Teoria dell'Informazione

Simone Alessandro Casciaro

22 Novembre 2024

Lezione 12: Secondo Teorema di Shannon

Nota: i seguenti appunti sono rielaborazioni degli appunti di Sofia Zanelli e Nicolas Lampreda, che ringrazio. Gli appunti potrebbero non essere precisi a causa della mia comprensione sull'argomento.

Riepilogo delle lezioni precedenti

Abbiamo visto che calcolare $C = \max_{p(x)} I(X, Y)$ non è scontato e non esiste un processo meccanico da seguire, che invece è dipendente dalle distribuzioni di X e Y e dalle probabilità della matrice di canale.

Per rendere più facile lavorare con il canale, consideriamo la sua estensione n-esima.

Estensione del Canale

Ricordando la definizione di un canale

$$< \mathbb{X}, \mathbb{Y}, \mathbb{P}(y|x) >$$

consideriamo la sua estensione n-esima definita come

$$< \mathbb{X}^n, \mathbb{Y}^n, \mathbb{P}(y^n|x^n) >$$

dove $p(y^n|x^n)$ è la probabilità di ricevere in uscita una stringa y di n bit sapendo che in input è stata spedita la stringa x di n bit.
Nel caso di un canale con assenza di memoria,

$$p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i)$$

Codice di tipo (M, n)

Definiamo un codice di tipo (M, n) su un canale, con queste caratteristiche:

- M è l'insieme dei messaggi da spedire. $|M| = m$
- n è il numero di volte che il canale $< \mathbb{X}, \mathbb{Y}, \mathbb{P}(y|x) >$ viene utilizzato per trasmettere
- x^n è la funzione di codifica che mappa i messaggi in M in parole del codice appartenenti a $\mathbb{X}^n$.
  $x^n : M \to \mathbb{X}^n$
- g è la funzione di decodifica che prende l'output ricevuto dal canale e lo mappa in uno dei possibili messaggi in M
  $g : \mathbb{Y}^n \to M$

A causa del rumore sul canale, è possibile che la decodifica del messaggio porti ad un errore, dunque introduciamo λ_i che indica la probabilità che g produca un messaggio m_i diverso da quello originariamente spedito.

$$\lambda_i = \mathbb{P}(g(y_i) \neq m_i | X^n = x^n(m_i))$$

Definiamo anche la probabilità massima d'errore

$$\lambda^{(n)} = \max_{i=1,\ldots,m} \lambda_i$$

e la probabilità media d'errore

$$p_e^{(n)} = \frac{1}{m} \sum_{i=1}^{m} \lambda_i$$

Vale la seguente disuguaglianza:

$$\lambda^{(n)} \geq p_e^{(n)}$$

ovvero la probabilità massima d'errore è maggiore o uguale della probabilità media d'errore.

Tasso di Trasmissione

Il tasso di trasmissione di un codice di tipo (M, n) è definito come

$$R = \frac{\log_d m}{n}$$

In questo corso, consideriamo $d = 2$

Il massimo di messaggi che si può spedire usando base 2 è 2^n, dunque l'upper bound per R è

$$R = \frac{\log_2 2^n}{n} = \frac{n}{n} = 1$$

$R = 1$ rappresenta una situazione idilliaca in cui si spedisce n volte il massimo dei messaggi su un canale privo di rumore.
Nella realtà, si calcola il tasso di trasmissione raggiungibile che è generalmente minore di 1.

Il tasso di trasmissione raggiungibile si ottiene se esiste una sequenza di codici di tipo $(2^{\lceil nR \rceil}, n)$ per $n = 1, 2, 3, \dots$ dove n è l'utilizzo del canale e $2^{\lceil nR \rceil} = M$ tale che $\lim_{n \to \infty} \lambda^{(n)} = 0$

R influisce, dunque, sul massimo numero di messaggi che è possibile inviare sul canale.

Legge dei Grandi Numeri

Per ogni sequenza di variabili aleatorie indipendenti e identicamente distribuite $X_1, \dots, X_n$ con valore atteso μ finito, si ha che $\forall \epsilon > 0$:

$$\lim_{n \to \infty} \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \epsilon \right) = 0$$

La legge dei grand numeri indica che la media dei valori attesi delle variabili aleatorie si avvicina sempre più a μ man mano che n cresce.

Proprietà di Equipartizione Asintotica

Per ogni sequenza di variabili aleatorie indipendenti e identicamente distribuite $X_1, \ldots, X_n$ con valore atteso μ finito, si ha che $\forall \epsilon > 0$:

$$\lim_{n \to \infty} \mathbb{P}\left(\left|\frac{1}{n} \log \frac{1}{p(x_1) \ldots p(x_n)} - H(X)\right| > \epsilon\right) = 0$$

Le entropie delle variabili aleatorie sono uguali poiché sono indipendenti e identicamente distribuite.

