

# Procesamiento de Lenguaje Natural

---

Clasificación multi-etiqueta

# Clasificación Multi-etiqueta

## Definición

**Multi-etiqueta** (multi-label) se refiere a las tareas de aprendizaje donde cada ejemplo o instancia se le asignan una o más clases (etiquetas)

etiqueta 1	✓
etiqueta 2	

Binaria

etiqueta 1	
etiqueta 2	
etiqueta 3	
etiqueta 4	✓
etiqueta ...	
etiqueta ...	
etiqueta L	

Multi-clase

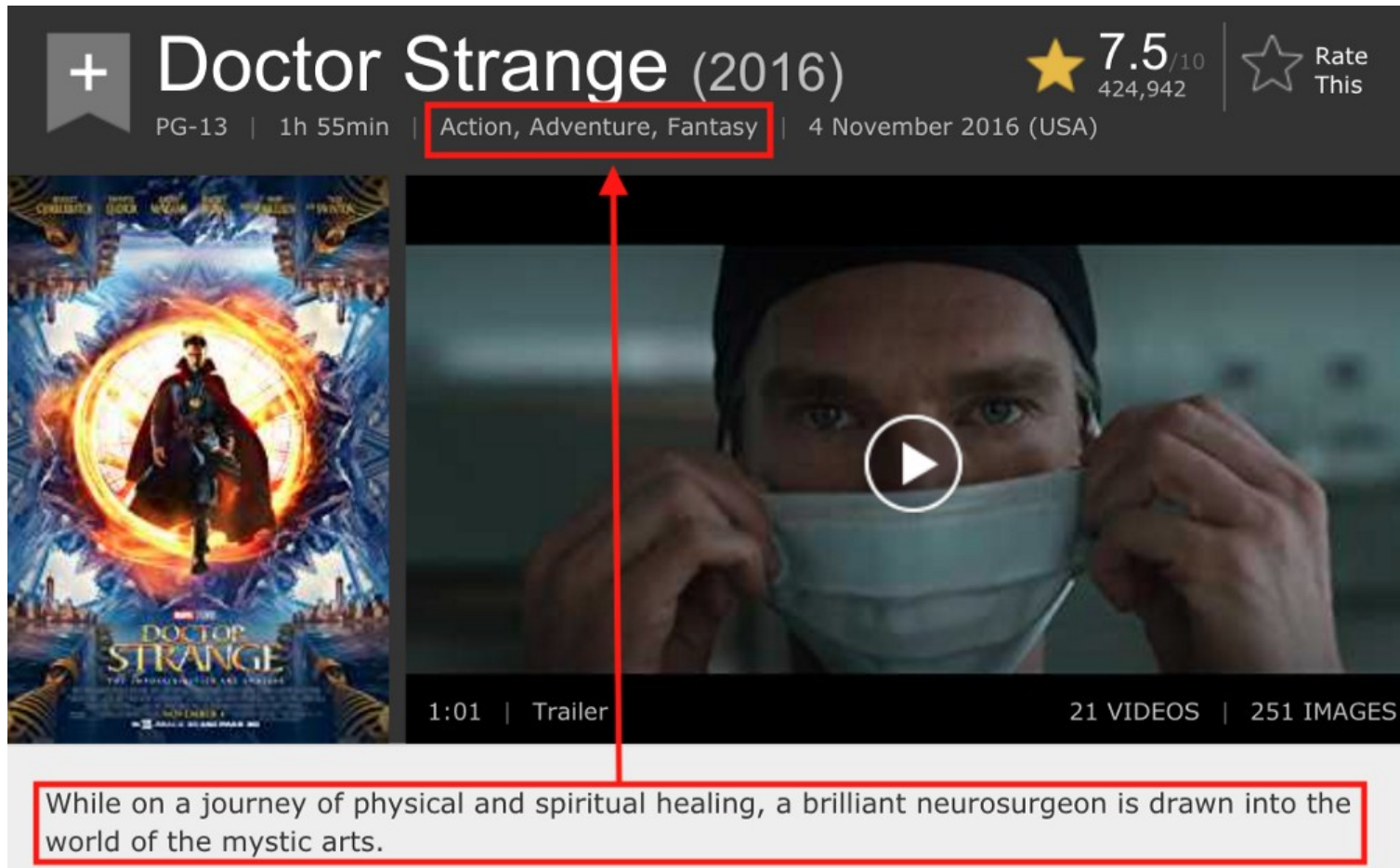
etiqueta 1	
etiqueta 2	✓
etiqueta 3	
etiqueta 4	✓
etiqueta ...	
etiqueta ...	
etiqueta L	✓

Multi-etiqueta

# Clasificación Multi-etiqueta

## Ejemplos

Es una forma de clasificación común en colecciones de textos



The image shows a movie page for "Doctor Strange (2016)". At the top, the title "Doctor Strange (2016)" is displayed next to a rating of 7.5/10 and a "Rate This" button. Below the title, the genres "Action, Adventure, Fantasy" are listed and highlighted with a red box. A red arrow points from this box down to a red box containing a movie description: "While on a journey of physical and spiritual healing, a brilliant neurosurgeon is drawn into the world of the mystic arts." The page also features a movie poster on the left, a video player in the center, and a trailer thumbnail at the bottom left.

