

Conjuntos de datos multietiqueta

**MÁSTER EN INTELIGENCIA COMPUTACIONAL E
INTERNET DE LAS COSAS -
Universidad de Córdoba**

Contenidos

- **INTRODUCCIÓN**
- **DEFINICIÓN ML**
- **ESTADÍSTICOS DATASETS**
- **APLICACIONES**
- **CATEGORÍAS ML**
- **MÉTRICAS EVALUACIÓN**
- **GENERALIZANDO: MULTIOUTPUT**

Introducción

- **¿Qué es un problema ML?**
- **Categorías de clasificación**
- **Ponte a prueba**

Introducción:

¿Qué es un problema Multilabel?

- ¿Qué es un problema Multilabel?
 - Plantea ejemplos, ¿qué los diferencia de otros problemas?
 - ¿Te aventuras a dar una definición?



Introducción:

¿Qué es un problema Multilabel?

- Analiza los diferentes retos que plantean:



Introducción:

¿Qué es un problema Multilabel?

- Binary classification problem:



- Cada imagen solamente puede ser etiquetada en una de dos clases

Introducción:

¿Qué es un problema Multilabel?



- **Binary VS Multiclass:**
- Número de valores que puede tomar la etiqueta
- Todo problema multiclase puede ser convertido en varios problemas binarios independientes.

■ Ej. gato: sí, no, perro: sí, no; conejo: sí, no; pájaro: sí, no

Introducción:

¿Qué es un problema Multilabel?

- Multiclass classification problem:



- Cada imagen contiene únicamente un objeto/etiqueta (de una de cuatro clases) y sólo puede ser clasificado en una de ellas

Introducción:

¿Qué es un problema Multilabel?

- ¿Y ahora qué?



Introducción:

¿Qué es un problema Multilabel?

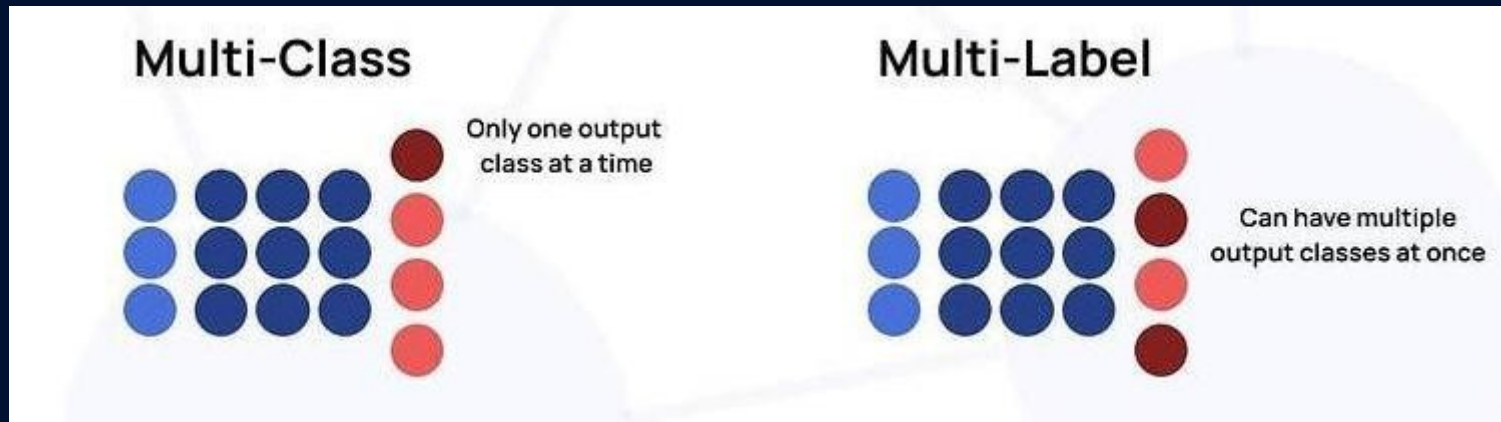
- Multilabel classification problem:



- Cada imagen puede contener más de un objeto (de una de cuatro categorías) y por tanto cada imagen puede ser clasificada en más de una categoría / clase
- ¿Relación entre las clases?
 - ¿Es más habitual ver un perro y gato juntos que un gato con un pájaro?...

Introducción:

¿Qué es un problema Multilabel?



- **Multiclass VS Multilabel:**
- Número de etiquetas que puede tener asignada la instancia como salida
- Todo problema multilabel puede ser **convertido** en varios problemas multiclase independientes.
 - Pierde la información de la relación entre las etiquetas

Introducción:

¿Qué es un problema Multilabel?

□ Otra vuelta de tuerca más: afinemos...

□ Nos replanteamos las etiquetas:

- “Tipo de animal”:
perro, gato, conejo, pájaro
- “Color de animal”:
marrón, azul, gris, negro



Introducción:

¿Qué es un problema Multilabel?

□ Otra vuelta de tuerca más: afinemos...

□ Nos replanteamos las etiquetas:

- “Tipo de animal”:
perro, gato, conejo, pájaro
- “Color de animal”:
marrón, azul, gris, negro

- Cada imagen tiene asociadas dos etiquetas (más de una salida por instancia)
 - Cada etiqueta pertenece a una de varias clases determinadas
 - Relación entre las etiquetas: ¿perros azules?



Introducción:

¿Qué es un problema Multilabel?



- **Multilabel VS Multioutput / Multitask:**
- Número de valores que puede tener asignada cada etiqueta
- Es una **generalización** tanto de los problemas multiclase (sólo generan una propiedad, generan única etiqueta por salida) como de los problemas multilabel (etiquetan atributos binarios)

Introducción:

¿Qué es un problema Multilabel?

□ Y más...



Introducción:

¿Qué es un problema Multilabel?

- Multitud de configuraciones de problemas



Introducción:

Categorías de clasificación

- Pensemos en los diferentes retos que plantea cada problema. ¿Qué características son clave?



Introducción:

Categorías de clasificación

- Pensemos en los diferentes retos que plantea cada problema. ¿Qué características son clave?



- Número de valores que toma cada etiqueta
- Número de etiquetas que se generan como salida de una instancia

Introducción:

Categorías de clasificación

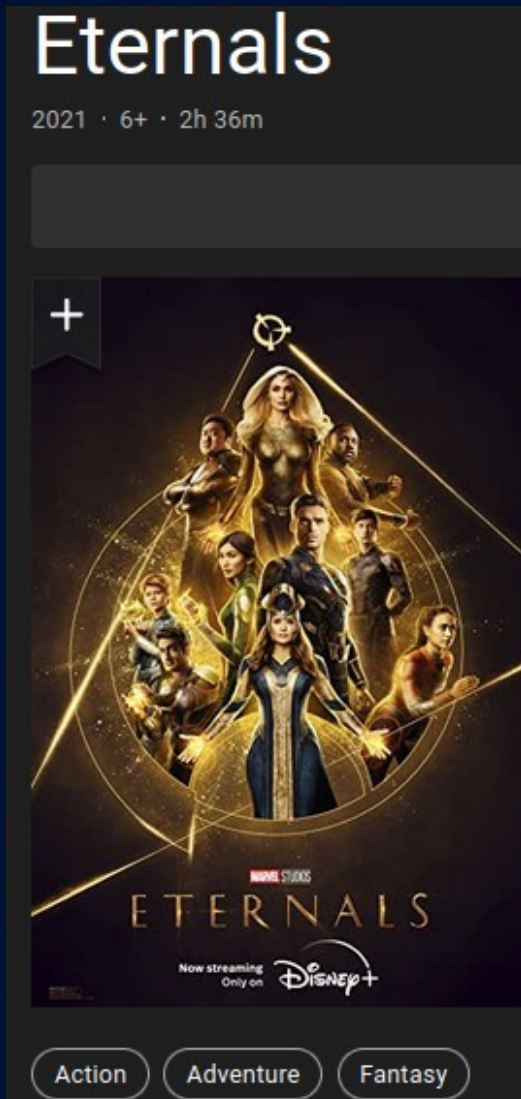
	$K = 2$	$K > 2$
$L = 1$	binary	multi-class
$L > 1$	multi-label	multi-output [†]

[†] also known as multi-target, multi-dimensional.

Figure: For L target variables (labels), each of K values.

- **K**: Número de valores que toma cada etiqueta (atributos binarios o multiclase)
- **L**: Número de etiquetas que puede tener asignada la instancia como salida
 - **Casting**: tanto de multiclase a binario como de multisalida a multilabel

Introducción: Ponte a prueba



Storyline

[Edit](#)

Following the events of [Avengers: Endgame \(2019\)](#), an unexpected tragedy forces the Eternals, ancient aliens who have been living on Earth in secret for thousands of years, out of the shadows to reunite against mankind's most ancient enemy, the Deviants.

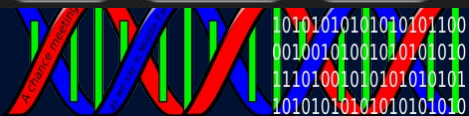
[superhero](#)[banned in the middle east](#)[eternals](#)[banned film](#)[spaceship](#)[560 more](#)[Plot summary](#) · [Plot synopsis](#)

Taglines In the beginning... [>](#)

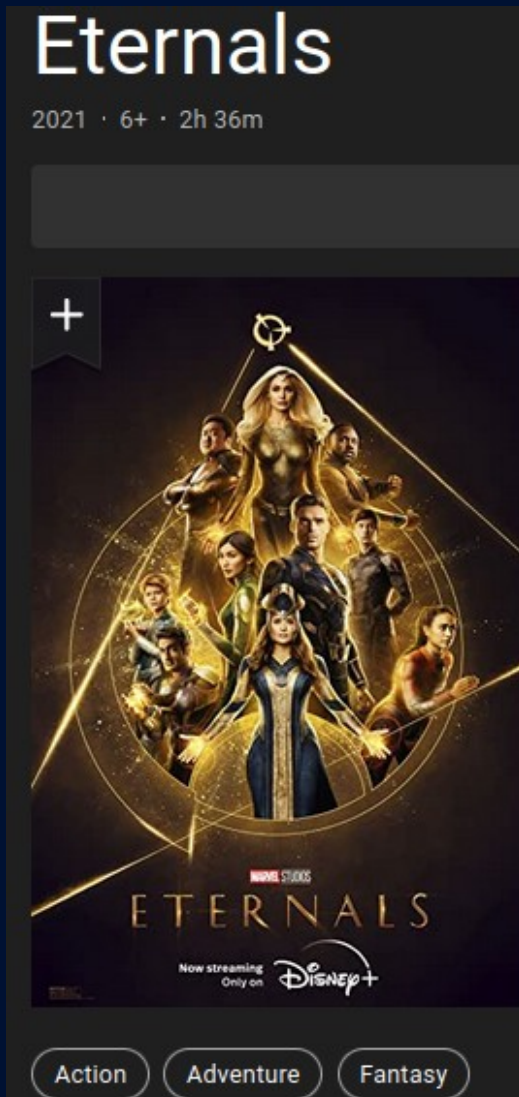
Genres [Action](#) · [Adventure](#) · [Fantasy](#) · [Sci-Fi](#)

Certificate 6+ [>](#)

- ¿De qué va la película?
- ¿A quién va dirigida la película?
- ¿Qué tipo de película es?
- ¿Qué palabras claves seleccionarías?



Introducción: Ponte a prueba



Storyline

[Edit](#)

Following the events of [Avengers: Endgame \(2019\)](#), an unexpected tragedy forces the Eternals, ancient aliens who have been living on Earth in secret for thousands of years, out of the shadows to reunite against mankind's most ancient enemy, the Deviants.

