QUÉ ES DATAMINING?

CECILIA RUZ

RUZ.CECILIA@GMAIL.COM

AGENDA

- Qué es Data Mining?
- Cómo se integra en el proceso de Descubrimiento del conocimiento?
- Funcionalidades del Data Mining
- Técnicas
  - Supervisadas
    - Redes neuronales
    - Árboles
    - Regresión
  - No supervisadas
    - Clustering
    - Reglas de Asociación

AGENDA

- **Qué es Data Mining?**
- Cómo se integra en el proceso de Descubrimiento del conocimiento?
- Funcionalidades del Data Mining
- Técnicas
  - Supervisadas
    - Redes neuronales
    - Árboles
    - Regresión
  - No supervisadas
    - Clustering
    - Reglas de Asociación

QUÉ ES DATA MINING?

"Es la extracción de patrones o información interesante (no trivial, implícita, previamente desconocida y potencialmente útil) de (grandes) bases de datos"

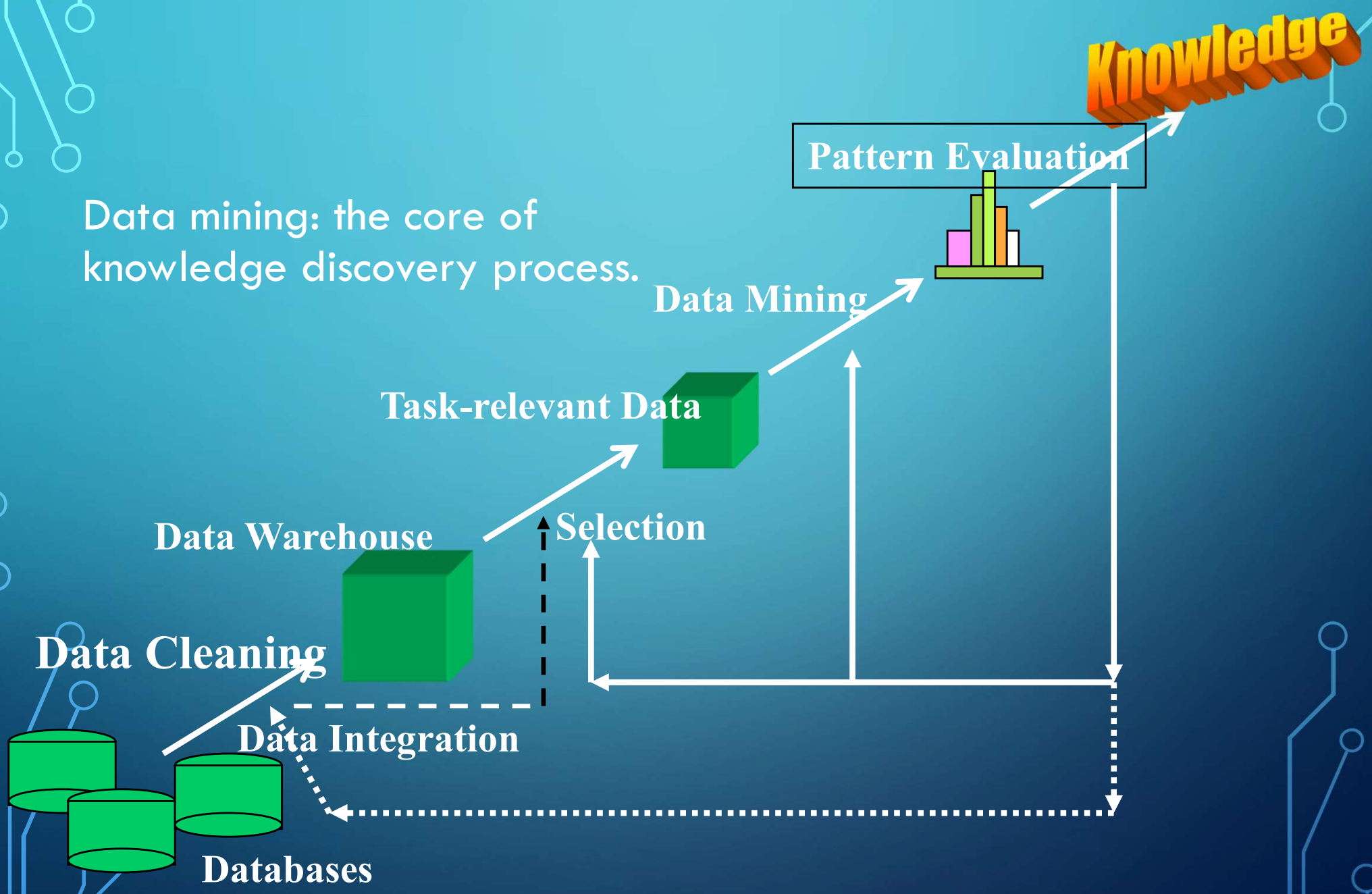
- Esta definición tiene numerosas cosas a definir:
  - ¿Qué quiere decir no-trivial?
  - ¿Útiles para quién?

AGENDA

- Qué es Data Mining?
- **Cómo se integra en el proceso de Descubrimiento del conocimiento?**
- Funcionalidades del Data Mining
- Técnicas
  - Supervisadas
    - Redes neuronales
    - Árboles
    - Regresión
  - No supervisadas
    - Clustering
    - Reglas de Asociación

DATA MINING: A KDD PROCESS

Data mining: the core of knowledge discovery process.



AGENDA

- Qué es Data Mining?
- Cómo se integra en el proceso de Descubrimiento del conocimiento?
- **Funcionalidades del Data Mining**
- Técnicas
  - Supervisadas
    - Árboles
    - Redes neuronales
    - Regresión
  - No supervisadas
    - Clustering
    - Reglas de Asociación

FUNCIONALIDADES DEL DM (1)

- Descripción de conceptos: Caracterización y discriminación

  Generalizar, resumir y contrastar las características de la información (por ejemplo las regiones secas vs. Las regiones húmedas)

- (Reglas) Asociación (correlación y causalidad)
  - Multi-dimensionales vs. única dimensión
  - age(X, “20..29”) ^ income(X, “20..29K”) → buys(X, “PC”) [support = 2%, confidence = 60%]
  - contains(T, “computer”) → contains(x, “software”) [1%, 75%]

FUNCIONALIDADES DEL DM (2)

- Classificación y Predicción
  - Encontrar modelos o funciones que describan y distingan clases para futuras predicciones

  Este posteo en Facebook, ¿corresponde a una noticia *falsa*, intento de *manipulación*, *cyberbullying*, trolling, o es inocente?
  - Enfoques: árboles de decisión, reglas de clasificación, redes neuronales
  - Predicción (inferencia): Predecir valores numéricos desconocidos o faltantes.

¿Cuánto tendría que pagar de seguro si me *compro* un auto 0Km?

FUNCIONALIDADES DEL DM (2)

- Cluster analisis
  - No se sabe a que clase pertenecen los datos : se agrupan datos para formar clases,
  - El Clustering se basa en el principio de maximizar la similitud dentro de la clase y minimizar la misma entre clases
- Análisis de Outliers
  - Outlier: un dato (o un objeto) que no respeta el comportamiento general.
  - Puede ser ruido o excepciones, pero son muy útiles en la detección de fraudos o eventos raros.

FUNCIONALIDADES DEL DM(3)

- Análisis de tendencias y evolución
  - Tendencia y Desvíos: análisis de regresión
  - Análisos de patrones secuenciales
  - Análisis de similitudes
- Otros análisis estadísticos o de patrones

AGENDA

- Qué es Data Mining?
- Cómo se integra en el proceso de Descubrimiento del conocimiento?
- Funcionalidades del Data Mining
- Técnicas
  - **Supervisadas**
    - **Redes neuronales**
    - Árboles
    - Regresión
  - No supervisadas
    - Clustering
    - Reglas de Asociación

