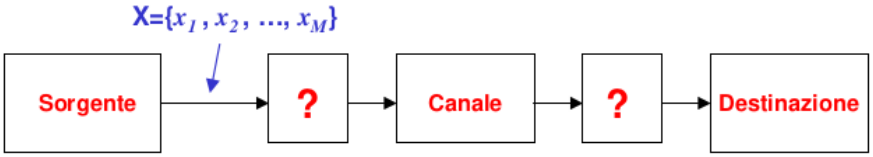
Teoria dell'*Informazione*

La teoria che studia l'informazione dal punto di vista matematico fu formulata per la prima volta da Claude Shannon che la formalizzò nel **1948** pubblicando l'articolo scientifico **"A Mathematical Theory of Communication"**

Le motivazioni che la rendono interessante per la didattica STEM

- La teoria fornisce interessanti collegamenti con il **calcolo delle probabilità** per quanto riguarda la matematica
- Il concetto di **Entropia dell'informazione** è del tutto simile all'analogo concetto studiato in fisica
- Viviamo nell'epoca in cui l'informazione riveste un ruolo di primaria importanza, soprattutto per i nuovi sviluppi concernenti l' **Intelligenza Artificiale Generativa**. Possedere delle nozioni di tipo quantitativo oltre che qualitativo può rivelarsi molto utile per comprendere meglio determinati meccanismi complessi
- **Per gli studenti dell'ultimo anno del liceo Scienze Applicate**, un modo per affrontare con più semplicità questiti riguardanti calcolo combinatorio e della probabilità



Le lezioni riguarderanno i modelli più semplici della teoria, e saranno sviluppate non solo a livello teorico ma mediante applicazioni pratiche utilizzando il **lingiuaggio di programmazione Python**

Abbiamo iniziato a delineare alcune proprietà che la misura cercata da Shannon per misurare l'informazione, dovesse essere in qualche modo legata, **all'incertezza del verificarsi di un determinato evento di cui veniamo a conoscenza.**

Il **calcolo delle probabilità** definisce una misura dell'incertezza e delle regole per calcolarla e manipolarla.

Tra le poprietà che desideriamo dare ad una misura quantitativa dell'informazione, vi è l'aspettativa abbastanza intuitiva, che **l'informazione ricevuta da un evento di cui veniamo a conoscenza, sarà tanto più grande quanto più piccola sarà la probabilità che questo evento si verifichi.** Per contro se venissimo a conoscenza di qualcosa che non ci porta alcuna sorpresa, allora significa che non abbiamo ricevuto nessuna informazione. Viene quindi spontaneo utilizzare la probabilità per defonire l'informazione. Ma prima di procedere ai passi logici che hanno portato Shannon a scegliere una particolare espressione matematica per definire la base della teoria cerchiamo di sintetizzare alcun proprietà fondamentali della probabilità

LA PROBABILITA'

La probabilità è una misura associata ad un "evento" ossia qualcosa che potrà accadere o non accadere in futuro.

$$0 \leq p \leq 1$$

Se consideriamo come esempio concreto il lancio di un dado a 6 facce, e la rilevazione della faccia mostrata dal dado dopo il lancio, questa potrà assumere uno dei seguenti valori: {1,2,3,4,5,6}

In questo caso gli eventi elementari sono costituiti dai 6 risultati possibili, escludendo per semplicità che possano accadere cose come il dado che rimane in bilico su uno spigolo.

Possiamo indicare quindi come **EVENTO CERTO** il fatto che uno di questi 6 risultati sicuramente accadrà anche se non sappiamo quale.

Un **EVENTO IMPOSSIBILE** è invece quell'evento che sicuramente non accadrà mai, come ad esempio che il dado mostri il valore 7

Associamo il valore 1 ossia probabilità 1 all'evento certo e 0 all'evento impossibile.

Ma come si misura la probabilità in tutti gli altri casi ? Non esiste un solo e unico approccio a questo problema.

DEFINIZIONE CLASSICA

La definizione classica di probabilità è applicabile quando possiamo in qualchemodo contare il numero di elementi elementari di un evento e cnfrontarlo con la numerosità dell'evento certo. Ad esempio consideriamo l'evento **E = uscirà un numero pari**.

