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

Lezione 12

Ugo Vaccaro

Nelle applicazioni che vedremo nel seguito del corso avremo bisogno di considerare relazioni tra due o più variabili casuali. In questa lezione introdurremo misure di informazione su più variabili casuali e ne studieremo le proprietà formali.

Siano X e Y due variabili casuali, che assumono valori negli insiemi $\mathsf{X} = \{x_1, \ldots, x_n\}$ e $\mathsf{Y} = \{y_1, \ldots, y_m\}$, rispettivamente. Per ogni $x \in X$ e $y \in \mathsf{Y}$ denoteremo con $p(xy)$ il valore della probabilità congiunta che X assuma valore x e che Y assuma valore y, ovvero $p(xy) = P\{X = x, Y = y\}$. Ovviamente varrà che

$$p(x) = \sum_y p(xy), \quad p(y) = \sum_x p(x,y) \ \text{ e } \ p(x|y) = \frac{p(xy)}{p(y)}, \ \text{ per tutti i valori } y \text{ per cui } p(y) > 0. \tag{1}$$

Nella lezione scorsa abbiamo introdotto una nuova misura, la mutua informazione $I(X;Y)$ tra due vc X e Y, definita come

$$I(X;Y) = D(\mathbf{p}(xy)||\mathbf{p}(x)\mathbf{p}(y)) = \sum_{x\in\mathsf{X}, y\in\mathsf{Y}} p(xy) \log \frac{p(xy)}{p(x)p(y)} \tag{2}$$

dove $D(\cdot||\cdot)$ denota la divergenza informazionale. Abbiamo argomentato sul perchè $I(X;Y)$ rappresenta una ragionevole misura dell'informazione che Y fornisce su X (e viceversa). Abbiamo altresì notato che la mutua informazione possiede le seguenti importanti proprietà:

1. $I(X;Y) \geq 0$ con uguaglianza a 0 se e solo se X e Y sono indipendenti
2. $I(X;Y) = I(Y;X)$
3. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

dove

$$H(X|Y) = \sum_{y\in\mathsf{Y}} p(y) \sum_{x\in\mathsf{X}} p(x|y) \log \frac{1}{p(x|y)} \qquad H(Y|X) = \sum_{x\in\mathsf{X}} p(x) \sum_{y\in\mathsf{Y}} p(y|x) \log \frac{1}{p(y|x)}.$$

Come conseguenza delle proprietà di sopra, abbiamo anche notato che vale la seguente proprietà:

$$H(X|Y) \leq H(X) \tag{3}$$

con uguaglianza se e solo se X e Y sono indipendenti.

Se consideriamo ora la coppia di vc XY come un'unica variabile casuale Z, che assume come valori le coppie $z = (x,y) \in \mathsf{X} \times \mathsf{Y}$ con probabilità $p(z) = p(xy) = p(yx)$, avremo che la sua entropia sarà

$$H(Z) = \sum_z p(z) \log \frac{1}{p(z)} = \sum_{x\in\mathsf{X}, y\in\mathsf{Y}} p(xy) \log \frac{1}{p(xy)} = H(XY). \tag{4}$$

Usando la (1) e note proprietà dei logaritmi, potremo anche scrivere

$$H(XY) = \sum_{x\in\mathsf{X},y\in\mathsf{Y}} p(xy)\log\frac{1}{p(xy)} \tag{5}$$

$$= \sum_{x\in\mathsf{X},y\in\mathsf{Y}} p(xy)\log\frac{1}{p(x)p(y|x)} \tag{6}$$

$$= \sum_{x\in\mathsf{X},y\in\mathsf{Y}} p(xy)\log\frac{1}{p(x)} + \sum_{x\in\mathsf{X},y\in\mathsf{Y}} p(xy)\log\frac{1}{p(y|x)} \tag{7}$$

$$= \sum_{x\in\mathsf{X},y\in\mathsf{Y}} p(xy)\log\frac{1}{p(x)} + \sum_{x\in\mathsf{X}} p(x)\sum_{y\in\mathsf{Y}} p(y|x)\log\frac{1}{p(y|x)} \tag{8}$$

$$= H(X) + H(Y|X). \tag{9}$$

dove abbiamo denotato con $H(Y|X)$ la quantità $\sum_{x\in\mathsf{X},y\in\mathsf{Y}} p(xy)\log\frac{1}{p(y|x)}$. Essa verrà chiamata entropia di Y, condizionata da (la conoscenza) della v.c. X. Osserviamo che per essa ovviamente vale che $H(Y|X) \geq 0$. In più, poichè vale ovviamente