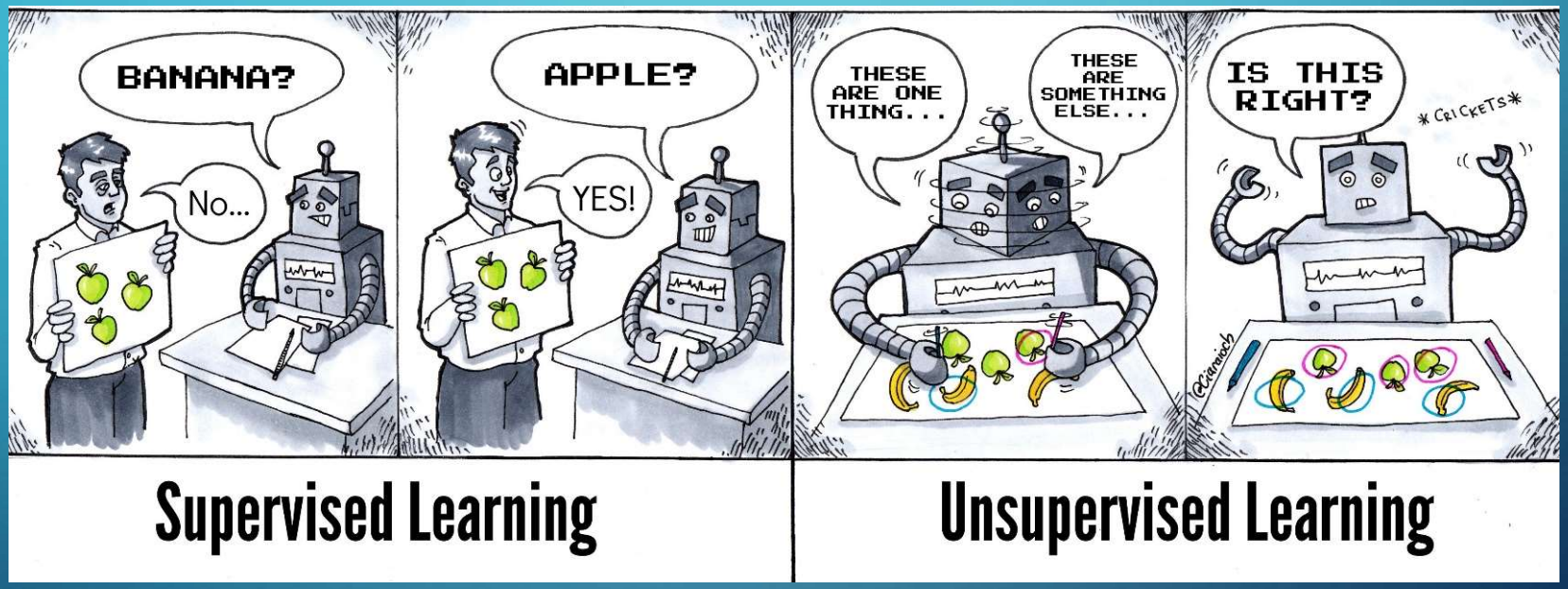
MÉTODOS SUPERVISADOS

- Los datos de entrada vienen *etiquetados* con sus correspondientes salidas.
- Se llama "supervisado" porque simula que un "tutor" provee los resultados *correctos*.
- La meta es *aprender un modelo* que produce salidas a partir de entradas.
- Ejemplos:
  - Fotos de gatos
  - Transacciones fraudulentas
  - Posteos maliciosos en Facebook

MÉTODOS NO SUPERVISADOS

- Los datos de entrada *no vienen* asociados con resultados.
- Los algoritmos deben encontrar *por sí solos* la estructura subyacente en los datos de entrada.
- La meta puede ser encontrar esta *estructura* en sí, o como *paso intermedio* para otros procesos (como aprendizaje representacional).

SUPERVISADOS VRS NO SUPERVISADOS



AGENDA

- Qué es Data Mining?
- Cómo se integra en el proceso de Descubrimiento del conocimiento?
- Funcionalidades del Data Mining
- Técnicas
  - Supervisadas
    - **Redes neuronales**
    - Árboles
    - Regresión
  - No supervisadas
    - Clustering
    - Reglas de Asociación

REDES NEURONALES (1)

Son sistemas:

- Capaces de aprender patrones sobre datos
- Adaptarse a condiciones variantes
- Adaptarse al ruido
- Predecir el estado futuro
- Enfrentar problemas que eran resueltos sólo por el cerebro humano

REDES NEURONALES (2)

No son algorítmicas:

- No se programan haciéndoles seguir una secuencia predefinida de instrucciones.
- Las RNA generan ellas mismas sus propias "reglas", para asociar la respuesta a su entrada (quedan implícitas en el modelo, no tenemos acceso a esas reglas);
- Aprenden por ejemplos y de sus propios errores.
- Utilizan un procesamiento paralelo mediante un gran numero de elementos altamente interconectados.

REDES NEURONALES – APLICACIONES

La clase de problemas que mejor se resuelven con las redes neuronales son los mismos que el ser humano resuelve mejor pero a gran escala.

- Asociación,
- Evaluación
- Reconocimiento de Patrones.

Las redes neuronales son ideales para problemas que son muy difíciles de calcular

- No requieren de respuestas perfectas,
- Sólo respuestas rápidas y buenas.

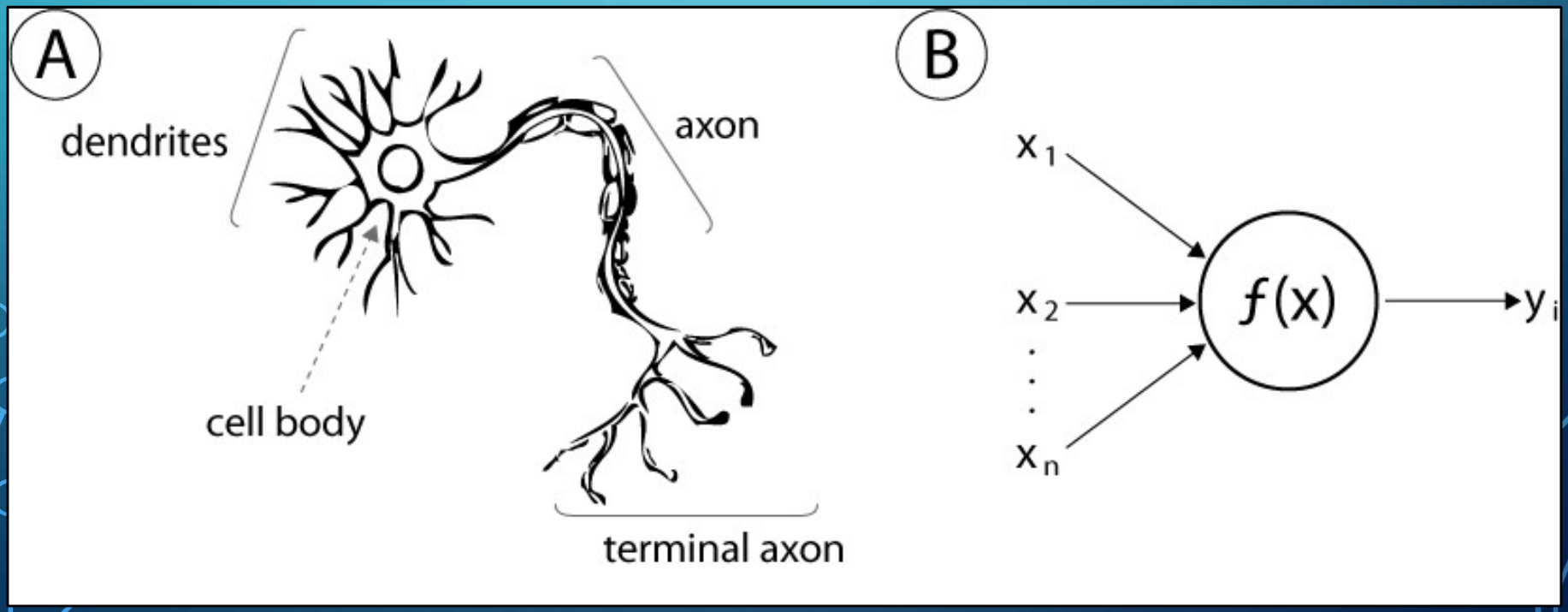
Ejemplos

- Escenario bursátil: ¿Compro? ¿Vendo? ¿Mantengo?
- Reconocimiento: ¿se parece? ¿es lo mismo con una modificación?

REDES NEURONALES

Inspiradas (lejanamente) en las *neuronas biológicas*.

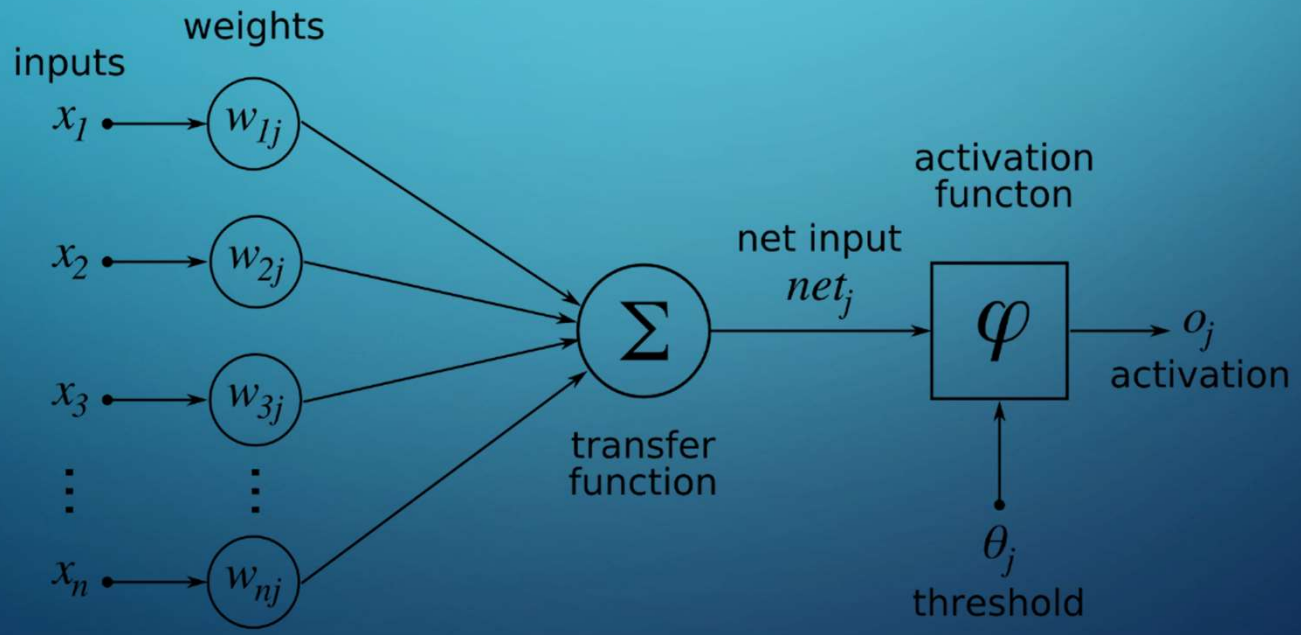
La unidad básica es la neurona *artificial*:



REDES NEURONALES: UNIDAD BÁSICA

Neurona artificial: *Función* construida al multiplicar cada *entrada* por un *peso*, y combinar los resultados con una función de transferencia.

