

ANÁLISIS DE DATOS



Clase 5. Pipelines,
Transformación de
variables y Teoría de la
información

Temario



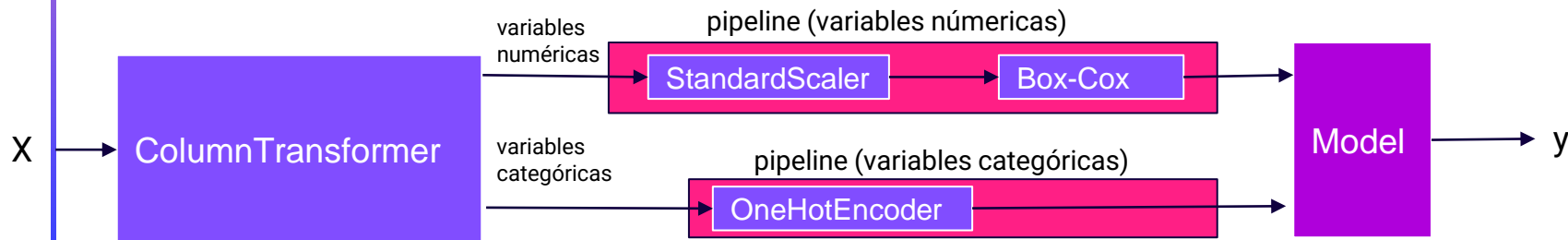
1. Cadenas de procesamiento.
2. Transformación de variables.
3. Conceptos básicos de Teoría de la información
4. Entropía y entropía conjunta
5. Entropía relativa
6. Información mutua
7. Test estadísticos

1. CADENAS DE PROCESAMIENTO



Implementación de cadenas de procesamiento con SKLearn

- Arquitectura de SKlearn:
 - **Transformer**: clase para transformar datos. Debe implementar fit() y transform().
 - **Predictor**: clase para realizar predicciones. Debe implementar fit() y predict().
 - **Pipeline**: clase que ejecuta secuencia de transformers y/o predictors.
 - Cada elemento de la lista es un paso (step).
 - El único elemento de la lista que puede ser predictor es el último (los predecesores deben ser transformers).
 - **Columntransformer**: clase que permite aplicar transformaciones específicas para cada columna de un dataset.



Implementación de cadenas de procesamiento con SKLearn

- Ejemplos de uso de clases **Pipeline** y **ColumnTransformer** para problemas de clasificación y regresión.
- Referencias:
 - <https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html>
 - <https://scikit-learn.org/stable/modules/compose.html#columntransformer-for-heterogeneous-data>
 - Sección de transformaciones de datos de documentación de SKLearn: https://scikit-learn.org/stable/data_transforms.html

Ejemplos en jupyter

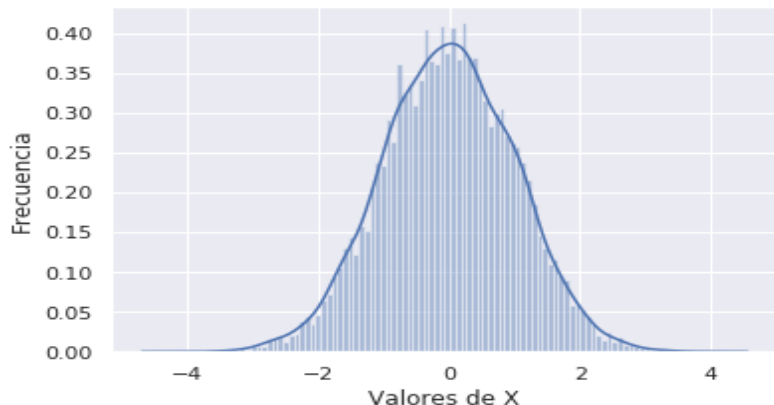
Clase 4.8 - Preparación de datos - Cadenas de procesamiento.ipynb

