Machine Learning aplicado a las Ciencias Sociales

Clase 1. Introducción y análisis no supervisado 2 (MCA)

factor~data
EIDAES_UNSAM

Tipos de problemas

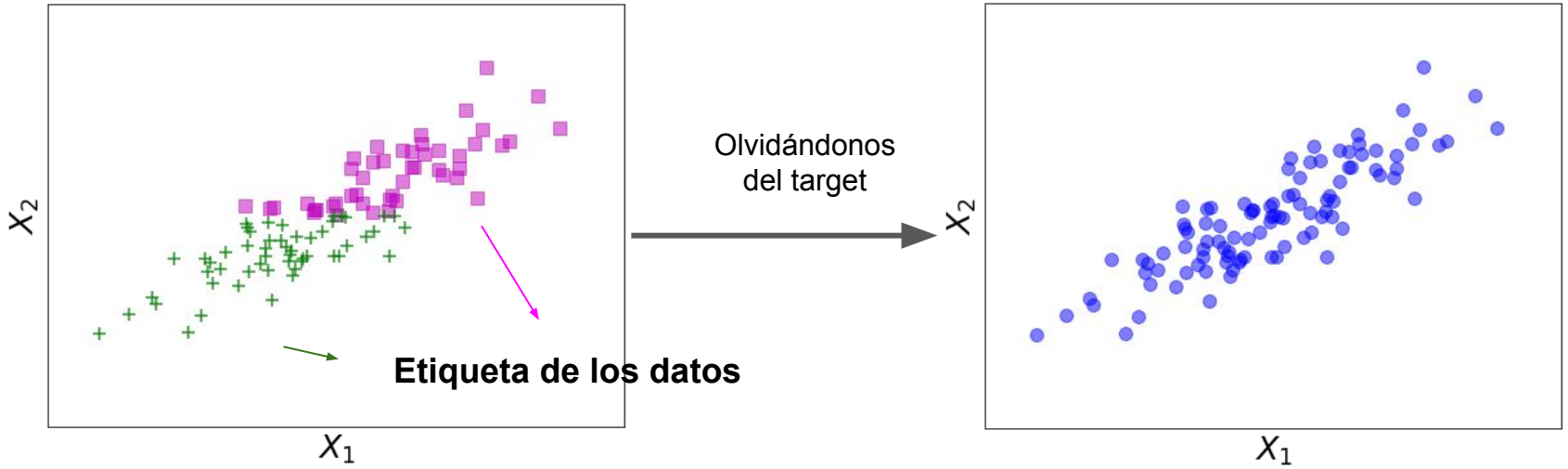
- Aprendizaje Supervisado
  - Variable dependiente => Y, resultado, target
  - Matriz de predictores (p), X, features, variables independientes, etc.
  - Problemas de regresión: Y es cuantitativa
  - Problemas de regresión: Y es cualitativa
  - Tenemos datos de entrenamiento (conjuntos de Xi,Yi), observaciones
  - Podemos definir una (o varias) métricas para evaluar los modelos

- **Aprendizaje no Supervisado**
  - **No hay variable target (Y)**
  - **Solo hay X**
  - **Es más difícil evaluar qué tan bien funciona el modelo**

Aprendizaje No Supervisado

¿Qué es el aprendizaje no supervisado?

- El estudio o la exploración de cómo están representados datos y qué conclusiones podemos sacar de dicha representación.



Aprendizaje no supervisado es todo lo que podemos hacer con solo la representación de los datos en el espacio de features.

¿Qué podemos hacer con solo el conocimiento de la representación de los datos en el espacio de features?

Exploración de los datos y extracción de conocimiento! Por ejemplo, aplicando los siguientes conceptos:

- Clustering (próxima clase): agrupar instancias que tengan una descripción similar en el espacio de features (próxima clase). Básicamente respondemos si hay subconjuntos de datos muy parecidos entre si.

- Reducción de la dimensionalidad (clase de hoy): encontrar combinaciones de features que reemplacen a los originales para reducir la dimensión del problema.

¿Por qué reducir la dimensionalidad del problema?

- ¿Todos los features aportan información relevante? ¿Hay redundancia en la información que aportan? ¿Es necesario trabajar con todos?
- Reduciendo la dimensión podemos:
  - Visualizar los datos en el espacio de dimensión reducida, más fácil de interpretar (ojalá siempre pudiéramos llevar todo a 2D)
  - Comprimir la información: nos permite separar la señal del ruido (pérdida de información = abstracción)
  - Tener un punto de partida para clustering: instancias parecidas en un espacio multidimensional son más parecidas en un espacio reducido (emergencia de estructuras).