La función de *activación* y el *umbral* determinan si la neurona se *activa* o *no* en este caso.



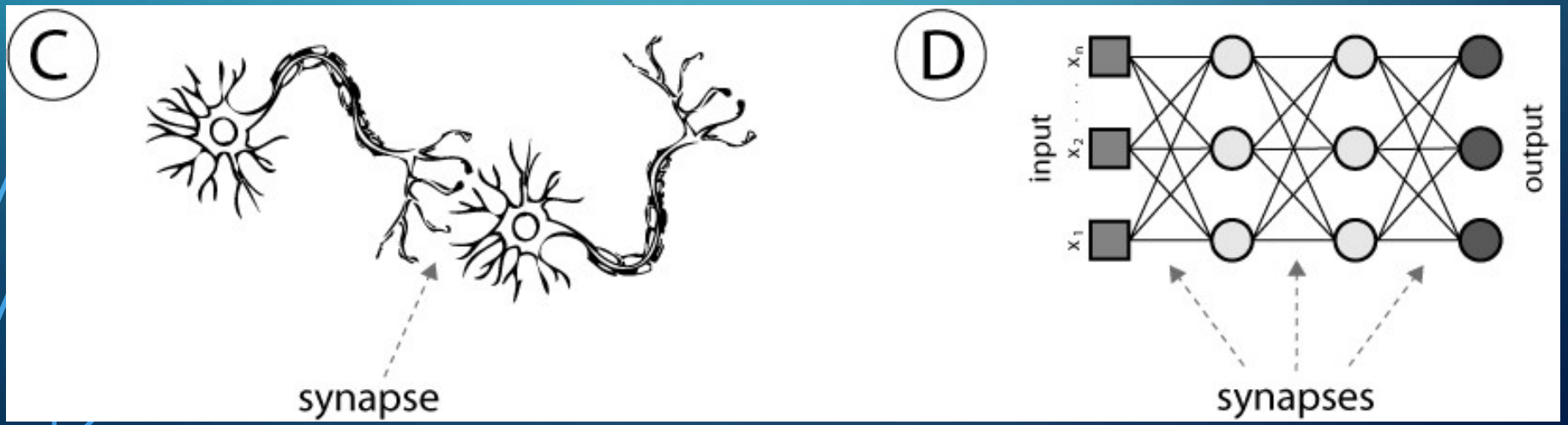
Fuente: www.packetpub.com
Fuente: https://medium.com/@ivanliljeqvist/

REDES NEURONALES: CONEXIONES ("SINAPSIS")

La *combinación* de estas funciones resulta en una *red neuronal*: las salidas de unas neuronas son entradas de otras.

Cada *capa* tiene *pesos* asociados (parámetros).

El aprendizaje es el proceso de *ajuste* de parámetros e hiperparámetros (tipos de función, topología, etc.).



REDES NEURONALES - APRENDIZAJE

Regla Delta Generalizada o Back Propagation

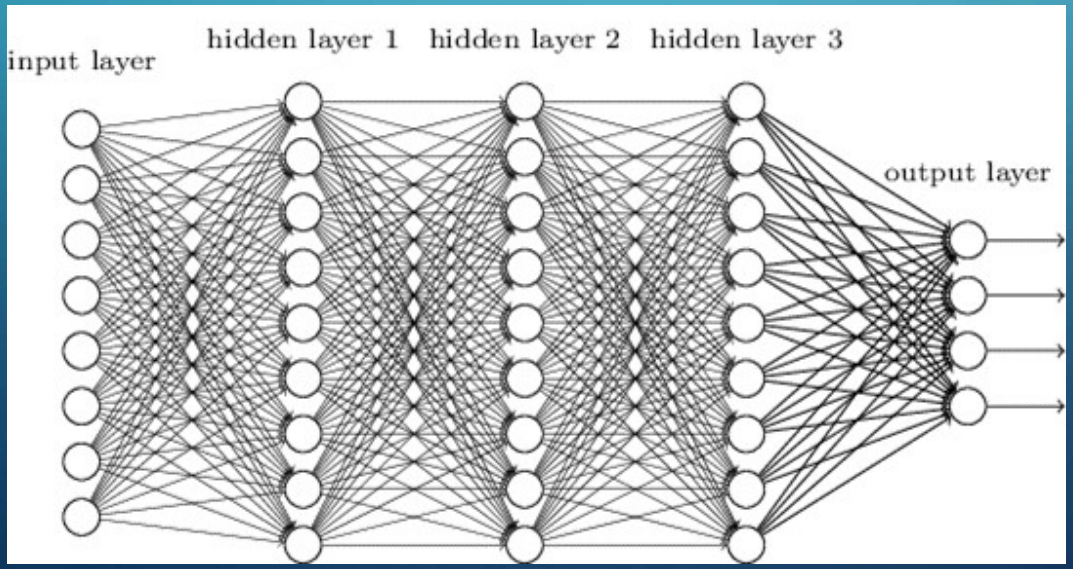
Para que una RNA aprenda o se Entrene se deben hacer pasar a todos los valores de entrenamiento por el siguiente proceso, según la topología de la red este ciclo puede repetirse varias veces y con los datos en diferente orden.

- Calcular la diferencia de la salida con la esperada
- Corregir los valores de los W que intervienen en esa salida de modo que se achique esa diferencia
- Se utiliza una constante muy pequeña (Delta)
- No se busca que la diferencia tienda a cero sino que se minimice de a poco
- Si la constante es muy grande o se minimiza la diferencia muy de golpe se corre el riesgo de que cada vez que se aprende algo nuevo se modifique demasiado lo que aprendió anteriormente.

REDES NEURONALES: PROFUNDAS (DEEP)

El adjetivo “profundo” (o “deep”) aplicado a muchas herramientas de IA hace referencia a la existencia de una RN con *muchas capas*.

Desafío: La *cantidad de parámetros* crece con cada capa, y se torna muy costoso computacionalmente explorar el espacio.



Fuente: www.packetpub.com
Fuente: https://medium.com/@ivanliljeqvist/

REDES NEURONALES: CONVOLUCIONALES

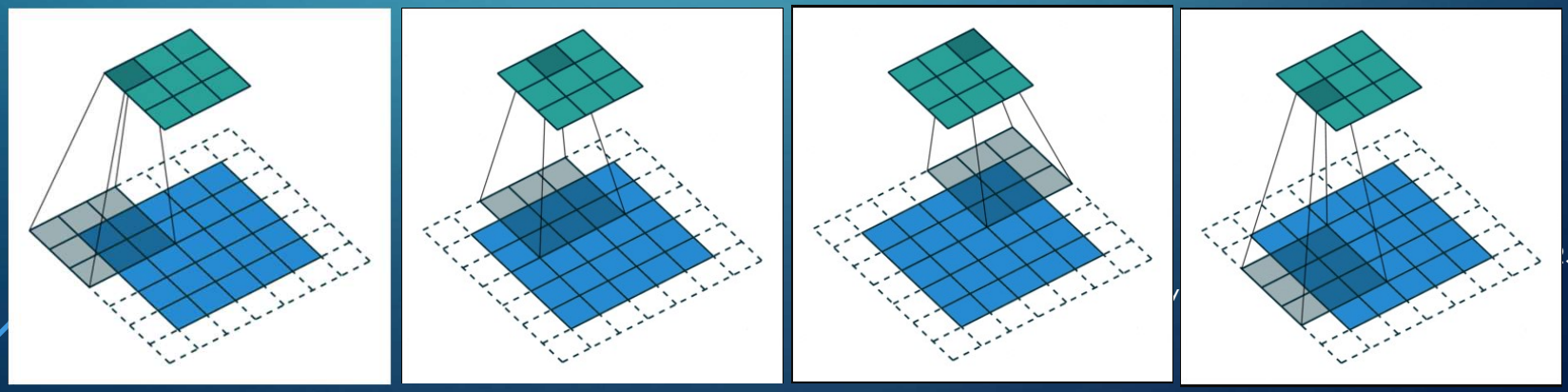
Tipo particular de RN en la cual al menos una de las capas realiza una *operación matemática* llamada "convolución".