2. TRANSFORMACI ÓN DE VARIABLES



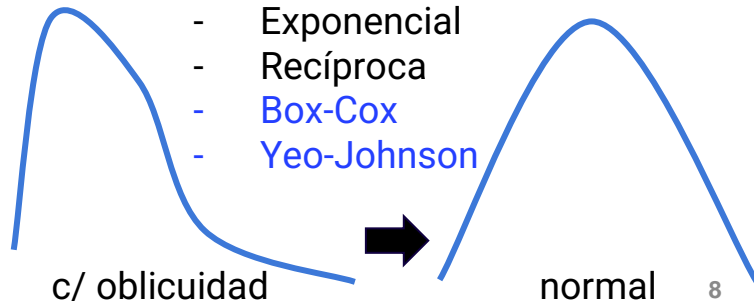
Distribución normal en modelos lineales

- Es deseable que los valores de cada variable independiente (X) tengan una distribución normal.



- Es muy común que esto no se cumpla en los datos originales. En este caso, puede intentarse obtenerse una distribución más semejante a una normal luego de aplicar una transformación.

- Logarítmica
- Exponencial
- Recíproca
- Box-Cox
- Yeo-Johnson



Transformaciones matemáticas

- Logarítmica: $\log(x)$, $x > 0$
- Recíproca: $\frac{1}{x}$, $\forall x \in \mathbb{R} - \{0\}$
- Potencia/exponencial
 - $X e^{\lambda}$
 - $X^{1/2}/X^3$
 - No definido para todo X.
- Exponencial (casos especiales):
 - *Box-Cox*, $X > 0$
 - *Yeo-Johnson*

Transformación de Box Cox

- La transformación de Box-Cox estima un valor de lambda que minimiza la desviación estándar de una variable transformada estandarizada.
- El método de Box-Cox busca entre muchos tipos de transformaciones.

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_i) & \text{if } \lambda = 0, \end{cases}$$

L	Y'
-2	$Y^{-2} = 1/Y^2$
-1	$Y^{-1} = 1/Y^1$
-0.5	$Y^{-0.5} = 1/(\text{Sqrt}(Y))$
0	$\log(Y)$
0.5	$Y^{0.5} = \text{Sqrt}(Y)$
1	$Y^1 = Y$
2	Y^2

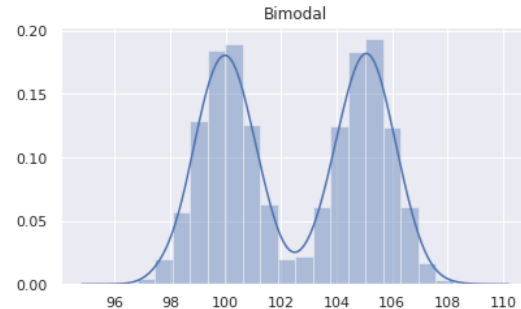
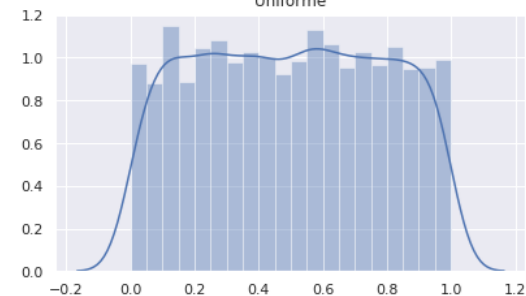
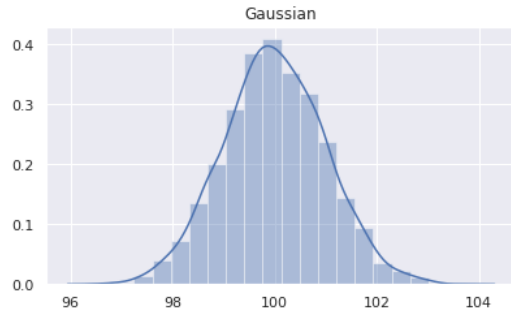
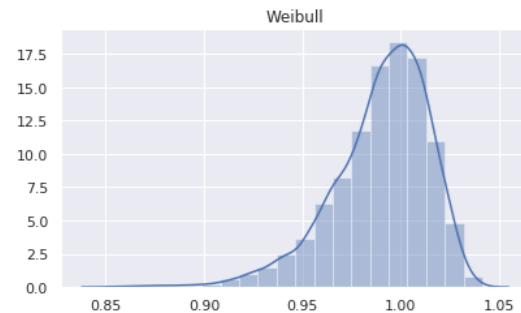
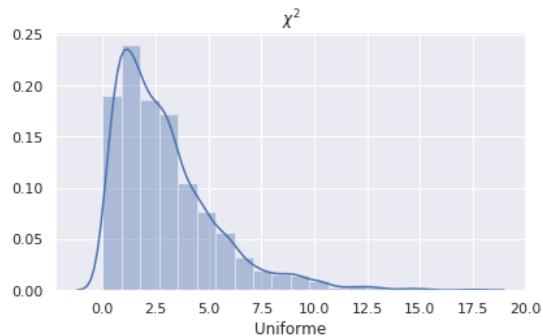
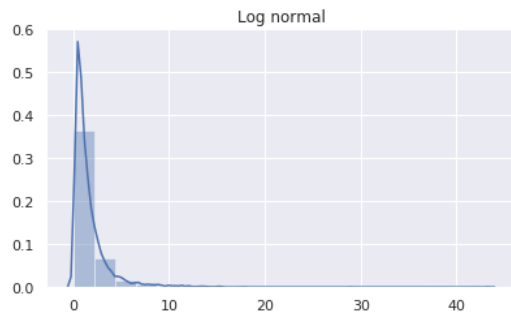
Source: Box and Cox (1964). Where, 'Y' is the transformation of t
Note that for Lambda = 0, the transformation is NOT Y^0 (because t