+ Doctor Strange (2016) ★ 7.5<sub>424,942</sub> ☆ Rate This


PG-13 | 1h 55min | Action, Adventure, Fantasy | 4 November 2016 (USA)

1:01 | Trailer 21 VIDEOS | 251 IMAGES

While on a journey of physical and spiritual healing, a brilliant neurosurgeon is drawn into the world of the mystic arts.

# Clasificación Multi-etiqueta

## Ejemplos



WIKIPEDIA  
The Free Encyclopedia

[Main page](#)  
[Contents](#)  
[Featured content](#)  
[Current events](#)  
[Random article](#)  
[Donate to Wikipedia](#)  
[Wikipedia store](#)

Interaction  
[Help](#)  
[About Wikipedia](#)  
[Community portal](#)  
[Recent changes](#)  
[Contact page](#)

Tools  
[What links here](#)

Article

Talk

Read

Edit

View history

Search

## Bharata Natyam

From Wikipedia, the free encyclopedia


**Bharathanatyam** (Tamil: பரதநாট्यம்) is a form of [Indian classical dance](#) that originated in the temples of [Tamil Nadu](#).<sup>[[citations needed](#)]</sup> It was described in the treatise *Natya Shastra* by *Bharata* around the beginning of the common era. Bharata Natyam is known for its grace, purity, tenderness, expression and sculptural poses. *Lord Shiva* is considered the God of this dance form. Today, it is one of the most popular and widely performed dance styles and is practiced by male and female dancers all over the world, although it is more commonly danced by women.<sup>[[B](#)]</sup>

Contents

hide

- Etymology
- Dance tradition
- Essential ideas
  - Spiritual symbolism
- Medieval decline
- Modern rebirth

**Bharathanatyam**



Dances by name, Indian culture, Performing arts in India, South India, Tamil culture



# Clasificación Multi-etiqueta

## Ejemplos

Las noticias de Reuters tienen 103 códigos de tópicos



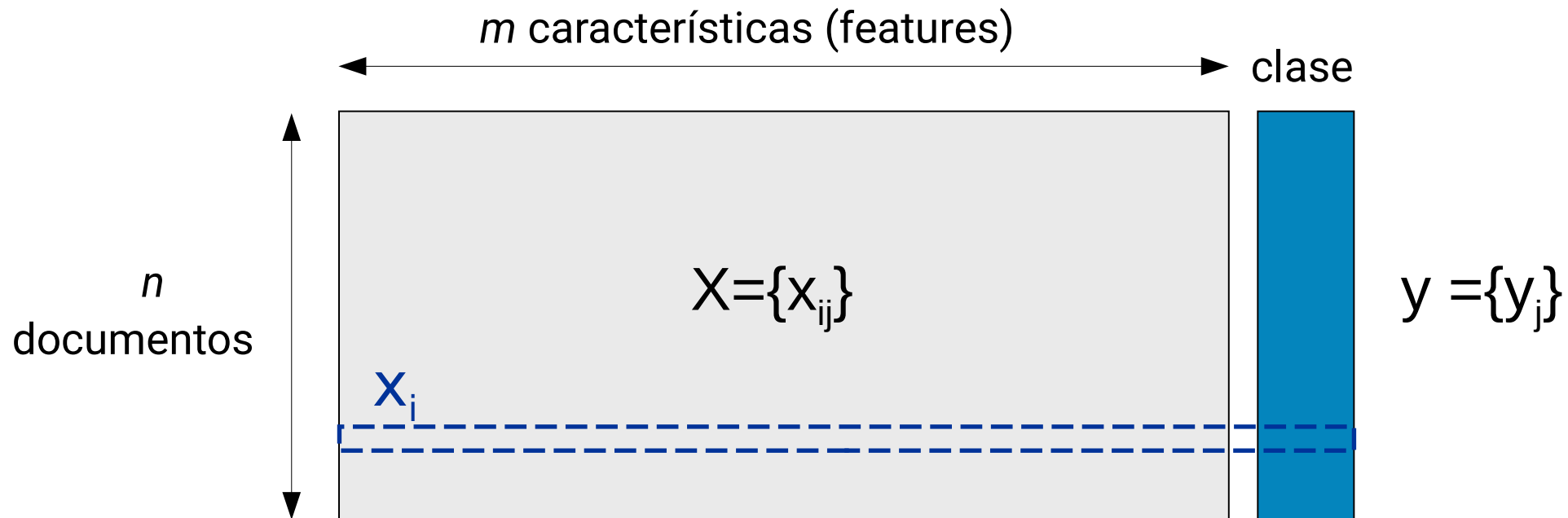
The image is a screenshot of the BBC News website. At the top, there is a navigation bar with the BBC logo, a 'Sign in' button, and links to News, Sport, Weather, Shop, Earth, Travel, and a 'More' dropdown menu. A search bar is located on the right. Below this is a red banner with the word 'NEWS' in white. Under the banner, there is a row of category links: Home, Video, World, UK, Business, Tech, Science, Magazine, Entertainment & Arts, Health, In Pictures, and a 'More' dropdown menu. The 'World' and 'Business' links are highlighted with red boxes. Below this row, there is another row of regional links: World, Africa, Asia, Australia, Europe, Latin America, Middle East, and US & Canada. The 'Europe' link is highlighted with a red box. The main content area shows a news article titled 'Novo Banco: Portugal bank sell-off hits snag'. The article text states: 'Portugal's central bank has missed its deadline to sell Novo Banco, a bank created after the collapse of the country's second-biggest lender.'

