

Ingeniería de Características

(Feature Engineering)

¿Por qué?

Curse of
dimensionality

Reducción del
volumen de datos

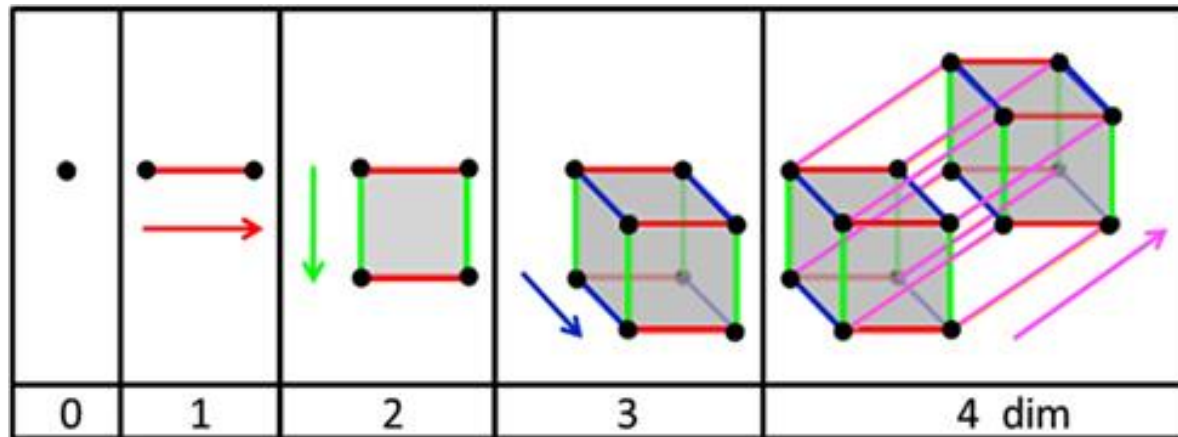
Enfocarse en los
datos relevantes

Mejorar la calidad
de los datos

Mejorar el
desempeño de los
algoritmos en
cuanto a tiempo o
la precisión

Curse of Dimensionality

- Añadir **demasiados** rasgos puede paradójicamente **empeorar** los resultados
 - El tamaño del espacio de características crece con el número de estas
 - Es conveniente que existan puntos en cada región de este espacio
 - El número de ejemplos de entrenamiento puede crecer exponencialmente



Seleccionar un subconjunto de atributos que provea la información más útil

Selección de rasgos
(Feature selection)

Extracción de rasgos
(*Feature extraction*)

Construcción de rasgos
(Feature construction)

Pueden ser evaluados en cuanto a:

- Reducción alcanzada.
- Relevancia de los atributos seleccionados.
- Efecto en los algoritmos de aprendizaje.