Transformación de Yeo-Johnson

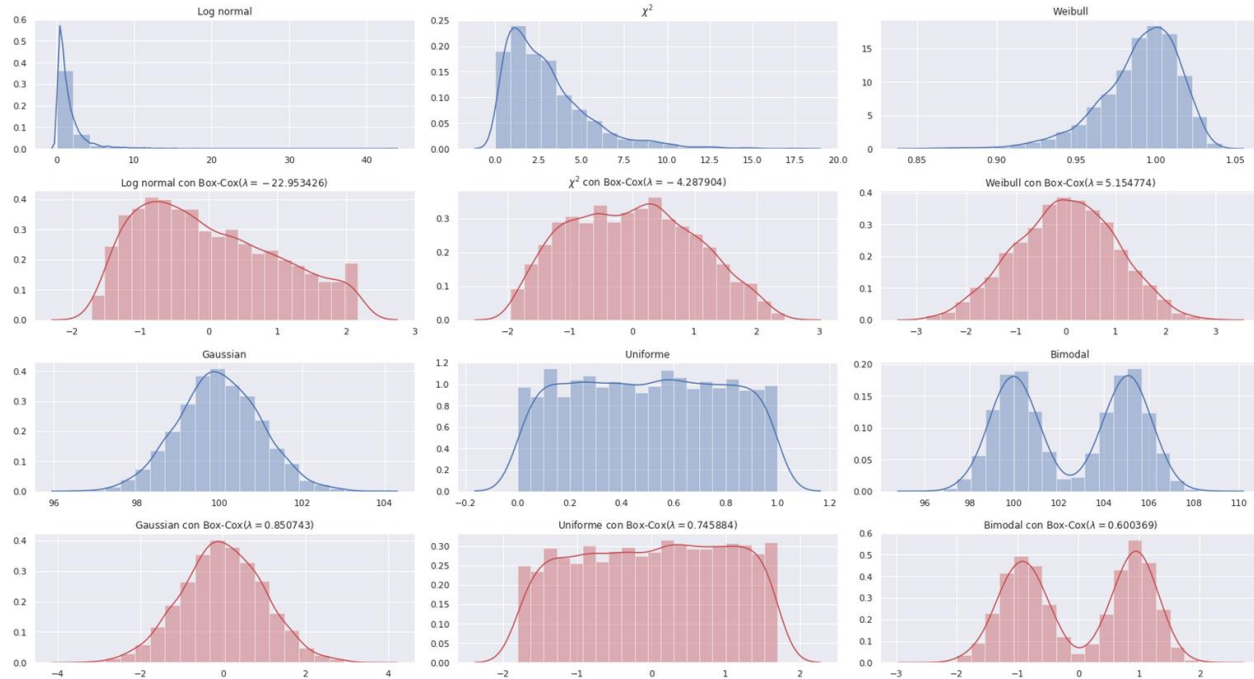
- Junto con Box-Cox, es otra de las transformaciones más utilizadas.
- Admite que X tome valores negativos.