Possiamo considerare questo evento costituito dai seguenti eventi elementari: {2,4,6}

La definizione classica di probabilità considera il rapoprto fra la numerosità (o cardinalità) di questi insiemi:

probabilità = $\frac{|\{2,4,6\}|}{|\{1,2,3,4,5,6\}|} = \frac{3}{6} = \frac{1}{2}$
(la barra verticale indica la cardinalità ossia il numero di elementi di un insieme)

ATTENZIONE: questo calcolo non solo si può fare quando è possibile contare il numero di casi favorevoli e quelli possibili, e in tal senso ci viene in soccorso il calcolo combinatorio, ma soprattutto:

QUANDO GLI EVENTI SONO EQUIPROBABILI Abbiamo fatto l'assunzione che il dado non sia truccato, ossia supponendo che le probabilità associate ad ogni risultato siano uguali. Che non ci siano motivi per ritenere che il mostrare una faccia piuttosto che un 'altra sia più o meno probabile.

La definizione classica di probabilità oltre a soffrire di una sorta di 'circolarità' nella definizione, (infatti assume che gli eventi conteggiati sia ugualmente probabili, quando non ha ancora definito come si calcola questa probabilità) non è applicabile quando non siamo in grado di contare il numero di eventi che entrano in gioco.

Nel modello che stiamo analizzando ossia di una trasmissione di simboli da una sorgenta ad una destinazione, infatti non sappiamo come associare delle probabilità a ciascun simbolo della sorgente.

E nemmeno è ragionevole supporre ceh la probabilità di trasmettere un simbolo, piuttosto che un altro, siano le stesse per tutti i smboli.

Nel nostro caso l'evento che prenderemo in considerazione è costituito dal fatto di decidere quali fra i diversi simboli trasemttere. Il ricevente verrà a conoscenza del risultato, dopo la trasmissione, ossia quando riceverà il simbolo o ne verrà a conoscenza. Prima può solo conoscere la probabilità associata a quel simbolo.

Nella trasmissione di testi ad esempio, non tutte le lettere dell'alfabeto hanno la stessa probabilità di essere utilizzate e quindi trasmesse. Nell'italiano, ma anche in altre lingue, le vocali hanno una maggior probabilià di essere utilizzate piuttosto che alcune consonanti come la 'q' ad esempio.

Come possiamo fare allora ?

APPROCCIO FREQUENTISTA

Un approccio potrebbe affidarsi a ciò che è accaduto in passato in circostanze analoghe. Supponendo che le condizioni non siano cambiate potremmo aspettarci che se raccogliamo un numero grande di rilevazioni del passato, queste possano fornirci una previsione significativa sul futuro.

A sostegno di questa assunzione (il calcolo delle probabilità è sempre permeato di assunzioni o ipotesi preliminari) ci viene in aiuto la **LEGGE DEI GRANDI NUMERI** che afferma che se siamo in grando di ripetere un esperimento un numero molto grande di volte, al crescere del numero di rilevazioni dei dati, se un evento aveva probabilità P allora la frequenza con con cui accadrà nel totale delle rilevazioni si avvicinerà sempre di più a P al crescere del numero di ripetizioni

Sebbene anche l'approccio frequentista offra il fianco a diversi problemi (ad esempio non tutti gli eventi possono essere ripetuti, e inoltre come si può stabilire quanto numerosi devono essere le rilevazioni perchè la stima delle probabilità sia significativa)

Nonostante ciò nel nostro caso, questo approccio può rivelarsi utile in pratica per arrivare ad una stima delle probabilità da associare ad ogni simbolo della nostra sorgente.

Cosa faremo

Adotteremo un approccio "frequentista" analizzando ad esempio un testo scelto come riferimento, scritto nella stessa lingua caratteristica dei messaggi che vorremo trasmettere. Andremo quindi a stimare per ogni simbolo della nostra sorgente, le probabilità associate.
Supponiamo di utilizzare come sorgente i caratteri alfabetici dell'alfabeto della lingua inglese, 26 caratteri compresi le lettere jkxyw non presenti nell'italiano. Contando quante volte ciascuna lettere compare nel testo che abbiamo scelto, possibilmente abbastanza lungo, potremmo calcolare le frequenze assolute di ciascuna lettera. Ossia quante volte abbiamo incontrato quella lettera nel testo.
Dividento questi valori per il numero totale di lettere incontrate, otteremo le frequenze relative.
Possiamo utilizzare questi valori come stima delle probabilità che ci servono per i futuri calcoli ? Dobbiamo solo porre attenzione ad una cosa, i valori analizzati sono frequenze non probabilità, si riferiscono ad eventi già accaduti.
Le probabilità sono misure di eventi futuri non ancora accaduti e che potrebbero accadere. La principale conseguenza pratica è che **se una certa lettere non compare nel testo che abbiamo scelto come modello, non è detto che non possa comparire nei messaggi che invieremo in futuro**. Una operazione indispensabile quindi è quella di **sostituire eventuali valori nulli delle frequenze in valori molto piccoli ma non nulli di probabilità.**