**Novo Banco: Portugal bank sell-off hits snag**

Portugal's central bank has missed its deadline to sell Novo Banco, a bank created after the collapse of the country's second-biggest lender.

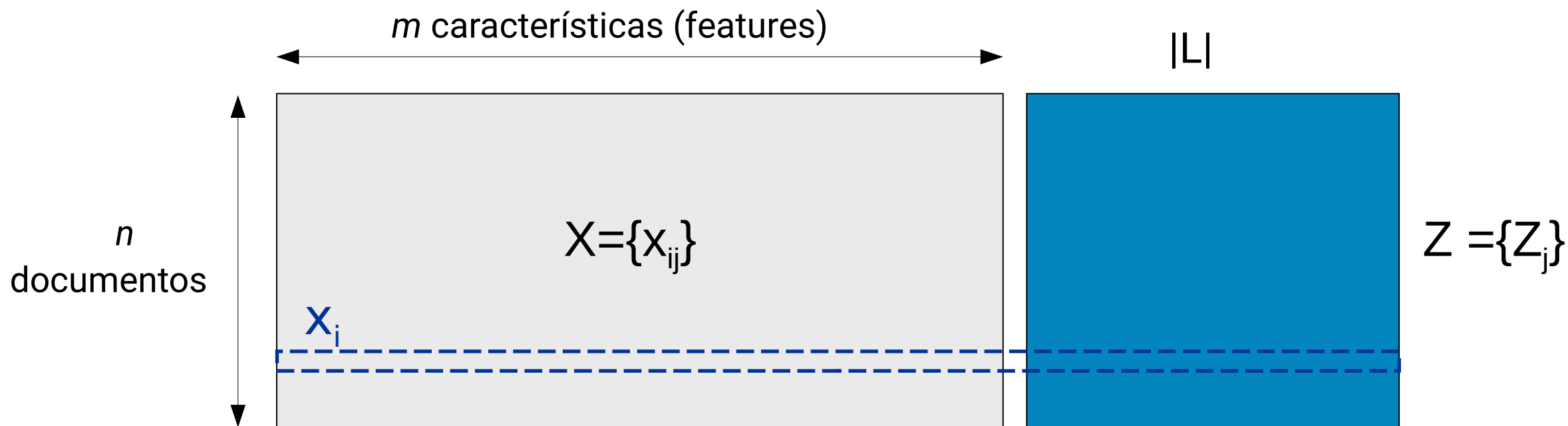
# Clasificación Multi-etiqueta

## Definición



# Clasificación Multi-etiqueta

## Definición



# Clasificación Multi-etiqueta

## Definición

Simple-etiqueta  
(single-label)

$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	Y
1.0	0.9	3.0	1.0	0.0	0
0.0	0.1	1.0	0.0	1.0	1
0.0	0.0	1.0	1.0	0.0	0
1.0	0.8	2.0	0.0	1.0	1
1.0	0.0	2.0	0.0	1.0	0
0.0	0.0	3.0	1.0	1.0	?