Insieme Tipico

Definiamo l'insieme tipico:

$$A_e^{(n)} = \left\{(x_1, \ldots, x_n) \in \mathbb{X}_n : 2^{-n(H(X)+\epsilon)} \leq p(x_1) \ldots p(x_n) \leq 2^{-n(H(X)-\epsilon)}\right\}$$

Estraendo in maniera casuale n oggetti, si moltiplicano le probabilità di ciascuno di loro. Se il prodotto è compreso tra il lower bound $A = 2^{-n(H(X)+\epsilon)}$ e l'upper bound $B = 2^{-n(H(X)-\epsilon)}$, allora $x^n \in A_e^{(n)}$

Dimostriamo che se un messaggio x^n appartiene all'insieme tipico, allora detiene la proprietà di equipartizione asintotica.

- Passo 1
  Passiamo al reciproco

$$p(x_1) \ldots p(x_n) = \left(\frac{1}{p(x_1) \ldots p(x_n)}\right)^{-1}$$

- Passo 2
  Si prende la disuguaglianza dell'insieme tipico e si applica a tutti i membri il logaritmo

$$\log_2 2^{-n(H(X)+\epsilon)} \leq \log_2 \left(\frac{1}{p(x_1) \ldots p(x_n)}\right)^{-1} \leq \log_2 2^{-n(H(X)-\epsilon)}$$

$$\log_2 2^{-n(H(X)+\epsilon)} \leq -\log_2 \frac{1}{p(x_1)\dots p(x_n)} \leq \log_2 2^{-n(H(X)-\epsilon)}$$

- Passo 3
  Si semplifica il logaritmo e si divide tutto per n

$$\frac{-n(H(X)+\epsilon)}{n} \leq -\frac{1}{n}\log_2 \frac{1}{p(x_1)\dots p(x_n)} \leq \frac{-n(H(X)-\epsilon)}{n}$$

$$-H(X)-\epsilon \leq -\frac{1}{n}\log_2 \frac{1}{p(x_1)\dots p(x_n)} \leq -H(X)+\epsilon$$

- Passo 4
  Portiamo $H(X)$ al centro della disuguaglianza

$$-\epsilon \leq -\frac{1}{n}\log_2 \frac{1}{p(x_1)\dots p(x_n)} + H(X) \leq \epsilon$$

  Invertiamo il segno della disuguaglianza

$$-\epsilon \leq \frac{1}{n}\log_2 \frac{1}{p(x_1)\dots p(x_n)} - H(X) \leq \epsilon$$

  Applichiamo il modulo

$$\left|\frac{1}{n}\log_2 \frac{1}{p(x_1)\dots p(x_n)} - H(X)\right| \leq \epsilon$$

Dunque, possiamo riscrivere l'insieme tipico come:

$$A_e^{(n)} = \left\{(x_1,\dots,x_n) \in \mathbb{X}^n : \left|\frac{1}{n}\log_2 \frac{1}{p(x_1)\dots p(x_n)} - H(X)\right| \leq \epsilon\right\}$$

Per $n \to \infty$ si ha che:
$\lim\limits_{n\to\infty} \mathbb{P}(x^n \in A_e^{(n)}) = 1$ e $\lim\limits_{n\to\infty} \mathbb{P}(x^n \notin A_e^{(n)}) = 0$
Asintoticamente, le sequenze $x_1,\dots,x_n$ sono tutte ugualmente probabili (infatti le variabili

aleatorie sono tutte i.i.d.) e appartengono all'insieme tipico.

Il lower bound A e l'upper bound B differiscono solo per ϵ, un valore estremamente piccolo, perciò sono fondamentalmente lo stesso valore, perciò la probabilità degli oggetti appartenenti all'insieme tipico sono pari a $2^{-nH(X)}$ per costruzione.

$$\mathbb{P}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i) = \begin{cases} 0 & x^n \notin A_e^{(n)} \\ 2^{-nH(X)} & x^n \in A_e^{(n)} \end{cases}$$

Teorema

Ipotesi

Siano $X_1, \ldots, X_n$ variabili aleatorie indipendenti e identicamente distribuite.
Sia $A_e^{(n)}$ l'insieme tipico ad esse associato

Tesi

$$\forall n \quad \left|A_e^{(n)}\right| \leq 2^{n(H(X)+\epsilon)} \tag{1}$$

$$\exists n_0 : \quad \forall n > n_0 \quad \left|A_e^{(n)}\right| \geq (1-\epsilon)2^{n(H(X)-\epsilon)} \tag{2}$$

In altre parole, l'insieme cresce esponenzialmente con la lunghezza del messaggio n, ma lo fa in modo controllato dall'esponente: ogni parte rimane una piccola frazione del totale delle sequenze possibili.
Anche se l'insieme tipico $A_e^{(n)}$, con l'aumentare di n, contiene la maggior parte delle sequenze possibili, non ha bisogno di includerle tutte. Infatti, dopo un certo numero di simboli n_0, l'insieme tipico diventa abbastanza grande da contenere quasi tutta la probabilità totale, ignorando quelle sequenze che quasi certamente non si verificheranno mai.

