

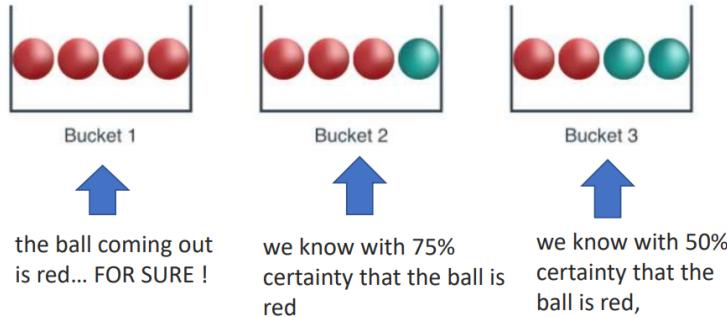
# Esercitazione 3 23/10/2023

Seconda parte dell'esercitazione, prima parte su carta

## Entropy & surprise

How much information we have on the colour of a ball drawn at random

- When something unlikely happens, we say it's a big news.
- When we say something predictable, it's not really interesting.



Nel primo caso la conoscenza è alta, nel secondo meno, nel terzo è la più bassa, perché è più difficile predire il risultato.

Quindi come quantifichiamo questo **interesse (interestingness)**?

Se la probabilità è 1 allora l'interesse è basso.

$$P(\text{event}) = 1 \rightarrow I(p) = 0$$

Stessa cosa se la probabilità è 0

$$P(\text{event}) \rightarrow 0 \rightarrow I(p) \rightarrow \infty$$

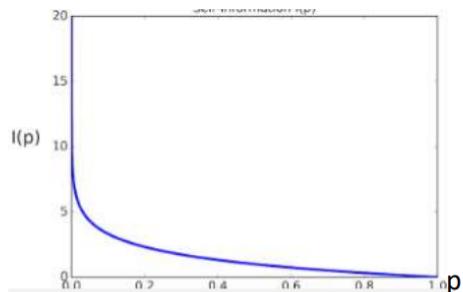
## Information

$$I(E) = -\log_2(P(E))$$

monotonically decreasing in p

$I(p) \geq 0$  : information is a non-negative quantity.

$$I(p_1, p_2) = I(p_1) + I(p_2) \quad \text{Additive for independent events}$$



## Entropia

L'entropia è grado di sorpresa media quando un evento si realizza

$$H[X] = - \sum_{i=1}^n p_i \log p_i = E_P[\log(p)]$$

Notation  $[P_X(x_i) = P[X = x_i]]$

## Esempio lancio di una moneta

$$\begin{aligned} H(p) &= -p(0) \log p(0) - p(1) \log p(1) \\ &= -2(0.5 \log 0.5) \\ &= 1 \end{aligned}$$

an unfair coin that always lands on heads.

$$\begin{aligned} H(p) &= -p(0) \log p(0) - p(1) \log p(1) \\ &= -\log 1 \\ &= 0 \end{aligned}$$

In questo caso sono più sorpreso quando la mia incertezza è maggiore, ovvero con la moneta bilanciata. Nel secondo caso l'entropia è 0 perché non sono sorpreso.

## Esercizio

Training set:

Example	A	B	C	D	T
$x_1$	0	0	1	1	1
$x_2$	0	1	1	1	1
$x_3$	0	1	0	0	0
$x_4$	0	1	0	1	1

Calcoliamo quindi l'entropia del target.

Cerchiamo innanzitutto la sua distribuzione:

We have:  $P_T = [1/4, 3/4]$ , then:

E poi

l'entropia:

$$H[P_T] = -P_T(0) \log(P_T(0)) - P_T(1) \log(P_T(1)) \\ = -1/4 \log(1/4) - 3/4 \log(3/4) = 1.81$$

$P_T(0)$  è la probabilità che il target assuma il valore 0.