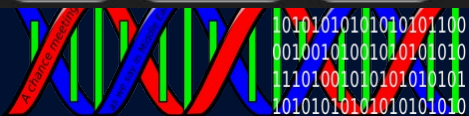
[superhero](#)[banned in the middle east](#)[eternals](#)[banned film](#)[spaceship](#)[560 more](#)[Plot summary](#) · [Plot synopsis](#)

Taglines In the beginning... [>](#)

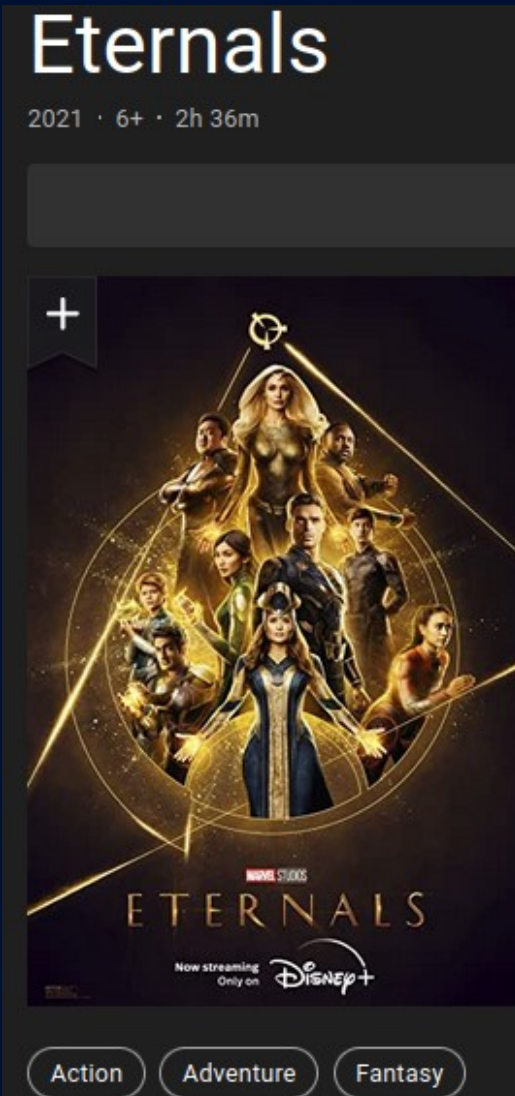
Genres [Action](#) · [Adventure](#) · [Fantasy](#) · [Sci-Fi](#)

Certificate 6+ [>](#)

- ¿De qué va la película? Sinopsis
- ¿A quién va dirigida la película? Edad
- ¿Qué tipo de película es?
- Más de un género
- ¿Qué palabras claves seleccionarías? No predefinido

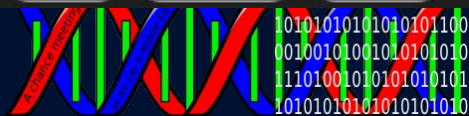


Introducción: Ponte a prueba

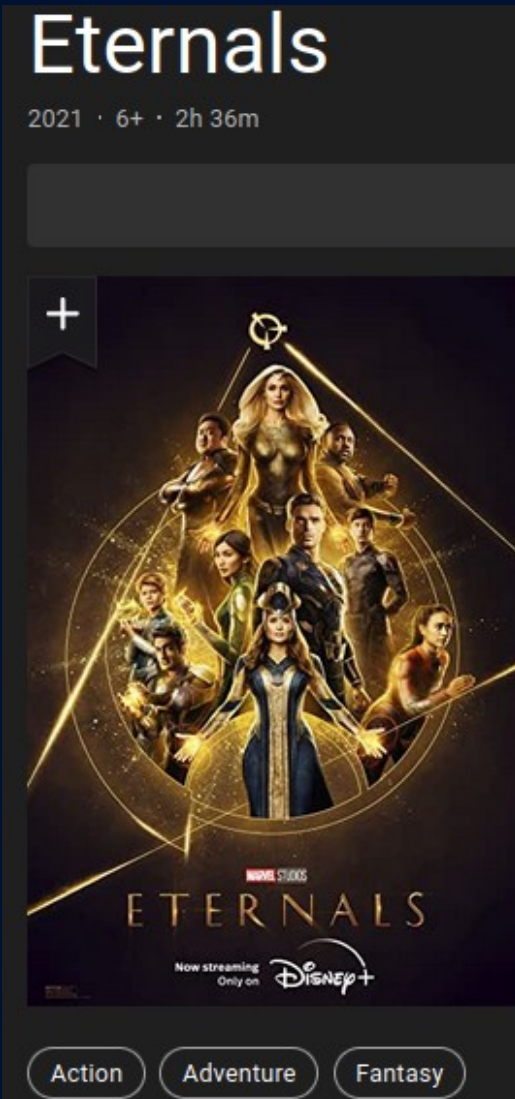


movie	Adventure	Comedy	Fantasy	Crime	children
Jumanji (1995)	1	0	1	0	1
Puccini for Beginners (2006)	0	1	0	0	0
How the Grinch Stole Christmas! (1966)	0	1	1	0	1

- ¿De qué va la película? Sinopsis
- ¿A quién va dirigida la película? Edad
- ¿Qué tipo de película es?
- Más de un género
- ¿Qué palabras claves seleccionarías? No predefinido

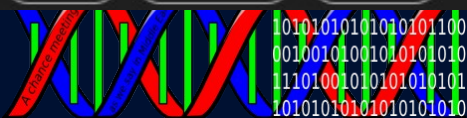


Introducción: Ponte a prueba

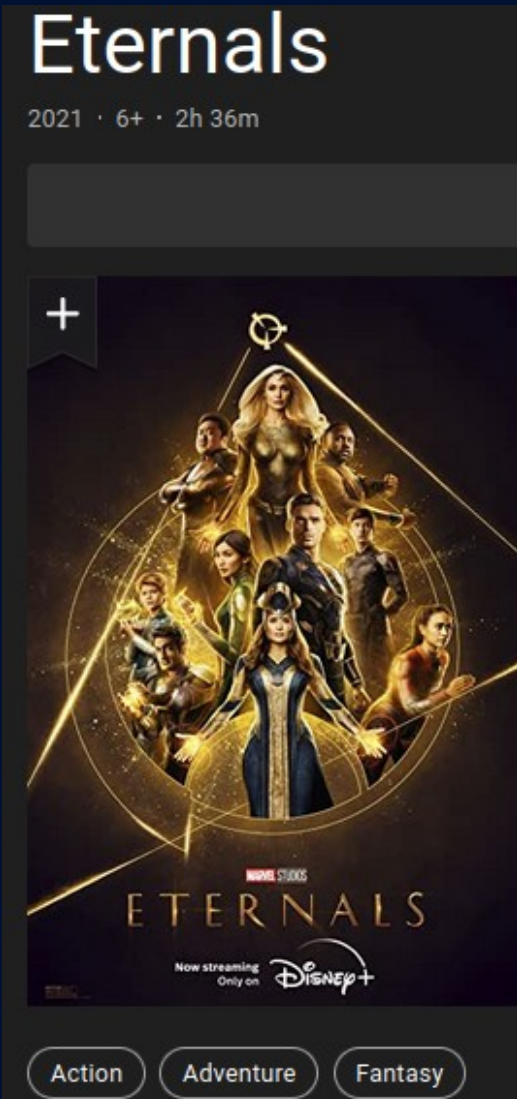


	<i>abandoned</i>	<i>accident</i>	<i>...</i>	<i>violent</i>	<i>wedding</i>	<i>horror</i>	<i>romance</i>	<i>...</i>	<i>comedy</i>	<i>action</i>
<i>i</i>	X_1	X_2	\dots	X_{1000}	X_{1001}	Y_1	Y_2	\dots	Y_{27}	Y_{28}
1	1	0	\dots	0	1	0	1	\dots	0	0
2	0	1	\dots	1	0	1	0	\dots	0	0
3	0	0	\dots	0	1	0	1	\dots	0	0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

- ¿A quién va dirigida la película? Edad
- ¿Qué tipo de película es?
- Más de un género
 - **IMDb dataset:** relaciona sinopsis de películas con géneros
- ¿Qué palabras claves seleccionarías? No predef.

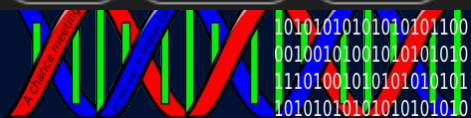


Introducción: Ponte a prueba



	<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	\dots	0	1	0	1	\dots	0	0
2	0	1	\dots	1	0	1	0	\dots	0	0
3	0	0	\dots	0	1	0	1	\dots	0	0
4	1	1	\dots	0	1	1	0	\dots	0	1
5	1	1	\dots	0	1	0	1	\dots	0	1
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
120919	1	1	\dots	0	0	0	0	\dots	0	1

- ¿A quién va dirigida la película? Edad (multiclase)
- ¿Qué tipo de película es? Más de un género
 - **IMDb dataset:** relaciona sinopsis de películas con géneros (multilabel)
- ¿Qué palabras claves seleccionarías? No predef.
 - Tagging



Contenidos

- INTRODUCCIÓN
- **DEFINICIÓN ML**
- ESTADÍSTICOS DATASETS
- APLICACIONES
- CATEGORÍAS ML
- MÉTRICAS EVALUACIÓN
- GENERALIZANDO: MULTIOUTPUT

Definición ML

- Base de Multilabel
 - Obtener mejores clasificaciones basándose en la correlación entre las etiquetas

Definition 2.1 Multi-label Learning (MLL):

Given a training set, $S = (\mathbf{x}_i, \mathbf{Y}_i)$, $1 \leq i \leq n$, consisting n training instances, ($\mathbf{x}_i \in \mathcal{X}$, $\mathbf{Y}_i \in \mathcal{Y}$) i.i.d¹ drawn from an unknown distribution D , the goal of the multi-label learning is to produce a multi-label classifier $h : \mathcal{X} \rightarrow \mathcal{Y}$ (in other words, $h : \mathcal{X} \rightarrow 2^{\mathcal{L}}$) that optimizes some specific evaluation function (i.e. loss function) [66].

Notations Let \mathcal{X} be an instance space, and $\mathcal{L} = \{\lambda_1, \lambda_2, \dots, \lambda_k\}$ be a finite set of class labels. An instance (or an example) $\mathbf{x} \in \mathcal{X}$, represented in terms of features vector $\mathbf{x} = (x_1, x_2, \dots, x_m)$, is (non-deterministically) associated with a subset of labels $L \in 2^{\mathcal{L}}$. Notice that if we call this set L be the set of relevant labels of \mathbf{x} , then we could call the complement $\mathcal{L} \setminus L$ to be the set of irrelevant

Definición ML

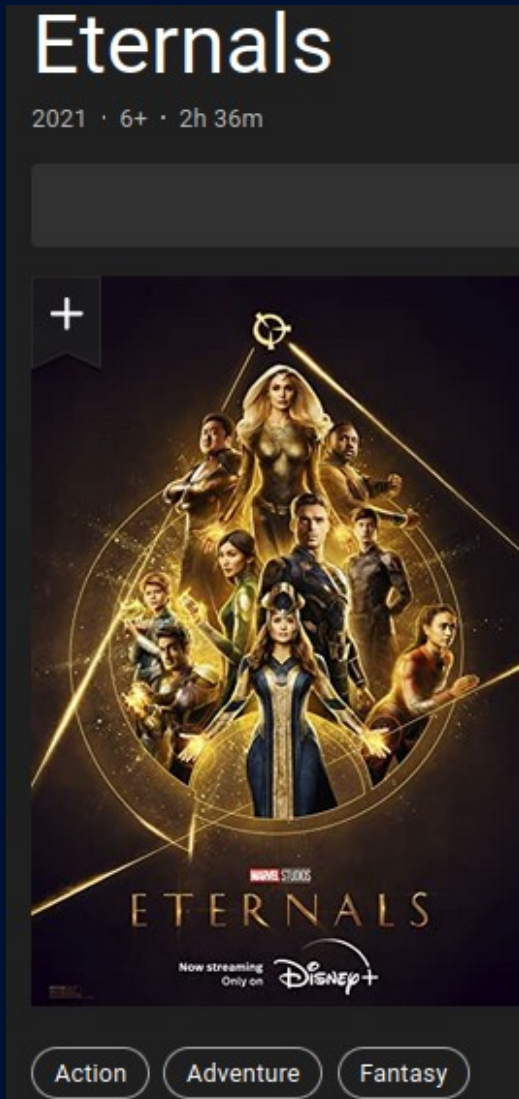
- Salida de Multilabel:
 - **Ranking:** Produce un ranking de relevancia de todas las etiquetas (total strict order) para la instancia en cuestión
 - **Clasificación:** Produce una división del conjunto de etiquetas en relevantes (conjunto positivo) e irrelevantes (negativo)