Quando si trasmette il messaggio x^n con un canale rumoroso, verrà mappato con messaggi appartenenti all'insieme tipico e che avranno poca distanza (a causa del rumore) dal messaggio originale, mentre si ignorano le sequenze troppo distanti da x^n.

$\star$, per comprendere quali messaggi y_n appartengono all'insieme si considera una dipendenza dal simbolo che è stato spedito:

$$|A_e^n| \approx 2^n H(Y|X)$$

L'obiettivo è che tutti gli insiemi generati in questo modo siano non sovrapposti, per poter permettere una decodifica univoca.
Per ogni insieme, non viene preso il numero totale di elementi $2^{nH(Y)}$, ma solo una sua parte:

$$|M| = \frac{2^n H(Y)}{2^n H(Y|X)} = 2^{n(H(Y)-H(Y|X))} = 2^{nI(X,Y)}$$

si trova dunque l'informazione mutua all'esponente.

In questo caso, si può riscrivere il tasso di trasmissione R in questo modo:

$$R = \frac{\log_2 |M|}{n} = \frac{\log_2 2^{nI(X,Y)}}{n} = \frac{nI(X,Y)}{n} = I(X,Y)$$

Insieme Congiuntamente Tipico

Ricordando le definizioni di insieme tipico per X e Y,

$$\left| \frac{1}{n} \log \frac{1}{p(x_1) \dots p(x_n)} - H(X) \right| < \epsilon$$

$$\left| \frac{1}{n} \log \frac{1}{p(y_1) \dots p(y_n)} - H(Y) \right| < \epsilon$$

L'insieme congiuntamente tipico si definisce come

$$B_e^{(n)} = \left\{ (x^n \times y^n) \in \mathbb{X}^n \times \mathbb{Y}^n : \left| \frac{1}{n} \log \frac{1}{p(x_1, y_1) \dots p(x_n, y_n)} - H(X,Y) \right| < \epsilon \right\}$$

Questo insieme ha due proprietà:

- $\lim_{n \to \infty} \mathbb{P}((x^n \times y^n) \in B_e^{(n)}) = 1$
- Se $\mathbb{X}^n$ ha distribuzione tale per cui $\mathbb{P}(\mathbb{X}^n = x^n) = \prod_{i=1}^{n} p(x_i)$ con $p(x_i)$ probabilità marginale di X e

se $\mathbb{X}^n$ ha distribuzione tale per cui $\mathbb{P}(\mathbb{Y}^n = y^n) = \prod_{i=1}^{n} p(y_i)$ con $p(y_i)$ probabilità marginale di Y
Allora, considerata una coppia $x^n \times y^n \in B_e^{(n)}$:

$$\forall n \geq 1 \quad \mathbb{P}(x^n \times y^n) \leq 2^{-n(I(X,Y)-3\epsilon)}$$

In poche parole l'insieme $B_e^{(n)}$ contiene tutte le sequenze $(x^n \times y^n)$ tali che:

- Le sequenze di $\mathbb{X}^n$ sono tipiche rispetto all'entropia $H(X)$
- Le sequenze di $\mathbb{Y}^n$ sono tipiche rispetto all'entropia $H(Y)$
- Le coppie di sequenze $(x^n \times y^n)$ sono tipiche rispetto all'entropia congiunta $H(X,Y)$

Quindi $B_e^{(n)}$ raccoglie le coppie di sequenze che rispettano sia le distribuzioni marginali per X e Y sia quella congiunta.
Le proprietà invece ci dicono che per n grande quasi tutte le coppie osservabili di sequenze appartengono all'insieme congiuntamente tipico.
Inoltre, viene mostrato quanto è improbabile che le sequenze escano dall'insieme tipico poiché la probabilità che succeda decresce esponenzialmente al crescere di n.

Si garantisce che lavorando con l'insieme tipico si abbia una rappresentazione affidabile e precisa del comportamento delle sequenze originali.

Secondo Teorema di Shannon

Ipotesi

Sia $< \mathbb{X}, \mathbb{Y}, \mathbb{P}(y|x) >$ un canale con capacità C.

Tesi

$\forall R < C \quad \exists k_1, \ldots, k_n$ con $k_i = (2^{nR_n})$ tali che:

$$\lim_{n\to\infty} R_n = R \quad \text{e} \quad \lim_{n\to\infty} \lambda_{k_n}^{(n)} = 0$$

dove $\lambda_{k_n}^{(n)}$ indica la massima probabilità d'errore del canale usando il codice k_n

In altre parole, è possibile scegliere un codice che si avvicina al massimo della capacità del canale (tramite R, ovvero avvicinando il codice al massimo valore di R). Inoltre, così facendo, si "spalma" l'errore sull'informazione trasmessa, perciò all'aumentare di n si riduce al minimo l'errore trasmesso (ricordando che all'aumentare di n le bolle dell'insieme tipico tendono a non sovrapporsi permettendo una decodifica univoca).