

ANÁLISIS DE DATOS



Clase 6.

Temario



1. Conceptos básicos de Teoría de la información
 - Entropía y entropía conjunta
 - Entropía relativa
 - Información mutua
2. Selección de variables
3. Reducción de dimensiones
4. Test estadísticos

TEORÍA DE LA INFORMACIÓN

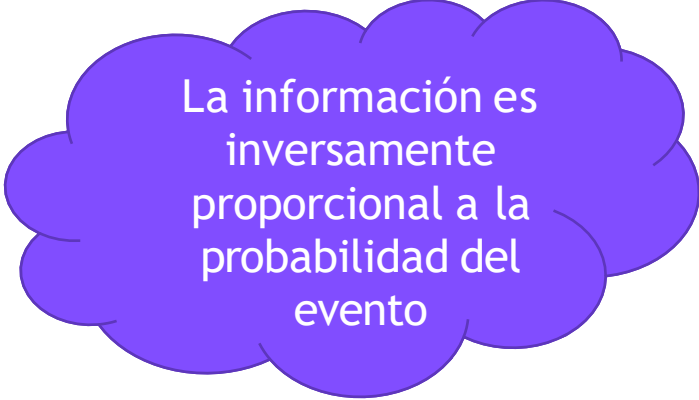


Teoría de la información

- 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.
- 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 :

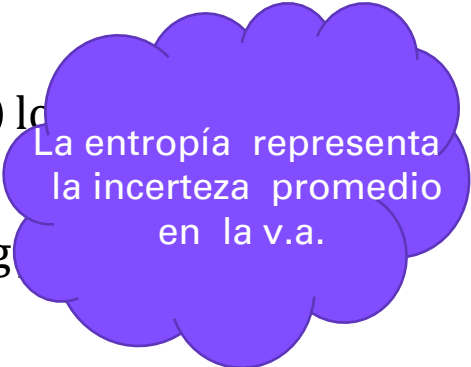
$$I(A) = I(B \cap C) = \log \frac{1}{P(B)P(C)} = \log \frac{1}{P(B)} + \log \frac{1}{P(C)}$$
$$I(A) = I(B) + I(C)$$

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

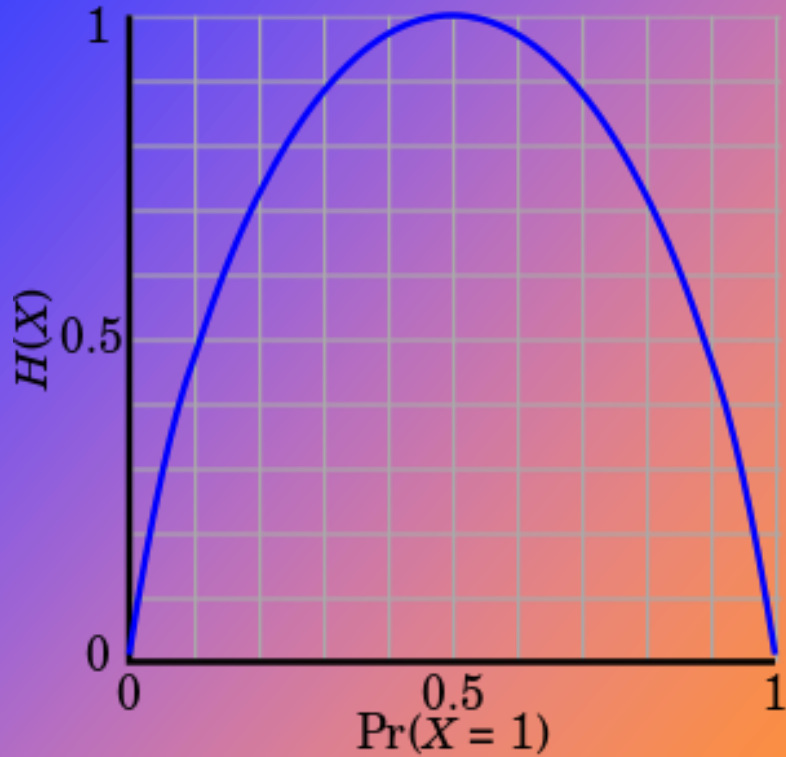
Entropía

- 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 desvios 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