El objetivo es obtener características (*features*) de más *alto nivel*.

En general resultan útiles en procesamiento de *imágenes*, por ejemplo para *detectar bordes*:

Las RNC filtran las imágenes antes de entrar a la RNN, extrayendo *features* de interés.

Fuente: https://medium.com/@ivanliljeqvist/



REDES NEURONALES - FALLAS

Las RNA no son buenas para:

- Cálculos precisos,
- Procesamiento en serie,
- Reconocer nada que no tenga inherentemente algún tipo de patrón.

REDES NEURONALES (3)

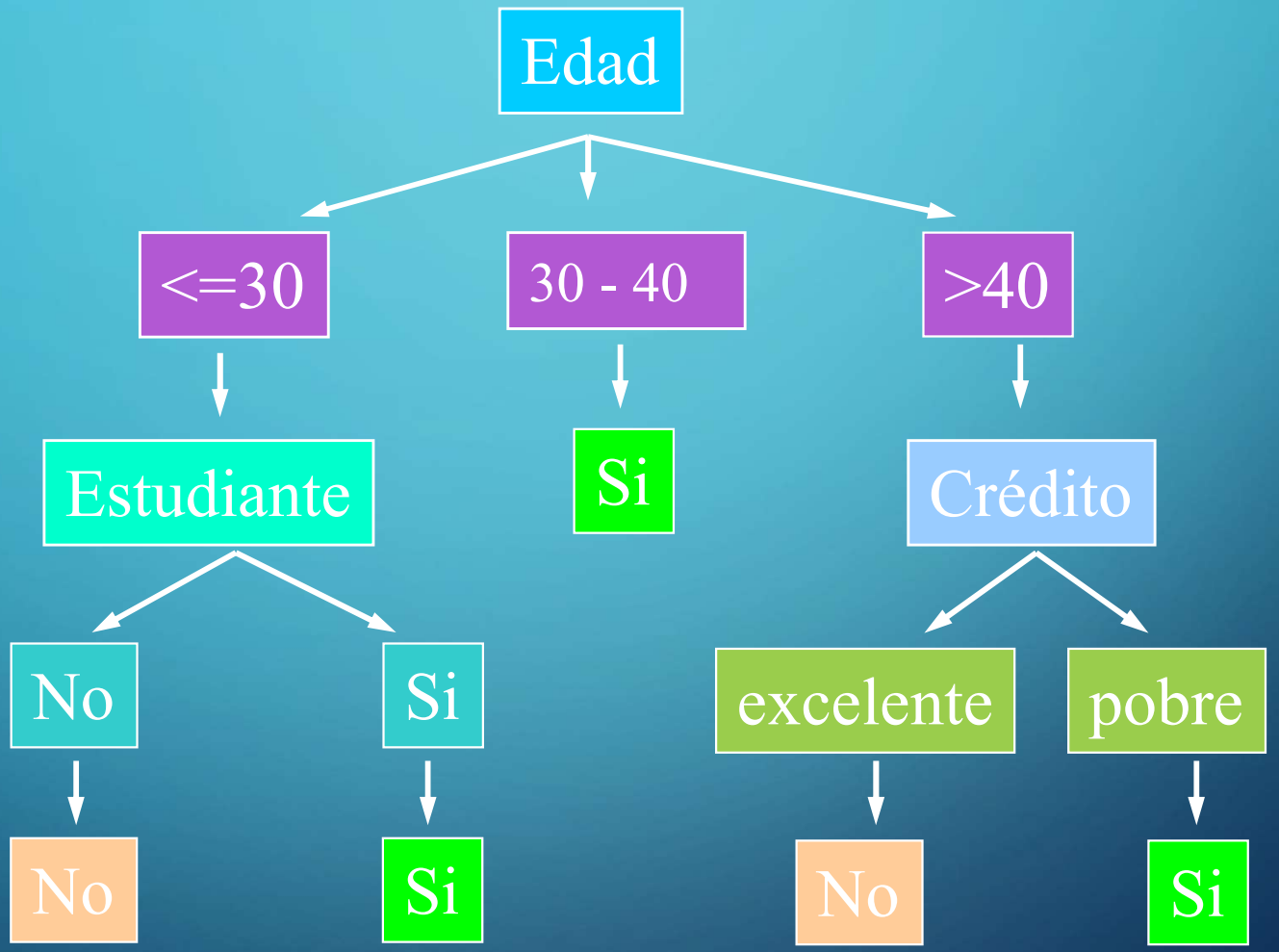
Para mejorar su performance las RNA pueden ser combinadas con otras herramientas

- Lógica Difusa (Fuzzy Logic)
- Algoritmos Genéticos
- Sistemas expertos (basados en reglas)
- Estadísticas
- Transformadas de Fourier
- Wavelets.

AGENDA

- Qué es Data Mining?
- Cómo se integra en el proceso de Descubrimiento del conocimiento?
- Funcionalidades del Data Mining
- Técnicas
  - Supervisadas
    - Redes neuronales
    - **Árboles**
    - Regresión
  - No supervisadas
    - Clustering
    - Reglas de Asociación

ÁRBOL DE DECISIÓN PARA VER QUIEN COMPRA UNA COMPUTADORA



CLASIFICACIÓN POR MEDIO DE ÁRBOLES DE DECISIÓN

- Árboles de Decisión
  - Los nodos internos son preguntas sobre los atributos
  - Las hojas representan las etiquetas o clases resultantes

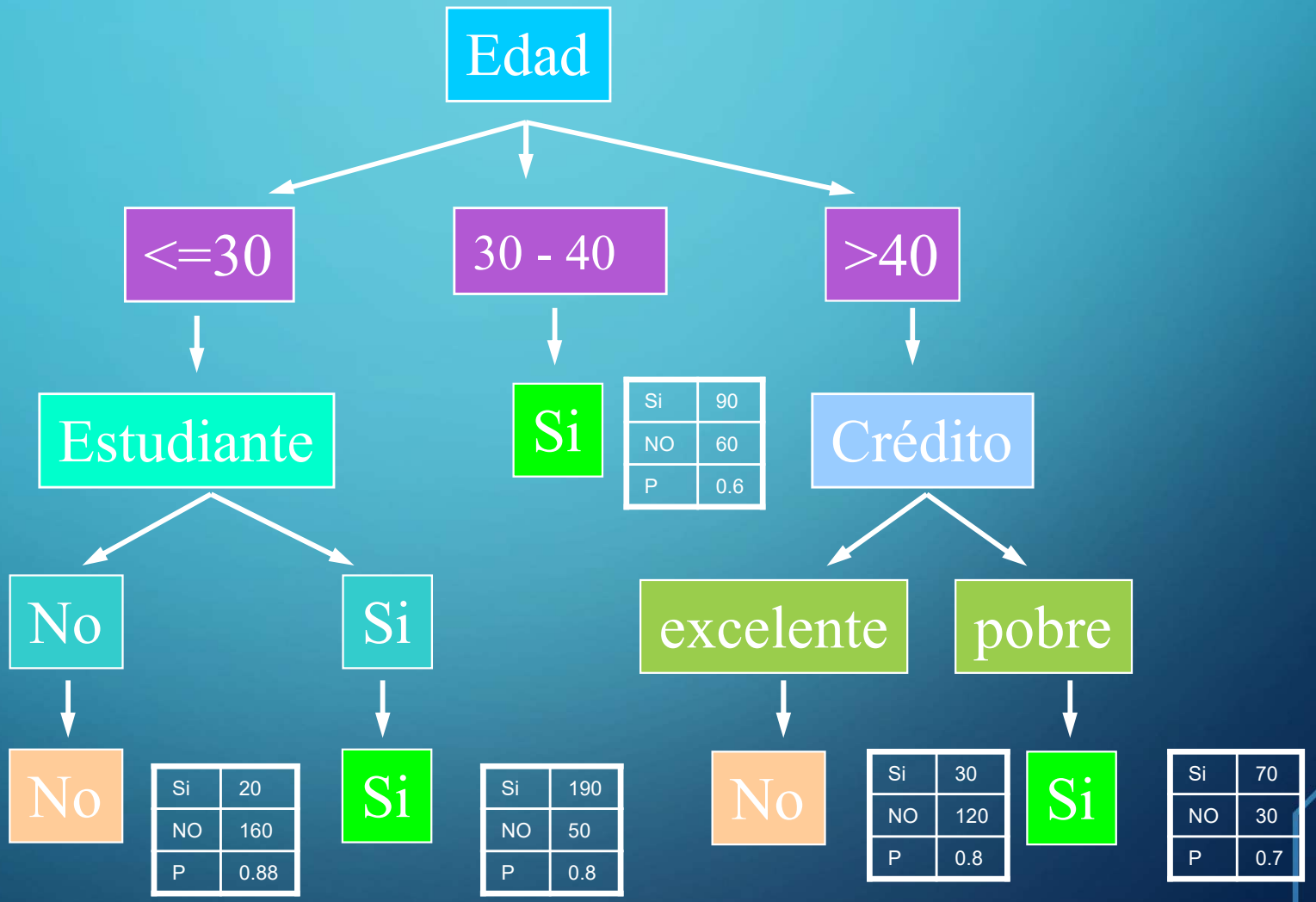
- La generación del árbol tiene fundamentalmente dos pasos
  - Construcción:
    - Al comienzo todos los ejemplos están en la raíz del árbol
    - Se dividen los ejemplos en forma recursiva basado en atributos elegidos
  - Prunning
    - Identificar y remover ramas que representan outliers o ruido

ÁRBOLES DE DECISIÓN

Clasificación de un ejemplo desconocido: se controlan los valores de los atributos del ejemplo para asignarle la clase

- Ventajas:
  - Fáciles de *entender* e *interpretar* por personas.
  - Capaces de manejar datos *numéricos* y *categóricos*.
  - Utilizan menor cantidad de *suposiciones* sobre los datos, y éstos requieren menos *preparación* que para otros modelos.
  - Estrechamente vinculados con las *reglas de asociación*, que a su vez son cercanas a modelos de *razonamiento* más complejo.