Soluzione in Python

```
alfabeto = "abcdefghijlmnopqrstuvwxyz"
```

```
with open('promessi_sposi.txt','r') as f:  # carichiamo il file di riferimento
   testo = f.read()
```

```
frequenze = {}            # inizializzo un dizionario vuoto
for lettera in alfabeto:
  frequenze[lettera] = 0
```

```
frequenze
```

```
lunghezza = 0                          # calcoliamo le frequenze assolute delle lettere dell'alfabeto dal testo di riferimento
for carattere in testo:
  if carattere.lower() in frequenze:    # se il carattere letto è tra quelli in dizionario
     lunghezza += 1                     # incrementa di 1 il numero di caratteri letti e la chiave corrispondente nel dizionario
     frequenze[carattere.lower()] += 1
```

```
for carattere in frequenze:             # trasformiamo le frequenze assolute in relative
   frequenze[carattere] = frequenze[carattere] / lunghezza
```

```
somma = 0                              # verifichiamo che la somma delle frequenze relative dia 1
for carattere in frequenze:
  somma += frequenze[carattere]
print(somma)
```

```
1.0
```

```
# prompt: salvami i dati della variabile frequenze che è dizionario in un file in formato Json

import json

with open("frequenze.json", "w") as f:
    json.dump(frequenze, f, indent=4)
```

```
# prompt: carica i dati contenuti nel file json chiamato frequenze.json nella variabile probabilita

import json

with open('frequenze.json', 'r') as json_file:
  probabilita = json.load(json_file)
```

```
probabilita
```

```
{'a': 0.11600588761471146,
 'b': 0.009474137563833765,
 'c': 0.045575059988357286,
 'd': 0.037518872076898684,
 'e': 0.11781950012068663,
 'f': 0.010220974864050054,
 'g': 0.01664623670672201,
 'h': 0.013093790080978366,
 'i': 0.09679162860699225,
 'j': 2.17709225665078e-05,
 'l': 0.056142476489770034,
 'm': 0.024321906754192085,
 'n': 0.07415554850892844,
 'o': 0.09383267626164862,
 'p': 0.030198162723447757,
 'q': 0.007577227614995433,
 'r': 0.06681306823291101,
 's': 0.05707862616012987,
 't': 0.061938274701714695,
 'u': 0.034261752748578976,
 'v': 0.022344539048042747,
 'w': 2.8396855521531916e-06,
 'x': 0.00012778584984689362,
 'y': 7.572494805741844e-06,
 'z': 0.008029684179638508}
```

Dopo aver verificato che ogni frequenza sia maggiore di zero (la lettera k non lo era e l'abbiamo eliminata) prendiamo le frequenze e consideriamole come stima delle probabilità associata alla nostra sorgente. Per un uso futuro di questi dati salviamo il contenuto del dizionario 'frequenze' in un file di testo in formato json

APPROCCIO BAYESIANO

Un terzo approccio al calcolo delle probabilità, molto interessante ma che non approfondiremo al momento è quello Bayesiano. Prende il nome dal matematico Thomas Bayes e dal suo teorema del calcolo dlle probabilità. Possiamo dire al momento (lo approfondiremo nelle lezioni successive) che la base di partenza è un approccio per cosi dire 'soggettivista' al problema di asseganre una probabilità ad un evento. Spesso lo sperimentatore ha una conoscenza poco formalizzabile del grado di incertezza che caratterizza un certo fenomeno, e la può esprimere numericamente con un valore compreso tra 0 e 1, senza altresì poter darene una giustificazione 'oggettiv' e analitica'. Sebbene questo valore sia soggettivo ossia possa differire da persona a persona, ci può considerare questo valore come una 'probabilità a priori' ossia valutata senza avere a disposizione altri elementi oggettivi, al di fuori della porpria esperienza che ha portato a determinare quella valutazione. Il fatto è che spesso si viene a conoscenza nel corso del tempo di indizi oggettivi, ossia eventi che possono confermare o attenuare la stima iniziale. Il teorema di Bayes consente in effetti di effettuare in modo matematico e secondo precise regole coerenti, questi aggiornamenti della probabilità a priori e ottenere quella probabilità chiamata a posteriori, ossia modificata in base all'osservazione dell'accadere di particolar eventi. Si dimostra che anche si parte da stime iniziale differenti della probabilità a priori, applicando correttamente la regola di Bayes per aggiornare la probabilità a posteriori, si convergerà ai medesimi risultati, man mano che le informazioni acquisite si accumuleranno per diminuire l'incertezza iniziale.

