Teoria dell'Informazione/RTI **Anno Accademico 2022–2023**

Lezione 1

Ugo Vaccaro

La Teoria dell'Informazione è una disciplina che sta a cavallo tra la Matematica e l'Informatica ed usa sia strumenti classici della Matematica che dell'Informatica (algoritmi).

La Teoria dell'Informazione è nata essenzialmente per risolvere tre principali problemi:

1. Come rappresentare "l'informazione" nella maniera più compatta possibile
2. Come rappresentare "l'informazione" nella maniera più sicura possibile (sicura verso malintenzionati)
3. Come rappresentare "l'informazione" nella maniera più affidabile possibile (in maniera tale che essa non sia corrotta da malfunzionamenti)

Perchè vorremo poter risolvere tali problemi?

Primo problema: come rappresentare "l'informazione" nella maniera più compatta possibile:

1. Per risparmiare spazio/memoria quando la dobbiamo memorizzare
2. Per risparmiare tempo quando la dobbiamo trasmettere/ricevere, dal momento che trasmettere un informazione compressa richiede meno tempo

Esempio 1 *Un minuto di video, non compatto/compresso, richiederebbe* $\approx 51GB$ *di memoria, mentre lo stesso minuto di video compresso richiede* $\approx 27MB$ *di memoria. In questo caso il rapporto tra informazione non compressa e informazione compressa è approssimativamento di 1855 (* $27MB \approx \frac{51}{1855}GB$*). Il vantaggio ottenuto è quindi notevole, infatti con lo stesso dispositivo di memorizzazione capace di memorizzare* $51GB$ *passiamo dal poter memorizzare un minuto di video a* ≈ 31 *ore di video in forma compressa.*

In altre parole la compressione di informazione fa sì che un problema non risolubile, ovvero quello di memorizzare 1 minuto di video in formato non compresso sul nostro cellulare, diventi risolubile.

Per le stesse ragioni, anche la trasmissione e la ricezione di informazione traggono beneficio dalla compressione di informazione. A parità di connessione internet una cosa è trasmettere $\approx 27MB$ per un minuto di video compresso un'altra e trasmettere $\approx 51GB$ per lo stesso minuto di video (ma in forma non compressa).

Secondo problema: come rappresentare "l'informazione" nella maniera più sicura possibile:

- Quando ci scambiamo informazione, persone malintenzionate potrebbero ascoltare ciò che trasmettiamo, ed usarlo per ricavarne vantaggio a nostro danno (es. carte di credito, password, ...)

Per tali ragioni vogliamo quindi essere in grado di poter scambiare informazioni in maniera tale che:

1. Eventuali malintenzionati non posso capire nulla di ciò che trasmettiamo
2. E al tempo stesso, i leggittimi destinari devono poter capire tutto

Terzo problema: come rappresentare "l'informazione" nella maniera più affidabile possibile:

Non tutto funziona sempre come dovrebbe!!

1. Quando memorizziamo delle informazioni e
2. Quando trasmettiamo delle informazioni

possono accadare dei malfunzionamenti/errori che corrompono l'informazione.

Per tali ragioni il problema è più precisamente il seguente; come possiamo trasmettere/memorizzare informazione in modo tale che:

- Eventuali errori occorsi durante la trasmissione/memorizzazione dei dati non impediscano la corretta ricostruzione di ciò che abbiamo trasmesso/memorizzato in origine

La situazione è quindi la seguente, noi vogliamo trasmettere/memorizzare un certo dato A e per via di errori e malfunzionamenti riceviamo B. Vogliamo essere in grado di poter, a partire da B, ricostruire esattamente A. Vederemo che ciò è possibile se rappresentiamo in maniera *opportuna* il dato A.

In un certo senso nel secondo problema avevamo come avversario un malintenzionato interessato a leggere i nostri messaggi, mentre nel terzo abbiamo come avversario la natura.

In sintesi questi sono problemi di grande importanza sia pratica che concettuale. Concettuale perchè: nello sforzo di risolvere questi tre problemi è stato necessario introdurre un ambito concettuale che ha portato alla formalizzazione del concetto di "informazione", della sua rappresentazione e manipolazione. Quindi a introdurre dei concetti/strumenti che hanno applicazioni anche in campi profondamenti lontani da quelli originale per cui la teoria dell'informazione è stata introdotta.