ÁRBOL DE DECISIÓN CON PROBABILIDAD



EXTRACIÓN DE REGLAS DE CLASIFICACIÓN A PARTIR DE LOS ÁRBOLES

- Representa el conocimiento en la forma de reglas de IF-THEN
- Se genera una regla para cada camino desde la raíz hasta las hojas.
- Cada par atributo – valor forma una conjunción
- La hoja tiene la clase a predecir
- Las reglas son fácilmente entendibles por los seres humanos
- Ejemplos

```
IF edad = "<=30" AND estudiante = "no"   THEN
  compra_PC = "no"
IF edad = "<=30" AND estudiante = "yes"  THEN
  compra_PC = "si"
IF edad = "31 - 40" THEN compra_PC = "si"
IF edad = ">40"   AND credito = "excelente"   THEN
  compra_PC = "si"
IF edad = ">40" AND credito = "pobre"  THEN
  compra_PC = "no"
```

OVERFITTING

- El árbol obtenido puede hacer overfitting sobre el conjunto de entrenamiento
  - Si hay demasiadas ramas algunas pueden reflejar anomalías
  - Como consecuencia de esto se tiene una performance muy mala sobre ejemplos nuevos

- Dos aproximaciones para evitar el overfitting
  - Prepruning: Interrumpir la construcción del arbol en forma anticipada. No partir un nodo si la mejora que esto produce está por debajo de un cierto umbral.
    - Es dificil encontrar el umbral adecuado
  - Postpruning: quitar ramas de un árbol ya contruido
    - Se puede usar un conjunto diferente del de entrenamiento para hacer esto.

AGENDA

- Qué es Data Mining?
- Cómo se integra en el proceso de Descubrimiento del conocimiento?
- Funcionalidades del Data Mining
- Técnicas
  - Supervisadas
    - Redes neuronales
    - Árboles
    - **Regresión**
  - No supervisadas
    - *Clustering*
    - Reglas de Asociación

REGRESIÓN LINEAL

Para poder crear un modelo de regresión lineal, es necesario que se cumpla con los siguientes supuestos:

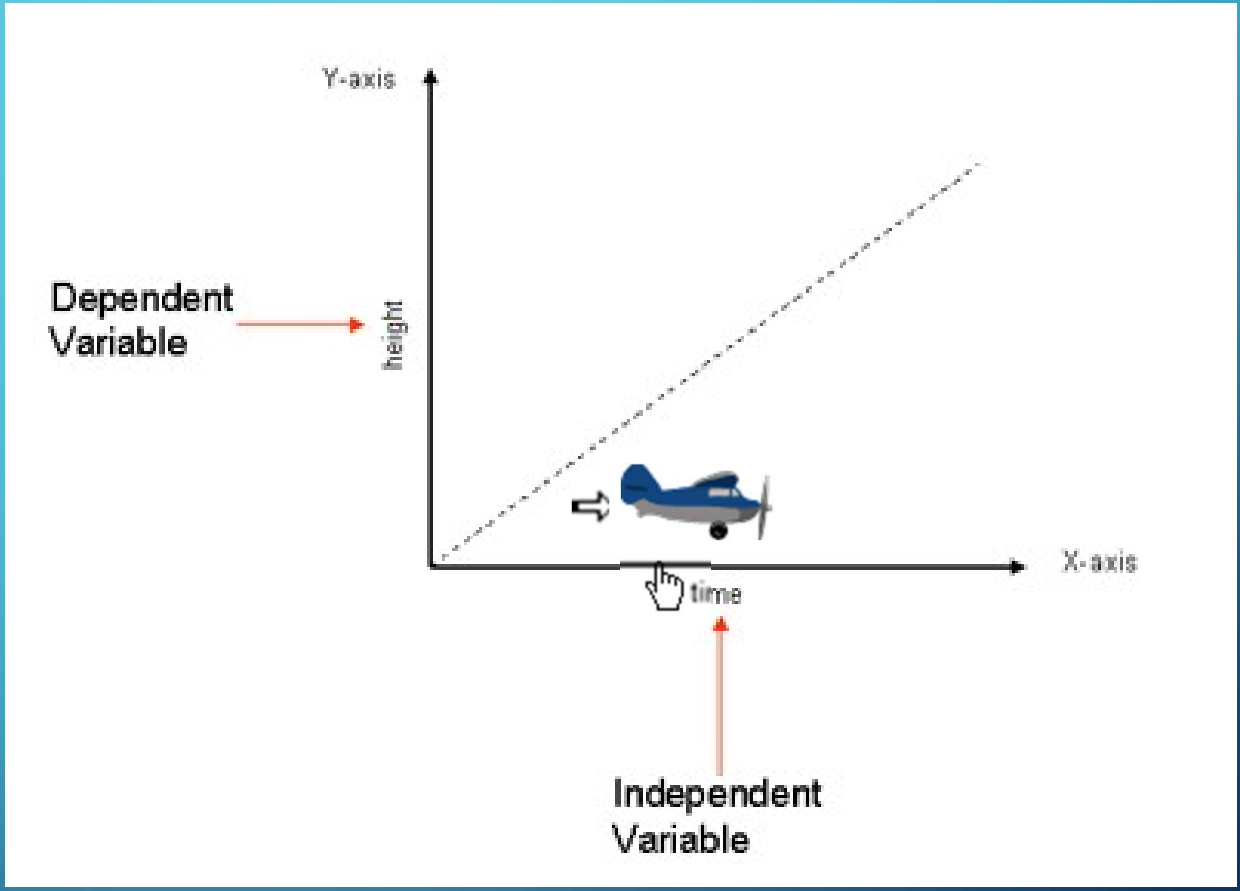
- La relación entre las variables es lineal.
- Los errores son independientes.
- Los errores tienen varianza constante.
- Los errores tienen una esperanza matemática igual a cero.
- El error total es la suma de todos los errores.

TIPOS DE REGRESIÓN LINEAL

- Regresión lineal simple. Sólo se maneja una variable independiente
- Regresión lineal múltiple. Maneja varias variables independientes.

REGRESIÓN LINEAL EJEMPLO

- **Variables dependientes:** Son las variables de respuesta que se observan en el estudio y que podrían estar influidas por los valores de las variables independientes.



Variables independientes: Son las que se toman para establecer agrupaciones en el estudio, clasificando intrínsecamente a los casos del mismo

REGRESIÓN LOGÍSTICA

- La regresión logística se aplica cuando la variable dependiente es dicotómica o politómica y no numérica.
- Para poder aplicar una regresión se asocia la variable dependiente a su probabilidad de ocurrencia.
- Por lo tanto el resultado de un regresión logística es la probabilidad de ocurrencia del suceso

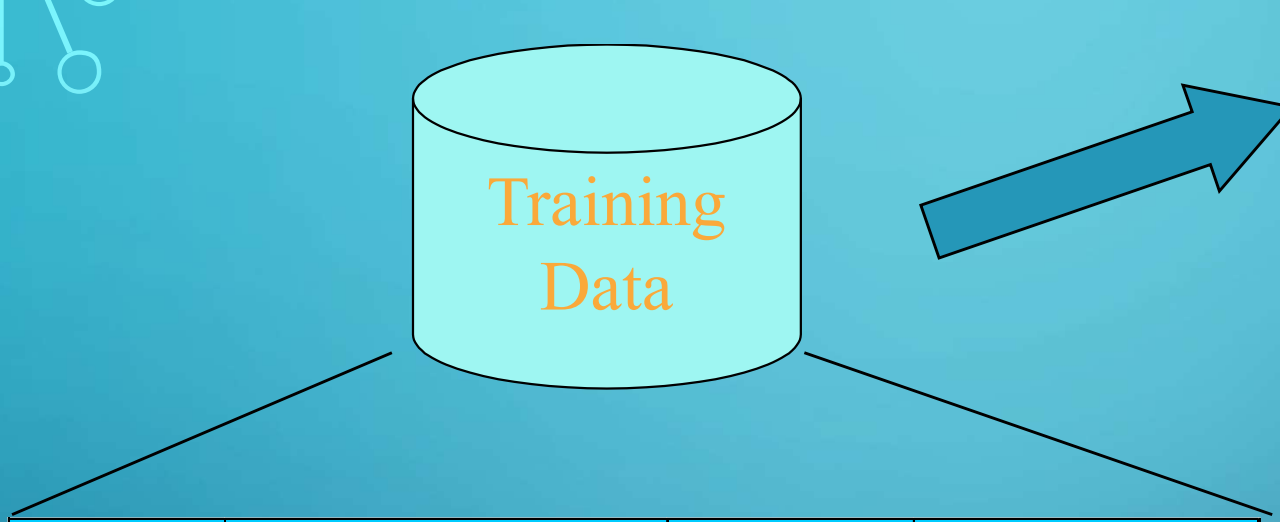
CLASIFICACIÓN—UN PROCESO DE DOS PASOS

- Construcción del modelo: descripción de las clases existentes
  - Cada ejemplo pertenece a una clase determinada
  - El training set es el conjunto de ejemplos que se usa para entrenar el modelo
  - El modelo se representa por medio de reglas de clasificación, árboles o fórmulas matemáticas

