(F1X) Teoria dell'Informazione e della Trasmissione

8 – Derivati dell'entropia

Docente: *Nicolò Cesa-Bianchi* versione 23 marzo 2016

Siano X e Y due variabili casuali con valori in insiemi finiti $\mathcal{X}$ e $\mathcal{Y}$. Detta p la loro distribuzione congiunta $p(x, y) = \mathbb{P}(X = x,\, Y = y)$, definiamo l'entropia congiunta $H(X, Y)$ come

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x, y) \ .$$

Possiamo pensare a $H(X, Y)$ come al numero medio di bit necessari per rappresentare una coppia di valori estratti da X e Y. In modo simile definiamo l'entropia condizionale $H(Y \mid X)$ nel modo seguente

$$\begin{aligned} H(Y \mid X) &= \sum_{x \in \mathcal{X}} p(x) H(Y \mid X = x) = \sum_{x \in \mathcal{X}} p(x) \left(-\sum_{y \in \mathcal{Y}} p(y \mid x) \log_2 p(y \mid x) \right) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(y \mid x) \end{aligned}$$

dove $p(x)$ è la marginale su X,

$$p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$$

e $p(y \mid x)$ è la condizionata di y su x,

$$p(y \mid x) = \frac{p(x, y)}{p(x)} \qquad p(x) > 0 \ .$$

Vediamo ora una semplice relazione fra entropia, entropia congiunta ed entropia condizionale.

Teorema 1 (Chain rule per entropia)

$$H(X, Y) = H(X) + H(Y \mid X) \ .$$

Dimostrazione.

$$\begin{aligned} H(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x, y) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \Big(\log_2 p(x) + \log_2 p(y \mid x) \Big) \\ &= -\sum_{x \in \mathcal{X}} \left(\sum_{y \in \mathcal{Y}} p(x, y) \right) \log_2 p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(y \mid x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) + H(Y \mid X) \\ &= H(X) + H(Y \mid X) \ . \end{aligned}$$

□

Notiamo, cosa importante, che questa e le prossime identità valgono anche in spazi condizionati, ovvero possiamo scrivere $H(X, Y \mid Z) = H(X \mid Z) + H(Y \mid X, Z)$ dove Z è una qualunque variabile casuale.

La prossima quantità che introduciamo è l'informazione mutua $I(X, Y)$ fra due variabili casuali X e Y,

$$I(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \ .$$

Si noti che l'informazione mutua corrisponde all'entropia relativa fra la congiunta e il prodotto delle marginali. Quindi sappiamo che $I(X, Y) \geq 0$.

Vediamo ora un risultato che lega l'informazione mutua all'entropia e all'entropia condizionale.

Teorema 2

$$I(X, Y) = H(X) - H(X \mid Y) \ .$$

Dimostrazione.

$$\begin{aligned} I(X, Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 \frac{p(y)p(x \mid y)}{p(x)p(y)} \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x) + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x \mid y) \\ &= H(X) - H(X \mid Y) \ . \end{aligned}$$

□

Questo risultato ci permette di interpretare l'informazione mutua come una decrescita di entropia, ovvero come il numero di bit che il valore di Y fornisce in media riguardo il valore di X. D'altra parte, la non negatività di $I(X, Y)$ implica anche che $H(X \mid Y) \leq H(X)$, ovvero il condizionamento su Y decresce l'entropia di X.

Vediamo ora il calcolo dell'informazione mutua in alcuni semplici casi particolari. Se X e Y sono indipendenti, allora $H(X \mid Y) = H(X)$ e quindi $I(X, Y) = 0$. Viceversa, se $X = g(Y)$ per una qualche funzione g, allora $H(X \mid Y) = 0$ e quindi $I(X, Y) = H(X)$. Per esempio, se g è la funzione identica, ovvero $X = Y$, allora $I(X, X) = H(X)$.

Infine, possiamo usare la chain rule per l'entropia, $H(X, Y) = H(Y) + H(X \mid Y)$, per dimostrare che

$$I(X, Y) = H(X) - H(X \mid Y) = H(X) + H(Y) - H(X, Y) \ .$$

Ricapitoliamo in forma rielaborata le relazioni dimostrate fra le entropie e l'informazione mutua

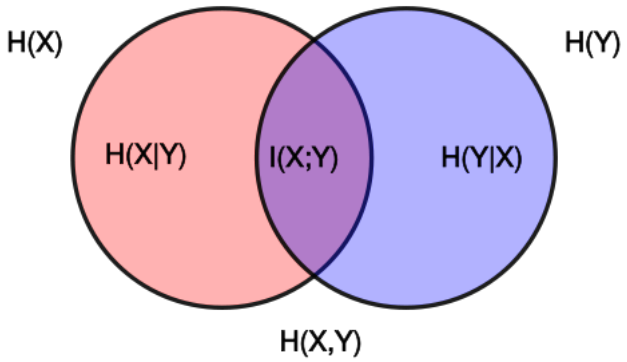

Figura 1: Relazioni fra entropie e informazione mutua

(si veda la Figura 1 per una rappresentazione grafica):

$$\begin{aligned} H(X) &= H(X \mid Y) + I(X,Y) \\ H(X,Y) &= H(X) + H(Y) - I(X,Y) \\ H(X,Y) &= H(X \mid Y) + H(Y \mid X) + I(X,Y) \ . \end{aligned}$$