Struttura delle lezioni:

1. In ogni lezione intrurremo un problema con motivazioni reali/pratiche
2. Formuleremo un modello formale/matematico del problema introdotto
3. Vedremo che la formulazione ottenuta è di tipo "Teorico-Informazionale"
4. Useremo quindi i concetti della Teoria dell'Informazione per ottenere la soluzione al problema in questione
5. Spesso, vedremo come la soluzione formale/matematica del problema ha un preciso senso intuitivo

Il concetto di informazione è strettamente dipende dall'osservatore nel senso che una "cosa" può rappresentare informazione per qualcuno, ma non per altri. Assumiamo quindi che vi sia un dispositivo/fenomeno che emetta qualcosa e che vi sia un osservatore interessato a tali emissioni, ovvero che le consideri informazioni. Il dispositivo/fenomeno che emette viene chiamato sorgente (di informazione) che emette simboli di un insieme X che chiameremo alfabeto sorgente. Il primo problema da affrontare è come dare un descrizione formale di una sorgente? Ovvero, come descrivere le regole con cui la sorgente emmette simboli dell'alfabeto sorgente X?

A prescindere da tutto è facile provare che tale descrizione deve essere necessariamente di tipo probabilistico e *non* può essere deterministico, vale a dire che il massimo che possiamo "fare" è dare le probabilità con la sorgente può emettere simboli $x \in X$. Infatti supponiamo che possiamo dare una descrizione deterministica e quindi essere capaci ad ogni istante (futuro) di determinare con esattezza ciò che la sorgente emette o emetterà. Banalmente se ciò fosse vero verrebbe meno in concetto di informazione, in quanto non ha senso ascoltare una sorgente di

cui sappiamo già il comportamento (ovvero ciò che emetterà); quindi, di fatto, di fatto non ci sta dando alcuna informazione che noi non sappiamo già!

Conclusione: possiamo ricevere informazione da una sorgente solo se abbiamo incertezza su quello che la sorgente può emettere. Possiamo quindi dire che l'informazione è conseguenza ad uno stato preliminare di incertezza. Ovverosia, l'informazione è ciò che non si sa a priori e che ci permette di passare da uno stato di incertezza (ignoranza) ad uno stato di non incertezza (di conoscenza). L'informazione è quindi l'opposto di incertezza. Perchè appunto grazie all'informazione passiamo da uno stato di incertezza ad uno stato di conoscenza.

Abbiamo prima osservato che non abbiamo alcuna incertezza apriori su fenomeni (sorgenti) la cui descrizione è deterministica, (e quindi non otteniamo informazione dalla osservazione del loro comportamento) siamo quindi naturalmente portati a considerare solo sorgenti la cui descrizione è probabilistica, perchè solo per esse siamo incerti sulle loro future emissioni.

Per i nostri scopi, una sorgente di informazione è una sequenza di variabili casuali $X_1, X_2, \ldots, X_i, \ldots$ le quali assumono valori in un di un insieme finito X (che chiameremo alfabeto sorgente). Il valore assunto dalla generica variabile casuale X_i corrisponde alla emissione che la sorgente effettuerebbe all'istante i-esimo, per $i = 1, 2, \ldots$.