## Entropia della distribuzione condizionata da un evento

Let  $X$  be a discrete random variable with outcomes,  $\{x_1, \dots, x_n\}$ , which occur with probabilities,  $p_X(x_i)$ . Consider the 1D distribution,

$$p_{Y|X=x_i}(y_j) = p_{Y|X}(y_j | x_i) \quad \leftarrow \text{the distribution of Y outcomes given that } X = x_i$$

The avg. information you gain when told the outcome of Y is:

$$H_{Y|X=x_i} = - \sum_{j=1}^m p_{Y|X}(y_j | x_i) \log p_{Y|X}(y_j | x_i).$$

In questo caso sto calcolando l'entropia della distribuzione condizionata, ovvero del ramo dove un attributo ha un valore prefissato.

## Entropia condizionata

$$\overbrace{H_{Y|X}} = \sum_{i=1}^n p_X(x_i) H_{Y|X=x_i}$$

Qui si applica l'entropia a tutti i valori che un attributo può avere.

L'information gain misura quanto l'incertezza di Y è ridotta quando abbiamo informazioni su Y.

### Esercizio

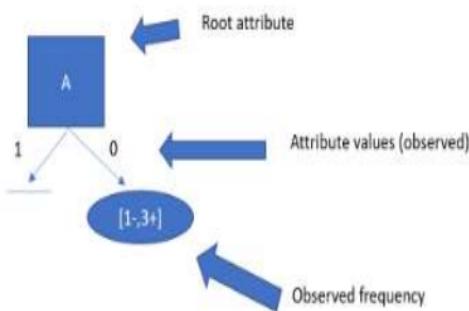
Ordiniamo l'attributo in base all'information gain.

Example	A	B	C	D	T
$x_1$	0	0	1	1	1
$x_2$	0	1	1	1	1
$x_3$	0	1	0	0	0
$x_4$	0	1	0	1	1

Iniziamo a calcolare l'entropia del target, andando a vedere la distribuzione dei valori del target. Applico la formula dell'entropia su questa distribuzione.

We have:  $P_T = [1/4, 3/4]$ , then:

$$\begin{aligned} H[P_T] &= -P_T(0) \log(P_T(0)) - P_T(1) \log(P_T(1)) \\ &= -1/4 \log(1/4) - 3/4 \log(3/4) = 0.81 \end{aligned}$$



Ora usiamo la formula dell'entropia per calcolare l'entropia condizionata da A quando A vale 0, dato che non ci sono casi in cui A è 1.

$$\begin{aligned} H[T|A=0] &= -P_{T|A}(0|0)\log(P_{T|A}(0|0)) - P_{T|A}(1|0)\log(P_{T|A}(1|0)) \\ &= -1/4(\log 1/4) - 3/4(\log 3/4) = 0.81 \\ H[T|A] &= P_A(0)H[T|A=0] + P_A(1)H[T|A=1] = 1 * 0.81 = 0.81 \end{aligned}$$

Quindi l'information gain è:

$$\text{Finally, } \text{IG}[T; A] = H[T] - H[T|A] = 0.81 - 0.81 = 0$$

Ovvero non ho guadagno informativo.

Proviamo ad usare l'attributo B:

We have:  $P_B(0) = 1/4$ ,  $P_B(1) = 3/4$   
Moreover,

$$\begin{aligned} H[T|B=0] &= -P_{T|B}(0|0)\log(P_{T|B}(0|0)) - P_{T|B}(1|0)\log(P_{T|B}(1|0)) \\ &= -0\log 0 - 1\log 1 = 0 \\ H[T|B=1] &= -P_{T|B}(0|1)\log(P_{T|B}(0|1)) - P_{T|B}(1|1)\log(P_{T|B}(1|1)) \\ &= -1/3\log 1/3 - 2/3\log 2/3 = 0.91 \\ H[T|B] &= P_B(0)H[T|B=0] + P_B(1)H[T|B=1] = 1/4 * 0 + 3/4 * 0.91 = 0.68 \end{aligned}$$

$$\text{Finally, } \text{IG}[T; A] = H[T] - H[T|A] = 0.81 - 0.68 = 0.13$$

L'entropia media è quindi 0,68 che è il valore di incertezza dovuto alla scelta dell'attributo B. Il delta è il guadagno informativo, ovvero 0,13.