$$H(Y|X) = \sum_{x\in\mathsf{X}} p(x)\sum_{y\in Y} p(y|x)\log\frac{1}{p(y|x)},$$

abbiamo che vale la seguente utile proprietà: $H(Y|X) = 0$ *se e solo se per ogni* $x \in \mathsf{X}$ *la somma*

$$\sum_{y\in Y} p(y|x)\log\frac{1}{p(y|x)} = 0$$

e questo ovviamente accade se e solo se per ogni $x \in \mathsf{X}$ *esiste un unico* $y \in \mathsf{Y}$ *per cui* $p(y|x) = 1$ Detto in altri termini, $H(Y|X) = 0$ se e solo se per ogni valore che la variabile casuale X assume esiste un (*unico*) valore y che la vc Y può assumere. per cui $p(y|x) = 1$. Ovvero, $H(Y|X) = 0$ se e solo se la conoscenza di X determina *univocamente* i valori che la Y può assumere. Più sinteticamente, $H(Y|X) = 0$ se e solo se $Y = f(X)$, per qualche funzione f (la f sarebbe quella funzione che otteniamo associando al generico $x \in \mathsf{X}$ quell'unico valore $y = f(x)$ per cui vale $p(y|x) = 1$). Di conseguenza, abbiamo anche che $H(XY) = H(X) + H(Y|X) = H(X)$ e se solo se la conoscenza di X determina *univocamente* i valori che la Y può assumere.

Facilmente, si può invece vedere che

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

se e solo se per ogni x, y vale che $p(xy) = p(x)p(y)$, ovvero se e solo se le v.c. X e Y sono statisticamente indipendenti. Poichè è immediato che, per definizione, $H(XY) = H(YX)$, le considerazioni di sopra si applicano anche quando i ruoli di X e Y sono invertiti.

La formula (3) ammette la seguente generalizzazione a più di 2 variabili casuali. Ovvero, per ogni tripla di variabili casuali X, Y e Z vale che

$$H(X|YZ) \leq H(X|Y) \text{ e } H(X|YZ) \leq H(X|Z) \tag{10}$$

il cui significato è altrettanto ovvio: più condizionamenti stabiliamo sui possibili valori che X può assumere, e minore (o tutt'al più uguale) è l'incertezza risultante su X. Ovviamente, il ragionamento si può iterare a tre, quattro, ... condizionamenti sulla variabile casuale X. La prova della (10) è analoga a quella della (3). Ne possiamo anche dare una differente prova diretta via la diseguaglianza di Jensen. Infatti, calcolando la differenza $H(X|YZ) - H(X|Y)$ otteniamo:

$$
\begin{aligned}
H(X|YZ) - H(X|Y) &= \sum_{x,y,z} p(xyz) \log \frac{1}{p(x|yz)} - \sum_{x,y} p(xy) \log \frac{1}{p(x|y)} \\
&= \sum_{x,y,z} p(xyz) \log \frac{1}{p(x|yz)} - \sum_{x,y,z} p(xyz) \log \frac{1}{p(x|y)} \\
&= \sum_{x,y,z} p(xyz) \log \frac{p(x|y)}{p(x|yz)} \\
&\leq \log \sum_{x,y,z} p(xyz) \frac{p(x|y)}{p(x|yz)} \quad \text{(applicando la diseguaglianza di Jensen alla funzione log} \\
&\qquad\qquad \text{ed alla vc. che assume valori } \frac{p(x|y)}{p(x|yz)} \text{ con prob. } p(xyz)) \\
&= \log \sum_{x,y,z} \frac{p(x|yz)p(yz)p(x|y)}{p(x|yz)} = \log \sum_{x,y,z} p(yz)p(x|y) \\
&= \log \sum_{x,y} p(x|y) \sum_{z} p(yz) = \log \sum_{x,y} p(x|y)p(y) = \log \sum_{x,y} p(xy) = \log 1 = 0.
\end{aligned}
$$

