)$.
Può essere interessante capire come si comporta l'entropia nel caso le variabili in considerazione siano dipendenti, per fare ciò definiremo l'*entropia condizionata*, prima però un pò di notazione: chiameremo con $p_j(k)$ la probabilità condizionata che $Y = k$ sapendo che $X = j$.

Definizione 1.4.2. La funzione $H_j(Y)$ sarà detta **entropia condizionale di** Y **data** $X = j$ dove

$$H_j(Y) := -\sum_{k=1}^{m} p_j(k)log(p_j(k)) \tag{1.4.2}$$

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

Definizione 1.4.3. definiamo l'**entropia condizionale di** Y **data** X, $H_X(Y)$ come:

$$H_X(Y) := \mathbb{E}[H.(Y)] = \sum_{j=1}^{n} p_j H_j(Y) \tag{1.4.3}$$

Lemma 1.4.1.

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

Dimostrazione. Sostituendo 1.4.2 in 1.4.3 otteniamo

$$H_X(Y) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_j p_j(k)log(p_j(k)) \tag{1.4.5}$$

Ricordando che

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

otteniamo che

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

e possiamo concludere. □

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

$$H_X(Y) = H(Y) \tag{1.4.6}$$

Dimostrazione. supponiamo che la legge di probabilità di Y sia $\{q_1...q_m\}$ allora ci basterà notare che nel caso in cui X e Y sono indipendenti $p_j(k) = \mathbb{P}(Y = k|X = j) = \mathbb{P}(Y = k) = q_k$ e dunque 1.4.5 diventa

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

□

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 $\mathbb{P}(A \cup B) = \mathbb{P}(A|B)\mathbb{P}(B)$ quindi $p_{jk} = p_j p_j(k)$ sostituendo direttamente nella definizione di entropia congiunta 1.4.1 otteniamo

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

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

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

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

Dimostrazione. basta applicare 1.4.6 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 usiamo 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 [5] Ora applicando la disuguaglianza con:

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

per k fissato, otteniamo quindi:

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

dove l'uguaglianza la ricaviamo da: $\sum_{j=1}^{n} p_j p_j(k) = \sum_{j=1}^{n}\Big(\mathbb{P}(X = j)\mathbb{P}(Y = k|X = j)\Big) = \mathbb{P}(Y = k) = q_k$. 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:

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

Questo risultato si può interpretare come il fatto che aggiungendo un informazione (il valore di X) l'entropia del sistema diminuisce. Spesso è preferibile usare la notazione $H(Y|X)$ anzichè $H_X(Y)$ soprattutto nei casi in cui X ha una definizione molto verbosa

Osservazione 1. Nel caso di processi stocastici (si veda 1.6) 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. *Per siano X e Y due variabili casuali con legge di probabilità rispettivamente legge di probabilità $\{p_1...p_n\}\{q_1...q_n\}$*

1. $I(X, Y) = \sum_{j=1}^{n} \sum_{k=1}^{m} p_{jk} log\left(\frac{p_{jk}}{p_j p_k}\right)$
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 1.4.4 otteniamo

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

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

□

1.5 Principio della Massima Entropia

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 massima entropia:

Definizione 1.5.1. Data una una variabile casuale X con legge di probabilità legge di probabilità $\{p_1...p_n\}$incognita il **principio di massima entropia** 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 $\mathbb{E}[X] = E$ troviamo il massimo dell'entropia $H(X)$ utilizzando i 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:

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

dove $\lambda\mu$ sono i moltiplicatori di Lagrange.
Imponendo le derivate parziali ugiali 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. possimo ricaviamo

$$\lambda' = -ln(Z(\mu')) \ dove \ Z(\mu') := \sum_{j=1}^{n} e^{\mu' x_j}$$

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 pre una famiglia di variabili casuali tutta definite sullo stesso spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$, $(X(t), t \geq 0)$ è detta **processto 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

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

Per i nostri scopi inoltre considereremo l'insieme del tempo come come un insieme discreto $\mathbb{N}$.
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 possibilie 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).

Definizione 1.6.3. 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. 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 1.6.4. Siano X e Y due variabili casuali della stessa dimensione. 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*

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

Definizione 1.6.5. 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_∞ 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 uniforma è l'unica distribuzione stazionaria, allora $lim_{n\to\infty} D(X_n, X_\infty) = 0$. Segue quindi che $H(X_n)$ è crescente. [4]

1.7 La Regola della Catena

Vediamo ora come cambia l'informazione in un processo stocastico. Lapprocio più naturale può sembrare quelo di considerare l'entropia come funzione del tempo, in questo modo però ci si dimentica della relazione tra due passaggi dal tempo t_n al tempo t_{n+1}. Estendiamo la definizione di entropia congiunta 1.4.1 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 1.4.2 multivatiata diventa:

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

Non ci rimane che generalizzare il teorema 1.4.1.

