

Análisis de Datos

Clase 5: Teoría de la información + Tests estadísticos



Teoría de la información

Teoría de la información

El área de Teoría de la Información surge originalmente en el ámbito de las telecomunicaciones, buscando responder dos preguntas: cuál es la máxima compresión de datos y cuál es la máxima tasa de transmisión de datos.

Sin embargo, los conceptos obtenidos tienen aplicación en muchos otros ámbitos, entre ellos Data Science y Machine Learning.

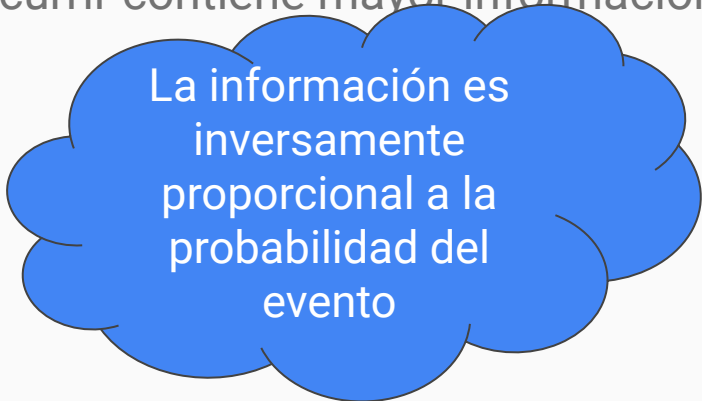
¿Qué es la información?

La información de un dato o evento está asociada a la incertidumbre del mismo. En otras palabras, está asociada a la probabilidad de ocurrencia que tenga.

Un evento con muy baja probabilidad de ocurrir contiene mayor información que uno con alta chance

Ejemplo:

- El sol va a salir mañana
- Mañana va a llover
- Mañana va a ocurrir un ciclón



La información es
inversamente
proporcional a la
probabilidad del
evento

Información

Formalmente, la información de un evento la podemos definir como

$$I(A) = \log \frac{1}{\mathbb{P}(A)} = -\log(\mathbb{P}(A))$$

Observar que si A resulta de la ocurrencia de dos eventos independientes, es decir que $A = B \cap C$ la información resultante es la suma de los eventos B y C :

$$I(A) = I(B \cap C) = \log \frac{1}{\mathbb{P}(B)\mathbb{P}(C)} = \log \frac{1}{\mathbb{P}(B)} + \log \frac{1}{\mathbb{P}(C)} = I(B) + I(C)$$

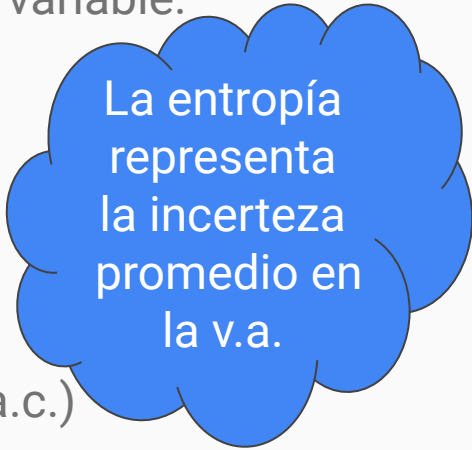
Si \log es en base 2 la unidad es bits, si es en base e la se mide en nats.

Entropía

La entropía de una variable aleatoria, se corresponde con la esperanza de la información que conlleva cada uno de los posibles valores de la variable.

Matemáticamente, la entropía se define como

$$\begin{aligned} H(X) = \mathbb{E}[I(x)] &= \sum_{x \in A} p_X(x) \log \frac{1}{p_X(x)} = - \sum_{x \in A} p_X(x) \log p_X(x) \text{ (v.a.d)} \\ &= \int f_X(x) \log \frac{1}{f_X(x)} dx = - \int f_X(x) \log(f_X(x)) dx \text{ (v.a.c.)} \end{aligned}$$



La entropía
representa
la incerteza
promedio en
la v.a.

Ejemplo

Calcular cómo se comporta la entropía al arrojar una moneda para los distintos valores posibles de éxito (p)

Entropía: propiedades

Propiedades de la entropía

1. $H(X) \geq 0$. Observar que si $H(X)=0$, entonces no hay incertidumbre, y la variable no era realmente aleatoria
2. $H_b(X) = (\log_b a) H_a(X)$. Esta fórmula nos permite encontrar la equivalente de la entropía en distintas unidades.