Definición ML

□ Notación:

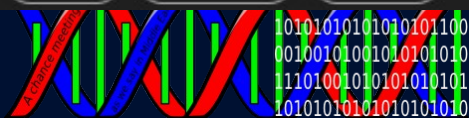
Definition	Symbol/Notation
Instance Space	\mathcal{X}
Label Set	\mathcal{L}
Instance	$\mathbf{x} = (x_1, x_2, \dots, x_m) \in \mathcal{X}$
Number of Features	m
Number of Labels	$k = \mathcal{L} $
Set of relevant labels (for an instance)	L
Correct Label Vector	$\mathbf{Y} = (y_1, y_2, \dots, y_k), y_i \in \{0, 1\}, 1 \leq i \leq k$
Label Vector Space	$\mathcal{Y} = \{0, 1\}^k$
Correct Label Set	$\mathbf{Y}^l = (y_1^l, y_2^l, \dots, y_p^l), y_i^l \in \mathcal{L}, 1 \leq i \leq p$
Label Presence	$Y^\lambda \in \{0, 1\}$
Number of labels for an instance	p
Training dataset	$S = (\mathbf{x}_i, \mathbf{Y}_i), 1 \leq i \leq n$
Number of Instances	n
Classifier predictions	$\mathbf{Z} = (z_1, z_2, \dots, z_k), z_i \in \{0, 1\}, 1 \leq i \leq k$
Predicted Label Set	$\mathbf{Z}^l = (z_1^l, z_2^l, \dots, z_p^l), z_i^l \in \mathcal{L}, 1 \leq i \leq p$
Rank of a label	$r(\lambda)$
Set of classifiers	H
Classifier	h

Definición ML: Ejemplo



	<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	\dots	0	1	0	1	\dots	0	0
2	0	1	\dots	1	0	1	0	\dots	0	0
3	0	0	\dots	0	1	0	1	\dots	0	0
4	1	1	\dots	0	1	1	0	\dots	0	1
5	1	1	\dots	0	1	0	1	\dots	0	1
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
120919	1	1	\dots	0	0	0	0	\dots	0	1

- ¿A quién va dirigida la película? Edad (multiclase)
- ¿Qué tipo de película es? Más de un género
 - **IMDb dataset:** relaciona sinopsis de películas con géneros (multilabel)
- ¿Qué género define mejor la película?
 - Ranking de etiquetas



Definición ML: Ejemplo

Conversión a etiquetas binarias

Table: Single-label $Y \in \{0, 1\}$

X_1	X_2	X_3	X_4	X_5	Y
1	0.1	3	1	0	0
0	0.9	1	0	1	1
0	0.0	1	1	0	0
1	0.8	2	0	1	1
1	0.0	2	0	1	0
0	0.0	3	1	1	?



Table: Single-label $Y \in \{0, 1\}$

X_1	X_2	X_3	X_4	X_5	Y
1	0.1	3	1	0	0
0	0.9	1	0	1	1
0	0.0	1	1	0	0
1	0.8	2	0	1	1
1	0.0	2	0	1	0
0	0.0	3	1	1	?

Table: Multi-label $Y \subseteq \{\lambda_1, \dots, \lambda_L\}$

X_1	X_2	X_3	X_4	X_5	Y
1	0.1	3	1	0	$\{\lambda_2, \lambda_3\}$
0	0.9	1	0	1	$\{\lambda_1\}$
0	0.0	1	1	0	$\{\lambda_2\}$
1	0.8	2	0	1	$\{\lambda_1, \lambda_4\}$
1	0.0	2	0	1	$\{\lambda_4\}$
0	0.0	3	1	1	?

Table: Multi-label $[Y_1, \dots, Y_L] \in 2^L$

X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Y_3	Y_4
1	0.1	3	1	0	0	1	1	0
0	0.9	1	0	1	1	0	0	0
0	0.0	1	1	0	0	1	0	0
1	0.8	2	0	1	1	0	0	1
1	0.0	2	0	1	0	0	0	1
0	0.0	3	1	1	?	?	?	?

Definición ML:

¿Houston tenemos un problema?

- Al plantear desarrollo experimental con diferentes clasificadores ML y discutir los resultados obtenidos:
 - Importante determinar **cómo de multietiqueta** es el conjunto de datos con el que se valida
 - A igual número de etiquetas: Influye en el rendimiento que haya grandes variaciones de número de etiquetas entre instancias [2]



Contenidos

- INTRODUCCIÓN
- DEFINICIÓN ML
- **ESTADÍSTICOS DATASETS**
- APLICACIONES
- CATEGORÍAS ML
- MÉTRICAS EVALUACIÓN
- GENERALIZANDO: MULTIOUTPUT

Estadísticos

• Multi-label Datasets and Statistics

- Distinct Label Set (DL) $DL = |\{Y|\}|\exists \mathbf{x} : (\mathbf{x}, Y) \in S|$

- Proportion of Distinct Label Set (PDL) $PDL = \frac{DL}{|S|}$

- Label Cardinality (Lcard) $LCard = \frac{1}{n} \sum_{i=1}^n |Y_i|$

- Label Density (LDen) $LDen = \frac{LCard}{k}$

Estadísticos

- **Distinct:** La más sencilla, es el número de combinaciones de distintas etiquetas presentes en un conjunto de datos
 - Una elevada medida implica trabajar con un gran espacio de etiquetas, cuyo límite superior es $2^{|L|} - 1$
- **Proportion of Distinct Label Set (PDL):** es la medida Distinct normalizada por el número total de ejemplos / training instances: $|S|$
- **Cardinalidad:** viene dada por el promedio de etiquetas asociadas a cada instancia
 - Distorsión: Por ejemplo podemos tener dos conjuntos de cardinalidades similares, pero puede ocurrir que el número de posibles etiquetas máximo que pueda tener asignado una instancia varíe enormemente
- **Densidad:** cardinalidad entre el número total de etiquetas $|L|$

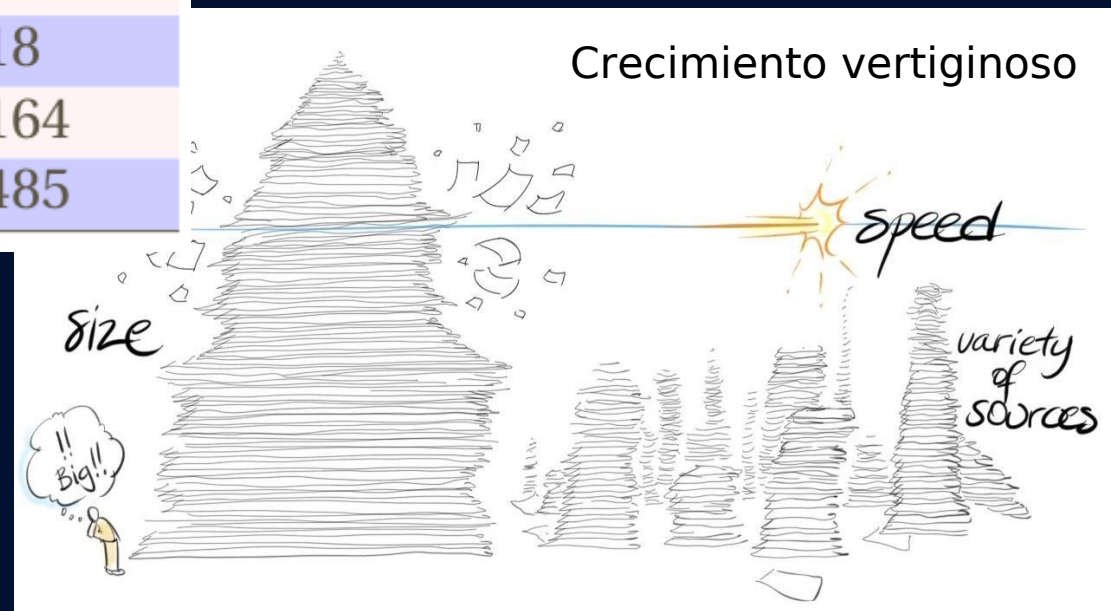
Contenidos

- INTRODUCCIÓN
- DEFINICIÓN ML
- ESTADÍSTICOS DATASETS
- **APLICACIONES**
- CATEGORÍAS ML
- MÉTRICAS EVALUACIÓN
- GENERALIZANDO: MULTIOUTPUT

Aplicaciones: Interés creciente

- **Google Scholar:** número de artículos que contienen las palabras “Multilabel classification”

year	in text	in title
1996-2000	23	1
2001-2005	188	18
2006-2010	1470	164
2011-2015	4550	485



Introducción: Aplicaciones



- **Imágenes: scene dataset**
 - Imágenes se etiquetan con etiquetas conceptuales
 $\subseteq \{\text{beach, sunset, foliage, field, mountain, urban}\}$
 - Múltiples conceptos
 - Múltiples objetos
 - Múltiples personas

Introducción: Aplicaciones

- **Vídeos: mediamill dataset**
 - Pertenecen al conjunto TRECVID 2005/2006
 - Contiene 85 horas de vídeo de noticias
 - Se categorizan en 101 clases

Introducción: Aplicaciones



Introducción: Aplicaciones

Boarding Pass Confirmation



Inbox

x

DOC

x

UNI

x

- **Texto: enron dataset**
 - 517431 emails (sin adjuntos)
 - 3500 carpetas
 - 151 usuarios (directivos de Enron Corp)
 - Tras preprocesamiento de 1702 emails
 - 53 clases

Introducción: Aplicaciones



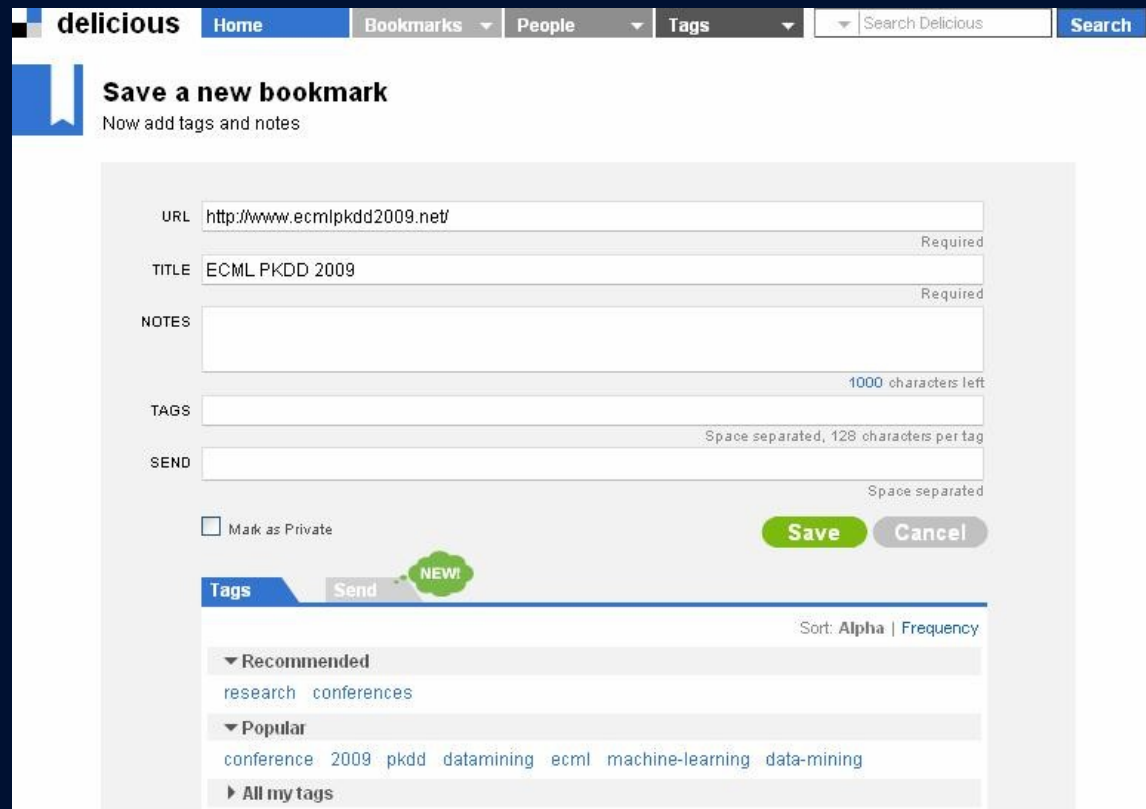
- **Texto: categorización de páginas web**
 - Nº de páginas webs crece vertiginosamente
 - Requiere: colección de doc webs etiquetados