CLASIFICACIÓN—UN PROCESO DE DOS PASOS

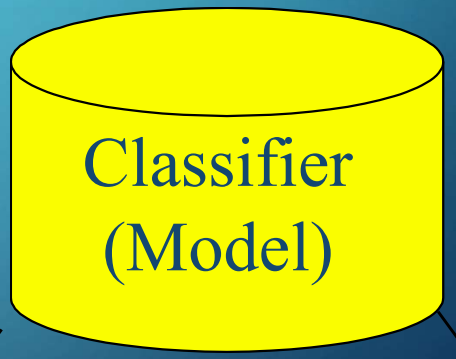
- Uso del modelo: para clasificar ejemplos futuros o desconocidos
  - Estimar la precisión del modelo
    - Para esto se aplica el modelo sobre un conjunto de test y se compara el resultado del algoritmo con el real.
    - **Precisión** es el porcentaje de casos de prueba que son correctamente clasificados por el modelo
    - El conjunto de entrenamiento debe ser independiente del de test para evitar "overfitting"

PROCESO DE CLASIFICACION (1): CONSTRUCCIÓN DEL MODELO



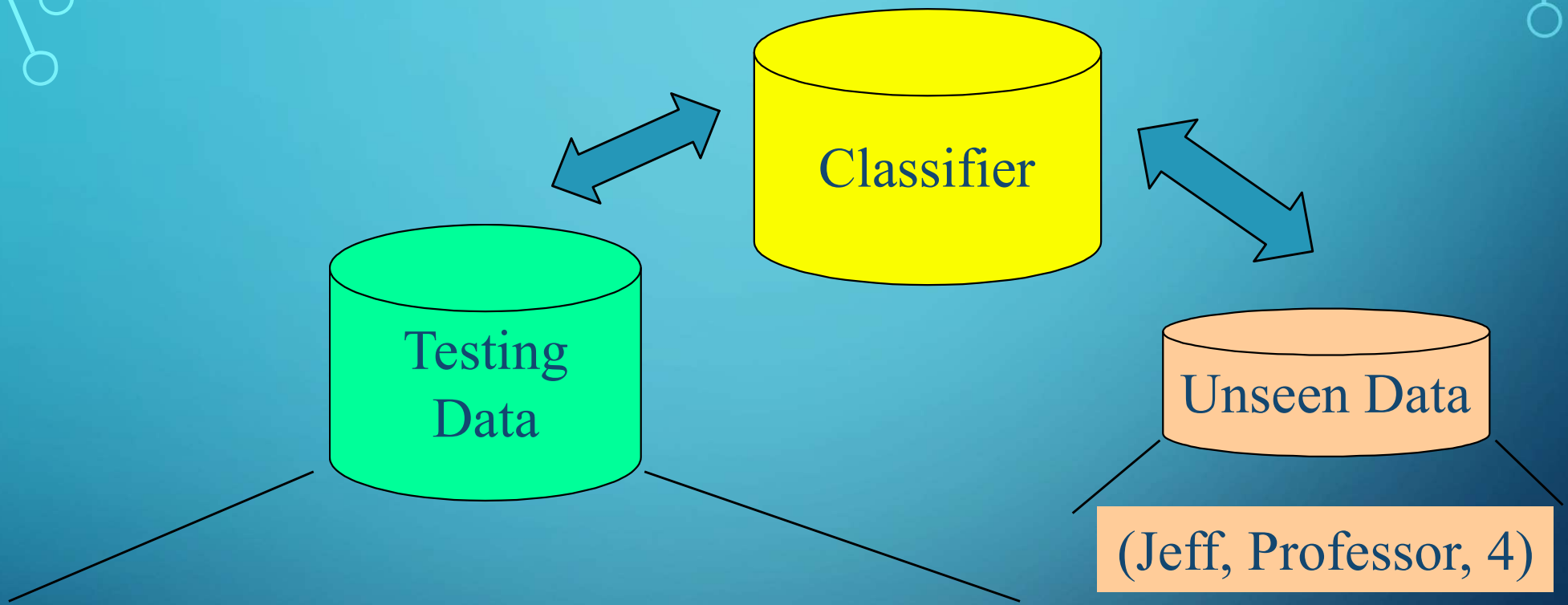
NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Classification Algorithms



IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

PROCESO DE CLASIFICACION (2): USO DEL MODELO PARA PREDECIR



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

EVALUACION DE PROCESOS DE CLASIFICACION

- Cuando se utilizan procesos de clasificación es necesario dividir el conjunto de datos de los que se conoce el resultado en dos conjuntos disjuntos: una de entrenamiento y una de test.
- La parte de entrenamiento es la que el algoritmo va a "ver" para armar su modelo y la de test es la que se va a usar para evaluar su performance.
- Uno de los elementos mas utilizados para evaluar la perfomance de un modelo es la matriz de confusión.

MATRIZ DE CONFUSIÓN 1 / 2

		Clase Real	
		T	F
Clase predicha	T	10	2
	F	5	15

- Esta matriz muestra que el algoritmo predijo acertadamente 25 casos de un total de 32. Esta medida se llama **acuraccy**
- Se debe armar una matriz de confusión para el set de entrenamiento y otra para el set de test. Si la diferencia de la tasa de error entre ambas es importante seguramente se produjo overfitting

MATRIZ DE CONFUSIÓN 2/2

		Clase Real	
		T	F
Clase predicha	T	TP : True positive	FP: False positive (error de tipo I)
	F	FN : Falsos negativos (error de tipo 2)	TN : True negative

- Con estos valores se pueden calcular varias métricas
- TPR (True Positive Rate) , también llamada sensibilidad, se calcula como TP/(TP + FN) . Representa el porcentaje de los casos positivos predichos
- TNR (True Negative Rate) , especificidad, TN / (TN + FP) , representa el porcentaje de los casos negativos predichos
- **Precision** , es el porcentaje de los positivos predichos que son correctos, TP / (TP + FP)

DETECCIÓN DE VALORES EXTREMOS, OUTLIERS

Los conjuntos de datos que analizamos generalmente proporcionan un subconjunto de datos en el que existe una variabilidad y/o una serie de errores.

Estos datos siguen un comportamiento diferente al resto del conjunto ya sea en una o varias variables. Muchas veces es útil estudiarlos para detectar anormalidades, mientras que otras veces es mejor descartarlos de los análisis porque ensucian o influyen en los resultados (por ejemplo en los promedios).

ORÍGENES DE LA VARIACIÓN

Variabilidad de la fuente. Es la que se manifiesta en las observaciones y que se puede considerar como un comportamiento natural de la población en relación a la variable que se estudia.

Errores del medio. Son los que se originan cuando no se dispone de la técnica adecuada para valorar la variable sobre la población, o cuando no existe un método para realizar dicha valoración de forma exacta. En este tipo de errores se incluyen los redondeos forzosos que se han de realizar cuando se trabaja con variables de tipo continuo.

ORÍGENES DE LA VARIACIÓN

Errores del experimentador. Son los atribuibles al experimentador, y que fundamentalmente se pueden clasificar de la siguiente forma:

- **Error de Planificación**. Se origina cuando el experimentador no delimita correctamente la población , y realiza observaciones que pueden pertenecer a una población distinta.
- **Error de Realización**. Se comete al llevar a cabo una valoración errónea de los elementos. Aquíı se incluyen, entre otros, transcripciones erróneas de los datos, falsas lecturas realizadas sobre los instrumentos de medida, etc.

DEFINICIONES

A la vista de lo anterior, podemos clasificar las observaciones atípicas o anómalas como:

- Observación **atípica**: Es aquel valor que presenta una gran variabilidad de tipo inherente.
- Observación **errónea:** Es aquel valor que se encuentra afectado de algún tipo de error, sea del medio, del experimentador, o de ambos.