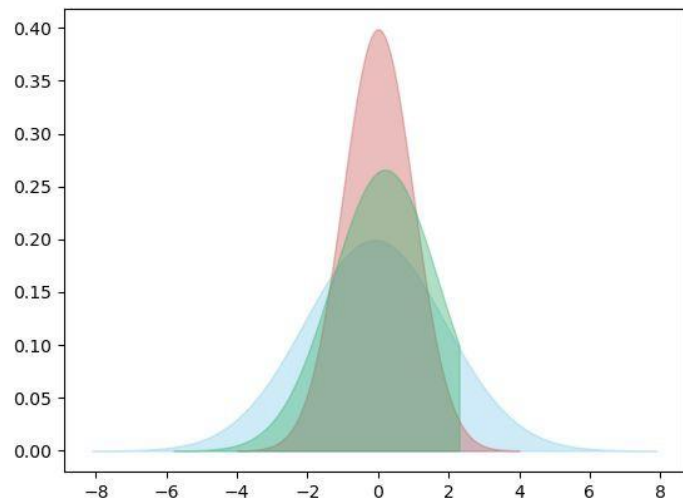
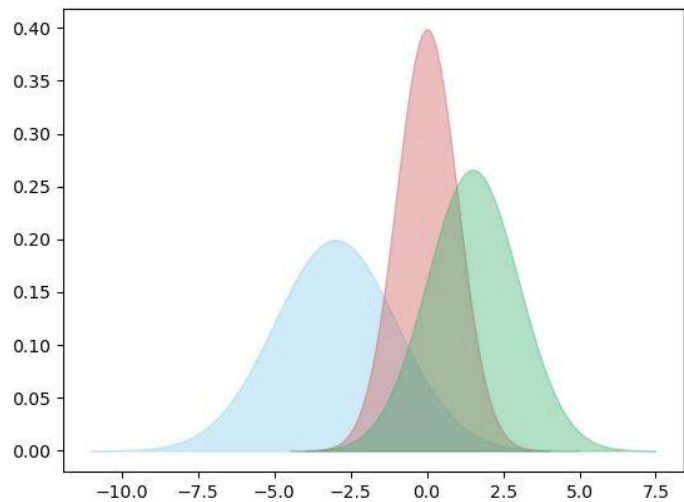
ANOVA

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

SELECCIÓN DE FEATURES



Métodos de proyección

Estos métodos de reducción de dimensiones se basan en aplicar transformaciones, en principio lineales, a los datos originales. De este modo esperamos capturar las direcciones de mayor importancia en los datos.

En esta materia vamos a estudiar tres métodos

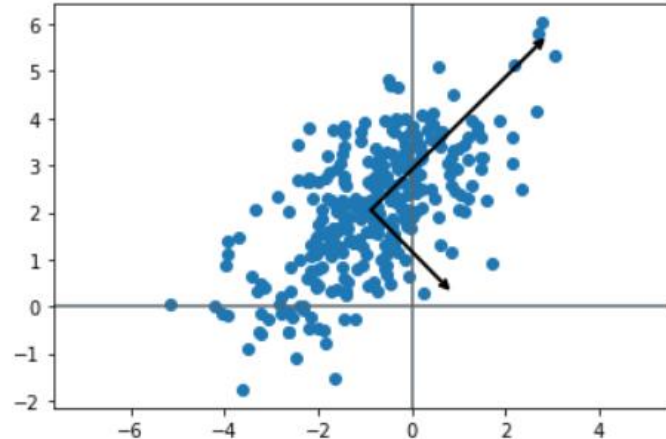
- PCA
- ICA
- SVD

pero existen muchos otros como KPCA, FA, etc.

Análisis de componentes principales (PCA)

Motivación

PCA busca proyectar los datos en un espacio lineal (de menor dimensión), llamado *subespacio principal*, tal que la varianza en los datos proyectados sea máxima.



Análisis de componentes principales (PCA)

Cómo hallar las direcciones de máxima varianza

Supongamos que tenemos una matriz $X \in \mathbb{R}^{n \times p}$ una matriz de n observaciones con p variables.

Buscamos hallar las $m \leq p$ direcciones que maximicen la varianza de las muestras.

