Entropia

Gianluca Amato

Corso di Laurea Specialistica in Economia Informatica
Università “G. D'Annunzio” di Chieti-Pescara
anno accademico 2005-2006

Il Concetto di Entropia

- L'entropia è un concetto legato al grado di “disordine” in un sistema.
- Il concetto è usato in vari settori delle scienze:
  - fisica (in particolare nella termodinamica)
  - teoria dell'informazione

Entropia e teoria dell'informazione

Teoria dell'Informazione (1)

- Entropia legata al concetto di “misura dell'informazione”
- Esperimento X1:
  - 4 possibili risultati: a, b, c, d equiprobabili
  - vogliamo memorizzare il risultato su un elaboratore: che codice utilizzare?

Probabilità	***Risultati***	***Codice binario***
¼	a	00
¼	b	01
¼	c	10
¼	d	11

  - 2 bit per risultato

Teoria dell'Informazione (2)

- Esperimento X2:
  - 4 risultati **non** equiprobabili

Probabilità	*Risultati*	*Codice binario*
½	a	0
¼	b	10
1/8	c	110
1/8	d	1110

- codice con un numero di bit variabili
  - lunghezza media: 1/2*1+1/4*2+1/8*3+1/8*4=15/8
  - la lunghezza media è minore di 2!!
- questo codice per l'esperimento X1?
  - lunghezza media 5/2

Teoria dell'Informazione (3)

- Alcune domande che vengono fuori:
  - Perché per X1 sono necessari due bit mentre per X2 si può fare di meglio?
    - X2 contiene “meno informazione” di X1
  - E` possibile fare ancora meglio per l'esperimento X2 ?
- A queste domande risponde la teoria dell'informazione.

Entropia di un esperimento finito

- Sia X un esperimento con un numero finito di possibili risultati $e_1, \ldots e_q$.
- La probabilità che l'evento e_i si verifichi è p_i.
- Entropia dell'esperimento X:

$$H(X) = H(p_{1,} p_{2,\ldots}, p_q) = -\sum_{i=1}^{q} p_i \log(p_i)$$

- Perché questa definizione?

Proprietà della funzione H (1)

- H è continua sulle p_i
  - piccole modifiche delle probabilità causano piccole modifiche della incertezza dell'esperimento
- se X e X' hanno q e q' risultati equiprobabili e q < q', allora:

  $H(X)=H(1/q,...,1/q) < H(1/q',...,1/q')=H(X')$

  - più risultati possibili abbiamo maggiore è l'incertezza dell'esperimento

Proprietà della funzione H (2)

- Se l'esperimento X è scomposto in due esperimenti successivi, il risultato non cambia:

e_1

1/2

1/3 e_2

1/6 e_3

$H(1/2,1/3,1/6)$

e_1

1/2

2/3 e_2

1/2

1/3 e_3

$H(1/2,1/2) + 1/2 * H(1/3,2/3)$

  - i due valori sono uguali

Caratterizzazione di H

- Le uniche funzioni che soddisfano le proprietà di cui ai lucidi precedenti sono (al variare di C):

$$H_C(p_{1,} p_{2,\ldots,} p_q) = -C \sum_{i=1}^{q} p_i \log(p_i)$$

- Fissiamo C=1 e chiamiamo bit l'unità di misura corrispondente a questa scelta.
- Se X è un esperimento, un qualunque codice per X che sia unicamente decifrabile avrà una lunghezza media n ≥ H(X).
  - nell'esperimento X1 l'entropia è 2 per cui tutti i codici possibili avranno lunghezza media ≥ 2.
  - nell'esperimento X2 l'entropia è 14/8, il codice fornito ha lunghezza media 15/8: forse si può migliorare
    - basta cambiare il codice del risultato C in 111.

Entropia e codici

- Se X è un esperimento, un qualunque codice per X che sia unicamente decifrabile avrà una lunghezza media $n \geq H(X)$.
  - nell'esperimento X1 l'entropia è 2 per cui tutti i codici possibili avranno lunghezza media ≥ 2.
  - nell'esperimento X2 l'entropia è 14/8, il codice fornito ha lunghezza media 15/8: forse si può migliorare
    - basta cambiare il codice del risultato C in 111.