Introducción: Aplicaciones

- **Texto: categorización de páginas web**
 - Directorio de yahoo.com, webs linkadas:
 - Estructura jerárquica de clases de documentos:
 - 14 categorías en nivel superior (ej: “Arts & Humanities”, “Business & Economy”, “Computers & Internet” etc.) +subcategorías

Introducción: Aplicaciones

- **Texto: sugerencia de etiquetas en web 2.0**
 - **Delicious**



The screenshot shows the 'Save a new bookmark' form on the Delicious website. The form includes fields for URL, Title, Notes, Tags, and a Send button. The URL field contains 'http://www.ecmlpkdd2009.net/'. The Title field contains 'ECML PKDD 2009'. The Notes field is empty. The Tags field is empty. The Send button is green and labeled 'NEW!'. Below the form, there is a section for 'Tags' with a 'Send' button and a 'NEW!' badge. The 'Tags' section shows a list of tags: 'research', 'conferences', 'conference', '2009', 'pkdd', 'datamining', 'ecml', 'machine-learning', 'data-mining'. The 'Tags' section also has a 'Sort: Alpha | Frequency' dropdown menu.

Introducción: Aplicaciones

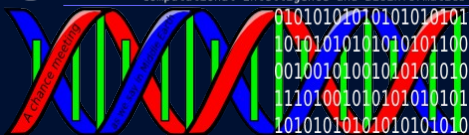


□ Audio

- 593 tracks
- 6 etiquetas:
 - {happy, calm, sad, angry, quiet, amazed}
 - 1.9 etiquetas de media
- Aplicaciones:
 - Selección de música en dispositivos móviles
 - Terapia musical
 - Sistemas de recomendación, TV, programas de radio

CIB

Computational Intelligence and Bioinformatics



Introducción: Aplicaciones

□ **Biología:**

- Anotación automática genes
 - Una proteína
 - Múltiples funciones
- Descubrimiento de medicinas
 - Múltiples estructuras químicas
 - Objetivo: funciones biológicas
- Diagnóstico de pacientes
 - Múltiples enfermedades (interrelacionadas)



Introducción: Aplicaciones y sus datasets

- **Formato de datasets:** Hay multitud de formatos
 - Restric: ser capaces de distinguir atributos y etiquetas
- XML, JSON...etc
- CSV (Comma Separated Values): los valores de distintas variables aparecen separados por comas, y las diferentes entradas se separan mediante un salto de línea
 - Formato estándar de pandas (lib. muy usada python)
- ARFF (Attribute-Relation Format File): es parecido al formato CSV, pero añade información sobre el dataset al comienzo del fichero, principalmente información relativa a los atributos que aparecen representados.
 - El más extendido: usado en Weka y MULAN

Introducción: Aplicaciones y sus datasets

```
@relation MultiLabelExample

@attribute feature1 numeric
@attribute feature2 numeric
@attribute feature3 numeric
@attribute label1 {0, 1}
@attribute label2 {0, 1}
@attribute label3 {0, 1}
@attribute label4 {0, 1}
@attribute label5 {0, 1}

@data
2.3,5.6,1.4,0,1,1,0,0
```

```
@relation weather.symbolic ← Dataset name

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no} ← Attributes

@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no ← Target / Class variable

← Data Values
```

```
<labels xmlns="http://mulan.sourceforge.net/labels">
<label name="label1"></label>
  <label name="label2"></label>
  <label name="label3"></label>
  <label name="label4"></label>
  <label name="label5"></label>
</labels>
```

Ejemplo dataset **ARFF**

Introducción: Aplicaciones y sus datasets

- Ejemplo carga dataset **ARFF** en **scikit-multilearn**

Loading both dense and sparse ARFF files is simple in scikit-multilearn, just use `:func:load_from_arff`, like this:

```
>> from skmultilearn.dataset import load_from_arff
```

Loading multi-label ARFF files requires additional information as the number or placement of labels, is not indicated in the format directly.

```
>> path_to_arff_file = '_static/example.arff'  
>> label_count = 7  
>> label_location="end"
```

Contenidos

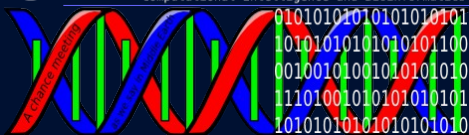
- INTRODUCCIÓN
- DEFINICIÓN ML
- ESTADÍSTICOS DATASETS
- APLICACIONES
- **CATEGORÍAS ML**
- MÉTRICAS EVALUACIÓN
- GENERALIZANDO: MULTIOUTPUT

Contenidos

- INTRODUCCIÓN
- DEFINICIÓN ML
- ESTADÍSTICOS DATASETS
- APLICACIONES
- **CATEGORÍAS ML**
 - Métodos de transformación de problemas
 - Métodos de adaptación de algoritmos
- MÉTRICAS EVALUACIÓN
- GENERALIZANDO: MULTIOUTPUT

CIB

Computational Intelligence and Bioinformatics



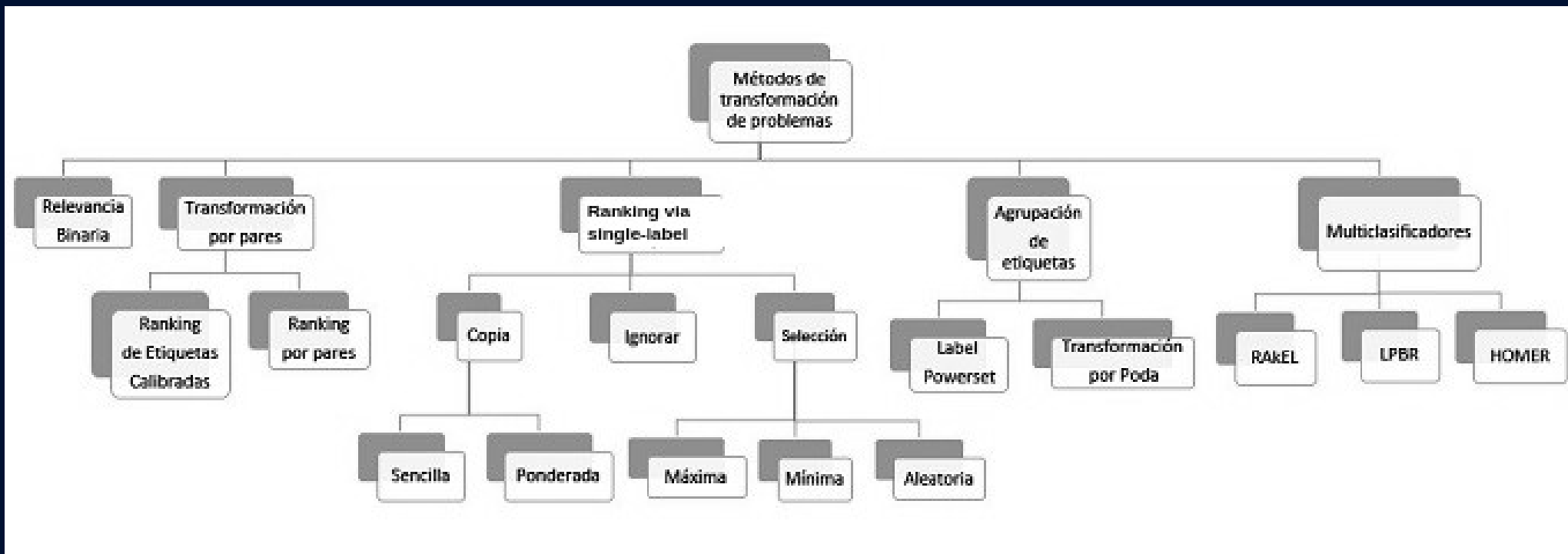
Categorías ML

- Los **métodos de transformación de problemas**: convierten un problema multietiqueta en uno o varios conjuntos de datos de una sola etiqueta
- Los **métodos de adaptación de algoritmos**: adaptan técnicas de clasificación clásicas para tratar con datos multietiqueta de forma directa
 - SVM, árboles de dec., redes N., métodos probabilísticos, alg.bioinspirados, k-vecinos
- **Diferencia fundamental**: Métodos de transformación son independientes del algoritmo de clasificación que se use después

Categorías ML

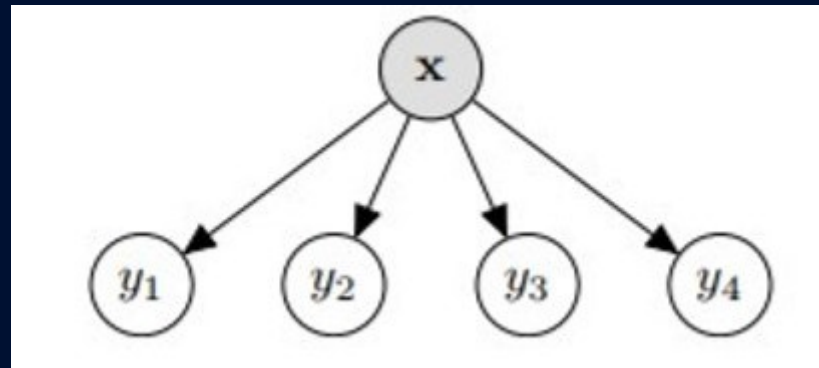
- **Métodos de transformación de problemas**
- **Métodos de adaptación de algoritmos**

Categorías ML: Métodos de transformación de problemas



Métodos de transformación de problemas: familia Binary Relevance

- **Binary Relevance:**
- Separar en L problemas binarios y entrenar con clasif. binario base
- Salida: unión de las predicciones de cada clasificador (o ranking si proporciona los scores)



Métodos de transformación de problemas: familia Binary Relevance

- **Ventajas:**
 - Utiliza técnicas de clasificación clásicas (muy verificadas)
 - Generaliza más allá de las etiquetas del entrenamiento
 - Escala linealmente en el nº de etiquetas (paralelizable)
- **Desventajas:**
 - Complejidad $O(kn)$
 - Muy lento si el espacio de etiquetas es grande
 - Asume independencia entre etiquetas: pérdida de resultados más reales y óptimos
 - Hay métodos basados en BR pero contabilizan dependencia: Classifier chain, Meta-BR, BRplus(BR+)

Métodos de transformación de problemas: familia Binary Relevance

□ Binary Relevance:

- Separar en L problemas binarios y entrenar cada uno con clasif. binario
- Salida: unión de las predicciones o ranking si los clasificadores proporcionan los scores

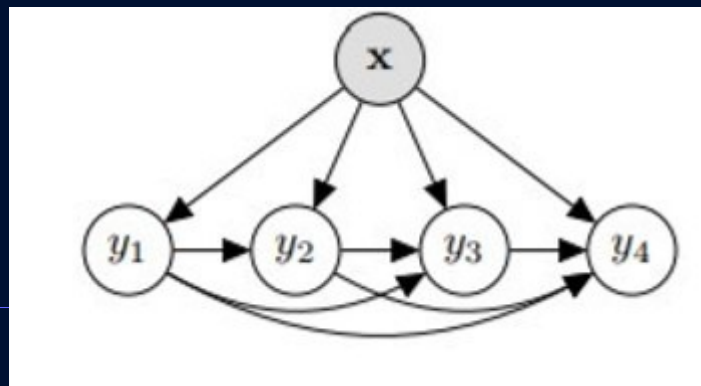
X	Y_1	Y_2	Y_3	Y_4
$\mathbf{x}^{(1)}$	0	1	1	0
$\mathbf{x}^{(2)}$	1	0	0	0
$\mathbf{x}^{(3)}$	0	1	0	0
$\mathbf{x}^{(4)}$	1	0	0	1
$\mathbf{x}^{(5)}$	0	0	0	1