Consideriamo C:

We have:  $P_C(0) = 1/2$ ,  $P_C(1) = 1/2$   
Moreover,

$$\begin{aligned} H[T|C=0] &= -P_{T|C}(0|0)\log(P_{T|C}(0|0)) - P_{T|C}(1|0)\log(P_{T|C}(1|0)) \\ &= -1/2\log 1/2 - 1/2\log 1/2 = 1 \\ H[T|C=1] &= -P_{T|C}(0|1)\log(P_{T|C}(0|1)) - P_{T|C}(1|1)\log(P_{T|C}(1|1)) \\ &= -0\log 0 - 1\log 1 = 0 \\ H[T|C] &= P_C(0)H[T|C=0] + P_C(1)H[T|C=1] = 1/2 * 1 + 1/2 * 0 = 1/2 \end{aligned}$$

$$\text{Finally, } \text{IG}[T; C] = H[T] - H[T|C] = 0.81 - 0.5 = 0.31$$

Infine consideriamo D:

We have:  $P_D(0) = 1/3$ ,  $P_D(1) = 2/3$

Moreover,

$$\begin{aligned} H[T|D=0] &= -P_{T|D}(0|0)\log(P_{T|D}(0|0)) - P_{T|D}(1|0)\log(P_{T|D}(1|0)) \\ &= -1\log 1 - 0\log 0 = 0 \\ H[T|D=1] &= -P_{T|D}(0|1)\log(P_{T|D}(0|1)) - P_{T|D}(1|1)\log(P_{T|D}(1|1)) \\ &= -0\log 0 - 1\log 1 = 0 \\ H[T|D] &= P_D(0)H[T|D=0] + P_D(1)H[T|D=1] = 1/3 * 0 + 2/3 * 0 = 0 \end{aligned}$$

Finally,  $IG[T; D] = H[T] - H[T|D] = 0.81 - 0 = 0.81$

Come vediamo dalla tabella, i valori di D sono gli stessi del target. Ovvero in questo caso la dipendenza è piena, la correlazione è forte.

L'incertezza condizionata è quindi nulla, e l'information gain sarà massimo.

(errore nella prima riga, sarebbe 1/4 e 3/4)

Ora andiamo a prendere l'attributo con information gain maggiore:

$$\begin{aligned} IG[T; A] &= H[T] - H[T|A] = 0.81 - 0.81 = 0 \\ IG[T; B] &= H[T] - H[T|B] = 0.81 - 0.68 = 0.13 \\ IG[T; C] &= H[T] - H[T|C] = 0.81 - 0.5 = 0.31 \\ IG[T; D] &= H[T] - H[T|D] = 0.81 - 0 = 0.81 \end{aligned}$$

## Esempio esercizio da compito

Example	A	B	C	D	T
$x_1$	0	0	1	...	1
$x_2$	0	1	1	...	1
$x_3$	0	1	0	...	0
$x_4$	0	1	0	...	1

Fill the values for D, so that the information gain obtained by applying D is maximum. Then complete the following expressions.

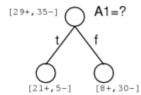
- $H(T|A=0) =$
- $H(T|B) =$
- $H(T|D) =$

## Altro esempio da compito

Example: What is the info gain of T|A1

### ID3 - Iterative Dichotomiser 3

A classification algorithm that follows a greedy approach of building a decision tree by selecting a best attribute that yields maximum Information Gain (IG) or minimum Entropy (H).



$f_1$

$f_2$

$f_3$

T

Training Set			
$f_1$	$f_2$	$f_3$	
1	1	1	1
0	1	1	0
0	0	1	1
0	0	0	0

T

0

1

0

0

1

0

0

0

1

0

1

0

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

1

0

1

0

0

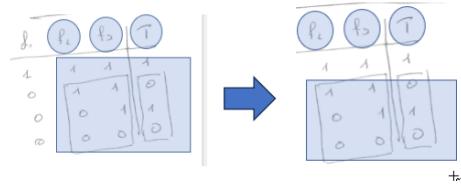
1

0

1

0

Ex: ID 3



iterate for the new table

$f_2$	$f_3$	T
1	1	0
0	1	1
0	0	0