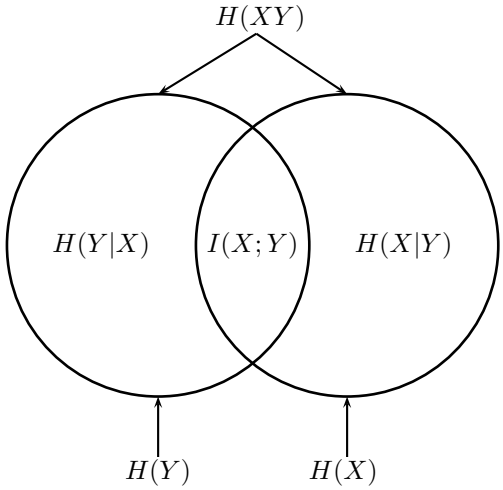
Otteniamo anche, quindi, che la mutua informazione condizionata $I(X;Z|Y) = H(X|Y) - H(X|YZ)$ è sempre non negativa. Le condizioni per l'eguaglianza sono semplici da derivare.

Dalla (9) e dalla (3) otteniamo anche

$$
H(XY) = H(X) + H(Y|X) \leq H(X) + H(Y). \tag{11}
$$

Anche la (11) è intuitiva: la incertezza che abbiamo sulla coppia di v.c. XY non è mai superiore alla somma delle singole incertezza, ed in generale è minore (infatti, è uguale alla somma se e solo se X e Y sono indipendenti, come è giusto che sia).

Il seguente diagramma rappresenta le relazioni tra le quantità $H(X), H(Y), I(X;Y), H(X|Y), H(Y|X), H(XY)$.



Dal diagramma si intuisce una "corrispondenza" tra l'entropia $H(XY)$ e l'unione insiemistica $\mathsf{X} \cup \mathsf{Y}$, tra la mutua informazione $I(X;Y)$ e l'intersezione $\mathsf{X} \cap \mathsf{Y}$, mentre le entropie condizionate $H(Y|X)$ e $H(X|Y)$ corrisponderebbero

alle differenze simmetriche tra insiemi $\mathsf{Y} \setminus \mathsf{X}$ e $\mathsf{X} \setminus \mathsf{Y}$, rispettivamente. La corrispondenza può essere resa precisa e formale in termini di funzioni di misure finite μ su famiglie di insiemi, ma non esploreremo quest'argomento in questa sede. Solo come esempio per illustrare la questione, osserviamo che la ovvia uguaglianza

$$|\mathsf{X} \cup \mathsf{Y}| = |\mathsf{X}| + |\mathsf{Y}| - |\mathsf{X} \cap \mathsf{Y}|$$

dove con $|\cdot|$ denotiamo la cardinalità di un insieme (che è una particolare misura di insiemi finiti) corrisponde alla uguaglianza tra misure di informazione

$$H(XY) = H(X) + H(Y) - I(X;Y).$$

Alcune delle considerazioni di sopra ammettono ovvia generalizzazione a più di due variabili casuali. Ad esempio, se $X_1, X_2, \ldots X_n$ sono generiche v.c. finite allora la (11) si generalizza a

$$H(X_1 X_2 \ldots X_n) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|H_{n-1} \ldots X_1) \leq H(X_1) + \ldots + H(X_n) \tag{12}$$

con ovvia solita interpretazione. La prova della (12) è immediata una volta che ci si ricordi della proprietà $p(x_1 x_2 \ldots x_n) = p(x_1) \cdot p(x_2|x_1) \cdot \ldots \cdot p(x_n|x_{n-1} \ldots x_1)$, che ogni distribuzione di probabilità congiunta $p(x_1 x_2 \ldots x_n)$ gode. Ovviamente varrà anche

$$H(X_1 X_2 \ldots X_n|Z) = H(X_1|Z) + H(X_2|X_1 Z) + \ldots + H(X_n|H_{n-1} \ldots X_1 Z) \leq H(X_1|Z) + \ldots + H(X_n|Z) \tag{13}$$

Vediamo una semplice conseguenza delle relazioni tra le quantità prima introdotte. Sia X una variabile casuale che assume valori in $\mathsf{X} = \{x_1, \ldots, x_n\}$ ed $f : \mathsf{X} \mapsto \mathsf{Y} = \{y_1, \ldots, y_m\}$, con $m \leq n$, una funzione arbitraria. Sia $Y = f(X)$ la variabile corrispondente, ovvero la v.c. che assume valori y in Y con le seguenti probabilità

$$P\{Y = y\} = \sum_{x : f(x) = y} P\{X = x\}.$$

Allora possiamo provare

$$H(f(X)) = H(Y) \leq H(X) \tag{14}$$

con eguaglianza se e solo se la funzione f è biettiva. Infatti, dalla definizione di mutua informazione $I(X;Y)$ otteniamo che