Esempio 2 *Ad esempio,* $\mathsf{X} = \{a, b, c, \ldots, z\}$ *potrebbe essere l'alfabeto della lingua italiana e potremmo porre,* $\forall x \in \mathsf{X}$ *e* $\forall i \geq 1$ $P\{X_i = x\}$ =*probabilità di occorrenza della lettera* x *nella lingua italiana. Consultando la letteratura, scopriremmo che in tal caso avremmo* $P\{X_i = a\} \approx 11.74/100$*,* $P\{X_i = b\} \approx 0.92/100$*,* $P\{X_i = c\} \approx 4.50/100$*,* $P\{X_i = d\} \approx 3.73/100$*,* $P\{X_i = e\} \approx 11.79/100$*,* $P\{X_i = f\} \approx 0.95/100$*,* $P\{X_i = g\} \approx 1.64/100$*,* $P\{X_i = h\} \approx 1.54/100$*,* $P\{X_i = i\} \approx 11.28/100$*,* $P\{X_i = \ell\} \approx 6.51/100$*,* $P\{X_i = m\} \approx 2.51/100$*,* $P\{X_i = n\} \approx 6.88/100$*,* $P\{X_i = o\} \approx 9.83/100$*,* $P\{X_i = p\} \approx 3.05/100$*,* $P\{X_i = q\} \approx 0.51/100$*,* $P\{X_i = r\} \approx 6.73/100$*,* $P\{X_i = s\} \approx 4.98/100$*,* $P\{X_i = t\} \approx 5.62/100$*,* $P\{X_i = u\} \approx 3.01/100$*,* $P\{X_i = v\} \approx 2.10/100$*,* $P\{X_i = z\} \approx 0.49/100$*, dove il segno* $\approx$ *è da intendersi che i valori prima scritti sono in realtà le frequenze di apparizione dei caratteri della lingua italiana, ma ai fini pratici tali frequenze possono essere considerate, sotto opportune ipotesi, buone approssimazioni delle relative probabilità.*

Effettueremo l'assunzione simplificatrice che le variabili casuali X_i siano indipendenti, ovvero che $\forall k \geq 1$ e $\forall \mathbf{x} = x_1 x_2 \ldots x_k \in \mathsf{X}^k$ valga che

$$P\{X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k\} = \prod_{i=1}^{k} P\{X_i = x_i\},$$

dove X^k =insieme di tutte le sequenze di lunghezza k che si possono costruire sull'alfabeto sorgente X (tecnicamente, X^k corrisponde al prodotto cartesiano $\mathsf{X} \times \mathsf{X} \times \ldots \times \mathsf{X}$ dell'insieme X con se stesso, k volte) . In più, assumeremo che tutte le variabili casuali X_i abbiano la stessa distribuzione P, ovvero

$$\forall x \in X_i \text{ e } \forall i \geq 1 \text{ vale che } P\{X_i = x\} = P(x).$$

In gergo, la sorgente $X_1, X_2, \ldots, X_i, \ldots$ in cui le X_i sono tra di loro indipendenti e hanno tutte la stessa distribuzione $P = \{P(x_1), \ldots, P(x_m)\}$ (dove abbiamo implicitamente assunto che $\mathsf{X} = \{x_1, \ldots, x_m\}$) viene chiamata *sorgente discreta, stazionaria e senza memoria.*

1. Discreta: l'alfabeto sorgente X è un insieme discreto
2. Stazionaria: le probabilità di emissioni non dipendono dall'istante di tempo in cui avviene l'emissione
3. Senza memoria: le probabilità di emissioni non dipendono dalle emissioni precedenti

In queste ipotesi, useremo la seguente compatta notazione per descrivere la sorgente:

$$X = \begin{pmatrix} x_1 & x_2 & \cdots & x_m \\ P(x_1) & P(x_2) & \cdots & P(x_m) \end{pmatrix}$$

e quindi $X_i = X$ per ogni $i \geq 1$. A meno di menzione esplicita, tutte le sorgenti che studieremo in questo corso sono da intendersi discrete, stazionarie e senza memoria, ed useremo l'acronimo DSSS per intendere ciò.

Nozioni e richiami di calcolo delle probabilità

Sia X una variabile casuale (abbreviata con vc), che assume valori in un insieme finito $\mathsf{X} = \{x_1, \ldots, x_n\}$, in accordo alle probabilità $p_1, \ldots, p_n$, con

$$p_i \geq 0 \quad \forall i = 1, \ldots n \quad \text{e} \quad \sum_{i=1}^{n} p_i = 1.$$

In altri termini, la vc X assume valore x_1 con probabilità p_1, assume valore x_2 con probabilità $p_2, \ldots$, assume valore x_n con probabilità p_n. Useremo la seguente notazione per denotare quanto appena detto:

$$\forall i = 1, \ldots n \quad \Pr\{X = x_i\} = p_i,$$

oppure la seguente notazione, con analogo significato

$$X = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ & & & \\ p_1 & p_2 & \cdots & p_n \end{pmatrix} \tag{1}$$