Le principali proprietà del calcolo delle Probabilità

Unione di due eventi:

Se abbiamo due **eventi incompatibili** fra loro, ossia che non posso accadere contemporaneamente e conosciamo le loro probabilità:

esempio:
A = 'dal lancio del dato uscirà 3'
B = 'dal lancio del dado uscirà 5'

L'unione dei due eventi A OR B consiste nel verificarsi dell'evento: 'uscirà A oppure B'

In questo esempio è intuitivo commprendere che le probabilità seguono semplicemente questa regola:

$$P(A\ or\ B) = P(A) + P(B)$$

nell'esempio essendo le due probabilità pari a 1/6 la loro somma è 2/3. Corrispondente al fatto che il nuovo evento possiede due casi favorevoli su 6

Ma se gli eventi non fossero incompatibili ad esempio:

A = esce un numero <= 3
B = esce un numero pari

dovremmo togliere dalla somma dei due eventi la probabilità dell'evento congiunto ossia che i due eventi citati possano accadere contemporaneamente. Ad esempio se esce 2

quindi P(A) = 1/2 p(B) = 1/2 dobbiamo sottrarre alla somma dei due eventi la probabilità che esca 2 il caso in comune P(A or B) = 1/2 + 1/2 - 1/6 = 5/6 infatti i casi favorevoli sono {1,2,3,4,6}

La formula generale da applicare è allora la seguente

$$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$$

Eventi congiunti

La probabilità di eventi congiunti è quella probabilità legata all'evento che consiste nel verificarsi contemporaneamente di due eventi A e B e la indicheremo con P(A end B)

Qui bisogna fare una distinzione: ci dobbiamo chiedere se i due eventi sono **INDIPENDENTI** oppure no.
Due eventi sono indipendenti, quando il verificarsi di uno dei due, non modifica la probabilità associata al verificarsi dell'altro.

Ad esempio consideriamo il lancio due volte del dado. Il risulato del prino lancio non influenzerà la probabilità legata al secondo lancio.

Se invece condideriamo l'estrazione di una carta da un mazzo di 40 carte eseguita due volte. Ma facendo in modo che la carta estratta la prima volta non venga rimessa nel mazzo. In questo caso il risultato della prima estrazione potrà influenzare la probabilità legata alla seconda estrazione.

Ad esempio supponiamo di voler calcolare la probabilità di estrarre al aprima volta una carta rossa. Questa probabilità è 1/2 avendo il mazzo metà carte rosse e metà nere. Ma estraiamo una carta rossa, e ci chiediamo quale sarà la probabilità alla seconda estrazione di estrarre di

nuovo una carta rossa, questa non sarà più 1/2. Ma 19/39 essendo rimaste nel mazzo 19 carte rosse su 39 in totale.
Se invece rimettiamo la prima carta estratta nel mazzo, la seconda estrazione avrà le stesse probabilità di verificarsi ossia 1/2. Riassumento A= carta rossa alla prima estrazione B= carta rossa alla seconda estrazione

P(A and B) = 1/2 * 19/39
se la carta estratta non viene rimessa nel mazzo

P(A and B) = 1/2 * 1/2 = 1/4 se la carta estratta viene rimessa nel mazzo

La formula da applicare in generale è quindi: Nel caso di eventi indipendenti

P(A and B) = P(A) * P(B)

nel caso non siano indipendenti, introduciamo il concetto di probabiltà condizionata

P(A and B) = P(A) P(B|A) oppure in modo equivalente

P(A and B) = P(B) P(A|B) esssendo i nomi assegnati ai due eventi interscambiabili

PROBABILITA CONDIZIONATA La barra verticale si chiama probabilità condizionata e si interpreta come la probabilità del primo evento, calcolata sapendo che il secondo evento si è verificato.

Definizione di Informazione

Abbiamo affermato che l'idea base di misurazione di informazione è che **tanto è piccola la probabilità tanto più grande deve essere la misura di informazione ad essa associata**.