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(Y). \tag{15}$$

L'ultima uguaglianza discende dal fatto che conoscendo X (ovvero i valori input della funzione f ne conosciamo ovviamente i valori $f(X) = Y$, per cui l'incertezza $H(Y|X)$ su Y, nota la vc X è nulla, come prima provato. Dalla non negatività della entropia condizionata $H(X|Y)$ otteniamo quindi che $H(X) = H(Y) + H(X|Y) \geq H(Y)$. L'uguaglianza varrà se e solo se anche $H(X|Y)$ è nulla, e ciò accadra se e solo conoscendo Y la nostra incertezza su X è nulla, ovverosia se e solo se conoscendo i valori della funzione $f(X)$ possiamo risalire ai valori di X. In altri termini, se f è biettiva, e quindi invertibile.

Allo stesso modo possiamo provare che

$$H(X|Y) = 0 \Rightarrow H(Y) \geq H(X). \tag{16}$$

Prima di provare la (16) osserviamo che il suo contenuto intuitivo è perfettamente ragionevole. Ricordiamo che $H(X|Y) = 0$) se e solo se la v.c. Y determina completamente X, ovvero dalla conoscenza di Y possiamo risalire a X. In altri termini, la conoscenza di Y azzera la incertezza che abbiamo su X: Ma ciò può accadere solo se la informazione che riceviamo da Y (ovvero $H(Y)$) è maggiore della incertezza che abbiamo su X (ovvero $H(X)$). Formalmente abbiamo

$$I(X;Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) = H(X) \tag{17}$$

da cui

$$H(Y) = H(X) + H(Y|X) \geq H(X).$$

Vediamo un'applicazione delle considerazioni prima fatte. Innanzitutto diamo la seguente definizione.

Definizione 1 *Diremo che le variabili casuali X, Y, Z formano una catena di Markov in tale ordine (e scriveremo $X \to Y \to Z$) se se la distribuzione di probabilità condizionata di Z dipende solo da Y e non dipende da X. Detto in altri termini, le variabili casuali X, Y, Z formano una catena di Markov $X \to Y \to Z$ se per ogni x, y, z vale che*

$$p(xyx) = p(x)p(y|x)p(z|y). \tag{18}$$

Semplici conseguenze della definizione sono:

- $X \to Y \to Z$ se e solo se X e Z sono condizionatamente indipendenti, dato Y. Infatti, se la (18) vale, allora si ha

$$p(xz|y) = \frac{p(xyx)}{p(y)} = \frac{p(xy)p(z|y)}{p(y)} = p(x|y)p(z|y). \tag{19}$$

- $X \to Y \to Z$ implica $Z \to Y \to X$
- se $Z = f(Y)$, per qualche funzione f, allora vale che $X \to Y \to Z$.

Possiamo provare il seguente importante risultato.

Teorema 1 *(Data Processing). Se $X \to Y \to Z$ allora $I(X;Y) \geq I(X;Z)$.*

Dimostrazione. Sappiamo che la mutua informazione $I(X;YZ)$ si può scrivere nei due modi equivalenti

$$\begin{aligned} I(X;YZ) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y). \end{aligned}$$

Poichè vale che $X \to Y \to Z \Rightarrow p(xz|y)p(x|y)p(z|y)$, ovvero X e Z sono condizionatamente indipendenti dato Y, vale che $I(X;Z|Y) = 0$. Dal fatto che $I(X;Y|Z) \geq 0$ otteniamo che $I(X;Y) \geq I(X;Z)$. É facile vedere che vale l'uguaglianza se e solo se $I(X;Y|Z) = 0$, ovvero se $X \to Z \to Y$ formano una catena di Markov. Analogamente, si può provare che $I(Y;Z) \geq I(X;Z)$.

□

Corollario 1 *Se $Z = f(Y)$, allora abbiamo che $I(X;Y) \geq I(X;f(Y))$.*

Dimostrazione. Basta osservare che $X \to Y \to f(Y)$ forma una catena di Markov.