X	Y_1	X	Y_2	X	Y_3	X	Y_4
$\mathbf{x}^{(1)}$	0	$\mathbf{x}^{(1)}$	1	$\mathbf{x}^{(1)}$	1	$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	1	$\mathbf{x}^{(2)}$	0	$\mathbf{x}^{(2)}$	0	$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	0	$\mathbf{x}^{(3)}$	1	$\mathbf{x}^{(3)}$	0	$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1	$\mathbf{x}^{(4)}$	0	$\mathbf{x}^{(4)}$	0	$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	0	$\mathbf{x}^{(5)}$	0	$\mathbf{x}^{(5)}$	0	$\mathbf{x}^{(5)}$	1

Métodos de transformación de problemas: familia Binary Relevance

- **Classifier Chain:** de la categoría de BR
- Convierte problema ML en problemas binarios pero incluye como características las predicciones de las etiquetas anteriores
 - Podemos emplear cualquier clasificador
- Calidad de los resultados depende del orden en que se predicen las etiquetas (user setting)
 - $k!$ posibles ordenaciones: lento si espacio etiquetas grande



Métodos de transformación de problemas: familia Binary Relevance

- Ventajas:
 - Tiene en cuenta dependencias entre etiquetas
 - Capaz de generalizar a combinaciones de etiquetas que no haya visto al entrenar

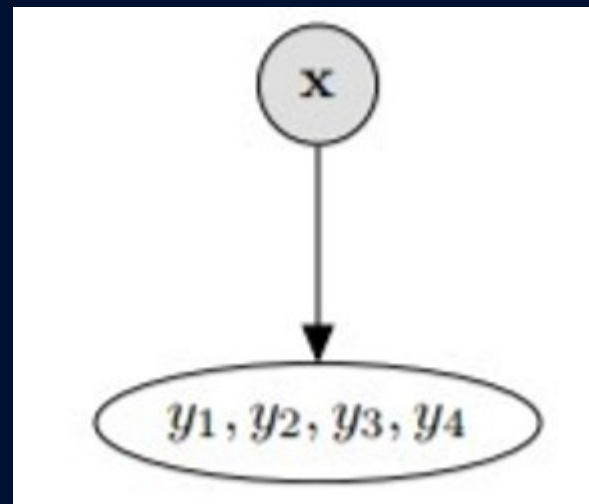
X	Y_1	X	Y_1	Y_2	X	Y_1	Y_2	Y_3	X	Y_1	Y_3	Y_3	Y_4
$\mathbf{x}^{(1)}$	0	$\mathbf{x}^{(1)}$	0	1	$\mathbf{x}^{(1)}$	0	1	1	$\mathbf{x}^{(1)}$	0	1	1	0
$\mathbf{x}^{(2)}$	1	$\mathbf{x}^{(2)}$	1	0	$\mathbf{x}^{(2)}$	1	0	0	$\mathbf{x}^{(2)}$	1	0	0	0
$\mathbf{x}^{(3)}$	0	$\mathbf{x}^{(3)}$	0	1	$\mathbf{x}^{(3)}$	0	1	0	$\mathbf{x}^{(3)}$	0	1	0	0
$\mathbf{x}^{(4)}$	1	$\mathbf{x}^{(4)}$	1	0	$\mathbf{x}^{(4)}$	1	0	0	$\mathbf{x}^{(4)}$	1	0	0	1
$\mathbf{x}^{(5)}$	0	$\mathbf{x}^{(5)}$	0	0	$\mathbf{x}^{(5)}$	0	0	0	$\mathbf{x}^{(5)}$	0	0	0	1

Métodos de transformación de problemas: Transformación pares

- **Ranking por comparación de pares:** sigue la filosofía uno VS uno (OVO), transforma un conjunto de datos con k etiquetas en $k(k - 1)/2$ conjuntos de datos binarios (uno por par de etiq.)
 - Clasificador binario: sobre cada conjunto de datos utiliza los ejemplos pertenecientes a una de las dos clases como positiva o negativa respectivamente
 - Elimina instancias pertenecientes a ambas etiquetas

Métodos de transformación de problemas: agrupación de etiquetas

- **Label Powerset:** genera un solo conjunto de datos multiclase, con tantas clases como combinaciones haya en el conjunto original.
 - Considera únicamente las combinaciones de clases presentes al entrenar
- **Ventajas:** considera las relaciones entre etiquetas



Métodos de transformación de problemas: agrupación de etiquetas

- Desventajas:
 - Elevada complejidad: lím. sup es $\min(n, 2^k)$
 - Suele ser menor que el rango superior pero aún así es mayor que k
 - No predice combinaciones no vistas antes
 - Riesgo de overfitting
 - Genera conjuntos desequilibrados
 - Enron dataset, 44% of labelsets son únicos (una sola instancia de entrenamiento o test)
 - del.icio.us dataset, 98% son únicos

Métodos de transformación de problemas: multclasificadores

- **Ensemble methods:** combinan las respuestas de varios clasificadores, desarrollados mediante la misma o distinta técnica, para obtener una respuesta en general más adecuada que la de cada uno de los clasificadores por separado.
- El algoritmo Random k Labelset (**RAkEL**)[5] ensemble de 'n' clasificadores LP, cada uno entrenado sobre un 'k'-labelset
 - **K-labelset:** subconjto de 'k' etiquetas aleatorias
 - **Fijar parámetros:** 'k' y 'n' ($k=3$, $n=2*\text{num_etiq}$)
 - **Threshold:** hace media de los difs clasificadores por etiqueta, un umbral determina si se asigna

Métodos de transformación de problemas: multclasificadores

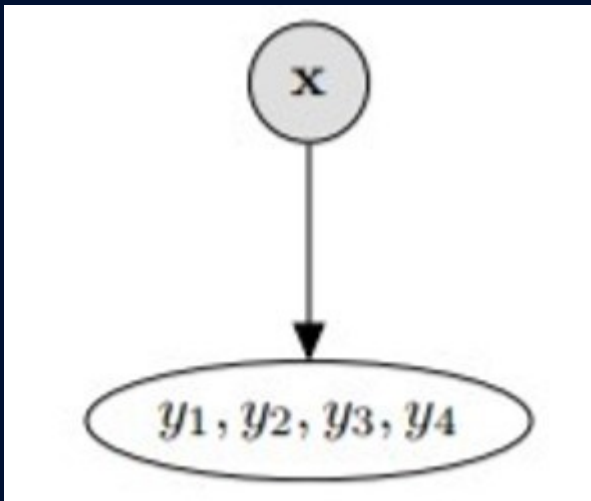
\mathbf{X}	$Y \in 2^L$
$\mathbf{x}^{(1)}$	0110
$\mathbf{x}^{(2)}$	1000
$\mathbf{x}^{(3)}$	0110
$\mathbf{x}^{(4)}$	1001
$\mathbf{x}^{(5)}$	0001



\mathbf{X}	$Y_{123} \in 2^k$
$\mathbf{x}^{(1)}$	011
$\mathbf{x}^{(2)}$	100
$\mathbf{x}^{(3)}$	011
$\mathbf{x}^{(4)}$	100
$\mathbf{x}^{(5)}$	000

\mathbf{X}	$Y_{124} \in 2^k$
$\mathbf{x}^{(1)}$	010
$\mathbf{x}^{(2)}$	100
$\mathbf{x}^{(3)}$	010
$\mathbf{x}^{(4)}$	101
$\mathbf{x}^{(5)}$	001

\mathbf{X}	$Y_{234} \in 2^k$
$\mathbf{x}^{(1)}$	110
$\mathbf{x}^{(2)}$	000
$\mathbf{x}^{(3)}$	110
$\mathbf{x}^{(4)}$	001
$\mathbf{x}^{(5)}$	001

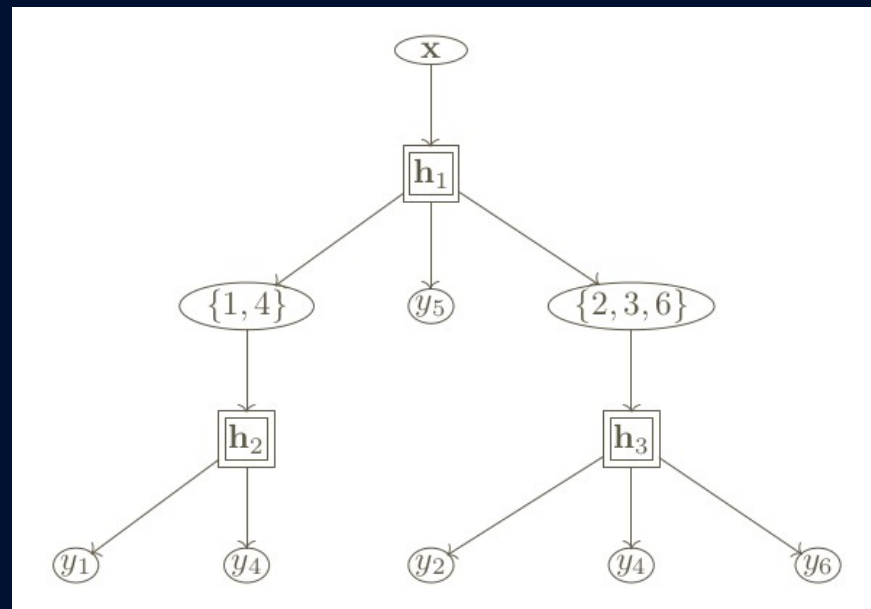


Ventajas:

- Subproblemas más simples
- Conjuntos de entrenamiento balanceados
- Predicen combinaciones no vistas en training

Métodos de transformación de problemas: multclasificadores

- **HOMER** (Hierarchy Of Multilabel classifiERs): hace clustering sobre las etiquetas (o parte de una jerarquía de etiquetas predefinida) y aplica problem-transformation



Categorías ML

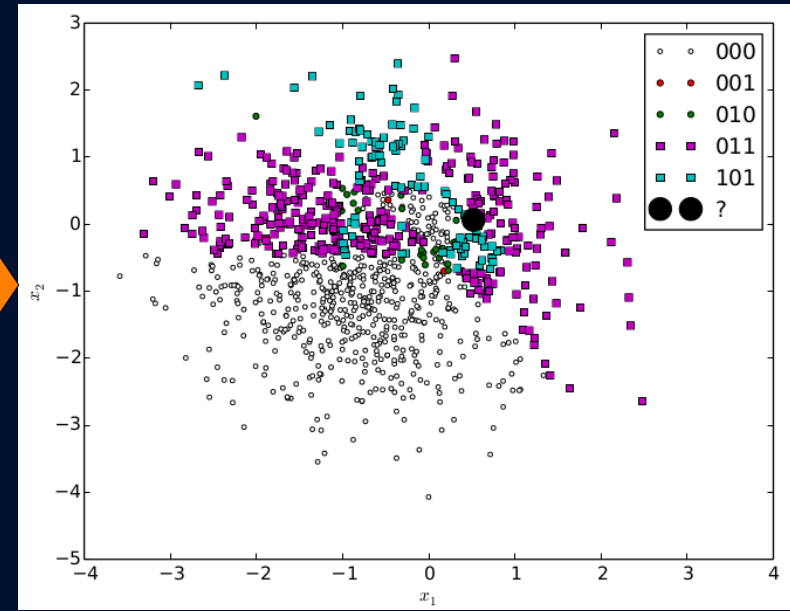
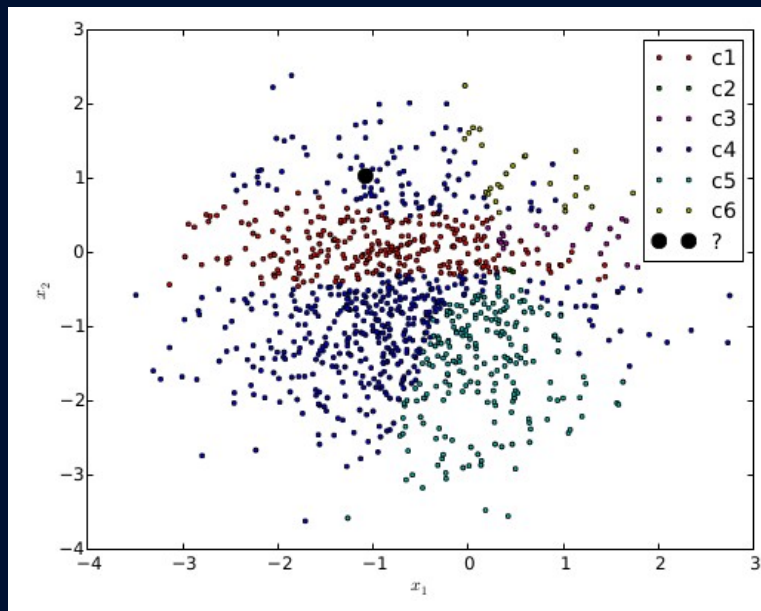
- Métodos de transformación de problemas
- **Métodos de adaptación de algoritmos**