1. Primero hallamos $\alpha_1 \in \mathbb{R}^n / \alpha_1^T X$ sea máximo sujeto a $\alpha_1^T \alpha_1 = 1$.
2. Luego buscamos $\alpha_2 \in \mathbb{R}^n / \alpha_2^T X$ sea máximo sujeto a $\alpha_2^T \alpha_2 = 1$ y además $\alpha_2^T X$ esté descorrelacionado con $\alpha_1^T X$. ($\alpha_1 \perp \alpha_2$)
3. Se prosigue de la misma forma hasta hallar los $\alpha_1, \dots, \alpha_m$ vectores

Notar que por definición la matriz $\alpha_m = [\alpha_1, \dots, \alpha_m]$ es ortonormal, y por lo tanto define una matriz de proyección.

Análisis de componentes principales (PCA)

Vinculación con los autovectores

Definiendo $\tilde{\Sigma} = (\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}})$, la matriz de covarianza muestral, los α_i están asociados a los m primeros autovectores de $\tilde{\Sigma}$.

Al ser $\tilde{\Sigma}$ simétrica, $\tilde{\Sigma} = \mathbf{V} \mathbf{S} \mathbf{V}^T$,

Donde $\mathbf{S} = \text{diag}(\lambda_1, \dots, \lambda_p)$, $\lambda_1 \geq \dots \geq \lambda_p$ son los autovalores de $\tilde{\Sigma}$, $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_p]$ y con \mathbf{v}_i el autovector asociado a λ_i .

Se concluye que $\alpha_m = \mathbf{V}_m$ donde $\mathbf{V}_m = [\mathbf{v}_1, \dots, \mathbf{v}_m]$

Los datos transformados resultan $\hat{\mathbf{X}} = \mathbf{X} \mathbf{V}_m$

Análisis de componentes principales (PCA)

Criterios de selección de orden

Una pregunta importante es cómo elijo la cantidad de features a retener. Existen dos enfoques comúnmente usados:

- Busco explicar el $k\%$ de la varianza de los datos: es tal que
- Método del codo ('elbow'). Gráfico los y tomo en el punto de inflexión de la curva.

Análisis de componentes principales (PCA)

Comentarios finales

- **Ventajas:**
 - Obtengo features descorrelacionados
 - No "tiro" información de ninguna variable
 - Explicable en términos de la matriz de correlación
 - No supervisado (sirve para mayor cantidad de problemas)
- **Desventajas:**
 - Si el dataset es muy grande puede ser muy costoso de computar
 - Pierdo explicabilidad de los features (ahora son una c.l. de las mediciones)
- **Observaciones:**
 - Las direcciones de los componentes principales se pueden ver afectadas por las unidades de medida, por ejemplo un feature es la altura en metros de una persona y otra el peso en gramos). Una práctica común estandarizar las variables para que tengan media 0 y varianza 1 antes de aplicar PCA.

Independent Component Analysis (ICA)

ICA, también conocido como “*blind source separation*” o “*Cocktail party problem*”, busca transformar el dataset en columnas independientes.

El modelo asume que cada señal se puede modelar como una combinación lineal de componentes independientes. Sean las fuentes independientes, s_1, \dots, s_k , luego cada señal se modela como}

$$\mathbf{x} = \alpha_1 \mathbf{s}_1 + \dots + \alpha_k \mathbf{s}_k$$

Si definimos $X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, podemos escribirlo como

$$X = A \begin{bmatrix} s_1 \\ \vdots \\ s_k \end{bmatrix} = AS$$

ICA

Suposiciones de ICA:

- El proceso de mezcla es lineal
- Todas las fuentes de señal son independientes
- Las señales **no** son gaussianas

Nota: a diferencia de lo que ocurre en PCA, el orden de las componentes no significa nada.