□

Il precedente corollario afferma, in altri termini, che una qualunque funzione/elaborazione $f(Y)$ dei dati Y non può incrementare l'informazione che Y stessa fornisce su X (e, in generale, la può solo diminuire).

Un'altra forma con cui questo fenomeno può essere quantificato è attraverso la divergenza informazionale. Si può provare (attraverso la Diseguaglianza di Jensen) che per ogni coppia di v.c. X e Y vale che

$$D(f(X)||f(Y)) \leq D(X||Y)),$$

dove f è una qualsiasi funzione, deterministica o stocastica. Essenzialmente, la diseguaglianza di sopra afferma che quando si osservano gli esiti delle v.c. X e Y attraverso una osservazione indiretta, fornita da f, si perde in potere "discriminatorio" rispetto a quando si osservano gli esiti delle v.c. X e Y *direttamente*.

Supponiamo che conosciamo una v.c. Y e vogliamo stimare il valore di una v.c. X ad essa correlata. Sappiamo che possiamo conoscere il valore esatto di X se e solo se X è una funzione deterministica di Y, ovvero se e solo se $H(X|Y) = 0$. In altri termini, avremo "zero" probabilità di errore nella stima di X se e solo se l'incertezza su X, noto Y, è pari a zero. Cosa accade se $H(X|Y) > 0$? Ragionevolmente, ci aspettiamo che la probabilità di errore nella stima del valore di X, noto Y, sia essa stessa maggiore di zero sarà più grande al crescere di $H(X|Y)$. Cerchiamo di rendere più precisa questa intuizione.

La nostra ipotesi è che vogliamo stimare i valori di una v.c. X (che assume valori $x \in \mathcal{X}$ con probabilità $p(x)$), osservando i valori di una v.c. Y che è correlata a X mediante le probabilità condizionate $p(x|y)$. Da Y calcoliamo una v.c. $\hat{X} = g(Y)$ che sarà la nostra stima di X. Notiamo che la g può essere una qualunque regola, sia deterministica che probabilistica. Vogliamo limitare la probabilità che $\hat{X} \neq X$. Osserviamo innanzitutto che $X \to Y \to \hat{X}$ forma una catena di Markov. Definiamo la probabilità di errore P_e come

$$P_e = \Pr\{\hat{X} \neq X\}.$$

Vale il seguente importante risultato, che và sotto il nome di Diseguaglianza Fano, dal nome dello scopritore

Teorema 2 *Per ogni stima $\hat{X}$ tale che $X \to Y \to \hat{X}$ è una catena di Markov, vale che*

$$\mathcal{H}(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y), \tag{20}$$

dove $\mathcal{H}(P_e) = -P_e \log P_e - (1 - P_e) \log(1 - P_e)$.

Prima di darne la prova, osserviamo che la (20) implica la seguente più debole (ma più semplice da utilizzare) formulazione seguente:

$$1 + P_e \log |\mathcal{X}| \geq H(X|Y), \tag{21}$$

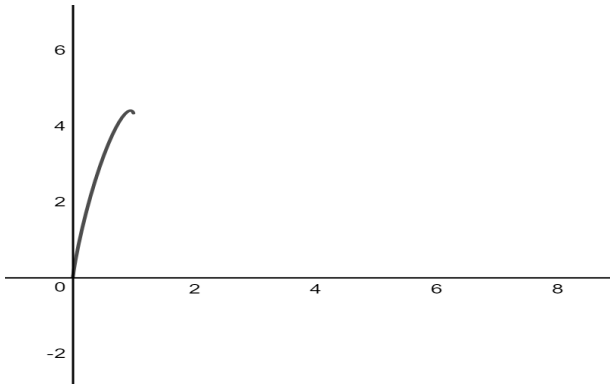
ovvero

$$P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}. \tag{22}$$

Nella figura seguente, si può vedere il grafico di $\mathcal{H}(P_e) + P_e \log 20$ in funzione di P_e, (per $|\mathcal{X}| \neq 20$ il comportamento è ovviamente simile) da cui si evince che $P_e = 0$ se e solo se $H(X|Y) = 0$ (e quindi necessariamente $H(Y) \geq H(X)$, come abbiamo già visto) e la probabilità di errore cresce al crescere di $H(X|Y)$, *qualunque* sia la regola utilizzata $\hat{X} = g(Y)$ per stimare X a partire da Y.