Categorías ML: Métodos de adaptación de algoritmos

- 1) Escoge tu clasificador favorito
- 2) Modifícalo para clasificación ML
 - **Ventajas:** un único modelo, escalable
 - **Desventajas:** rendimiento de la predicción dependerá del dominio del problema

Categorías ML: Métodos de adaptación de algoritmos

- **kNN**: asigna a la instancia la clase mayoritaria de los k vecinos (img izquierda)
- **MikNN**: asigna a las etiquetas más comunes de los k vecinos (img derecha)



Categorías ML: Métodos de adaptación de algoritmos

- **MLkNN**: aproximación lazy learning ML
 - Averigua los vecinos más cercanos y basándose en información estadística (maximum a posterior principle) determina la nueva combinación de etiquetas

Categorías ML: Métodos de adaptación de algoritmos

- **BrkNNaClassifier** (Binary Relevance k-Nearest Neighbours)
 - Se entrena un algoritmo k-Nearest Neighbour por cada etiqueta
 - Computacionalmente costoso si el espacio de etiquetas es grande

Categorías ML:

Houston, ¿tenemos un problema?

A small icon of a rocket ship, tilted upwards, with a flame-like shape below it, representing a problem or a launch.

□ ¿Cuál es *mejor*?



Categorías ML:

- **Depende del problema:**
 - Eficiencia: basados en árboles de decisión
 - Flexibilidad: métodos transf. problemas, esp. basados en BR
 - Capacidad predictiva: ensembles modernos
- **Estudios empíricos recientes [6] recomiendan:**
 - RT-PCT: Random Forest of Predictive Clustering Trees (Adaptación de algoritmos, basado en árboles)
 - HOMER: presentación original (Transformación de problemas, basado en LP)
 - CC: (Transformación de problemas, basado en BR)



Contenidos

- INTRODUCCIÓN
- DEFINICIÓN ML
- ESTADÍSTICOS DATASETS
- APLICACIONES
- CATEGORÍAS ML
- **MÉTRICAS EVALUACIÓN**
- GENERALIZANDO: MULTIOUTPUT

Métricas: Taxonomía

- **Basadas en instancias:** se calculan sobre cada ejemplo de test y luego se hace la media
- **Basadas en etiquetas:** se calculan sobre cada etiqueta independientemente y luego se hace media sobre todas las etiquetas
 - Se puede usar cualquier métrica de clasif. binaria
- **Implementadas en scikit:**
 - Hamming loss
 - Accuracy score
 - Coverage

Métricas evaluación

- **Basadas en instancias**
- **Basadas en etiquetas**

Métricas evaluación: Basadas en instancias

- **Precisión:** calcular la media de la precisión de los clasificadores binarios
 - Subset accuracy
- **Hamming loss:** Evalúa cuántas veces un par instancia-etiqueta se clasifica mal. Operación XOR

	y^i	\hat{y}^i
x_1	[1010]	[1001]
x_2	[0101]	[0101]
x_3	[1000]	[0100]
x_4	[1000]	[1001]

Métricas evaluación: Basadas en instancias

- **One-error:** evalúa cuántas veces la etiqueta mejor evaluada no se selecciona. No es muy apropiada para ML porque sólo contabiliza esa etiqueta
- **Coverage:** calcula cuántas etiquetas en media han de ser incluidas en la predicción final para que las etiquetas reales sean incluidas
 - Valor deseable para esta métrica: que sea el número medio de etiquetas reales

Métricas evaluación

- Basadas en instancias
- **Basadas en etiquetas**

Métricas evaluación: Basadas en etiquetas

- **Support:** número de ocurrencias de cada etiqueta en y_true
- **Accuracy:** la proporción de predichas correctamente respecto al total

		Actual relevancy	
		Yes	No
Classifier output	Yes	TP_i	FP_i
	No	FN_i	TN_i

$$ACC_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i}$$

$$F1_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}$$

Métricas evaluación: Basadas en etiquetas

- **Precision:** la fracción de etiquetas positivas que se predijeron que son realmente positivas. Mide la estabilidad de la medida frente a las repeticiones
- **Recall (TPR/sensitivity):** la fracción de etiquetas positivas reales que se consiguieron predecir
 - En imbalanced learning: se emplea recall para medir la cobertura de la clase minoritaria
- **F-measure:** en una única medida atiende a las preocupaciones tanto de precision como de recall

Métricas evaluación: Basadas en etiquetas

- Métricas clásicas, no considera relación entre etiquetas.
El cálculo de la media:
 - **Macro (per category averaging):** una medida para cada etiqueta individualmente y se hace media sin ponderar
 - No tiene en cuenta desequilibrio entre etiquetas
 - **Micro (per example averaging):** conteo total TP, FP... para todas las etiquetas, da igual peso a cada instancia
 - Dominada por el desempeño de clases más frecuentes

Métricas evaluación: Houston, ¿tenemos un problema?

- ¿Cuál es *mejor usar*?



Métricas evaluación: ¿Cuál es mejor usar?

Depende del problema [8]:

- Computer aided annotation by humans
Ej. tag suggestion en sistemas web 2.0
 - Ranking (basada en instancias)
- Automated annotation for retrieval
 - F-measure con macro-media
- Direct marketing
 - Precisión y ranking (basada en instancias)
- Query categorization
 - Precisión y ranking (basada en instancias)



Contenidos

- INTRODUCCIÓN
- DEFINICIÓN ML
- ESTADÍSTICOS DATASETS
- APLICACIONES
- CATEGORÍAS ML
- MÉTRICAS EVALUACIÓN
- **GENERALIZANDO: MULTIOUTPUT**

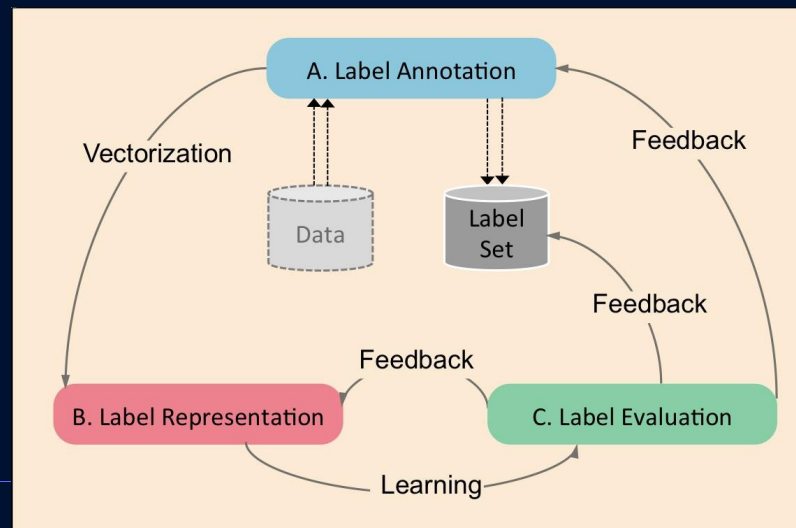
Generalizando: multioutput

- **¿Es fácil obtener las salidas?**
- **Boom de las salidas**
- **Subcampos de trabajo**
- **Nuevas métricas de evaluación**

Multitoutput:

¿Es fácil obtener las salidas?

- 1) **Anotación de etiquetas:** proceso intensivo que requiere tiempo y de un experto humano para anotar los datos semánticamente
 - Crucial para el entrenamiento en multitoutput
 - Hay varias maneras de obtener datos con etiquetas: crowdsourcing (Amazon Mechanical Turk), annotation tools (como LabelMe), redes sociales...



Multiooutput:

¿Es fácil obtener las salidas?

- 2) **Representación de las etiquetas:** cada tipo de anotación (tags, captions, masks...) puede tener diferentes tipos de representaciones de etiquetas
- Es muy importante seleccionar la representación más apropiada para la tarea en cuestión
 - Vectores binarios (pierden información en tareas más complejas, estructura, semántica), vectores de valores reales que indican la relevancia / grado, vectores de etiquetas jerárquicas, vectores de palabras semánticos...

Multiooutput:

¿Es fácil obtener las salidas?

- 3) **Evaluación de las etiquetas:** esencial para garantizar la calidad de las etiquetas y sus representaciones. Según:
 - 1) Si la anotación es de buena calidad
 - 2) Si la representación escogida representa adecuadamente a las etiquetas
 - 3) Si el conjunto de etiquetas cubre el conjunto de datos (Label Set)
- Después de la evaluación, se necesita que un experto humano analice los problemas subyacentes y que especifique los distintos aspectos a mejorar de las etiquetas

Generalizando: multioutput

- ¿Es fácil obtener las salidas?
- **Boom de las salidas**
- Subcampos de trabajo
- Nuevas métricas de evaluación