$$y_i^{(\lambda)} = \begin{cases} ((y_i + 1)^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y_i + 1) & \text{if } \lambda = 0, y \geq 0 \\ -[(-y_i + 1)^{(2-\lambda)} - 1]/(2 - \lambda) & \text{if } \lambda \neq 2, y < 0 \\ -\log(-y_i + 1) & \text{if } \lambda = 2, y < 0 \end{cases}$$

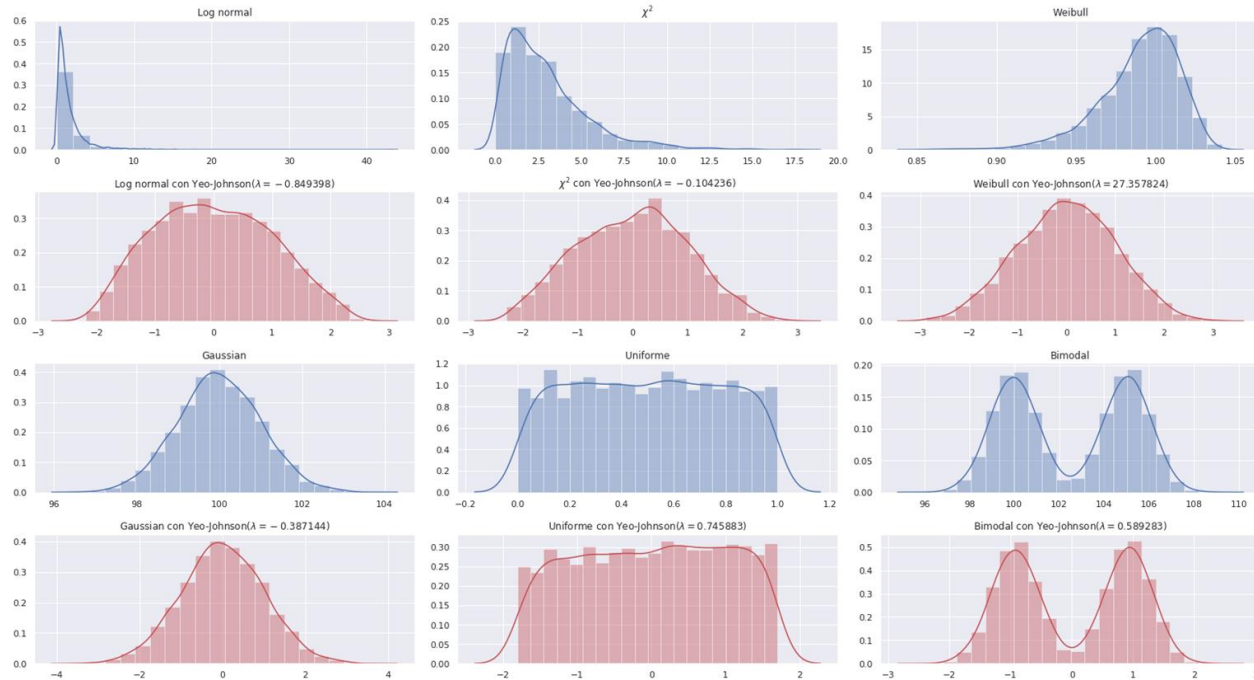
Ejemplos



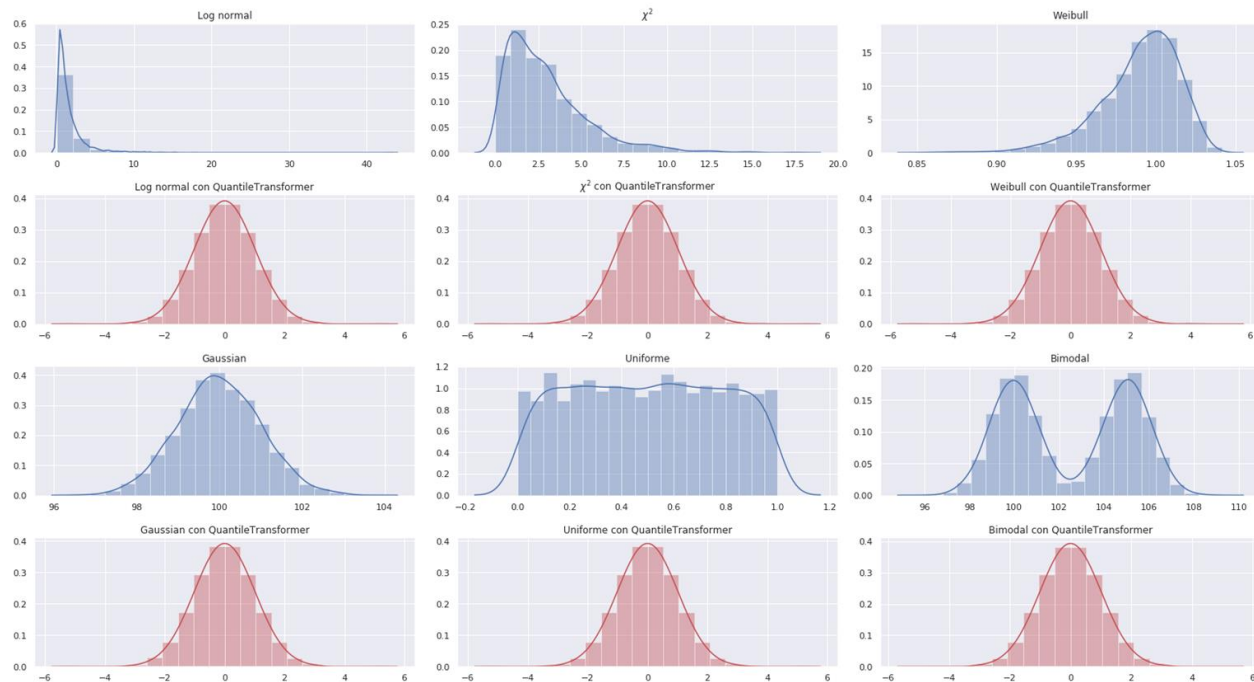
Ejemplos. Box-Cox



Ejemplos. Yeo-Johnson



Ejemplos. QuantileTransformer



Ejemplos en jupyter

Clase 4.5 - Preparación de datos - Transformación de variables.ipynb

Resumen

El objetivo de esta clase fue presentar algunas de las técnicas más utilizadas para la preparación de datasets y cómo combinarlas para construir cadenas de procesamiento en SKLearn.