Dimostrazione. Proviamo inannzitutto che

$$\mathcal{H}(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}). \tag{23}$$



Sia E la v.c. così definita:

$$E = \begin{cases} 1 & \text{se } X \neq \hat{X} \\ 0 & \text{se } X = \hat{X}. \end{cases}$$

É chiaro che $H(E) = \mathcal{H}(P_e)$. Espandendo l'entropia $H(EX|\hat{X})$ in due modi otteniamo allora

$$\begin{aligned} H(EX|\hat{X}) &= H(X|\hat{X}) + H(E|X\hat{X}) \\ &= H(X|\hat{X}) \quad \text{(in quanto conoscendo } X \text{ e } \hat{X} \text{ si conosce anche } E \text{ e quindi } H(E|X\hat{X}) = 0) \\ &= H(E|\hat{X}) + H(X|E\hat{X}) \end{aligned}$$

Osserviamo ora che $H(E|\hat{X}) \leq H(E) = \mathcal{H}(P_e)$. Inoltre

$$\begin{aligned} H(X|E\hat{X}) &= \Pr\{E=0\} \cdot H(X|\hat{X}E=0) + \Pr\{E=1\} \cdot H(X|\hat{X}E=1) \\ &= \Pr\{E=1\} \cdot H(X|\hat{X}E=1) \quad \text{(in quanto se } E=0 \text{ allora } X=\hat{X} \text{ e quindi } H(X|\hat{X}E=0)=0) \\ &= P_e H(X|\hat{X}E=1) \qquad \text{(in quanto } \Pr\{E=1\} = \Pr\{X \neq \hat{X}\} = P_e) \\ &\leq P_e \log|\mathcal{X}| \qquad \text{(in quanto } H(X|\hat{X}E=1) \leq H(X) \leq \log|\mathcal{X}|) \end{aligned}$$

Mettendo tutto insieme otteniamo che

$$H(X|\hat{X}) = H(E|\hat{X}) + H(X|E\hat{X}) \leq \mathcal{H}(P_e) + H(X|E\hat{X}) \leq \mathcal{H}(P_e) + P_e \log|\mathcal{X}|,$$

e quindi la (23) è provata.

Per provare che $H(X|\hat{X}) \geq H(X|Y)$, usiamo il Teorema 1. Infatti, $X \to Y \to \hat{X}$ è una catena di Markov, per cui

$$H(X) - H(X|\hat{X}) = I(X;\hat{X}) \leq I(X;Y) = H(X) - H(X|Y),$$

che implica, appunto, $H(X|\hat{X}) \geq H(X|Y)$.

□

Possiamo riscrivere la (20) nell'ipotesi che la v.c. X assuma valori in $\mathcal{X}$ con probabilità uniforme, ovvero sotto l'ipotesi che $H(X) = \log|\mathcal{X}|$. Osserviamo che la (20) implica che

$$P_e \geq \frac{H(X|Y) - \mathcal{H}(P_e)}{\log|\mathcal{X}|},$$

ovvero

$$\begin{aligned} 1 - P_e &\leq 1 - \frac{H(X|Y) - \mathcal{H}(P_e)}{\log|\mathcal{X}|} \\ &= \frac{\log|\mathcal{X}| - H(X|Y) + \mathcal{H}(P_e)}{\log|\mathcal{X}|} \\ &= \frac{H(X) - H(X|Y) + \mathcal{H}(P_e)}{\log|\mathcal{X}|} \\ &= \frac{I(X;Y) + \mathcal{H}(P_e)}{\log|\mathcal{X}|}, \end{aligned}$$

da cui

$$(1 - P_e)\log|\mathcal{X}| - \mathcal{H}(P_e) \leq I(X;Y). \tag{24}$$

La (24) ha la seguente intuitiva interpretazione: se vogliamo stimare X, conoscendo $\hat{X} = g(Y)$, con probabilità di errore $P_e = \Pr\{X \neq \hat{X}\} = \Pr\{X \neq g(Y)\}$ piccola, allora necessariamente $1 - P_e$ è grande. Di conseguenza anche $I(X;Y)$ deve essere grande, ovvero per avere P_e *piccola*, dobbiamo necessariamente disporre di una Y che deve fornire *molta* informazione su X.