Programmi di Compressione

- I programmi di compressione sfruttano il fatto che alcune combinazioni di caratteri sono più probabili di altre.
  - ovvero, l'esperimento "*prendo a caso un file di m bytes dal disco fisso*" ha una entropia inferiore ad 8m (come sarebbe se il file fosse generato in modo completamente casuale)
  - vuol dire anche, in ogni sistema di compressione, ci saranno dei file in cui la versione compressa è più grande di quella originaria.

Entropia e statistica

Misure di eterogeneità (1)

- Si chiamano misure di eterogeneità quegli indici numerici che tentato di stabilire quanto i possibili valori di un attributo in un insieme di dati sono distribuiti in maniera uniforme.
  - in termini statistici, si parla spesso di collettivo per insieme di dati, caratter per attributo e modalità per un possibile valore di un attributo.
- Si distinguono due casi:
  - eterogenità massima: quando tutte le istanze (unità statistiche) hanno la stessa modalità; .
  - omogeneità massima: quando le unità statistiche sono suddivise equamente tra le modalità.

Misure di eterogenità (2)

- Dato un insieme di dati S, indichiamo con $H_S(A)$ l'entropia riferita all'attributo A
  - in caso di eterogeinità massima, l'entropia è 0
  - in caso di omogenità massima, l'entropia è $\log_2(k)$ dove k è il numero di modalità di A.
- Quando S è chiaro dal contesto, lo ometteremo, scrivendo solo H(A) invece di $H_S(A)$.
- Può essere comodo rapportare l'entropia di un attributo al suo massimo valore possibile:
  - se A è un attributo con k modalità, definiamo l'entropia relativa di A come $H^r(A)=H(A) / \log_2(k)$.

Guadagno di Informazione (1)

- Supponiamo di avere un insieme di dati con due attributi A e C. C è di solito chiamato attributo classe, in quanto, nelle applicazioni, è la classe di una istanza che vorremmo predirre a partire dagli altri attributi.
- Vogliamo una misura di quanto A è importante nel predirre C.
- Se v è una modalità di A, definiamo l'entropia di C condizionata da A=v come l'entropia di C per i soli dati in cui A=v.
- Definiamo l'entropia di C condizionata da A come la media delle entropie condizionate per tutte le modalità di A, pesata con la frequenza relativa di tali modalità:

$$H(C|A)=\sum_{i=1}^{k} f_i \cdot H(C|A=x_i)$$

Guadagno di Informazione (2)

- La H(C|A) misura l'incertezza del valore C quando sono già a conoscenza del valore per A.
- Definisco guadagno di informazione di A il valore

$$IG(A) = H(A) - H(C|A)$$

- Più è alto questo valore, maggiore è l'importanza di A per determinare il valore di C
  - in quanto diminuisce di molto l'incertezza su C, quando conosco già il valore di A
- In maniera del tutto analoga si definisce il guadagno di informazione per più di un attributo.

Guadagno di Informazione (3)

- Questa definizione di guadagno di informazione va bene per valori categoriali o numerici discreti, ma non per valori continui.
  - In questo caso, spesso si calcola il guadagno di informazione che si ottiene dividendo l'intervallo di tutti i possibili valori per un attributo A in due sotto-intervalli, in corrispondenza del valore v.
- IG(A,*v*) definito come la differenza tra
  - l'entropia dell'insieme dei dati iniziali: H(C)
  - l'entropia media dopo il partizionamento.

$$H(C|A,v)=\frac{|S_1|}{|S|}H(C|A\leqslant v)+\frac{|S_2|}{|S|}H(C|A>v)$$

  dove S_1 ed S_2 sono gli insiemi delle istanze corrispondenti alle condizioni $A \leq v$ e $A > v$ rispettivamente.

  - IG(A,v)=H(C) – H(C|A,v)