Se llamará "outlier" a aquella observación que siendo atípica y/o errónea, tiene un comportamiento muy diferente respecto al resto de los datos, en relación al análisis que se desea realizar sobre las observaciones. Análogamente, se llamará "inlier" a toda observación no considerada como outlier.

AGENDA

- Qué es Data Mining?
- Cómo se integra en el proceso de Descubrimiento del conocimiento?
- Funcionalidades del Data Mining
- Técnicas
  - Supervisadas
    - Redes neuronales
    - Árboles
    - Regresión
  - No supervisadas
    - **Clustering**
    - Reglas de Asociación

MÉTODOS NO SUPERVISADOS

- Los datos de entrada *no vienen* asociados con resultados.
- Los algoritmos deben encontrar *por sí solos* la estructura subyacente en los datos de entrada.
- La meta puede ser encontrar esta *estructura* en sí, o como *paso intermedio* para otros procesos (como aprendizaje representacional).

QUÉ ES UN BUEN CLUSTERING?

- Un buen método de *clustering* produce *clusters* de alta calidad con
  - Alta similitud en la clase
  - Baja similitud entre clases
- La calidad de un *clustering* depende de la medida de "similitud" usada por el método y de la forma en que está implementado.

MEDICIÓN DE LA CALIDAD DE UN CLUSTER

- Medida de similitud: La similitud está expresada en base a una función de distancia
- Hay una función separada que mide la bondad del clustering
- Las funciones de distancia a utilizar son muy diferentes de acuerdo al tipo de dato.
- Algunas veces es necesario asignarle "peso" a las variables dependiendo del significado que tienen para el problema

DISTANCIAS

$$d_{ij} = \sum_{k=1}^{p} W_k \left| x_{ik} - x_{jk} \right|$$

City-Block (Manhatan)

$$d_{ij} = \sqrt{\sum_{k=1}^{p} W_k \left(x_{ik} - x_{jk} \right)^2}$$

Euclídea

$$d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{p} W_k \left(x_{ik} - x_{jk} \right)^{\lambda}} \quad \lambda > 0$$

Minkowski

Otras

$$d_{ij} = \frac{\sum_{k=1}^{p} x_{ik} \cdot x_{jk}}{\sqrt{\sum_{k=1}^{p} x_{ik}^2} \cdot \sqrt{\sum_{l=1}^{p} x_{jl}^2}}$$

$$d_{ij} = \frac{\sum_{k=1}^{p} \left(x_{ik} - \bar{x}_i \right) \cdot \left(x_{jk} - \bar{x}_j \right)}{\sqrt{\sum_{k=1}^{p} \left(x_{ik} - \bar{x}_i \right)^2} \cdot \sqrt{\sum_{l=1}^{p} \left(x_{jl} - \bar{x}_j \right)^2}}$$

DEFINICIÓN DE LA DISTANCIA: LA DISTANCIA EUCLÍDEA

D_{ij} distancia entre los casos *i* y *j*

x_{ki} valor de la variable X_k para el caso *i*

$$D_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ki} - x_{kj} \right)^2}$$

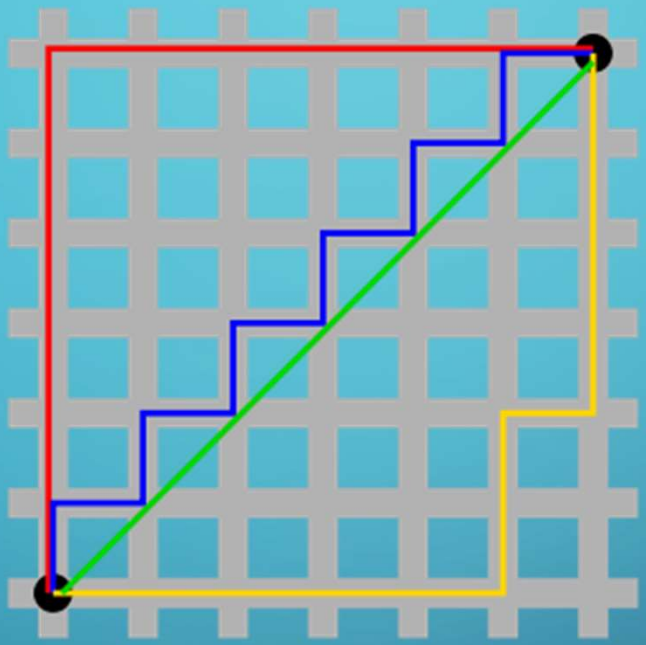
Problemas:

- Diferentes medidas = diferentes ponderaciones
- Correlación entre variables (redundancia)
- Variables faltantes (Missing Values)
- Variables de distinto tipo.
- Incompatibilidad en las Unidades de Medida

Soluciónes:

- *Análisis de Componentes Principales*
- *Normalización o Estandarización de las Variables*

MANHATTAN VERSUS EUCLIDEAN



El rojo, azul, y amarillo representan la distancia Manhattan, todas tienen el mismo largo(12),mientras que la verde representa la distancia Euclidia con largo de $6 \times \sqrt{2} \approx 8.48$.

VARIABLES NUMÉRICAS

- Estandarizar los datos
  - Calcular la desviación absoluta de la media

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

donde $m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf})$.

  - Normalizar (*z-score*)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

SIMILITUD ENTRE OBJETOS

- Las distancias se usan habitualmente para medir la similitud entre dos objetos
- Algunas de las más conocidas: distancia de *Minkowski*

$$d(i,j)=\sqrt[q]{(|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+...+|x_{ip}-x_{jp}|^q)}$$

Donde $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ y $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ son dos objetos de p dimensiones y q es un entero positivo

- Si $q = 1$, d es la distancia de Manhattan

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+...+|x_{ip}-x_{jp}|$$

SIMILITUD ENTRE OBJETOS (CONT)

- *Si q = 2, d* es la distancia euclideana:

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

- Propiedades de cualquier función de distancia
  - $d(i,j) \geq 0$
  - $d(i,i) = 0$
  - $d(i,j) = d(j,i)$
  - $d(i,j) \leq d(i,k) + d(k,j)$

VARIABLES BINARIAS

- Una tabla de contingencia

		Object j		
		1	0	*sum*
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

- Coeficiente simple

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

Coeficiente de Jaccard :

$$d(i,j) = \frac{b+c}{a+b+c}$$

VARIABLES NOMINALES

- Pueden tomar más de dos estados : estado civil
- Método1: Macheo Simple
  - *m*: # de coincidencias, *p*: # total de variables

$$d(i, j) = \frac{p - m}{p}$$

- Método 2: transformación de las variables en dummy

VARIABLES ORDINALES

- Puede ser discreta o continua, el orden es importante, por ejemplo nivel de educación
- Pueden ser tratadas como las numéricas comunes
  - Reemplazando por su lugar en el ranking

$$r_{if} \in \{1, \ldots, M_f\}$$

  - normalizar

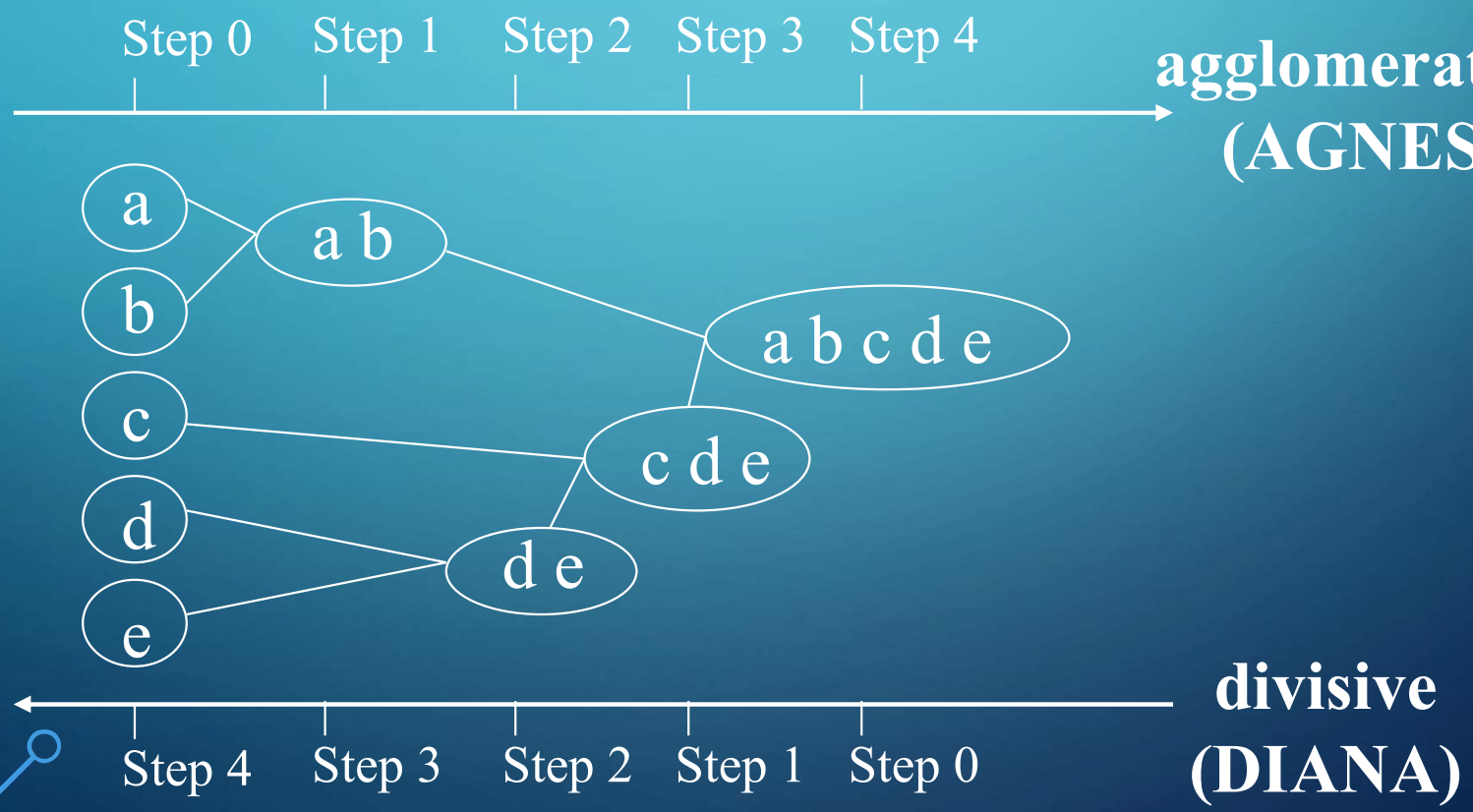
$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