Data la vc in (1), i cui i valori assunti x_i sono numeri reali, definiamo il suo valor medio $E[X]$ come

$$E[X] = \sum_{i=1}^{n} x_i p_i \tag{2}$$

Esempio 3 *Supponiamo di avere la vc*

$$X = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ & & & & & & & \\ 1/2 & 1/4 & 1/8 & 1/16 & 1/32 & 1/64 & 1/128 & 1/256 \end{pmatrix} \tag{3}$$

il suo valor medio sarà pari a

$$E[X] = \sum_{i=1}^{8} i \left(\frac{1}{2}\right)^i = 1.9609375...$$

In alcune situazioni ci sarà utile considerare funzioni di variabili casuali.

Supponiamo di avere una variabile casuale X che assume valori nell'insieme finito $\mathsf{X} = \{x_1, \ldots, x_n\}$, in accordo alle probabilità $p_1, \ldots, p_n$, ovvero per cui $\forall i \quad \Pr\{X = x_i\} = p_i$,. Sia $\mathsf{Y} = \{y_1, \ldots, y_m\}$ e sia f una funzione $f : x \in \mathsf{X} \mapsto f(x) \in \mathsf{Y}$. La vc X e la funzione f naturalmente inducono una nuova vc $Y = f(X)$ che adesso assume valori nell'insieme Y. Con che probabilità la vc $Y = f(X)$ assumerà i valori $y_j \in \mathsf{Y}$? Dipende. Assumiamo, per iniziare, che la funzione f sia iniettiva, ovvero che a valori distinti delle $x \in \mathsf{X}$ associ valori distinti in Y. Detto in simboli, assumiamo per il momento che

$$\forall x, \overline{x} \in \mathsf{X} \quad \text{con } x \neq \overline{x} \quad \text{vale che} \quad f(x) \neq f(\overline{x}). \tag{4}$$

In questo caso, per ogni $y \in \mathsf{Y}$ esiste *un unico* valore $x \in \mathsf{X}$ per cui $f(x) = y$; quindi, la vc $Y = f(X)$ assumerà un dato valore y *se e solo se* la vc X assumerà il corrispondente valore x per cui vale $y = f(x)$. Poichè la vc X assume valore x con probabilità pari a $\Pr\{X = x\}$, otteniamo che

$$\forall y \in \mathsf{Y} \quad \Pr\{Y = y\} = \Pr\{X = x\} \tag{5}$$

dove x è quell'unico valore per cui $f(x) = y$. Di conseguenza, la vc X e la vc $Y = f(X)$ hanno la *stessa* distribuzione di probabilità.

Esempio 4 *Sia*

$$X = \begin{pmatrix} 1 & 2 & 4 & 8 & 16 & 32 & 64 & 128 \\ \\ 1/2 & 1/4 & 1/8 & 1/16 & 1/32 & 1/64 & 1/128 & 1/256 \end{pmatrix} \tag{6}$$

e sia $f(x) = \log_2 x$. *Avremo che*

$$Y = \log_2 X = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \\ 1/2 & 1/4 & 1/8 & 1/16 & 1/32 & 1/64 & 1/128 & 1/256 \end{pmatrix} \tag{7}$$

in quanto Y *vale 0 se e solo se* X *assume valore 1 (e quindi* Y *assumerà il valore 0 con la stessa probabilità con cui* X *assume il valore 1),* Y *vale 1 se e solo se* X *assume valore 2 (e quindi* Y *assumerà il valore 1 con la stessa probabilità con cui* X *assume il valore 2),* Y *vale 2 se e solo se* X *assume valore 4 (e quindi* Y *assumerà il valore 2 con la stessa probabilità con cui* X *assume il valore 4), etc.*

Diversa è la situazione nel caso in cui la funzione $f : x \in \mathsf{X} \mapsto f(x) = y \in \mathsf{Y}$ *non* soddisfa la (4). Consideriamo il seguente