Selección de Rasgos

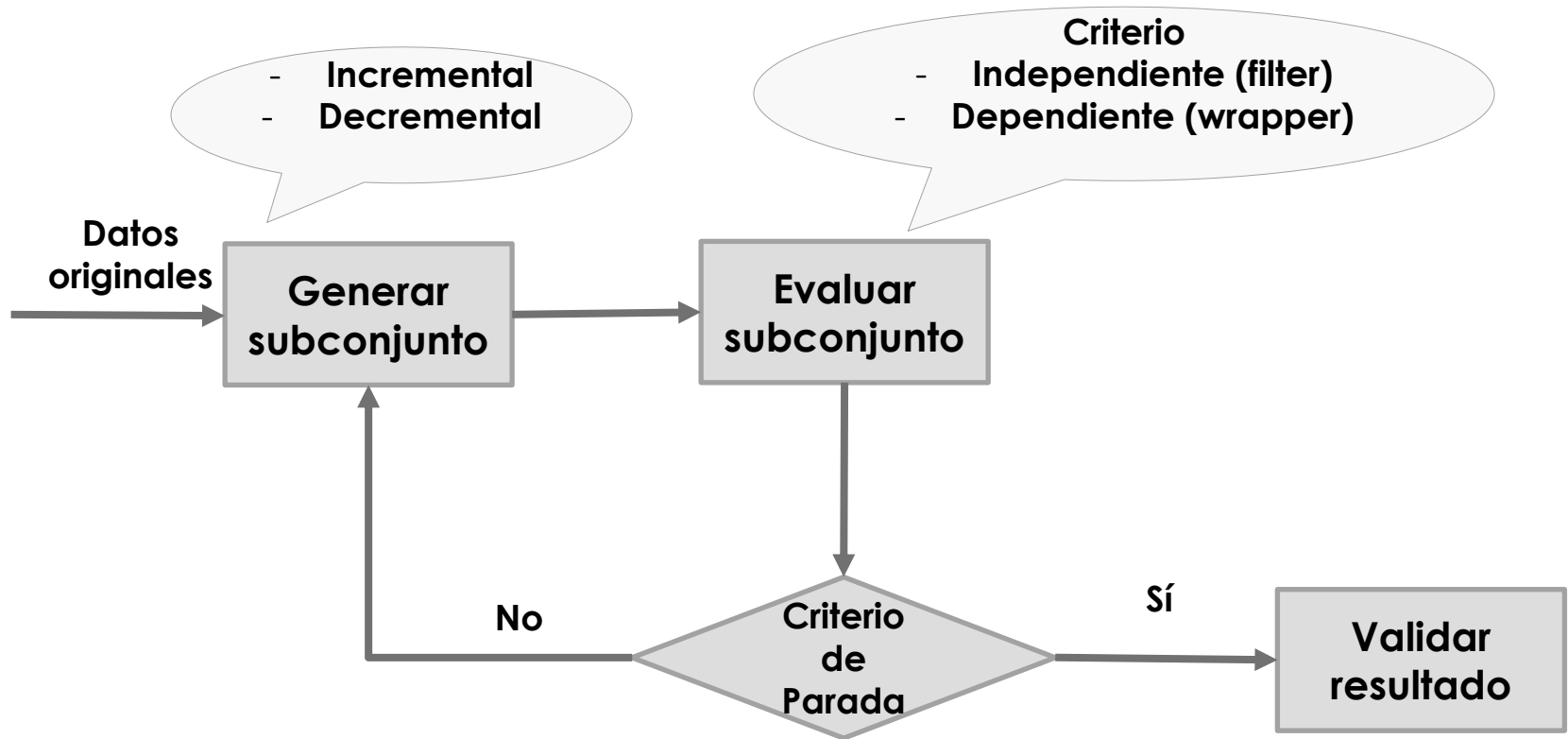
Es el proceso que elige a partir de los atributos originales $A_1, A_2, \dots, A_n$ un **subconjunto** $B_1, B_2, \dots, B_m$, donde $m < n$, de acuerdo con cierto criterio

¿qué criterio utilizar para evaluar los atributos?

- clasificación: precisión en la predicción, etc.
- clustering: similaridad inter/intra clúster, varianza, etc.

¿cuántos atributos nuevos utilizar?

- establecer un umbral (*threshold*)
- determinar automáticamente de acuerdo con alguna medida del desempeño, por ejemplo, si la adición de un atributo al conjunto actual no conduce a una mejora



● Selección Las Vegas (filter)

Sean $X = \{X_1, X_2, \dots, X_n\}$ un conjunto de instancias pertenecientes a diferentes clases θ donde $X_i = [x_i^1, x_i^2, \dots, x_i^m, \theta_i]$ y x_i^j es el valor del rasgo j en la instancia i , se definen:

- el par X_i, X_j como inconsistente si $[x_i^1, x_i^2, \dots, x_i^m] = [x_j^1, x_j^2, \dots, x_j^m]$ y $\theta_i \neq \theta_j$

Dado un grupo de instancias inconsistentes $Y = \{Y_1, Y_2, \dots, Y_m\} \subseteq X$

- $lc(Y) = |Y| - |Y_{\theta_i}|$: θ_i clase mayoritaria en Y

- $IR(x^1, x^2, \dots, x^m) = \frac{\sum_{Y \in I} lc(Y)}{|X|}$, I =conjunto de todos los subconjuntos inconsistentes en X