Entropía conjunta

Así como definimos la entropía para una única variable, podemos definir la entropía conjunta entre dos variables:

$$\begin{aligned} H(X, Y) &= -\mathbb{E}[\log p_{X,Y}(X, Y)] = - \sum_{x \in A_X} \sum_{y \in A_Y} p(x, y) \log p_{X,Y}(x, y) \quad (X, Y \text{ v.a.d.}) \\ &= -\mathbb{E}[\log f_{X,Y}(X, Y)] = - \int \int f(x, y) \log p_{X,Y}(x, y) dy dx \quad (X, Y \text{ v.a.c.}) \end{aligned}$$

Entropía condicional

Vamos a poder calcular también la entropía condicional. La misma se va a corresponder con la entropía de la variable condicionada $Y|X=x$:

$$\begin{aligned} H(Y|X) &= - \sum_{x \in A_X} p_X(x) H(Y|X=x) = - \sum_{x \in A_X} \sum_{y \in A_Y} p(x, y) \log p_{Y|X=x}(y|x) = -\mathbb{E}[\log p(Y|X)] \\ &= - \int f_X(x) H(Y|X=x) dx = - \int \int f(x, y) \log f_{Y|X=x}(y|x) dy dx = -\mathbb{E}[\log f(Y|X)] \end{aligned}$$

Otra forma de escribirlo:

$$\begin{aligned} H(X, Y) &= - \sum_{x \in A_X} \sum_{y \in A_Y} p(x, y) \log \frac{p(x, y)}{p(x)} \\ &= - \int \int f(x, y) \log \frac{f(x, y)}{f(x)} dy dx \end{aligned}$$

Propiedades

- Condicionar reduce la entropía: $H(X|Y) \leq H(X)$
- $H(X_1, X_2, \dots, X_n) \leq H(X_1) \dots H(X_n)$. La igualdad vale s.i.i X_i son independientes
- $H(X) \leq \log|Sop_X|$ La igualdad vale únicamente para v.a. uniformemente distribuidas
- $H(p)$ es cóncava en p .

Algunas relaciones

Regla de la cadena:

$$H(X, Y) = H(X) + H(Y|X)$$

Corolario:

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$$

Ejemplo

Sea (X,Y) un vector aleatorio con la siguiente función de probabilidad conjunta:

$Y \backslash X$	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

Calcular, $H(X)$, $H(Y)$, $H(X,Y)$, $H(Y|X)$, $H(X|Y)$

Entropía relativa o divergencia de Kullback-Leibler

Dadas dos funciones de probabilidad p , q , la divergencia de Kullback-Leibler se define como

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Propiedades:

- $D(p||q) \geq 0$ y $D(p||q)=0$ s.i.i. X e Y son independientes
- $D(p||q)$ es convexa en el par (p,q)
- $D(p||q) \neq D(q||p)$

Información mutua

La información mutua entre dos v.a X e Y se define como

$$I(X; Y) = \sum_x \sum_y p_{X,Y}(x, y) \log \frac{p_{X,Y}(x, y)}{p_X(x)p_Y(y)}$$

$$I(X; Y) = \int \int f_{X,Y}(x, y) \log \frac{f_{X,Y}(x, y)}{f_X(x)f_Y(y)}$$

Propiedades

- $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(X|Y) = H(X) + H(Y) - H(X,Y)$
- $I(X;Y) = D(p_{X,Y}(x,y) || p_X(x)p_Y(y)) \geq 0$. $I(X;Y) = 0$ si X, Y son independientes
- $I(X;Y) = I(Y;X)$
- $I(X;X) = H(X)$

Regla de la cadena

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1).$$

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1).$$

$$D(p(x, y) || q(x, y)) = D(p(x) || q(x)) + D(p(y|x) || q(y|x)).$$

Tests estadísticos

Test de comparación de medias

El objetivo es saber si la media de dos poblaciones son diferente.

Si tengo \underline{X}_n e \underline{Y}_m dos m.e. de tamaños n y m correspondientes a dos poblaciones $X \sim \mathcal{N}(\mu_X, \sigma^2)$ e $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$ respectivamente y , diseñamos el test

$$H_0 : \Delta = 0 \quad \text{vs.} \quad H_1 : \Delta \neq 0, \quad \Delta = \mu_X - \mu_Y$$

El estadístico utilizado será

$$U(\underline{X}, \underline{Y}, \Delta) = \frac{\bar{X} - \bar{Y} - \Delta}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}, \quad S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{n+m-2}$$

Rechazaremos el test cuando $|U(\underline{x}, \underline{y}, 0)| > t_{n+m-2, (1-\alpha/2)}$

ANOVA

- ANOVA:

Para comparar las medias de dos poblaciones con distribución normal podemos usar el test de t de Student. Si queremos comparar las medias de más de dos conjuntos usamos ANOVA.

Tenemos k categorías cuyas medias (reales) son μ_1, \dots, μ_k , cuyas medias muestrales son $\bar{x}_1, \dots, \bar{x}_k$ y los desvíos muestral estándar S_1, \dots, S_k .