Esempio 5 *Sia*

$$X = \begin{pmatrix} 1 & 2 & 4 & 8 & 16 & 32 & 64 & 128 \\ \\ 1/2 & 1/4 & 1/8 & 1/16 & 1/32 & 1/64 & 1/128 & 1/256 \end{pmatrix} \tag{8}$$

e sia $f(x) = \lfloor \sqrt{x} \rfloor$, *dove denotiamo con* $\lfloor \alpha \rfloor$ *il più grande numero intero che è* $\leq \alpha$. *Consideriamo la vc* $Y = f(X)$. *Avremo che* Y *assumerà valore* 1 *sia quando* X *assume valore* 1 *che quando* X *assume valore* 2 *(infatti* $1 = f(1) = \lfloor \sqrt{1} \rfloor = f(2) = \lfloor \sqrt{2} \rfloor = \lfloor 1.414... \rfloor$.

Di conseguenza

$$\Pr\{Y = 1\} = \Pr\{X = 1\} + \Pr\{X = 2\} = 1/2 + 1/4 = 3/4.$$

Analogamente, avremo che Y *assumerà valore* 2 *sia quando* X *assume valore* 4 *che quando* X *assume valore* 8 *(infatti* $2 = f(4) = \lfloor \sqrt{4} \rfloor = f(8) = \lfloor \sqrt{8} \rfloor = \lfloor 2.828... \rfloor$.

Di conseguenza

$$\Pr\{Y = 2\} = \Pr\{X = 4\} + \Pr\{X = 8\} = 1/8 + 1/16 = 3/16.$$

Continuando, avremo che Y *assumerà valore* 4 *se e solo se* X *assume valore* 16, Y *assumerà valore* 5 *se e solo se* X *assume valore* 32, Y *assumerà valore* 8 *se e solo se* X *assume valore* 64, Y *assumerà valore* 11 *se e solo se* X *assume valore* 128. *Possiamo quindi concludere che la vc casuale* $Y = f(X)$ *è in questo caso pari a*

$$Y = f(X) = \begin{pmatrix} 1 & 2 & 4 & 5 & 8 & 11 \\ \\ 3/4 & 3/16 & 1/32 & 1/64 & 1/64 & 1/256 \end{pmatrix} \tag{9}$$

Dall'esempio, comprendiamo che la regola generale per ottenere le probabilità con cui una vc $Y = f(X)$ assume dati valori è la seguente:

$$\forall y \in \mathsf{Y} \quad \Pr\{Y = y\} = \sum_{x \in \mathsf{X} \text{ tale che } f(x) = y} \Pr\{X = x\}.$$

Siano date due vc X e Y, con X che assume valori nell'insieme $\mathsf{X} = \{x_1, \ldots, x_n\}$ e Y che assume valori nell'insieme $\mathsf{Y} = \{y_1, \ldots, y_m\}$, con le rispettive probabilità di assumere valori date da

$$\forall x_i \in \mathsf{X} \quad \Pr\{X = x_i\} = p(x_i) \qquad \text{e} \qquad \forall y_j \in \mathsf{Y} \quad \Pr\{Y = y_j\} = p(y_j). \tag{10}$$

Denoteremo con XY la vc *congiunta*, che assume valori nell'insieme $\mathsf{X} \times \mathsf{Y} = \{x_i y_j | i = 1, \ldots, n, j = 1, \ldots, m\}$. Detto in altri termini, XY è la vc del tipo

$$XY = \begin{pmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_m & x_2y_1 & x_2y_2 & \cdots & x_2y_m & x_ny_1 & x_ny_2 & \cdots & x_ny_m \\ p(x_1y_1) & p(x_1y_2) & \cdots & p(x_1y_m) & p(x_2y_1) & p(x_2y_2) & \cdots & p(x_2y_m) & p(x_ny_1) & p(x_ny_2) & \cdots & p(x_ny_m) \end{pmatrix} \tag{11}$$

dove

$$\Pr\{XY = x_i y_j\} = p(x_i y_j) \tag{12}$$

denota la probabilità che la vc X assuma il valore x_i *e* la vc Y assuma valore y_j.