A1	A2	A3	A4	Class
1	2	2	1	1
4	3	2	2	2
4	3	2	2	1
4	3	2	2	1
9	3	8	1	2
9	3	8	2	2
9	3	8	2	1

$$Ic = 3 - 2 = 1 \quad Ic = 2 - 1 = 1$$

$$IR(A^1, A^2, A^3, A^4) = \frac{(Ic + Ic)}{|X|} = \frac{(1+1)}{|7|} = 0.285$$

- Selección Las Vegas (filter)

Sean $X = \{X_1, X_2, \dots, X_n\}$ un conjunto de instancias pertenecientes a diferentes clases θ donde $X_i = [x_i^1, x_i^2, \dots, x_i^m, \theta_i]$ y x_i^j es el valor del rasgo j en la instancia i :

$L = \emptyset$ /*Lista que almacena subconjuntos de igual calidad*/

T : umbral previamente definido

mientras ciclos < MaxCiclos

$S \subset \{x^1, x^2, \dots, x^m\}$ aleatorio

si $|S| < |Mejor| \ \&\& \ IR(S) \leq T$

$Mejor = S$

$L = S$

si no

si $|S| == |Mejor| \ \&\& \ IR(S) \leq T$

$L = L \cup S$

ciclos ++

finalizar

- Selección de acuerdo con la Ganancia de Información o criterios relacionados (filter)

$$GA(S, A) = E(S) - \sum_{v \in \text{Valores}(A)} \frac{|S_v|}{|S|} E(S_v)$$

$$\text{SplitInformation}(S, A) = \sum_{v \in \text{Valores}(Fecha)} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|}$$

$$\text{GainRatio}(S, A) = \frac{\text{Gain}(S, A)}{\text{SplitInformation}(S, A)}$$

donde:

- $\text{Valores}(A)$ es el conjunto de valores posibles del atributo A
- $S_v = \{s \in S | A(s) = v\}$
- $A(s)$ = valor del atributo A en la instancia s

- Generación de Subconjuntos

Búsqueda Exhaustiva

Random Mutation Hill Climbing

Algoritmos Genéticos

Simulated Annealing

Extracción de Rasgos

Es el proceso que a partir de los atributos originales $A_1, A_2, \dots, A_n$ obtiene un conjunto de **nuevos** rasgos $B_1, B_2, \dots, B_m$ a través de algún mapeo funcional F_i tal que $B_i = F_i(A_1, A_2, \dots, A_n)$ y $m < n$

¿qué criterio utilizar para evaluar los atributos?

- clasificación: precisión en la predicción, etc.
- clustering: similaridad inter/intra clúster, varianza, etc.

¿qué transformación realizar?

- lineales o no lineales

¿cuántos atributos nuevos utilizar?

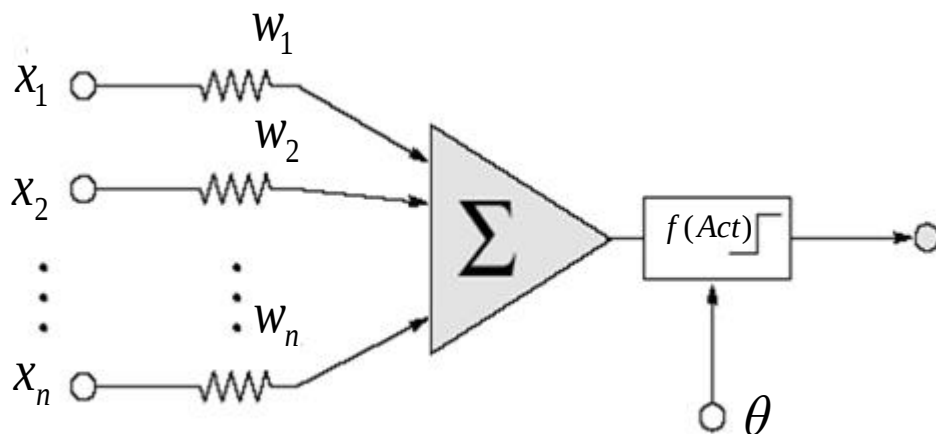
- establecer un umbral (*threshold*)
- determinar automáticamente de acuerdo con alguna medida del desempeño