Multi-etiqueta  
(multi-label)

$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$Y_1$	$Y_2$	$Y_3$	$Y_4$
1.0	0.9	3.0	1.0	0.0	0	1	1	0
0.0	0.1	1.0	0.0	1.0	1	0	0	0
0.0	0.0	1.0	1.0	0.0	0	1	0	0
1.0	0.8	2.0	0.0	1.0	1	0	0	1
1.0	0.0	2.0	0.0	1.0	0	0	0	1
0.0	0.0	3.0	1.0	1.0	?	?	?	?



# Clasificación Multi-etiqueta

## Definición

IMDb dataset: predicción del género basado en el plot de la película

	<i>abandoned</i>	<i>accident</i>	<i>...</i>	<i>violent</i>	<i>wedding</i>	<i>horror</i>	<i>romance</i>	<i>...</i>	<i>comedy</i>	<i>action</i>
<i>i</i>	$X_1$	$X_2$	$\dots$	$X_{1000}$	$X_{1001}$	$Y_1$	$Y_2$	$\dots$	$Y_{27}$	$Y_{28}$
1	1	0	...	0	1	0	1	...	0	0
2	0	1	...	1	0	1	0	...	0	0
3	0	0	...	0	1	0	1	...	0	0
4	1	1	...	0	1	1	0	...	0	1
5	1	1	...	0	1	0	1	...	0	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
120919	1	1	...	0	0	0	0	...	0	1

# Clasificación Multi-etiqueta

## Definición

Dado  $L$  etiquetas, con  $K$  valores posibles

		Cardinalidad	
		$K = 2$	$K > 2$
Targets	$L = 1$	Binaria	Multi-clase
	$L > 1$	Multi-label	Multi-output

# Clasificación Multi-etiqueta

## Enfoques

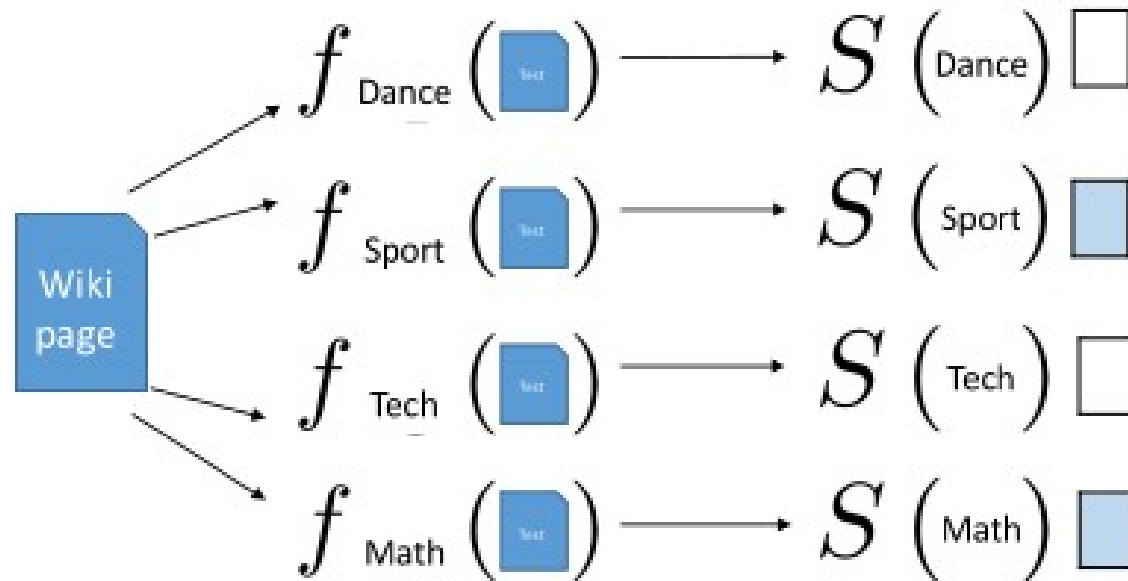
Enfoques de clasificación multi-etiqueta:

- **Métodos de transformación:** transforman el problema de clasificación multi-etiqueta en varios problemas de clasificación de una única etiqueta
- **Métodos de adaptación:** modifican algoritmos de aprendizaje para soportar problemas multi-etiqueta

# One-vs-rest

## Métodos de transformación

**One-vs-rest (o one-vs-all)** involucra entrenar un clasificador binario para cada clase (label). Cada clasificador predice si la instancia pertenece o no a la clase.