ANOVA analiza las hipótesis: $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ vs. H_1 : “no todas las medias son iguales”

- Supone: independencia entre observaciones, distribución normal de las variables numéricas, homocedasticidad
- Analiza relación lineal entre variables

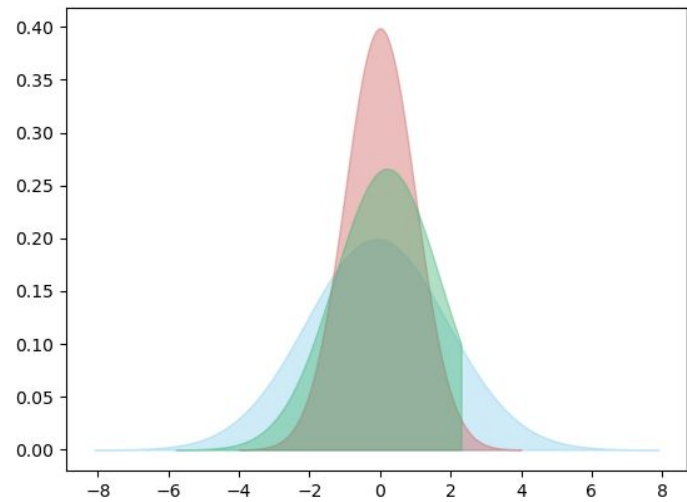
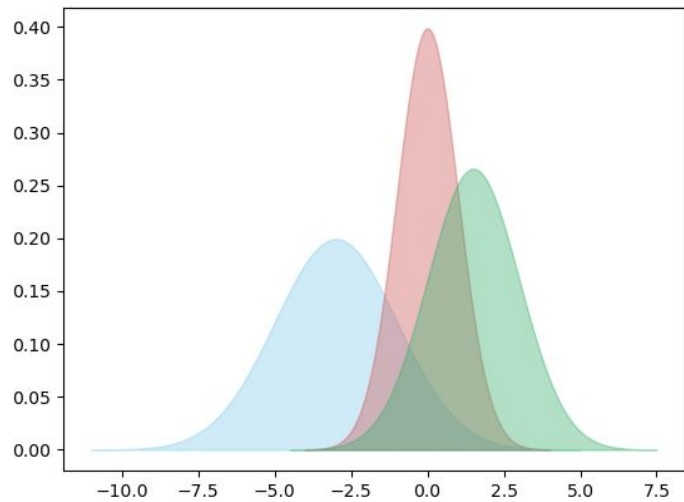
ANOVA

Cálculo del estadístico:

- Calculamos la media total $\bar{x} = \frac{\sum_{i=1}^k n_i \bar{x}_i}{N}$, donde $N = \#$ de muestras y $n_i = \#$ de muestras de clase i
- Estimamos la varianza entre grupos $S_e^2 = \frac{\sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2}{k-1}$
- Estimamos la varianza dentro de los grupos $S_d^2 = \frac{\sum_{i=1}^k (n_i - 1) S_k^2}{N-k}$
- Definimos el test $F = \frac{S_e^2}{S_d^2} \sim F_{k-1, N-k}$

F va a ser grande si la varianza entre clases es mucho mayor que var. dentro de las clases, lo cual es poco probable que ocurra si las medias son todas iguales.

ANOVA



Ejemplo

- Un grupo de amigos discute en un bar si Messi, Riquelme y Maradona rindieron igual de bien en la selección argentina de fútbol. Proponen usar como criterio la cantidad de goles por partido para describir un comportamiento más general del juego de cada jugador en la selección nacional. Usar un test de ANOVA con significancia de 5% para responder la duda planteada por el grupo de amigos.

	Maradona	Messi	Riquelme
No. Partidos en Selección	91	142	51
Goles Promedio en Selección	0.37	0.5	0.33
Desvío estándar Goles en Selección	4.6	5.9	3.4

Test Chi-cuadrado

- Test de Chi-Cuadrado (test de independencia de Pearson):

$$\chi = \sum_{i,j} \frac{O_{ij} - E_{ij}}{E_{ij}}$$

donde O_{ij} son la cantidad de observaciones pertenecientes a las categorías i, j de cada variable, y E_{ij} es el valor esperado observado si las variables fueran independientes.

- Se usa para rechazar la H_0 que las variables son independientes.
- $\chi \sim \chi^2_{r-1, k-1}$, r y k son la cantidad de factores de las variables de entrada y salida respectivamente.

Ejemplo

- Se quiere saber si algunos genios del fútbol rinden mejor que otros (meten más goles) en sus equipos que en la selección nacional. Usar un test de independencia con significancia de 5% para responder la pregunta.

Genio del Fútbol	Goles Selec. Nacional	Goles Equipos
Maradona	34	320
Messi	71	741

Datos verdaderos al 13 Mayo 2021.