Analogamente all'entropia, anche l'informazione mutua possiede una sua versione condizionata, $I(X,Y \mid Z)$, definita come segue

$$I(X,Y \mid Z) = \sum_{x,y,z} p(x,y,z) \log_2 \frac{p(x,y \mid z)}{p(y \mid z)p(x \mid z)} \ .$$

Dimostriamo ora un risultato che ha a che fare con la possibilità di acquisire informazione circa una variabile casuale X manipolando una variabile casuale Y che dipende da X. In concreto, supponiamo che una realizzazione di X è stata trasmessa in un canale con rumore. Rappresentiamo con la variabile casuale Y il risultato ottenuto. Vogliamo studiare la quantità di informazione che Y ci fornisce su X, che come sappiamo è misurata da $I(X,Y)$. Se non ci fosse rumore nel canale, allora $Y = X$ e quindi $I(X,Y) = H(X)$, ovvero Y mi fornisce tutti i $H(X)$ bit necessari in media per descrivere il valore di X. Se invece il canale introduce rumore, allora $I(X,Y)$ diminuisce, fino a diventare zero quando Y è indipendente da X. La domanda che ci poniamo è: utilizzando solo l'informazione fornita da Y, posso costruire un'altra variabile casuale Z tale che $I(X,Z) > I(X,Y)$? Il seguente risultato mostra che ciò non è possibile.

Teorema 3 (Data Processing Inequality) *Siano X, Y, Z delle variabili causali con codominio finito tali che la loro distribuzione congiunta $p(x,y,z)$ soddisfa $p(x,z \mid y) = p(x \mid y)p(z \mid y)$ per ogni x, y, z; ovvero, x e z sono indipendenti dato y. Allora $I(X,Y) \geq I(X,Z)$.*

DIMOSTRAZIONE. Cominciamo con introdurre l'informazione mutua fra la variabile casuale X e la coppia di variabili casuali (Y, Z),

$$\begin{aligned} I\big(X,(Y,Z)\big) &= \sum_{x,y,z} p(x,y,z)\log_2 \frac{p(x,y,z)}{p(x)p(y,z)} \\ &= \sum_{x,y,z} p(x,y,z)\log_2 \frac{p(y \mid x,z)p(x,z)}{p(x)p(y \mid z)p(z)} \\ &= \sum_{x,y,z} p(x,y,z)\log_2 \frac{p(x,z)}{p(x)p(z)} + \sum_{x,y,z} p(x,y,z)\log_2 \frac{p(y \mid x,z)}{p(y \mid z)} \ . \end{aligned}$$

Ora osserviamo che l'identità $p(y \mid x,z)p(x \mid z) = p(x,y \mid z)$ —che vale per ogni terna X, Y, Z— implica

$$\frac{p(y \mid x,z)}{p(y \mid z)} = \frac{p(x,y \mid z)}{p(y \mid z)p(x \mid z)} \ .$$

Quindi possiamo proseguire scrivendo

$$\begin{aligned} I\big(X,(Y,Z)\big) &= \sum_{x,y,z} p(x,y,z)\log_2 \frac{p(x,z)}{p(x)p(z)} + \sum_{x,y,z} p(x,y,z)\log_2 \frac{p(x,y \mid z)}{p(y \mid z)p(x \mid z)} \\ &= I(X,Z) + I\big(X,Y \mid Z\big) \ . \end{aligned}$$

Si noti che questa derivazione vale per qualunque terna X, Y, Z di variabili casuali. Quindi possiamo, in modo del tutto analogo, ricavare l'identità

$$I\big(X,(Y,Z)\big) = I(X,Y) + I(X,Z \mid Y) \ .$$

Mettendo insieme le due relazioni appena trovate otteniamo

$$I(X,Z) + I\big(X,Y \mid Z\big) = I(X,Y) + I(X,Z \mid Y) \ .$$

Ora notiamo che $I(X,Z \mid Y) = 0$, dato che per ipotesi x e z sono indipendenti dato y. Da ciò si deduce che $I(X,Y) = I(X,Z) + I\big(X,Y \mid Z\big)$. Dato che l'informazione mutua è non negativa, ricaviamo immediatamente la tesi, ovvero $I(X,Y) \geq I(X,Z)$. □

Un corollario immediato della Data Processing Inequality afferma che, quando Z è indipendente da X dato Y, conoscere Z non può aumentare l'informazione mutua fra X e Y. Infatti, dalla relazione $I(X,Y) = I(X,Z) + I\big(X,Y \mid Z\big)$ utilizzando ancora la non negatività dell'informazione mutua otteniamo $I(X,Y) \geq I\big(X,Y \mid Z\big)$.