ICA

Observación:

ICA no es exactamente una técnica de reducción de dimensiones, sin embargo puede usarse para tal fin. Supongamos que se tiene un dataset de n observaciones, con q variables cada una. Luego, si modelamos las observaciones como provenientes de $m < q$ fuentes, obtenemos una reducción de dimensiones.

```
import numpy as np
from sklearn.decomposition import FastICA

# Generate random Data of size (n x 5).
X = np.random.uniform(low=0, high=100, size=(20, 5))

# Number of sources wanted. The resulted sources are (n x 3).
ica = FastICA(n_components=3)
sources = ica.fit_transform(X)
```

ICA - Preprocesamiento

- **Centrado:** restamos la media de todas las señales

- **Blanqueado:** Buscamos descorrelacionar las variables y estandarizar las a varianza unitaria. Esto puede hacerse con PCA.
$$D = X - \mu = \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} = \begin{bmatrix} x_1 - \mu \\ \vdots \\ x_n - \mu \end{bmatrix}$$

Observación: Si las señales fueran gaussianas, entonces con PCA alcanza para obtener señales independientes. Por este motivo se pide que las señales no tengan distribución gaussiana para aplicar ICA.

¿Cómo funciona ICA?

El objetivo de ICA es a partir de las señales X , encontrar las componentes S , es decir que buscamos la matriz W tal que $S = WX$. A esto se lo conoce como *unmixing problem*.

Una vez hallada W , se proyectan los datos blanqueados sobre esa matriz para hallar las componentes independientes.

ICA: criterios de cómputo

Existen tres principales criterios de independencia que llevar a distintas :

1. Basados en la no-gaussianidad. Esto puede medirse usando medidas como ***negentropy*** o kurtosis. El objetivo es hallar las componentes que maximicen la no-gaussianidad.
2. Minimizando la información mutua entre las componentes
3. Usando estimación de máxima verosimilitud.

El preprocesamiento se calcula directo de los datos, pero la matriz se obtiene por aproximación numérica, mediante métodos de optimización. La solución óptima es difícil de hallar debido a la presencia de extremos locales en la función objetivo.

ICA: implementación de Scikit-Learn

La implementación de Scikit-Learn se basa en el algoritmo FastICA, basado en la negentropy.

Se define la negentropy como $J(y) = H(y_{\text{gauss}}) - H(y)$. Esta definición se basa en el hecho que para un nivel de varianza constante, las variables gaussianas con las que tienen máxima entropía.

Se utiliza $J(y) \propto [E\{G(y)\} - E\{G(\nu)\}]^2$ para estimar $J(y)$.

ICA: Implementation de Conkit Learn

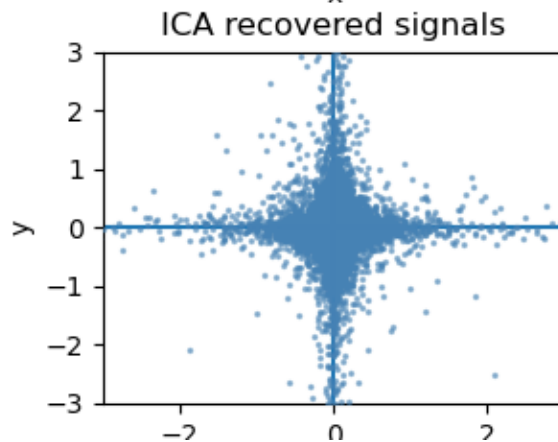
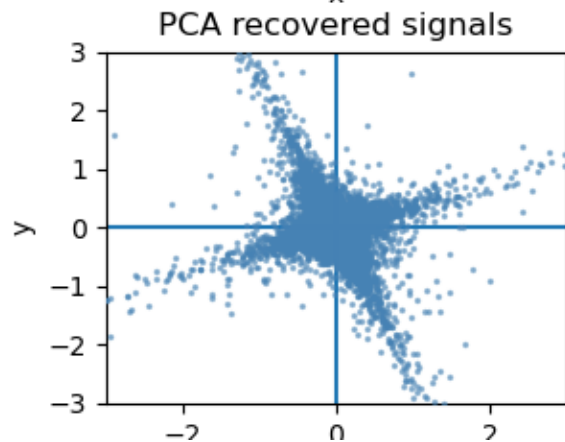
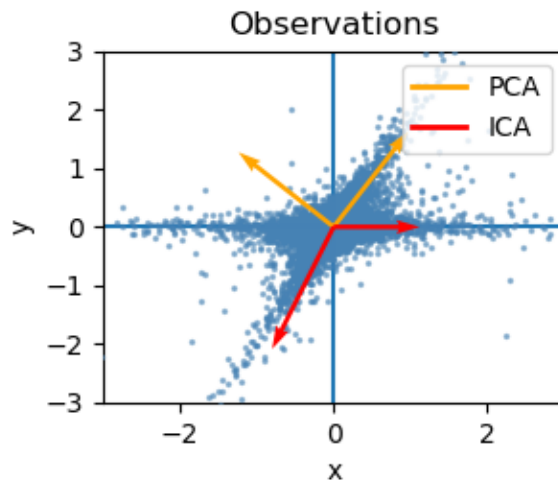
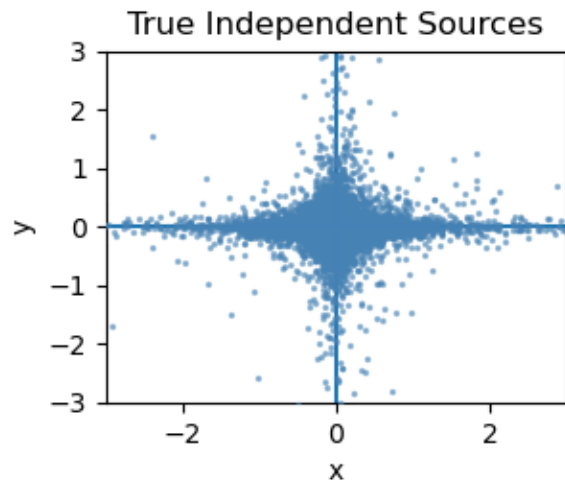
FastICA algorithm is as follows:

1. Choose an initial (e.g. random) weight vector \mathbf{w} .
2. Let $\mathbf{w}^+ = E\{\mathbf{x}g(\mathbf{w}^T \mathbf{x})\} - E\{g'(\mathbf{w}^T \mathbf{x})\}\mathbf{w}$
3. Let $\mathbf{w} = \mathbf{w}^+ / \|\mathbf{w}^+\|$
4. If not converged, go back to 2.

ICA

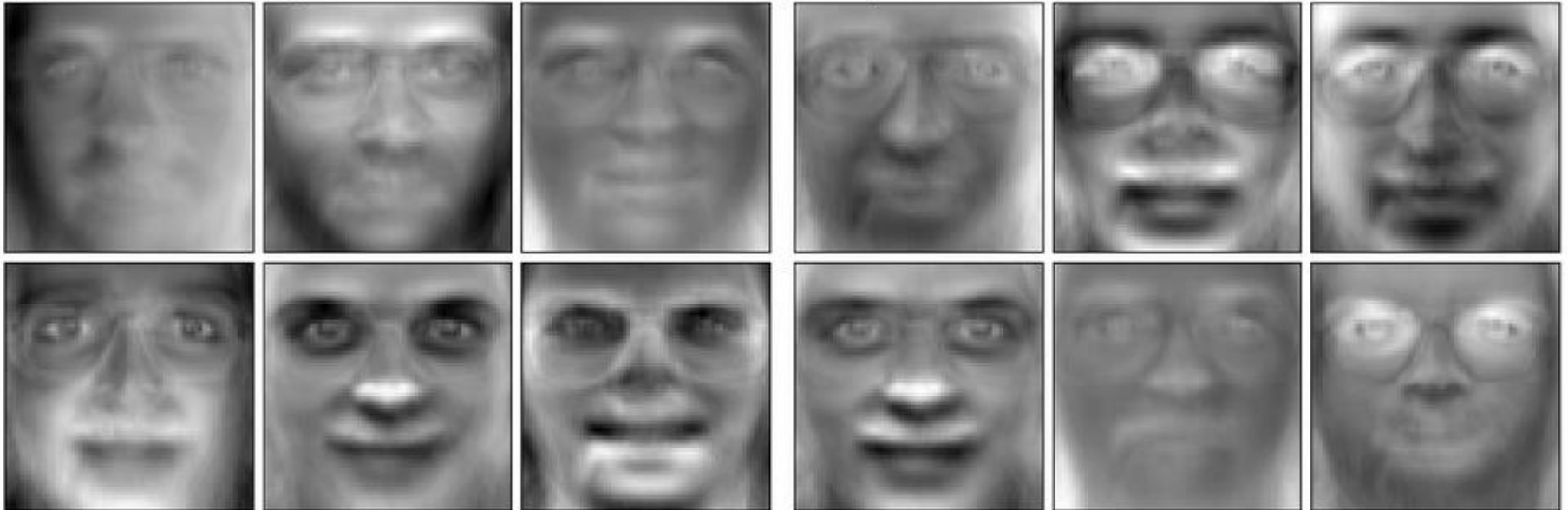
PCA

ICA



ICA vs. PCA - ejemplo 2

Eigenfaces - PCA using randomized SVD - Train time 0.0s Independent components - FastICA - Train time 0.1s