Come interessante applicazione delle formule (12) e (16), diamo una dimostrazione basata su entropia del ben noto fatto, noto già ad Euclide nel periodo 300 A.C., che esiste un'infinità di numeri primi. Per ogni intero positivo n, siano $p_1, \ldots, p_{\pi(n)}$ tutti e solo i primi nell'insieme $\{1, 2, \ldots, n\}$. Sia X la variabile casuale che assume valori interi compresi tra 1 ed n con probabilità pari a $1/n$. Ovviamente, $H(X) = \log n$. Dato un generico valore $a \in \{1, \ldots, n\}$ assunto dalla variabile casuale X, esprimiamo a nella sua unica rappresentazione nel modo seguente:

$$a = b^2 p_1^{x_1} p_2^{x_2} \cdots p_{\pi(n)}^{x_{\pi(n)}}, \tag{25}$$

dove b è il più grande intero tale che b^2 divide a e

$$p_1^{x_1} p_2^{x_2} \cdots p_{\pi(n)}^{x_{\pi(n)}} \tag{26}$$

è la (unica!) fattorizzazione dell'intero a/b^2 in prodotto dei primi $p_1, p_2, \cdots, p_{\pi(n)}$. Osserviamo che i valori $x_1, \ldots, x_{\pi(n)} \in \{0, 1\}$, in quanto, per definizione di b, non esiste alcun primo p_i tale che p_1^2 divide a/b^2.

Le espressione (25) e (25) definiscono una nuova collezione di variabili casuali $B, X_1, \ldots X_{\pi(n)}$. Detto in altri termini, per ogni valore $a \in \{1, \ldots, n\}$ (probabilisticamente) assunto dalla vc X, B assume come valore il più grande intero b per cui b^2 divide a, X_1 assume valore 1 se e solo se il primo p_1 divide a/b^2, assume valore 0 altrimenti; X_2 assume valore 1 se e solo se il primo p_2 divide a/b^2, assume valore 0 altrimenti; etc. etc.

Visto che, per ogni $i = 1, \ldots, \pi(n)$, vale che ogni variabile casuale X_i assume valori nell'insieme $\{0, 1\}$, otteniamo che $H(X_i) \leq \log 2 = 1$, per ogni $i = 1, \ldots, \pi(n)$, mentre la vc B assume valori in $\{1, 2, \ldots, \lfloor\sqrt{n}\rfloor\}$, per cui $H(B) \leq \log\lfloor\sqrt{n}\rfloor \leq \log\sqrt{n} = \frac{1}{2}\log n$.

Osserviamo ora, che se conosciamo i valori $b, x_1, \ldots, x_{\pi(n)}$ assunti dalle vc $B, X_1, \ldots, X_{\pi(n)}$, conosciamo *con certezza* il valore $a \in \{1, \ldots, n\}$ assunto dalla variabile casuale X. Infatti, a sarà uguale al prodotto $b^2 p_1^{x_1} p_2^{x_2} \cdots p_{\pi(n)}^{x_{\pi(n)}}$. Per quanto detto prima, ciò comporta che $H(X|B, X_1, \ldots, X_{\pi(n)}) = 0$.

Possiamo quindi dire che

$$\begin{aligned}\log n &= H(X)\\ &\leq H(BX_1X_2\dots X_{\pi(n)})\\ &\leq H(B)+H(X_1)+H(X_2)+\dots+H(X_{\pi(n)})\\ &\leq \frac{1}{2}\log n+\pi(n).\end{aligned}$$

La prima diseguaglianza la si ottiene dal fatto che le variabili casuali $B, X_1, X_{\pi(n)}$ determinano *univocamente* la variabile casuale X, per cui $H(X|BX_1X_2\dots X_{\pi(n)})=0$, e dalla (16) ciò implica $H(BX_1X_2\dots X_{\pi(n)})\geq H(X)$. La seconda diseguaglianza è conseguenza della (12). Quindi otteniamo che

$$\pi(n)\geq \frac{1}{2}\log n, \qquad \forall n\geq 2,$$

da cui ovviamente discende che $\pi(n)\to\infty$ al crescere di n.