Transformaciones en la etapa de pre-procesamiento

- Análisis de Componentes Principales (PCA)
- Linear Discriminant Analysis (LDA)

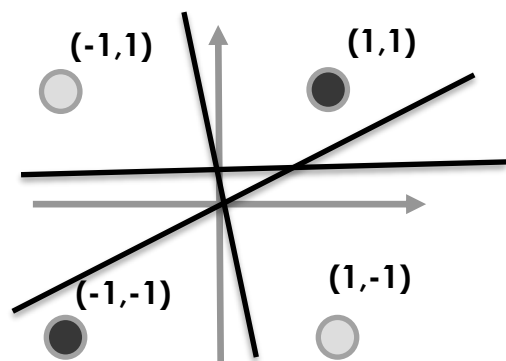
Transformaciones realizadas por el propio algoritmo de aprendizaje

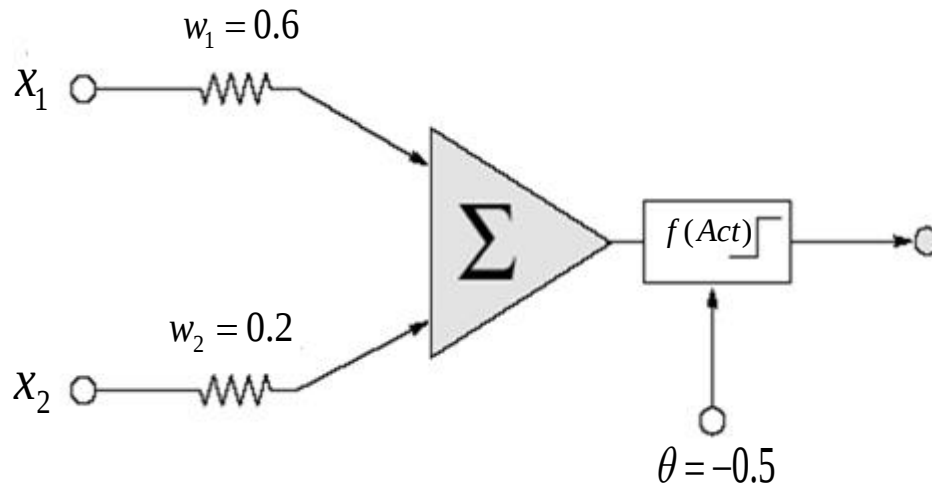
- Perceptrón Multicapa
- Máquinas de Vectores de Soporte