Descomposición en valores singulares (SVD)

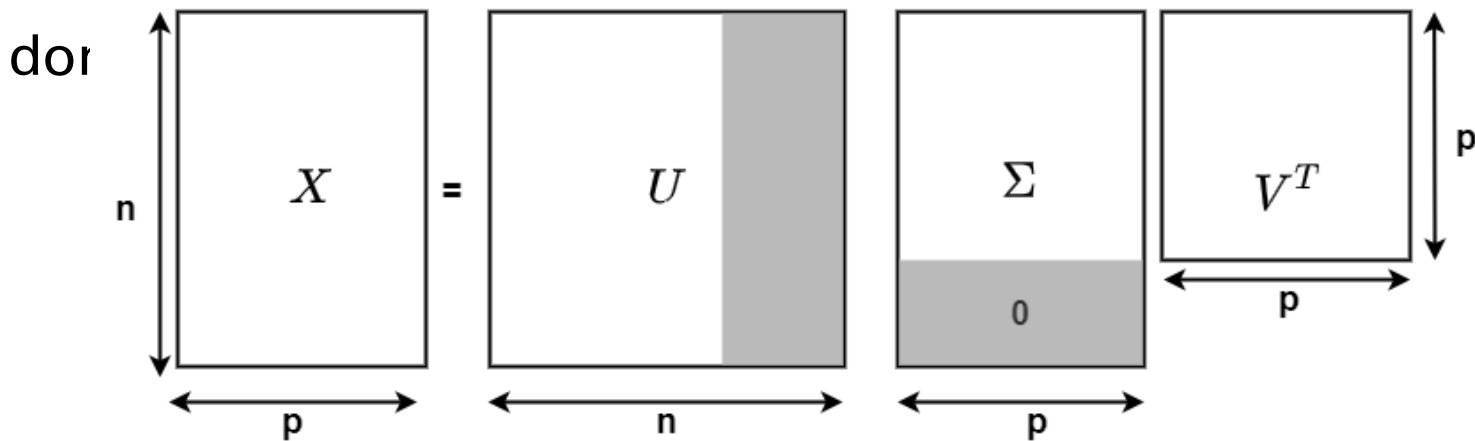
La SVD es muy popular para reducir dimensiones para datos *sparse*.

Sea $\mathbf{X} \in \mathbb{R}^{n \times p}$ una matriz de n observaciones de p variables. Luego, se puede descomponer como

donde $\mathbf{X} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$, $\mathbf{U} \in \mathbb{R}^{n \times n}$ es una matriz "diagonal", $\mathbf{V} \in \mathbb{R}^{p \times p}$ es una matriz

Descomposición en valores singulares (SVD)

Finalmente, los datos transformados resultan



$$\hat{X} = X V_k,$$

$$V_k = [v_1, \dots, v_k]$$

Descomposición en valores singulares (SVD)

¿Qué representan las matrices \mathbf{U} y \mathbf{V} ?

- \mathbf{U} se corresponde con los autovectores de $\mathbf{X}\mathbf{X}^T$ (correlación empírica entre muestras)
- \mathbf{V} está asociada a los autovectores de $\mathbf{X}^T\mathbf{X}$ (correlación empírica de los features)
- Σ es la raíz cuadrada de los autovalores de ambas matrices.