Análisis de Correspondencias Múltiples (ACM)

Generalización de AC

Individuos	Género	Años	Ingreso
1	Mujer	5	Medio
2	Mujer	3	Alto
3	Hombre	4	Bajo
4	Mujer	1	Bajo
5	Mujer	2	Medio
6	Hombre	5	Alto
7	Mujer	2	Medio
8	Hombre	3	Bajo
9	Hombre	1	Alto
10	Mujer	4	Medio

Análisis de Correspondencias Múltiples (ACM)

Generalización de AC

A partir de la tabla original se construye la *tabla disyuntiva (matriz Z)* con tantas columnas como categorías:

Género		Años					Ingresos		
Mujer	Hombre	1	2	3	4	5	Bajo	Medio	Alto
1	0	0	0	0	0	1	0	1	0
1	0	0	0	1	0	0	0	0	1
0	1	0	0	0	1	0	1	0	0
1	0	1	0	0	0	0	1	0	0
1	0	0	1	0	0	0	0	1	0
0	1	0	0	0	0	1	0	0	1
1	0	0	1	0	0	0	0	1	0
0	1	0	0	1	0	0	1	0	0
0	1	1	0	0	0	0	0	0	1
1	0	0	0	0	1	0	0	1	0

En la tabla disyuntiva completa (matriz Z), si hay alguna variable continua, debe transformarse en nominal, ordenándose en intervalos a los que se da un rango de valores.

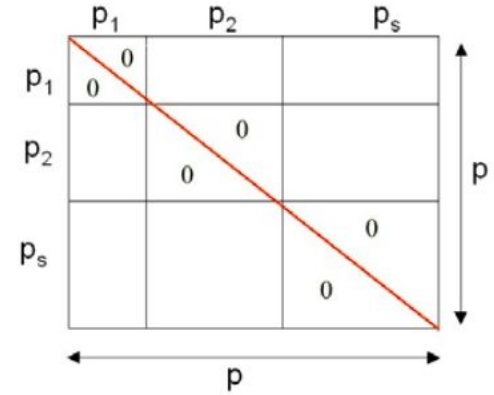
Análisis de Correspondencias Múltiples (ACM)

Generalización de AC

A partir de la tabla disyuntiva completa se puede construir la tabla de contingencia de Burt (B), que es una tabla simétrica de orden (p, p): $B = Z'.Z$

B es una yuxtaposición de tablas de contingencia y está formada de s^2 bloques de la forma:

$B = Z'Z =$



Cada bloque es una submatriz formada por tablas de contingencia de las variables dos a dos, salvo los bloques que se están en la diagonal que son las tablas de contingencia de cada variable consigo misma.

Análisis de Correspondencias Múltiples (ACM)

Tabla disyuntiva completa y Burt producen los mismos factores.

Z'.Z =

MATRIZ DE BURT

		Género		Años					Ingresos		
		M	H	1	2	3	4	5	B	M	A
Género	M	6	0	1	2	1	1	1	1	4	1
	H	0	4	1	0	1	1	1	2	0	2
Años	1	1	1	2	0	0	0	0	1	0	1
	2	2	0	0	2	0	0	0	0	2	0
	3	1	1	0	0	2	0	0	1	0	1
	4	1	1	0	0	0	2	0	1	1	0
	5	1	1	0	0	0	0	2	0	1	1
Ingresos	B	1	2	1	0	1	1	0	3	0	0
	M	4	0	0	2	0	1	1	0	4	0
	A	1	2	1	0	1	0	1	0	0	3

Análisis de Correspondencias Múltiples (ACM)

Se puede diagonalizar la tabla de Burt para obtener los factores

MATRIZ DE BURT

$Z'.Z =$

		Género		Años					Ingresos		
		M	H	1	2	3	4	5	B	M	A
Género	M	6	0	1	2	1	1	1	1	4	1
	H	0	4	1	0	1	1	1	2	0	2
Años	1	1	1	2	0	0	0	0	1	0	1
	2	2	0	0	2	0	0	0	0	2	0
	3	1	1	0	0	2	0	0	1	0	1
	4	1	1	0	0	0	2	0	1	1	0
	5	1	1	0	0	0	0	2	0	1	1
Ingresos	B	1	2	1	0	1	1	0	3	0	0
	M	4	0	0	2	0	1	1	0	4	0
	A	1	2	1	0	1	0	1	0	0	3

Resumen de MCA

- Análogo a PCA para variables ordinales.
- Se computan a partir de la Tabla Disyuntiva Completa o la Matriz de Burt
- Las componentes están ordenadas de mayor a menor, en el sentido de la información (inercia) que se llevan. Si tiramos las últimas componentes, estamos reduciendo la dimensión de nuestro problema.

Vamos al Notebook