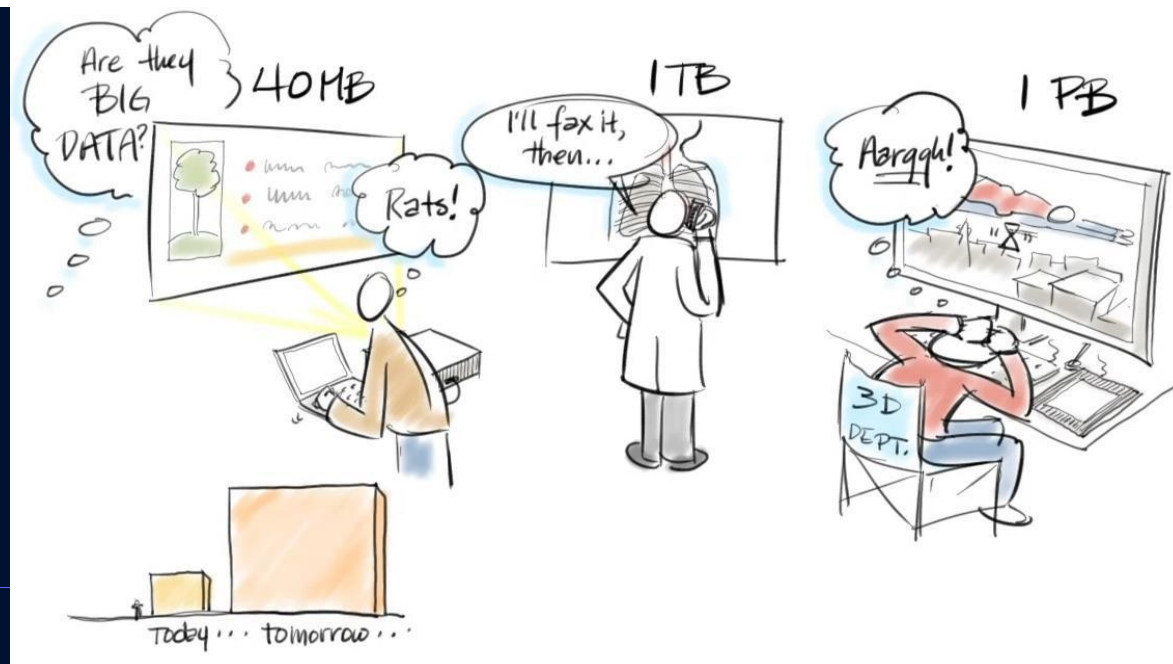
Multiooutput: Boom de las salidas



Big? data

- Capacidad ordenadores
- Medios disponibles

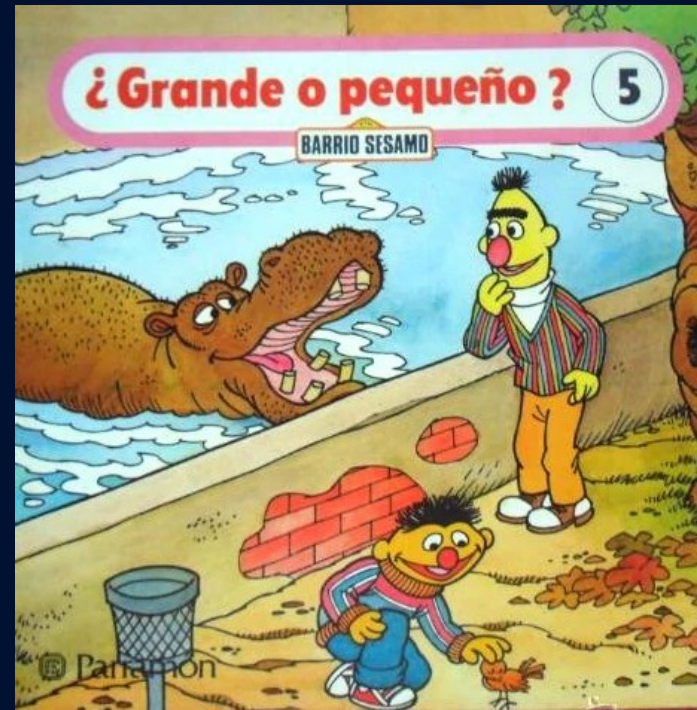
¿Demasiadas salidas?



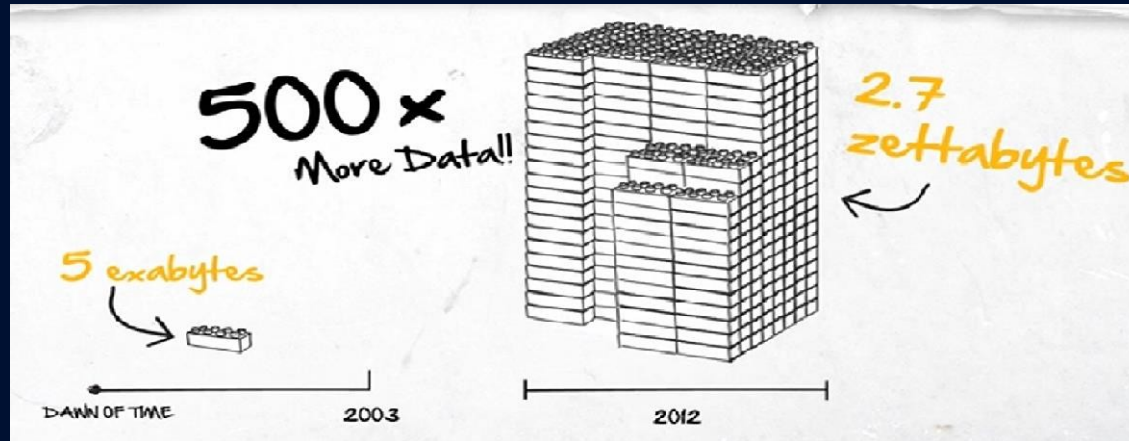
Multiooutput: Boom de las salidas

- Las tecnologías avanzan rápidamente y los datos se acumulan a una velocidad inalcanzable por la capacidad de procesamiento de datos humana y de los algoritmos

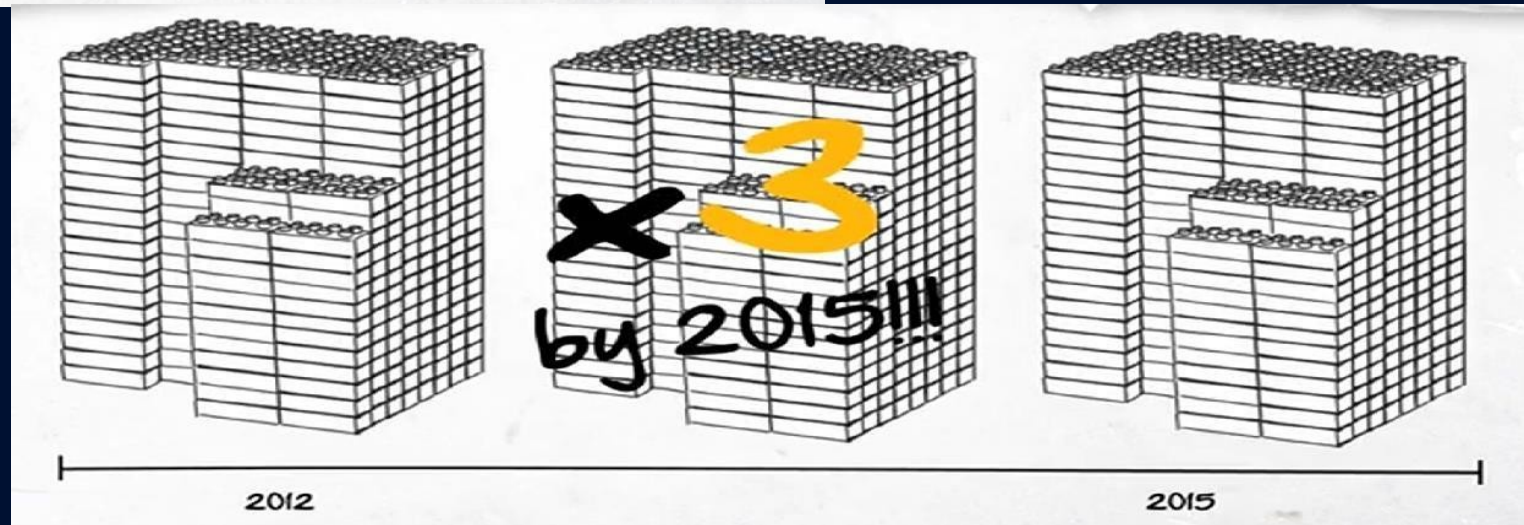
Lo que es grande hoy,
¡no lo será mañana!



Multiooutput: Boom de las salidas

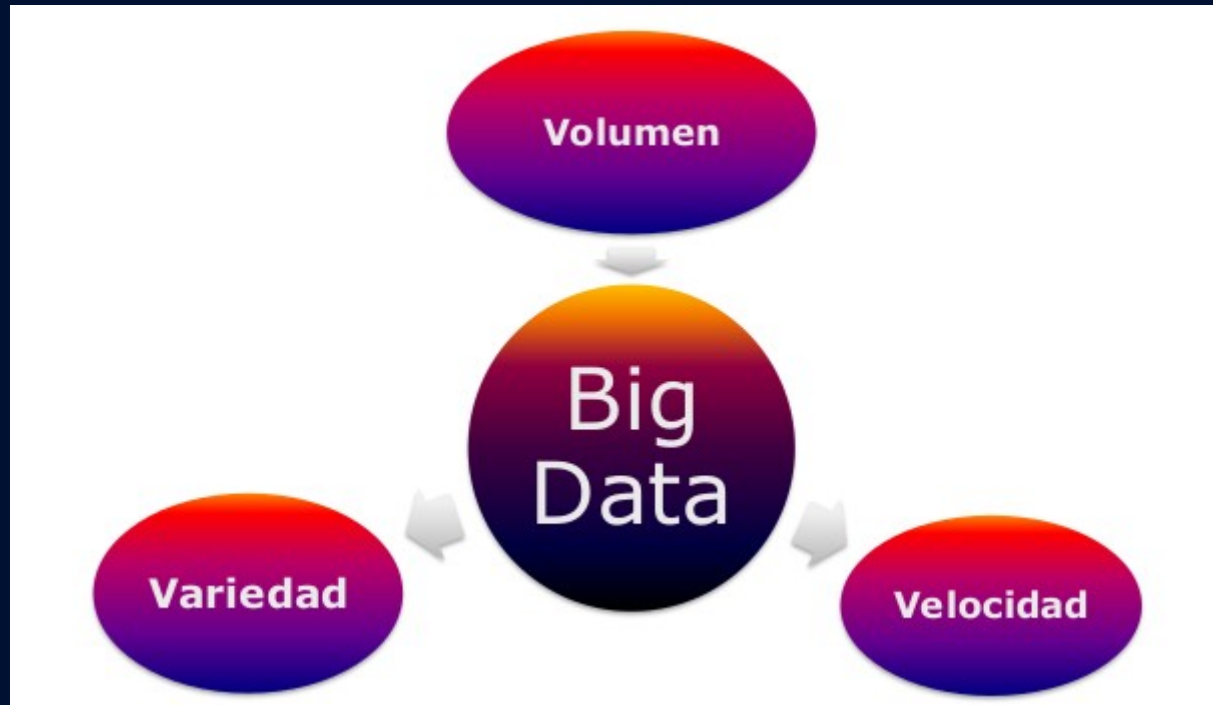


... será **más grande**
mañana



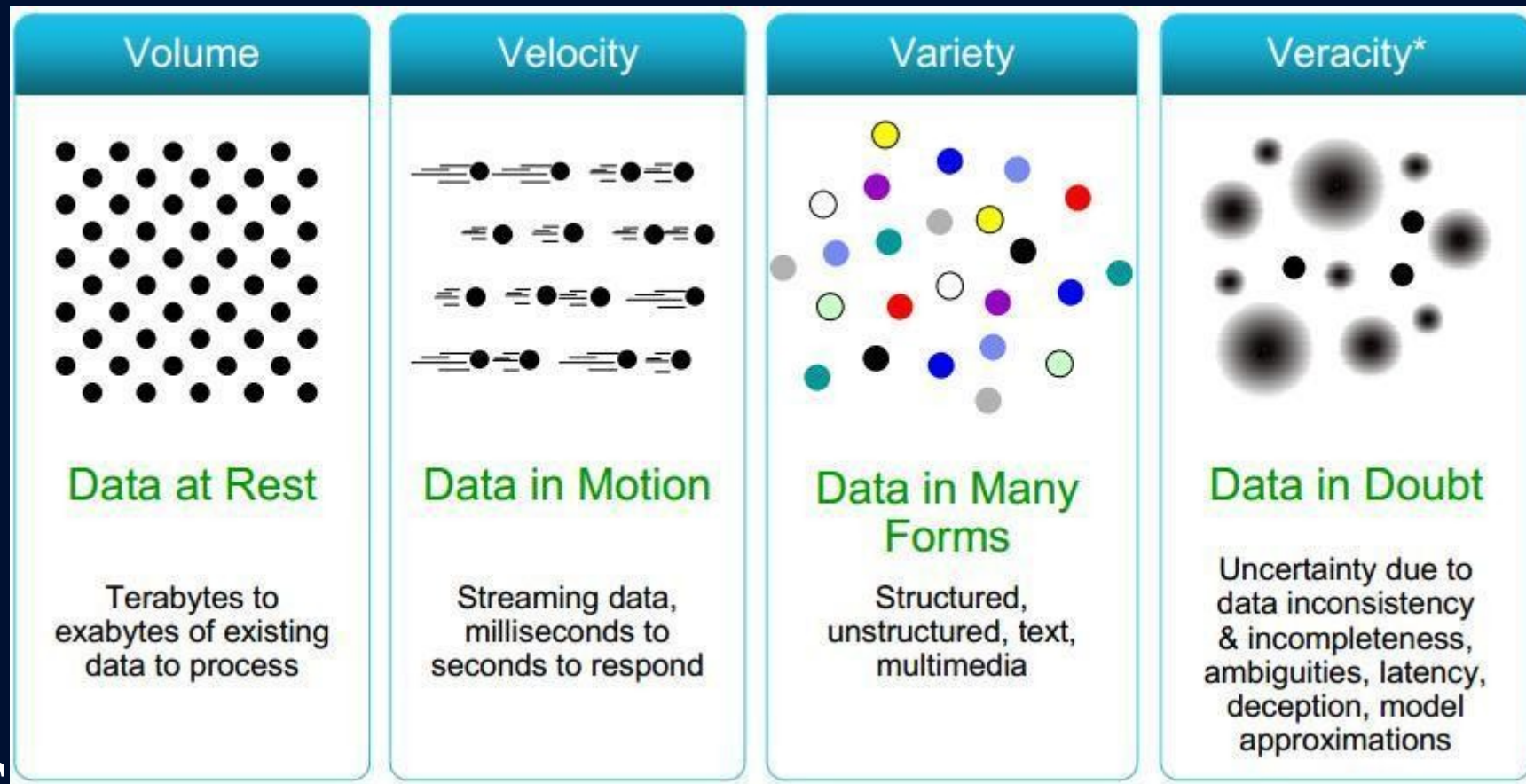
Multioutput: Boom de las salidas

- Volumen, diversidad y complejidad [9]
 - Aplicable a big data y también multioutput