Σ

Descomposición en valores singulares (SVD)

¿Qué representan las matrices **U** y **V**? Ejemplo

Puntuación de distintos usuarios a 5 películas^[1]

SVD

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix} = \underbrace{\begin{bmatrix} 0.14 & 0 \\ 0.42 & 0 \\ 0.56 & 0 \\ 0.7 & 0 \\ 0 & 0.6 \\ 0 & 0.75 \\ 0 & 0.3 \end{bmatrix}}_{\text{temas}} \underbrace{\begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}}_{\text{películas}}^T$$

[1] <http://infolab.stanford.edu/~ullman/mmds/ch11.pdf>

Descomposición en valores singulares (SVD)

Comentarios finales

- Realizar la descomposición SVD sobre los datos centrados es equivalente a hacer PCA
- SVD funciona sobre datos sparse, sin necesidad de "redensificarlos" que puede ocupar mucha memoria
- Según el problema, las matrices U y V de la SVD pueden dar información útil acerca de las correlaciones entre las muestras y variables.

Bibliografía

- "Mining of Massive Datasets", Leskovec J, Rajaraman A., Ullman J.D., Stanford University. Capítulo 11.
<http://infolab.stanford.edu/~ullman/mmds/ch11.pdf>
- <https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c>
- "Pattern Recognition and Machine Learning", Bishop, Christopher M. New York, Springer, 2006
- [A. Hyvarinen and E. Oja, Independent Component Analysis: Algorithms and Applications, Neural Networks, 13\(4-5\), 2000, pp. 411-430](#)
- [Independent component analysis: An introduction](#)

REDUCCIÓN DE DIMENSIONES



¿Por qué es importante reducir la cantidad de features?

Maldición de la dimensión (curse of dimensionality):

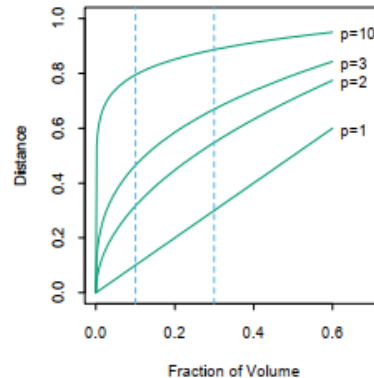
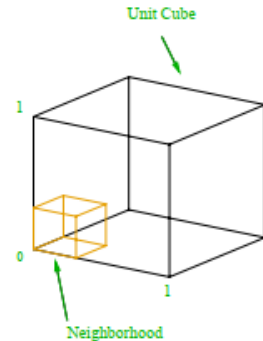
↑ la cantidad de
dimensiones del
espacio de features



↑ en volumen del
espacio



las muestras que
tenemos se vuelven
poco representativas
(muestra pequeña)



Distintas formas de reducir dimensiones

- Selección de features: elegimos, con cierto criterio, un subconjunto de los variables originales. Existen tres enfoques:
 - **Métodos de filtrado:** se realiza un análisis supervisado de los features para determinar cuáles son los más relevantes, y sólo luego se procede al modelado. Ej. selección basada en test estadísticos.
 - **Métodos embebidos:** la selección de features se encuentra naturalmente incorporada al proceso de modelado. Ej: árboles de decisión, LASSO
 - **Métodos Wrapper :** emplean un método iterativo de búsqueda, donde en cada paso se da al predictor un subconjunto distinto de features, y utiliza la performance del predictor para guiar la selección del siguiente subconjunto de variables. Ej: eliminación recursiva de features (*recursive feature elimination* - RFE)
- Métodos de proyección de variables: busco transformar mis variables para llevarlas a un espacio de menor dimensión. Ejemplo: PCA, ICA, SVD, etc.

Métodos básicos de selección

1. Eliminar variables constantes: Si existe algún feature que toma siempre el mismo valor para todas las mediciones, debemos quitarlo
2. Eliminar variables cuasi-constantes: una buena idea puede ser eliminar variables cuya varianza sea muy pequeña.
3. Eliminar variables duplicadas
4. Eliminar variables muy correlacionadas. ¿Cómo elegir entre todas las variables correlacionadas?
 - La que tenga menos # de datos faltantes
 - Elegir la más correlacionada con la variable de salida
 - Entrenar algún algoritmo de ML con las variables correlacionadas y elegir la más informativa

Observación: Estos métodos son no paramétricos, ya que no dependen de la variable de salida

Métodos de filtrado

Los métodos de filtrado disponibles dependen de los tipos de las variables de entrada y salida.