Inoltre **l'informazione legata ad una probabilità 1**, ossia l'evento certo ci deve fornire informazione nulla, **ossia pari a 0**

Vorremmo anche che valesse un'altra proprietà, **quella additiva**, ossia se riceviamo in momenti successivi due quantità di informazione, desideriamo che **il totale ricevuto sia la somma delle due quantità**

Domanda: quale funzione matematica ci garantisce le ultime due condizioni ? risposta la funzione **Logaritmo**

Proprietà della funzione logaritmo

Infatti ricordiamo che il logaritmo di un numero ad esempio N è quell'esponente che occorre formire ad una base per ottenere quel numero. In altre parole se

$$a^x = N$$

allora

$$x = log_a N$$

Qualunque base intera scegliamo il logaritmo gode di queste interessanti e utili proprietà:

$$log(1) = 0$$

infatti qualsiasi base elevata alla 0 da 1

inoltre sappiamo che vale la proprietà addittiva per un prodotto

$$log(a * b) = log(a) + log(b)$$

cioè il logaritmo trasforma una moltiplicazione in una somma (e anche un elevqmento a potenza in una moltiplicazione, ma per il momento questa proprietà ci interessa di meno)

Allora definiziamo l'informazione legata ad un certo evento E che ha un probabilità p con la sgiente formula:

$$I_E = log(\frac{1}{p})$$

Questa misura rispetta tutte le proprietà desiderate: infatti se p = 1 l'informazione è 0 poichè 1/1 = 1 e il suo logaritmo è 0

Inoltre se due eventi E_1 E_2 accadono uno dopo l'altro con probabilità p e q rispettivamente e sono indipendneti, la loro probabilità congiunta ossia che si verificahino entrambi è p*q
L'informazione totale ricevuta sarà

$$log\frac{1}{p*q} = log(\frac{1}{p} * \frac{1}{q}) = log(\frac{1}{p}) + log(\frac{1}{q}) = I_1 + I_2$$

La definizione scelta non solo è appropriata, ciè rispetta le priprietà desdierate ma si dimosstra che sia l'unica possinìbile

Chiameremo questa misura **Autoinformazione**

Definizione di Inofrmazione e di Autoinformazione

$$I_x = log(\frac{1}{p}) = -log(p)$$

possiamo notare che le due forme sono equivalenti visto che la proprietà addittiva si applica anche alla divisione dove questa volta la divisione fra due valori, e una frazione è di fatto una divisione, viene trasformata in una sottrazione. E poichè abbiamo visto che log(1) = 0, di conseguenza 0 - log(p) = -log(p)

NOTA: l'autoinformazione è legata ad un evento, nellla nostra trattazione quindi ad un simbolo della sorgente ed è un caso particolare, in cui una volta ricevuto o scoperto il v alore del carattere la sua probabilità da p diventa 1 ossia certezza.
Questo perché abbiamo ipotizzato, che non vi siano errori di trasmissione, ossia rumore sul canale.

In generale infatti

$$I = log(\frac{probabilità\ a\ posteriori}{probabilità\ a\ priori})$$

Dopo che abbiamo ricevuto una informazione, la probabilità associata a quell'evento cambia e diventa probabilità a posteriori, ossia il valore di incertezza che avremo dopo che l'evento si è verificato.
L'evento quindi ci fornirà una informazione quantificata dalla formula precedente.

Entropia dell'Informazione

Entropia dell'informazione è l'informazione media legata all'intera sorgenza

$$H(x) = \sum_{i=1}^{n} I(x_i)p(x_i) = \sum_{i=1}^{n} log_2(\frac{1}{p_i})p(x_i) = -\sum_{i=1}^{n} log_2(p(x_i))\ p(x_i)$$

```
sorgente = {'a' : 1/2, 'b': 1/4, 'c':1/8, 'd':1/8}
```

L'entropia ha un valore sempre positivo, al minimo 0 che rappresenta una completa conoscenza e di conseguenza informazione nulla. Il massimo si ottiente quando la distribuzione di probabilità è uniforme e vale log(n) dove n è il numero di simboli della sorgente

```
import math
def informazione(carattere):
  p = sorgente[carattere]
  return (-math.log2(p))
```

+ Codice + Testo

```
informazione('c')
```

```
3.0
```

```
def entropia(sorgente):
  somma = 0
  for carattere in sorgente:
    somma = somma + informazione(carattere) * sorgente[carattere]
  return somma
```

```
entropia(sorgente)
```

```
1.75
```

```
Inizia a programmare o genera codice con l'IA.
```