Multiooutput: Boom de las salidas

□ ...y veracidad



Multiooutput: Boom de las salidas

- ...y muchas más **uves**



"The answer to the ultimate question of life, the universe and everything is **42**"

Douglas Adams,

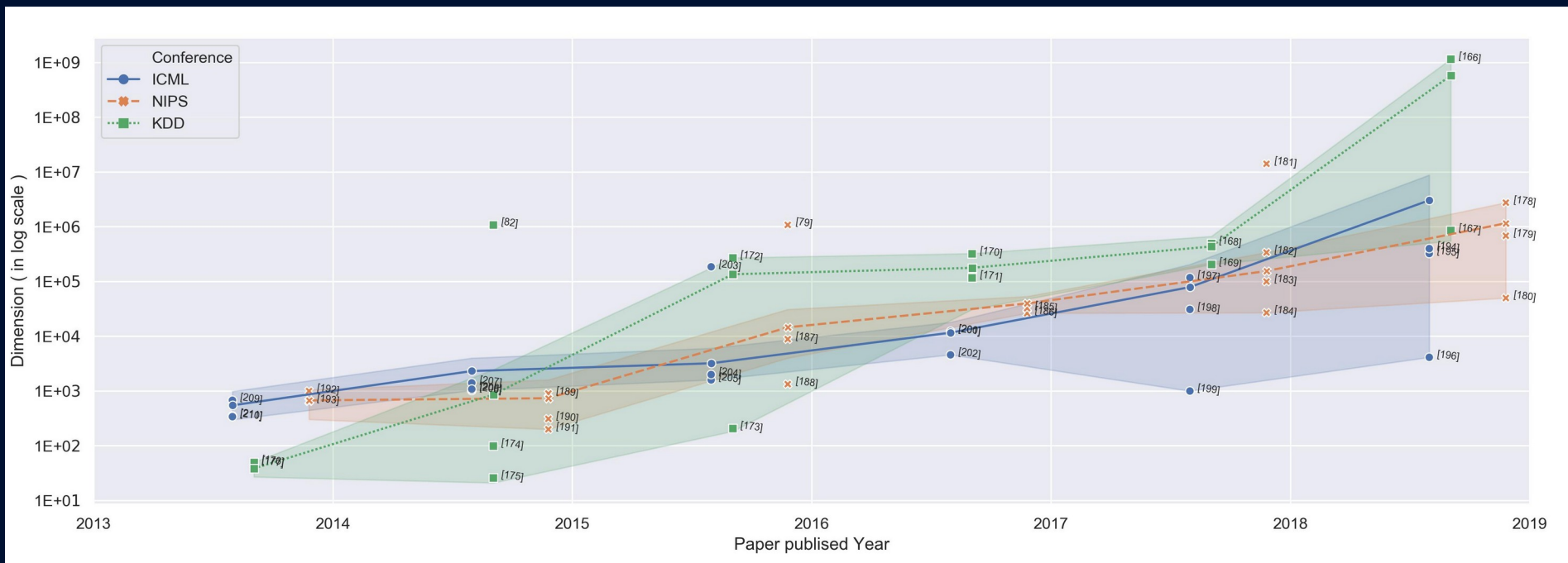
"The Hitchhiker's Guide to the Galaxy"

Multiooutput: Boom de las salidas

- **Volumen:** crecimiento exponencial de las etiquetas. Plantea retos:
 - Problemas de escalabilidad
 - Más carga de trabajo para los anotadores, que no haya suficientes en el entrenamiento y se muestren instancias nuevas durante el test
 - Un gran volumen de salidas suele estar relacionado con desequilibrio (no todas las etiquetas tienen suficientes instancias)

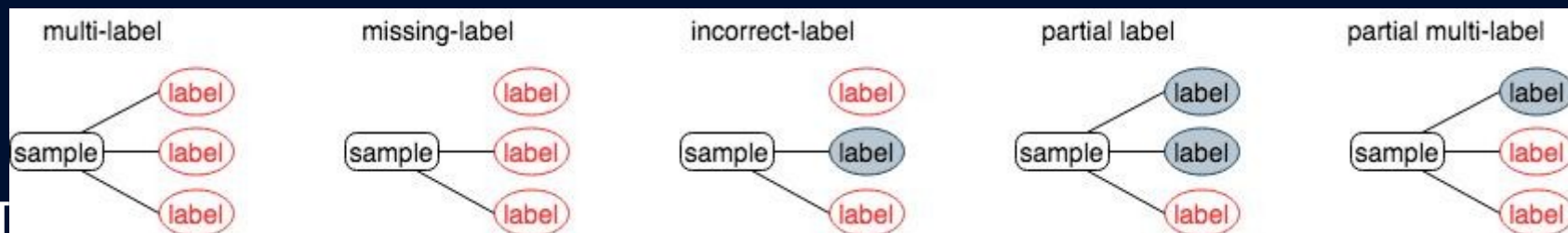
Multiooutput: Boom de las salidas

- **Volumen:** crecimiento exponencial de las etiquetas en papers y congresos (millones y billones de salidas)



Multitoutput: Boom de las salidas

- **Velocidad:** velocidad con la que se generan etiquetas. Streams de datos que pueden dar lugar a *concept drift* (la distribución de las salidas se modifique a lo largo del tiempo)
- **Variedad:** la naturaleza diversa de las salidas. Reto de encontrar el método de modelado apropiado que refleje estructuras complejas de etiquetas: relaciones de dependencia, correlaciones entre las etiquetas...
- **Veracidad:** la calidad de las etiquetas de salida
 - Ej: ruido, valores que faltan, anomalías, incompletitud



Generalizando: multioutput

- ¿Es fácil obtener las salidas?
- Boom de las salidas
- **Subcampos de trabajo**
- Nuevas métricas de evaluación

Multiooutput: subcampos de trabajo

- **Clasificación multilabel:** aprende una función que determina para cada instancia qué etiquetas son relevantes (vector bin. disperso)
 - Salidas: son valores nominales
- **Regresión multiobjetivo:** predice para cada instancia el nivel de pertenencia a cada etiqueta (vector de reales)
 - Salidas: son valores reales
- **Distribución de etiquetas:** determina el grado en el que representa cada etiqueta a la instancia
 - Suma de los grados para cada instancia es 1
- **Ranking de etiquetas:** cada instancia se asocia con rankings de varias etiquetas

Multiooutput: subcampos de trabajo

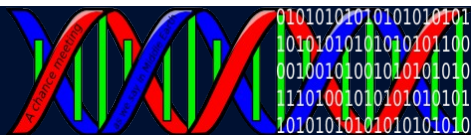
- **Alineación de secuencias:** aspira a identificar las regiones en las que se relacionan dos o más secuencias
 - Dada una instancia, la función de alineación predice un conjunto de etiquetas de salida que forman la secuencia
- **Network Analysis:** explora las relaciones entre objetos y entidades en una estructura de red, predice los links
 - Existencia o no de lado entre cada par de nodos (vector bin. salida)
- **Generación de datos:** crean como salida datos estructurados que siguen una determinada distribución
 - Datos de salida múltiples en forma de texto (palabras de vocabulario), imágenes (valores de píxeles) o sonidos (tonos de audio)

Multiooutput: subcampos de trabajo

- **Recuperación de significado semántico:** recuperación de significado de un dato de entrada, mediante etiquetas semánticas de salida que pueden usarse como ayuda
 - Salidas: vector real intermedio que se emplea para recuperar una lista de instancias similares de la base datos (emplea métodos de recuperación basados distancia)
- **Predicción de series temporales:** basándose en una serie de observaciones previas
 - Entrada: vectores de datos que comprenden un periodo de tiempo determinado
 - Salidas: vectores de datos para valores de tiempo posteriores

Multitask: subcampos de trabajo

Subfield	Output Structure	Application	Discipline
Multi-label Learning	Independent Binary Vector	Document Categorization [19]	Natural Language Processing
		Semantic Scene Classification [20]	Computer Vision
		Automatic Video Annotation [21]	Computer Vision
Multi-target Regression	Independent Real-valued Vector	River Quality Prediction [22]	Ecology
		Natural Gas Demand Forecasting [23]	Energy Meteorology
		Drug Efficacy Prediction [24]	Medicine
Label Distribution Learning	Distribution	Head Pose Estimation [25]	Computer Vision
		Facial Age Estimation [26]	Computer Vision
		Text Mining [27]	Data Mining
Label Ranking	Ranking	Text Categorization Ranking [28]	Information Retrieval
		Question Answering [29]	Information Retrieval
		Visual Object Recognition [30]	Computer Vision
Sequence Alignment Learning	Sequence	Protein Function Prediction [31]	Bioinformatics
		Language Translation [32]	Natural Language Processing
		Named Entity Recognition [33]	Natural Language Processing
Network Analysis	Graph	Scene Graph [34]	Computer Vision
	Tree	Natural Language Parsing [35]	Natural Language Processing
	Link	Link Prediction [36]	Data Mining
Data Generation	Image	Super-resolution Image Reconstruction [37]	Computer Vision
	Text	Language Generation	Natural Language Processing
	Audio	Music Generation [38]	Signal Processing
Semantic Retrieval	Independent Real-valued Vector	Content-based Image Retrieval [39]	Computer Vision
		Microblog Retrieval [40]	Data Mining
		News Retrieval [41]	Data Mining
Time-series Prediction	Time Series	DNA Microarray Data Analysis [42]	Bioinformatics
		Energy Consumption Forecasting [43]	Energy Meteorology
		Video Surveillance [44]	Computer Vision



[10] Xu D. et al. "Survey on Multi-Output Learning," IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 7, pp. 2409-2429 (2020)

Generalizando: multioutput

- ¿Es fácil obtener las salidas?
- Boom de las salidas
- Subcampos de trabajo
- **Nuevas métricas de evaluación**

Multiooutput: nuevas métricas de evaluación

- **Clasificación multilabel:** medidas basadas en instancias y basadas en etiquetas, que ya hemos estudiado
- **Regresión:** las métricas habituales de regresión que se extienden a multiooutput haciendo la media sobre todas las salidas
 - Mean Absolute Error (MAE), Mean Squared Error (MSE), ... etc

Multiooutput: nuevas métricas de evaluación

- **Nuevas métricas específicas multiooutput:** el subcampo de la Generación de datos se evalúa habitualmente según:
 - 1) Si los datos generados siguen la distribución deseada
 - Average log-likelihood, coverage, maximum mean discrepancy (MMD), geometry score
 - 2) La calidad de las instancias generadas
 - Inception scores (IS), mode score (MS), Frchet inception distance (FID) y kernel inception distance (KID)
- Además, se emplean precision, recall y F1 score para medir el overfitting

¡Muchas gracias por vuestra atención!

□ ¿Alguna pregunta?