La distribuzione di probabilità di una vc congiunta XY soddisfa la regola delle probabilità marginali, ovvero

$$\forall y_j \in \mathsf{Y} \text{ vale che } \sum_{i=1}^{n} \Pr\{XY = x_i y_j\} = \Pr\{Y = y_j\} = p(y_j) \tag{13}$$

e

$$\forall x_i \in \mathsf{X}, \quad \text{e} \quad \sum_{j=1}^{m} \Pr\{XY = x_i y_j\} = \Pr\{X = x_i\} = p(x_i). \tag{14}$$

Spesso, saremo interessati ai valori che una vc X possa assumere, *dato che già sappiamo* che un'altra vc Y abbia assunto determinati valori. Siamo quindi interessati alla vc X, *condizionata* dal fatto che $Y = y_j$, per qualche $y_j \in \mathsf{Y}$. Denoteremo tale vc con il simbolo $X|Y = y_j$. Per ogni fisssato $y_j \in \mathsf{Y}$, per cui $\Pr\{Y = y_j\} > 0$, le probabilità con cui $X|Y = y_j$ assume i valori in $\mathsf{X} = \{x_1, \ldots, x_n\}$ sono notoriamente date da

$$\forall x_i \in \mathsf{X} \quad \Pr\{X = x_i | Y = y_j\} = \frac{\Pr\{XY = x_i y_j\}}{\Pr\{Y = y_j\}}, \tag{15}$$

equivalentemente, possiamo dire che

$$\forall y_j \in \mathsf{Y} \text{ per cui } \Pr\{Y = y_j\} > 0, \ \forall x_i \in \mathsf{X} \quad \text{vale che } \Pr\{XY = x_i y_j\} = \Pr\{Y = y_j\} \Pr\{X = x_i | Y = y_j\}. \tag{16}$$

Dalla (15) è evidente che, per ogni fisssato $y_j \in \mathsf{Y}$, vale

$$\sum_{i=1}^{n} \Pr\{X = x_i | Y = y_j\} = 1.$$

Dalla (15) possiamo dire, per esempio, che il valor medio di X, condizionato dal fatto che $Y = y_j$ (per qualche $y_j \in \mathsf{Y}$) è pari a

$$E[X|Y = y_j] = \sum_{i=1}^{n} x_i \times \Pr\{X = x_i | Y = y_j\}.$$

Introduciamo ora l'importante concetto di *indipendenza* (statistica) di vc. Date le vc X e Y, diremo che esse sono indipendenti se e solo se vale che

$$\forall x_i \in \mathsf{X}, \forall y_j \in \mathsf{Y} \quad \text{vale che} \quad \Pr\{XY = x_i y_j\} = \Pr\{X = x_i\} \Pr\{Y = y_j\}. \tag{17}$$

Analogamente, dalla (15), sotto l'ipotesi che $\Pr\{X = x_i\} > 0$ e $\Pr\{Y = y_j\} > 0$, possiamo dire che X e Y sono indipendenti se

$$\forall x_i \in \mathsf{X}, \forall y_j \in \mathsf{Y} \quad \Pr\{X = x_i | Y = y_j\} = \Pr\{X = x_i\} \quad (\text{ovvero} \quad \Pr\{Y = y_j | X = x_i\} = \Pr\{Y = y_j\}). \tag{18}$$

Intuitivamente, la (18) dice che se X e Y sono indipendenti, la conoscenza del fatto che la vc Y assuma un generico valore y_j *non* influisce sulle probabilità con cui la vc X assume i valori x_i. Se la (18) non vale, allora

diremo che le vc X e Y sono dipendenti (ed in questo caso la conoscenza del fatto che la vc Y abbia assunto un valore y_j *può* influire sulle probabilità con cui la vc X assume i valori x_i). Ovviamente, i concetti sopra esposti di vc congiunte si possono generalizzare a più di due variabili casuali.