Caso	Variable de Entrada	Variable de Salida	Método
1	Númerica	Numérica	Pearson, Spearman's, Información Mutua
2	Númerica	Categórica	ANOVA, Kendall's, Información Mutua
3	Categórica	Numérica	Poco frecuente.
4	Categórica	Categórica	χ^2 , Información Mutua

Numérica-Numérica

Coeficiente de correlación de Pearson

- Coeficiente de correlación de Pearson:

- Asume que las variables siguen una distribución normal
- Es un estimador de $r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$ mide relación lineal entre variables
- Bajo la hipótesis nula que las variables están descorrelacionadas

(independientes) $\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \Rightarrow$

■

■ $t = r \sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$

(transformación de Fisher)

■ $z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) = \text{arctanh}(r) \approx \mathcal{N}\left(0, \frac{1}{\sqrt{n-3}}\right)$

. Representa la proporción de la varianza explicada por una función lineal de la variable X

$f = \frac{r^2}{1-r^2} (n-1) \sim F_{1,n-2}$

Numérica-Numérica

Coeficiente de Spearman

- **Coeficiente de Spearman**

, si no hay valores repetidos:

- $\rho = \frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}$ Es un método no paramétrico, basado en estadísticas de orden
- Mide la relación monotónica entre variables las variables
- Bajo la hipótesis de variables independientes, $\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$, $d_i = (rg(x)_i - rg(y)_i)$
- Menos sensible a outliers

Numérica - Numérica

Información Mutua

- Información mutua:

Recordemos que

- Debemos estimar las funciones de densidad (probabilidad). El algoritmo de Scikit-learn lo hace basándose en el principio de vecinos más cercanos (A. Kraskov, H. Stogbauer and P. Grassberger, "Estimating mutual information". Phys. Rev. E 69, 2004.).
- No paramétrico (no hace suposiciones de la distribución de las variables)
- Permite identificar relaciones no lineales entre variables. Si $I(X,Y) = 0 \Rightarrow X,Y$ son independientes

Numérica-Categorica 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 $S_d^2 = \frac{\sum_{i=1}^k (n_i - 1) S_k}{N - k}$
- 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.

Numérica - Categórica

Coeficiente de correlación de Kendall

- Coeficiente de Kendal b (considera empates):

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}},$$

$$n_0 = \binom{n}{2} \quad n_1 = \sum_i t_i(t_i - 1)/2 \quad n_2 = \sum_j u_j(u_j - 1)/2$$

- $n_c = \# \text{ pares concordantes}$, $n_d = \# \text{ pares discordantes}$
- Test no paramétrico basado en rangos (estadísticas de orden)
- $t_i = \# \text{ valores empatados del grupo } i \text{ de } x$, $u_j = \# \text{ valores empatados del grupo } j \text{ de } y$
- Mide la correlación de rangos
- Asume que la variable categórica tiene ordinalidad
- Costo computacional orden

$$\frac{\tau_b}{n^2}$$

Categórica - Categórica

Test Chi-cuadrado

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

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.
- r y k son la cantidad de factores de las variables de entrada y salida respectivamente.

- Criterio de Información mutua

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

Categórica - Numérica

Es el caso menos frecuente, pero si ocurriera se puede tratar con los mismos criterios que Numérica categórica con los roles intercambiados.

Comentarios finales

- Ventajas:
 - Son simples y suelen ser rápidos de computar,
- Desventajas:
 - Propensos a la sobre selección de variables,
 - Puede haber desconexión entre lo que el test reconoce como importante y lo que necesita el modelo.

Bibliografía

- "Python Machine Learning Cookbook, practical solutions from preprocessing to deep learning", Albon, Cris. O'Reilly Media, Inc., 2018.
- "Feature Engineering and Selection, A Practical Approach for Predictive Models", Max Khun and Kjell Johnson. CRC Press, 2020.
- "Measures of Association How to Choose?" Harry Khamis, PhD. Journal of Diagnostic Medical Sonography May/June 2008 VOL. 24, NO. 3
(<https://journals.sagepub.com/doi/pdf/10.1177/8756479308317006>)
- "The Kendall Rank CorrelationCoefficient", Hervé Abdi (<https://personal.utdallas.edu/~herve/Abdi-KendallCorrelation2007-pretty.pdf>)
- W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," Journal of the American Statistical Association, vol. 47, no.260, pp. 583–621, 1952.

+



○



•



DUDAS?

ENCUESTA