En las siguientes clases, se estudiarán otros aspectos de la **preparación de datos**:

La **selección de variables** de entrada: ¿Qué variables son más relevantes para el modelo de AA?

Ingeniería de variables (como generar variables que aporten mayor información, a partir de las existentes).

Reducción de dimensiones. Creación de un nuevo espacio de variables de entrada, a partir de proyecciones compactas.

TEORÍA DE LA INFORMACIÓN

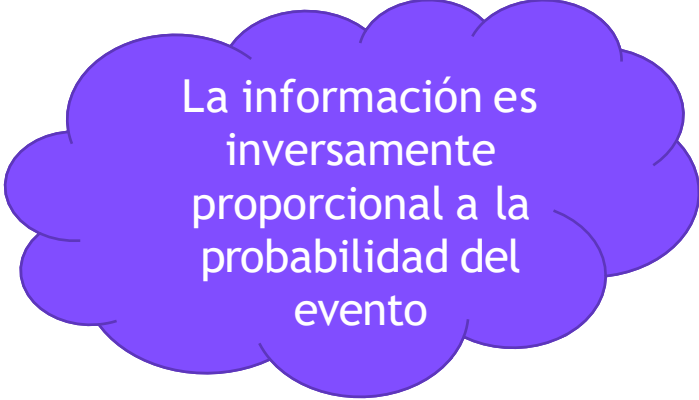


Teoría de la información

- 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.
- Ejemplos:
 - 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:

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

- 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 :

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

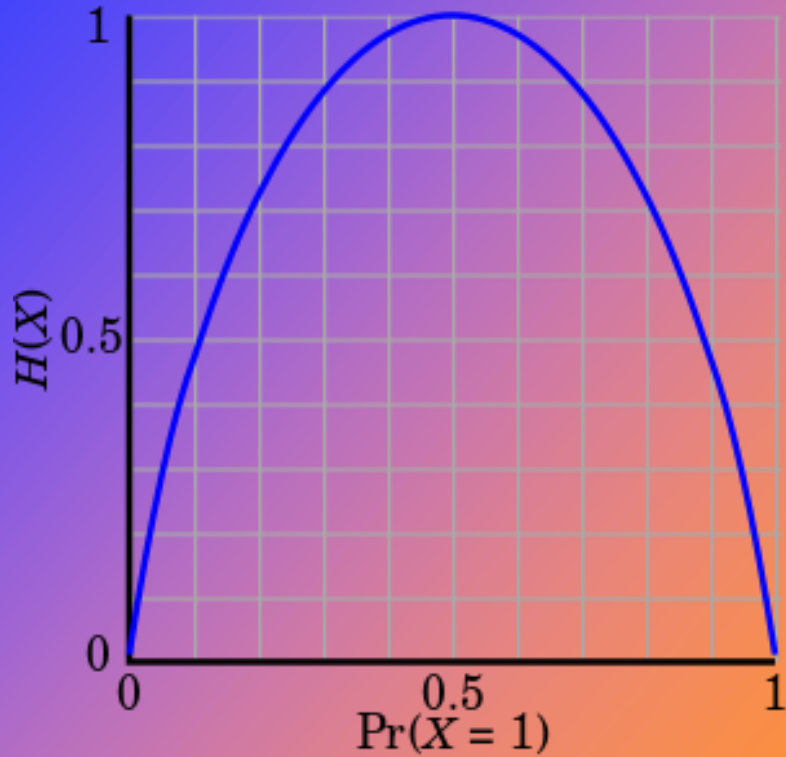
- Si \log es en base 2 la unidad es bits, si es en base e la se mide en *nats*.

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:

$$H(X) = \mathbb{E}[I(x)] = \begin{cases} \sum_{x \in A} p_X(x) \log \frac{1}{p_X(x)} = - \sum_{x \in A} p_X(x) \log p_X(x) \\ \int_A f_X(x) \log \frac{1}{f_X(x)} dx = - \int_A f_X(x) \log f_X(x) dx \end{cases}$$

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



Ejercicio

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

Entropía: propiedades

- Propiedades de la entropía:
 - $H(X) \geq 0$.
 - Observar que si $H(X) = 0$, entonces no hay incertidumbre, y la variable no era realmente aleatoria
 - $H_b(X) = (\log_b a) H_a(X)$.
 - 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:

$$H(X, Y) = -\mathbb{E}[\log p_{X,Y}(X, Y)]$$

$$H(X, Y) = \begin{cases} -\sum_{x \in A_x} \sum_{y \in A_y} p(x, y) \log p_{X,Y}(x, y) \\ -\iint_{A_x \times A_y} f(x, y) \log p_{X,Y}(x, y) dx dy \end{cases}$$

Entropía condicional

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

$$H(Y|X) = \begin{cases} -\sum_{x \in A_x} p_X(x) H(Y|X = x) = -\sum_{x \in A_x} \sum_{y \in A_y} p(x, y) \log p_{Y|X=x}(y|x) \\ \int_{A_x} f_X(x) H(Y|X = x) dx = -\iint_{A_x \times A_y} f(x, y) \log f_{Y|X}(y|x) dx dy \end{cases}$$

- Puede escribirse también como:

$$H(Y|X) = \begin{cases} \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)} \\ \int_{\mathbb{R}^2} f(x, y) \log \frac{f(x, y)}{f(x)} dy dx \end{cases}$$

Propiedades

- $H(X|Y) \leq H(X)$. Condicionar reduce la entropía.
- $H(X_1, \dots, X_n) \leq H(X_1) \cdot \dots \cdot H(X_n)$. Esta igualdad se cumple sii X_i son independientes.
- $H(x) \leq \log|sop_X|$. Esta igualdad vale unicamente cuando la variable aleatoria es uniforme.
- $H(p)$ es cóncava en p .