$$Act = \sum_{i=1} w_i x_i - \theta$$

$$f(Act) = \begin{cases} 1 & \text{si } Act > 0 \\ -1 & \text{eoc} \end{cases}$$





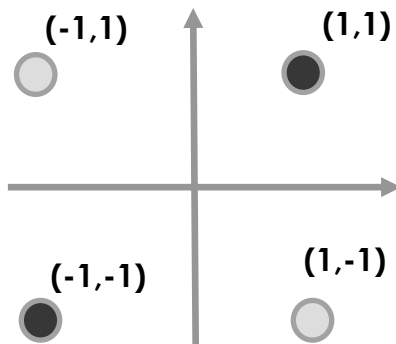
$$Act = \sum_{i=1} w_i x_i - \theta$$

$$f(Act) = \begin{cases} 1 & \text{si } Act > 0 \\ -1 & \text{eoc} \end{cases}$$

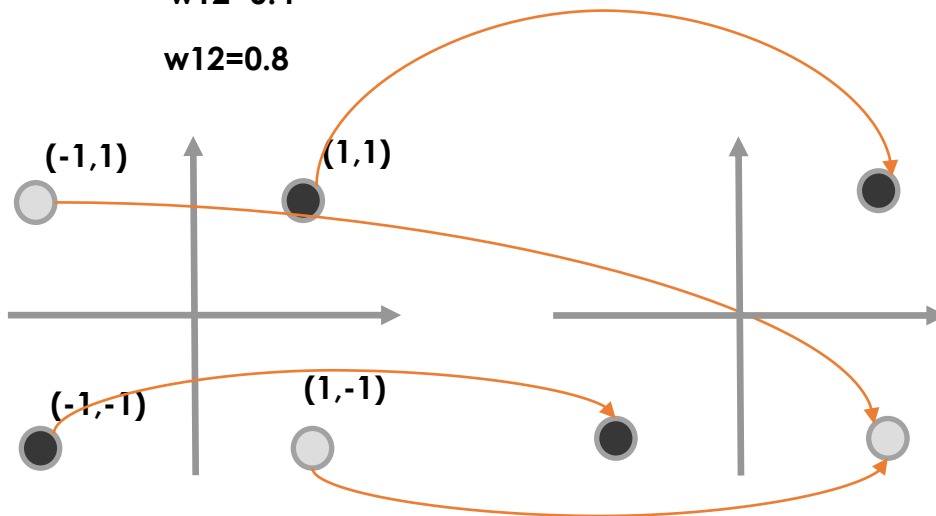
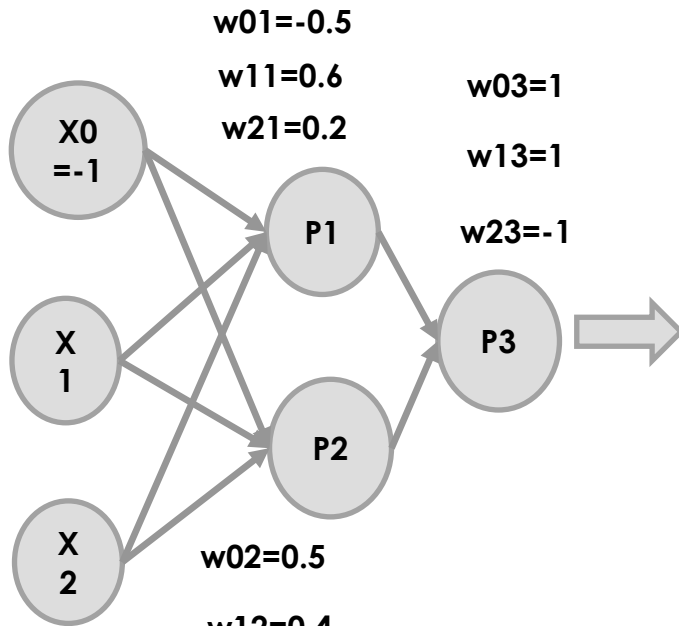
$$x = \langle -1, 1 \rangle$$