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

$$H(X_0 ... X_n) = H(X_0) + \sum_{i=1}^{n} H(X_i | X_0, ... X_{i-1}) = H(X_0) + H(X1|X_0) + ... + H(X_n | X_0 ... X_{n-1}) \tag{1.7.3}$$

Inoltre l'entropia 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. Uindi assumiamo che valga per n, dimostriamo che vale per $n + 1$.

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

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

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

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

Applicando l'ipotesi induttiva otteniamo il risultato. Inoltre da 1.7.4 e dal fatto che l'entropia è sempre maggiore di zero otteniamo che l'entropia cogniunta 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:*

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

Dimostrazione. Applicando la regola della catena 1.7.1 otteniamo subito che

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

Passiamo ora a dimostrare l'esistenza del secondo membro di 1.8.1:
dall'osservazione 1 otteniamo:

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

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:

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

Qundi 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 $a := \lim_{n\to\infty} a_n$:

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

E da qui possiamo concludere scegliendo N_0 tale che $\frac{1}{n}|a_i - a|$ sia piccolo a piacere cosa sempre possibilie 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 inifine 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 preedente 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 Comunicazione

In questo capitolo sarà proposto una modellizzazione della trasmissione di informazione attraverso canali comunicanti.

2.1 Trasmissione di informazione

Il modello più semplice sarà costituito da una sorgenta, un canale di comunicazione, ed un ricevente.
La sorgente sarà modellata con una variabile aleatoria S con valori $\{a_1...a_n\}$detti alfabeto sorgente e legge di probabilità $\{p_1...p_n\}$. Il fatto che la sorgente S sia una variabile casuale va interpretata come l'incertezza su quale sarà il messaggio inviato. in questo contesto un messaggio sarà una serie di simboli da $\{a_1...a_n\}$uno di seguito all'altro. il ricevente sarà un'altra variabile casuale R con valorin $\{b_1...b_m\}$detti alfabeto ricevente e legge di probabilità $\{q_1...q_m\}$. Solitamente avremo che $m \geq n$. Infine l'effetto di distorsione del canale sarà modellato dalla famiglia di probabilità condizionatate $\{p(j|i); 1 \leq i \leq n,\ 1 \leq j \leq m\}$dove $p(j|i) := \mathbb{P}(R = b_j|S = a_i)$ (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.

Definizione 2.1.1. viene detta **mutua informazione** tra due eventi $E(S = a_j)$ ed $F(R = b_k)$ il valore:

$$I(a_j, b_k) = -log(q_k) + log(p(k|j)) \tag{2.1.1}$$

se $p_j = 0$ allora diremo $I(a_j, b_k) = 0$.

È importante notare che questa definizione di mutua informazione è diversa da 1.4.12 data che si riferisce a due variabili casuali.
Dato che $-log(q_k)$ è l'informazione dell'evento $R = b_k$, mentre $-log(p(k|j))$ è l'informazione aggintiva 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. Notiamo che se non ci fosse rumore ($p(i|i) = 1$) avremmo che:

$$I(a_j, b_k) = -log(q_k) = I(q_k)$$

Teorema 2.1.1. *Per ongi* $1 \leq j \leq n, 1 \leq k, \leq m$ *si ha:*

1. $I(a_j, b_k) = -log(\frac{p_j k}{p_j q_k})$
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. Il punto 5. invece esprime la mutua informazione tra due variabili casuali definita in 1.4.12 come la media di tutte le possibili trasmissioni dei singoli simboli. Si può dimostrare che $I(S,R) \geq 0$ sempre.

Supponiamo ora preso un canale, 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 sono $\{p_1 ... p_n\}$.

Definizione 2.1.2. viene definita **capacità del canale C** la quantità:

$$C := maxI(S,R) \tag{2.1.2}$$

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

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

$$C = max(H(R) - H_s(R)) \tag{2.1.3}$$

ottenuta utilizzando la definizione 1.4.12.

2.2 Codici

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

Definizione 2.2.1. L'**alfabeto di un codice, C** è un insieme $\{c_1 ... c_r\}$i cui elementi c_i sono chiamati **simboli**.
Una **parola-codice** 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 quel linguaggio con una parola-codice.
Un esempio pratico di codice che poi utilizzeremo lungo tutto il capitolo è dato dal codice binario. Si ha:

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

Se ad esempio domandassimo che le nostre parole siano tutte di lunghezza 6 o meno allora è facile verificare che ci sono 126 possibili parole-codice.
Il nostro obiettivo sarà ora capire cosa succede all'imformazione trasmessa ora che il percosro sarà:

$$SORGENTE \rightarrow codificatore \rightarrow CANALE \rightarrow decodificatore \rightarrow RICEVENTE$$

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

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

[5] Wikipedia. Jensen's inequality.

Allegato A Titolo primo allegato

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.

A.1 Titolo

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.

A.1.1 Sottotitolo

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.

Allegato B Titolo secondo allegato

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.

B.1 Titolo

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.

B.1.1 Sottotitolo

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.