È facile costruire un esempio dove $I(X,Y \mid Z) > I(X,Y)$ quando non è vero che X e Z sono indipendenti dato Y. A questo scopo consideriamo due variabili casuali X, Y Bernoulliane indipendenti di parametro $\frac{1}{2}$ e definiamo $Z = X + Y$. Chiaramente, X e Z non sono indipendenti dato Y.

Osserviamo quindi che $I(X,Y)=0$ per l'indipendenza di X e Y, mentre

$$\begin{aligned} I(X,Y \mid Z) &= H(X \mid Z) - \underbrace{H(X \mid Y,Z)}_{=0} \\ &= \mathbb{P}(Z=0)\underbrace{H(X \mid Z=0)}_{=0} + \mathbb{P}(Z=1)H(X \mid Z=1) + \mathbb{P}(Z=2)\underbrace{H(X \mid Z=2)}_{=0} \\ &= \mathbb{P}(Z=1)H(X \mid Z=1) \\ &= \mathbb{P}(Z=1) = \frac{1}{2}\ . \end{aligned}$$

Le tre entropie sono zero in quanto X è determinato da $Z-Y$ e poi $Z=0$ se e solo se $X=0$, $Z=2$ se e solo se $X=1$. Inoltre, $H(X \mid Z=1)=1$ in quanto se $Z=1$ allora X assume i valori 0 o 1 con uguale probabilità. Quindi X condizionata su $Z=1$ è una Bernoulliana di parametro $\frac{1}{2}$.

Supponiamo di conoscere la distribuzione congiunta di due variabili casuali discrete X e Y. Qual è la probabilità di sbagliare a predire il valore di X in funzione del valore di Y? La disuguaglianza di Fano mostra che questa probabilità di errore è minorata dalla quantità di informazione necessaria a descrivere il valore di X dato il valore di Y, ovvero dall'entropia condizionata $H(X \mid Y)$. Quindi Fano lega direttamente l'entropia alla probabilità di errore.

Interpretando X come il simbolo trasmesso da una sorgente, Y come il simbolo ricevuto dal ricevente, e g come la funzione che ricostruisce il simbolo inviato sulla base del simbolo ricevuto, possiamo utilizzare Fano per quantificare l'effetto del rumore nel canale di trasmissione (rappresentato dall'entropia condizionata) sull'errore di decodifica del simbolo inviato.

Teorema 4 (Disuguaglianza di Fano) *Siano X e Y due variabili casuali con valori in $\mathcal{X}$ e $\mathcal{Y}$ entrambi finiti e sia $g:\mathcal{Y}\to\mathcal{X}$ una funzione qualunque che mappa i valori di Y in quelli di X. Sia p_e la probabilità di errore quando uso $g(Y)$ per predire X, $p_e=\mathbb{P}\big(g(Y)\neq X\big)$. Allora*

$$p_e \geq \frac{H(X \mid Y)-1}{\log_2 |\mathcal{X}|}\ .$$

Dimostrazione. Introduciamo la variabile casuale Bernoulliana

$$E = \begin{cases} 1 & \text{se } g(Y)\neq X, \\ 0 & \text{altrimenti.} \end{cases}$$

Ora usiamo la chain rule per l'entropia per esprimere $H(E,X \mid Y)$ in due modi diversi,

$$\begin{aligned} H(E,X \mid Y) &= \underbrace{H(E \mid X,Y)}_{=0} + H(X \mid Y) \\ H(E,X \mid Y) &= H(X \mid E,Y) + \underbrace{H(E \mid Y)}_{\leq H(E)} \end{aligned}$$

dove $H(E \mid X,Y)=0$ in quanto E è nota dati X e Y, mentre $H(E \mid Y)\leq H(E)$ dato che il condizionamento non aumenta l'entropia. Inoltre, usando la definizione di entropia condizionata,

possiamo scrivere

$$H(X \mid E,Y) = (1-p_e)\underbrace{H(X \mid E=0,Y)}_{=0} + p_e H(X \mid E=1,Y)$$
$$\leq p_e \log_2(|\mathcal{X}|-1) \ .$$

L'identità $H(X \mid E=0,Y)=0$ vale in quanto se $E=0$ allora $X=g(Y)$, ovvero determino X come funzione di Y. Invece $H(X \mid E=1,Y) \leq \log_2(|\mathcal{X}|-1)$ in quanto dato $E=1$ so che $X \neq g(Y)$ e quindi X varia tra al più $|\mathcal{X}|-1$ valori possibili.

Combinando ciò che abbiamo ottenuto ricaviamo

$$H(E) + p_e \log_2(|\mathcal{X}|-1) \geq H(X \mid Y) \ .$$

Usando la maggiorazione $H(E) \leq 1$, che vale in quanto E assume al più due valori, e la maggiorazione $\log_2(|\mathcal{X}|-1) \leq \log_2 |\mathcal{X}|$, otteniamo il risultato desiderato. □