$$Act = 0.6 * (-1) + 0.2 * (1) - (-0.5) = 0.1$$

$$f(Act) = f(0.1) = 1$$



Algoritmos



$A_1, A_2, \dots, A_n$

X_0	X_1	X_2	XOR
-1	1	1	-1
-1	1	-1	1
-1	-1	1	1
-1	-1	-1	-1

$B_1, B_2, \dots, B_m$

P_1	P_2	XOR
1	1	-1
1	-1	1
1	-1	1
-1	-1	-1

Construcción de Rasgos

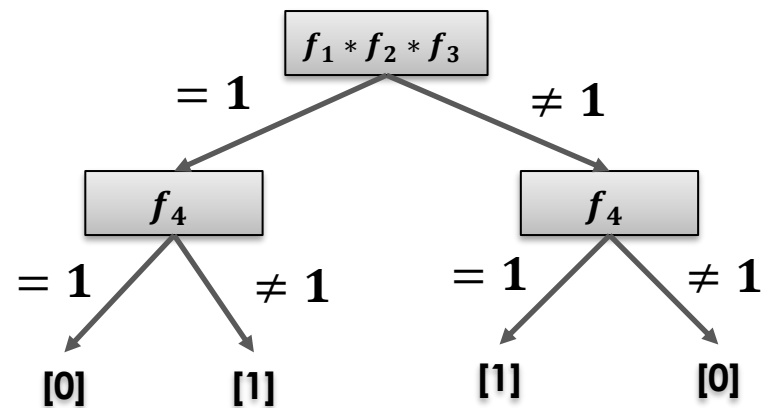
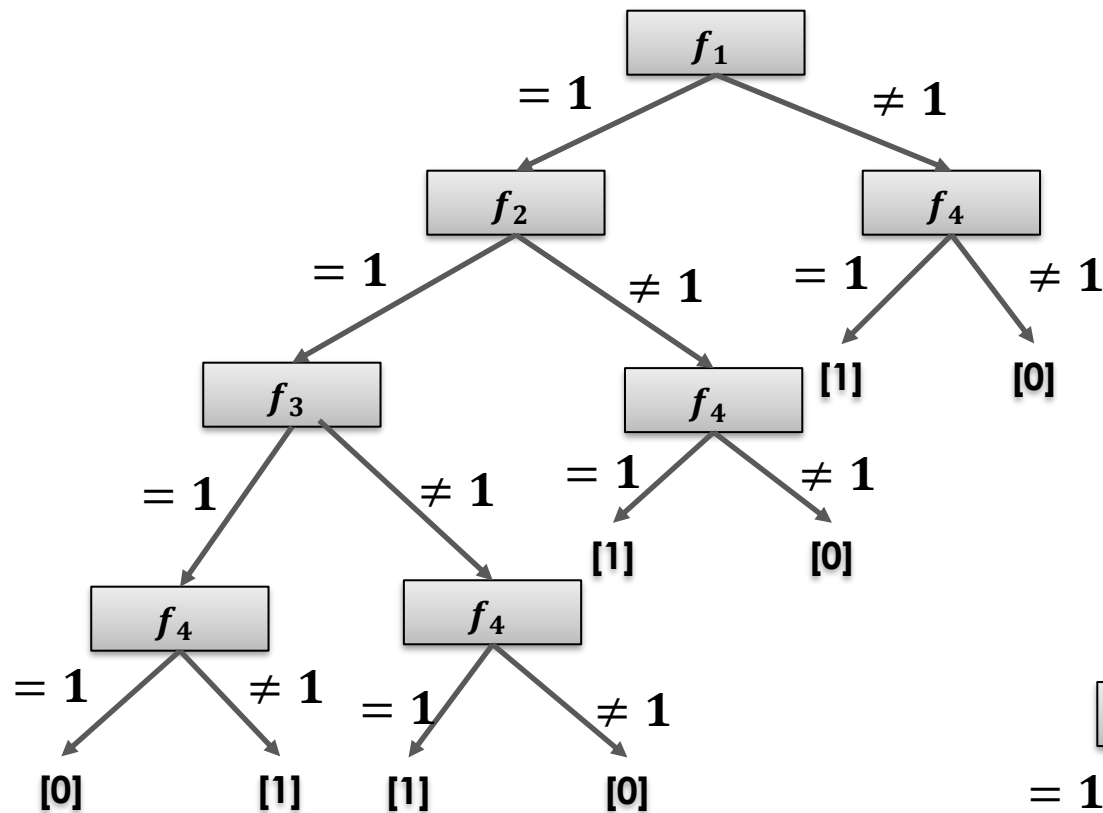
Es el proceso que a partir de los atributos originales $A_1, A_2, \dots, A_n$ construye un conjunto de nuevos rasgos $A_{n+1}, A_{n+2}, \dots, A_{n+m}$ mediante diferentes operadores que permiten combinar uno o más rasgos en otros nuevos.

¿cómo construir los nuevos atributos?

- operadores aritméticos, etc.

¿qué criterio utilizar para evaluar los atributos?

- clasificación: precisión en la predicción, etc.
- clustering: similaridad inter/intra clúster, varianza, etc.



Fin