Supponiamo di avere n vc $X_1, \ldots, X_n$, ciascuna di esse assume valore nell'insieme X con le stesse probabilità (in altri termini, *è come se* $X_1, \ldots, X_n$ fossero n copie di una stessa vc X che prende valori in X con date probabilità $p(x)$). In tal caso, diremo che $X_1, \ldots, X_n$ sono identicamente distribuite (nel senso che hanno la stessa ditribuzione di probabilità). Denotiamo con X^n l'insieme composto da *tutte* le possibili sequenze lunghe n che possiamo formare con elementi dell'insieme X. Chiaramente, $|\mathsf{X}^n| = |\mathsf{X}|^n$. Diremo che $X_1, \ldots, X_n$ sono indipendenti se e solo se

$$\forall x_{i_1} x_{i_2} \ldots x_{i_n} \in \mathsf{X}^n \quad \text{vale che } \Pr\{X_1 X_2 \ldots X_n = x_{i_1} x_{i_2} \ldots x_{i_n}\} = \Pr\{X_1 = x_{i_1}\} \Pr\{X_2 = x_{i_2}\} \ldots \Pr\{X_n = x_{i_n}\}. \tag{19}$$

Nel caso generale (ovvero, nel caso in cui non necessariamente valga la indipendenza) avremmo che $\forall x_{i_1} x_{i_2} \ldots x_{i_n} \in \mathsf{X}^n$

$$\Pr\{X_1 X_2 \ldots X_n = x_{i_1} x_{i_2} \ldots x_{i_n}\} = \Pr\{X_1 = x_{i_1}\} \Pr\{X_2 = x_{i_2} | X_1 = x_{i_1}\} \ldots \Pr\{X_n = x_{i_n} | X_1 X_2 \ldots X_{n-1} = x_{i_1} x_{i_2} \ldots x_{i_{n-1}}\}. \tag{20}$$

Siano $X_1, \ldots, X_n$ vc indipendenti, identicamente distribuite e con *stesso* valor medio

$$E = \sum_{x \in \mathsf{X}} x \Pr\{X = x\} = E[X_i], \quad \forall i = i, \ldots, n.$$

Vale il seguente importante risultato.

Legge debole dei grandi numeri:

$$\forall \epsilon > 0 \quad \lim_{n \to \infty} \Pr\left\{ \left| \frac{X_1 + \ldots + X_n}{n} - E \right| < \epsilon \right\} = 1, \tag{21}$$

equivalentemente

$$\forall \epsilon > 0 \quad \lim_{n \to \infty} \Pr\left\{ \left| \frac{X_1 + \ldots + X_n}{n} - E \right| > \epsilon \right\} = 0. \tag{22}$$

In altri termini, *la media aritmetica dei valori assunti* dalle vc $X_1, \ldots, X_n$ tende, con alta probabilità, al valore della (comune) media probabilistica. Quindi, la (19) è equivalente a dire che

$$\forall \epsilon > 0 \quad \lim_{n \to \infty} \Pr\left\{ x_{i_1} x_{i_2} \ldots x_{i_n} \in \mathsf{X}^n \text{ per cui vale che: } \left| \frac{x_{i_1} + x_{i_2} + \ldots + x_{i_n}}{n} - E \right| < \epsilon \right\} = 1.$$

ovvero

$$\forall \epsilon > 0 \quad \lim_{n \to \infty} \left(\sum_{\text{su tutti gli } x_{i_1} x_{i_2} \ldots x_{i_n} \in \mathsf{X}^n \text{ per cui vale che: } \left| \frac{x_{i_1} + x_{i_2} + \ldots + x_{i_n}}{n} - E \right| < \epsilon} \Pr\{X_1 X_2 \ldots X_n = x_{i_1} x_{i_2} \ldots x_{i_n}\} \right) = 1.$$

Informalmente, le due ultime espressioni ci dicono che, sotto le ipotesi che $X_1, \ldots, X_n$ siano vc indipendenti, identicamente distribuite e con *stesso* valor medio, per ogni ϵ arbitrariamente piccolo i possibili valori $x_{i_1}, x_{i_2}, \ldots, x_{i_n}$ che le vc $X_1, \ldots, X_n$ tendono ad assumere, *soddisfano* (con alta probabilità al crescere di n) la condizione:

$$\left| \frac{x_{i_1} + x_{i_2} + \ldots + x_{i_n}}{n} - E \right| < \epsilon.$$