Algunas relaciones

Regla de la cadena:

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

Corolario:

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

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

Ejemplo

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:

- $D(p||q) > 0$ y $D(p||q) = 0$ s.i.i $P=Q$
- $D(p||q)$ 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:**

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

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 X_n, Y_n dos muestras aleatorias de tamaño n, m respectivamente ambas distribuidas de manera normal con varianza σ y medias μ_x, μ_y respectivamente, diseñamos el test:

$$H_0: \Delta = 0 \text{ vs } H_1: \Delta \neq 0; \Delta = \mu_x - \mu_y$$

Bajo esta condición el estadístico será de la forma:

$$U(X, Y, \Delta) = \frac{X - Y - \Delta}{S_p \sqrt{n^{-1} + m^{-1}}} \sim t_{n+m-2}; \quad S_p^2 = \frac{(m-1)S_x^2 + (n-1)S_y^2}{n+m-2}$$

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

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 y sus medias muestrales $\bar{x}_1, \dots, \bar{x}_k$ y desvíos s_1, \dots, s_k

ANOVA propone el test:

$$\begin{cases} H_0: \mu_1 = \dots = \mu_k = \mu \\ H_1: \exists i \in [1, k] \setminus \mu_i \neq \mu \end{cases}$$

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

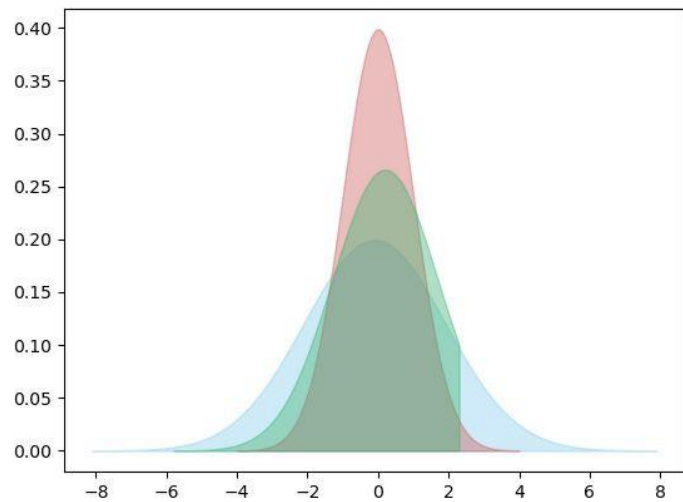
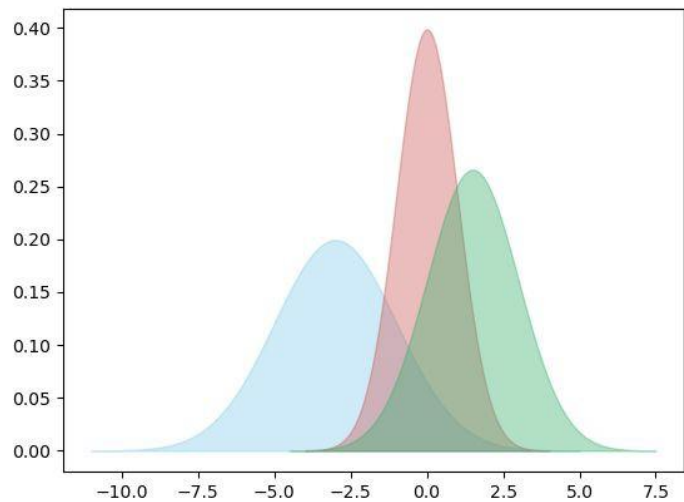
ANOVA

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.
- $\chi \sim \chi^2_{r-1k, -1}$, 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.

+



o



.



DUDAS?

ENCUESTA