# Binary Relevance (BR)

Métodos de transformación

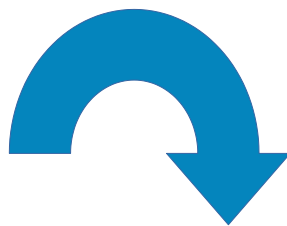
## Binary Relevance (BR):

- Transforma el problema de clasificación multi-etiqueta en múltiples problemas simple-etiqueta
- Aprende  $L$  clasificadores binarios independientes

# Binary Relevance (BR)

Métodos de transformación

X	Y <sub>1</sub>	Y <sub>2</sub>	Y <sub>3</sub>	Y <sub>4</sub>
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	1	0	0	1
x5	0	0	0	1



- Simple, eficiente, paralelizable
- Ignora dependencias entre las etiquetas

X	Y <sub>1</sub>
x1	0
x2	1
x3	0
x4	1
x5	0

X	Y <sub>2</sub>
x1	1
x2	0
x3	1
x4	0
x5	0

X	Y <sub>3</sub>
x1	1
x2	0
x3	0
x4	0
x5	0

X	Y <sub>4</sub>
x1	0
x2	0
x3	0
x4	1
x5	1

En el conjunto de películas:  $p(y_{\text{romance}} | x) = p(y_{\text{romance}} | x, y_{\text{horror}})$  ?



# Classifier Chains (CC)

Métodos de transformación

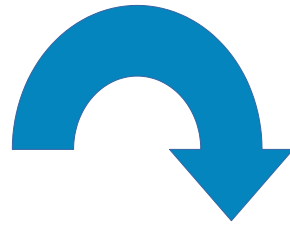
## Classifier Chains (CC):

- En forma similar a BR construye  $L$  clasificadores binarios
- Incluye la etiqueta previa como una nueva característica para el siguiente clasificador

# Classifier Chains (CC)

Métodos de transformación

X	Y <sub>1</sub>	Y <sub>2</sub>	Y <sub>3</sub>	Y <sub>4</sub>
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	1	0	0	1
x5	0	0	0	1



- Requiere búsqueda para determinar el orden de los clasificadores en la cadena

X	Y <sub>1</sub>
x1	0
x2	1
x3	0
x4	1
x5	0

X	Y <sub>1</sub>	Y <sub>2</sub>
x1	0	1
x2	1	0
x3	0	1
x4	1	0
x5	0	0

X	Y <sub>1</sub>	Y <sub>2</sub>	Y <sub>3</sub>
x1	0	1	1
x2	1	0	0
x3	0	1	0
x4	1	0	0
x5	0	0	0

X	Y <sub>1</sub>	Y <sub>2</sub>	Y <sub>3</sub>	Y <sub>4</sub>
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	1	0	0	1
x5	0	0	0	1

# Label Powerset (LP)

Métodos de transformación

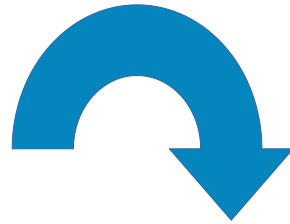
## Label Powerset (LP):

- Transforma cada combinación de etiquetas a un valor de clase
- Aprende un clasificador multi-clase con  $2^L$  valores posibles

# Label Powerset (LP)

## Métodos de transformación

X	$Y_1$	$Y_2$	$Y_3$	$Y_4$
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	1	0	0	1
x5	0	0	0	1



- El número de etiquetas puede ser exponencial
- Aprender un clasificador multi-clase con muchas clases es costoso
- La distribución de clases resultantes va a ser desbalanceada y dispersa

X	$Y \in 2^L$
x1	0110
x2	1000
x3	0100
x4	1001
x5	0001

**MLkNN:** adaptación del algoritmo  $k$ NN para soportar multi-etiqueta

- Recupera los  $N$  vecinos más cercanos a una instancia
- Calcula la frecuencia de ocurrencia de cada etiqueta
- Asigna una probabilidad a cada etiqueta y selecciona las mejores para un ejemplo

# Evaluación

## Métricas

En un problema simple-etiqueta, se compara la etiqueta real  $y$  con la que predice el clasificador  $\hat{y}$ . En multi-etiqueta:

	$y^{(i)}$	$\hat{y}^{(i)}$
x1	1 0 1 0	1 0 0 1
x2	0 1 0 1	0 1 0 1
x3	1 0 0 1	1 0 0 1
x4	0 1 1 0	0 1 0 0
x5	1 0 0 0	1 0 0 1

- **Accuracy:** proporción de etiquetas clasificadas correctamente sobre el total
- **Subset Accuracy:** porcentaje de instancias cuyas etiquetas predichas son exactamente las mismas que en el ground truth
- **Hamming Loss:** cuántas veces se predice una etiqueta incorrectamente