Vediamo un'altra applicazione della (12) e (16) Supponiamo di avere una popolazione di n individui $\mathsf{X}=\{x_1,x_2,\dots,x_n\}$, ed al più d di essi sono "infetti". Sia $P\subset\mathsf{X}$, $|P|\leq d$ tale insieme *incognito* di elementi infetti. Possiamo tentare di scoprire l'insieme incognito $P\subseteq\mathsf{X}$ degli individui infetti modo seguente: scegliamo a nostro piacimento un insieme $A\subseteq\mathsf{X}$, ed effettuiamo un test $T(A)$ su di esso che ci restituisce la seguente risposta:

$$T(A)=\begin{cases}1 & \text{se } A\cap P\neq\emptyset\\ 0 & \text{altrimenti}\end{cases} \tag{27}$$

Il problema consiste nel determinare esattamente chi è P, effettuando "pochi" test del tipo descritti in (27). Ovvero, possiamo scegliere una sequenza di sottoinsiemi $A_1,\dots,A_t\subset\mathsf{X}$, possiamo effettuare i test $T(A_1),\dots,T(A_t)$ e dalla conosceza dei risultati (secondo la (27)) dobbiamo essere in grado di scoprire chi è l'insieme incognito di elementi infetti P.

Situazioni del genere sorgono in moltissimi campi, dall'analisi di malattie al controllo di qualità industriale, dallo scrrening del DNA alla comunicazione via radio, e tante altri. Ad esempio, possiamo considerare la seguente (reale!) situazione. Per scoprire chi sono le persone infette, si può prelevare da ciascuna di esse un campione di sangue. Invece che fare, successivamente, l'analisi individuale di ciascun campione di sangue (per scoprire chi è malato e chi no), nel caso in cui il numero di individui n è molto grande, si preferisce usare uno schema di questi tipo: si riunisce in un'unica provetta un pò di sangue di un gruppo $A\subseteq\mathsf{X}$ di persone, e si testa tale unica provetta. Se il risultato è negativo, sappiamo con certezza che *nessuna* delle persone in A è malata, e ciò lo possiamo stabilire con un *unico* test, il che rappresenta un bel pò di risparmio rispetto alla situazione in cui testavamo, sigolarmente, tutte le persone una ad una. Ovviamente, se invece il risultato dell'analisi è postivo, occorre continuare a cercare, ovvero scegliere un altro insieme di persone $B\subseteq\mathsf{X}$, raggruppare un pò di sangue da ciascuna persona in B, analizzarlo, etc. etc. In ogni caso, si può provare che questo metodo di testare "insieme" gruppi di persone porta ad un risparmio sul numero di test totali da fare. Se ci si riflette un pò, questo metodi di test di "gruppi" corrisponde esattamente alla situazione descritta dalla formula (27), una volta che si interpreti l'1 della (27) come "test positivo" (ovvero A contiene almeno un infetto di P), e con lo 0 della (27) come "test negativo" (ovvero A non contiene al suo interno nessuno degli infetti di P). Vediamo una limitazione a tale modo di procedere.

Sia X la variabile casuale che assume come valori i possibili sottoinsiemi di elementi infetti, ciascheduno con equale probabilità. Essendo tali insiemi in numero di $\sum_{k=0}^{d}\binom{n}{k}$, ne segue che X è una v.c. che assume $\sum_{k=0}^{d}\binom{n}{k}$ distinti valori, ciascuno con probabilità $1/\sum_{k=0}^{d}\binom{n}{k}$, per cui $H(X)=\log\sum_{k=0}^{d}\binom{n}{k}$. Siano ora $Y_1,Y_2,\dots,Y_t$ le variabili casuali che assumono come valori i risultati di t test che un *arbitrario* algoritmo effettua. Se la conoscenza dei risultati di tali t test ci permette di scoprire l'insieme incognito $P\subseteq\mathsf{X}$ degli individui infetti, deve necessariamente valere che $H(X|Y_1Y_2\dots Y_t)=0$, (ovvero l'incertezza su chi è l'insieme degli infetti deve essere nulla) da cui, per

la (16), sappiamo che $H(Y_1 Y_2 \dots Y_t) \geq H(X)$. Usando la (12) ed il fatto che le Y_i sono variabili casuali binarie (e che pertanto $H(Y_i) \leq \log 2 = 1$) otteniamo che

$$\log \sum_{k=0}^{d} \binom{n}{d} = H(X) \leq H(Y_1 Y_2 \dots Y_t) \leq H(Y_1) + H(Y_2) + \dots + H(Y_t) \leq t.$$

Abbiamo quindi provato che *ogni* algoritmo che risolve il problema sopra descritto deve necessariamente effettuare un numero di test t tale che

$$t \geq \log \sum_{k=0}^{d} \binom{n}{k}.$$