FORMAS DE OBTENER UN CLUSTER

- Jerárquicas
- No jerárquicas

CLUSTERING JERÁRQUICO

- Usa la matriz de distancia como criterio. No requiere que el número de cluster sea uno de los parámetros de input



AGRUPAMIENTO AGLOMERATIVO

Criterio de enlace

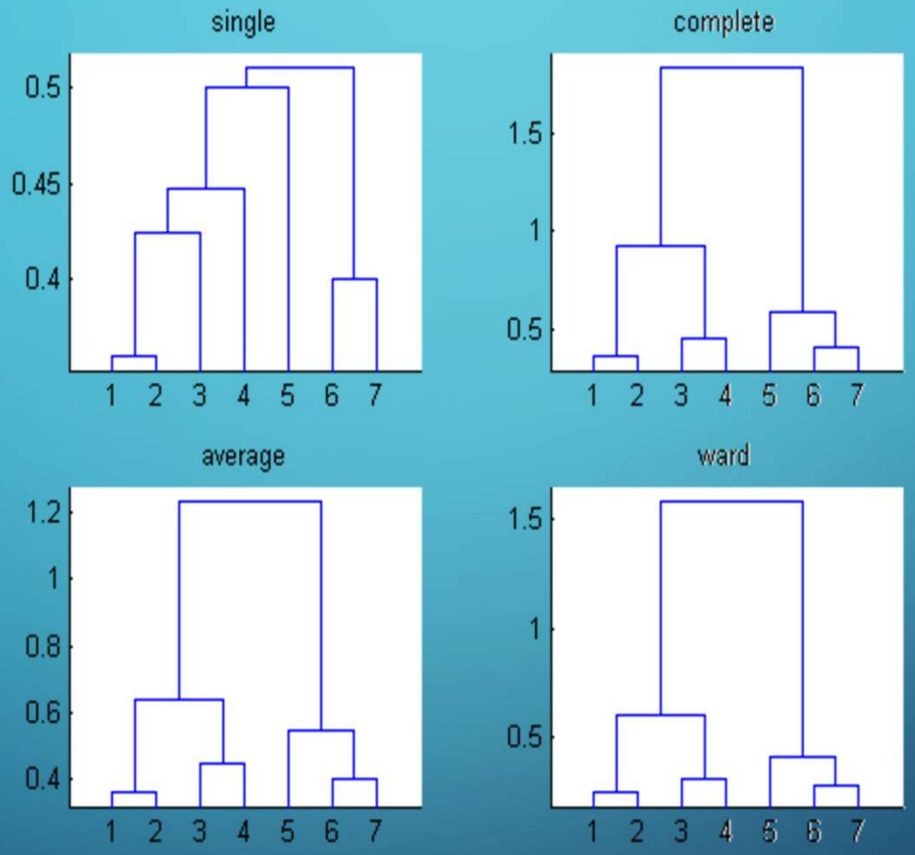
- Enlace simple (distancia mínima)
- Enlace Completo (distancia máxima)
- Enlace promedio

- **Método de Ward**

1. Calcular la suma de las distancias al cuadrado dentro de los clusters
2. Juntar clusters con incremento mínimo en la suma de cuadrados total

- **Método del centroide**
  - La distancia entre dos clusters se define como la distancia entre los centroides (medias de los cluster)

DENDROGRAMAS: OTROS MÉTODOS



NO JERÁRQUICAS: ALGORITMO BÁSICO

- Método de partionamiento: Construir una partición de la base de datos D de n objetos en k clusters
- Dado k encontrar una partición de k clusters que optimice el criterio de partición usado
  - Optimo Global: enumerar todas las particiones posibles
  - Métodos heurísticos:
    - *k-means* (MacQueen'67): cada cluster esta representado por el centro del cluster
    - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): cada cluster está representado por uno de los objetos del cluster

MÉTODOS JERARQUICOS VS NO JERARQUICOS

Agrupamiento jerarquico

- No hay decisión acerca del número de clusters
- Existen problemas cuando los datos contienen un alto nivel de error
- Puede ser muy lento

Agrupamiento no jerarquico

- Más rapido y más fiable
- Es necesario especificar el numero de clusters (arbitrario)
- Es necesario establecer la semilla inicial

AGENDA

- Qué es Data Mining?
- Cómo se integra en el proceso de Descubrimiento del conocimiento?
- Funcionalidades del Data Mining
- Técnicas
  - Supervisadas
    - Redes neuronales
    - Árboles
    - Regresión
  - No supervisadas
    - Clustering
    - **Reglas de Asociación**

PROPÓSITO DE MBA

- Generar reglas del tipo:
  - IF (SI) condición ENTONCES (THEN) resultado
- Ejemplo:
  - **Si** producto B ENTONCES producto C
- Minado de reglas de asociación:
  - "Encontrar patrones, asociaciones, correlaciones o estructuras causales frecuentes entre conjuntos de items u objetos in bases de datos transaccionales, relacionales u otro tipo de repositorios de información."

TIPOS DE REGLAS SEGÚN SU UTILIDAD

- Utiles / aplicables : reglas que contienen buena calidad de información que pueden traducirse en acciones de negocio.
- Triviales : reglas ya conocidas en el negocio por su frecuente ocurrencia
- Inexplicables : curiosidades arbitrarias sin aplicación práctica

¿CUÁN BUENA ES UNA REGLA?

- Medidas que califican a una regla:
  - Soporte
  - Confianza
  - Lift (Improvement)

EJEMPLO SOPORTE

T1 = {A, B, C, D}

T2 = {B, C}

T3 = {A, B, C}

T4 = {B, C, D}

T5 = {A, D}

T6 = {A, B}

- Es la cantidad (%) de transacciones en donde se encuentra la regla.
  - Ej : “Si B entonces C” está presente en 4 de 6 transacciones.
  - Soporte (B/C) : 66.6%

CONFIANZA

- Cantidad (%) de transacciones que contienen la regla referida a la cantidad de transacciones que contienen la cláusula condicional
  - Ej : Para el caso anterior, B está presente en 5 transacciones (83.33%)
  - Confianza (B/C) = 66.6/83.3 = 80%

T1 = {A, B, C, D}

T2 = {B, C}

T3 = {A, B, C}

T4 = {B, C, D}

T5 = {A, D}

T6 = {A, B}

MEJORA (IMPROVEMENT)

- Capacidad predictiva de la regla:
  - Mejora = p(B/C) / p(B) * p(C)
  - Ej:

p(B/C) = 0,67 ; p(B) = 0,833; p(C) = 0,67

Improv (B/C) = 0,67(0,833*0,67) = 1.2

Mayor a 1 : la regla tiene valor predictivo

T1 = {A, B, C, D}

T2 = {B, C}

T3 = {A, B, C}

T4 = {B, C, D}

T5 = {A, D}

T6 = {A, B}

TIPOS DE REGLAS

- Booleanas o cuantitativas (de acuerdo a los valores que manejan)
  - buys(x, “SQLServer”) ^ buys(x, “DMBook”) → buys(x, “DBMiner”) [0.2%, 60%]
  - age(x, “30..39”) ^ income(x, “42..48K”) → buys(x, “PC”) [1%, 75%]
- Una dimensión o varias dimensiones
- Con manejo de jerarquías entre los elementos (taxonomías que vienen dadas) o con elementos simples

ALGUNAS TÉCNICAS/ APLICACIONES MAS NOVEDOSAS

- Text mining, análisis de texto, que puede servir para clasificar textos y también para hacer análisis de sentimientos.
- Análisis de comunidades, permite detectar grupos de individuos que están mas estrechamente vinculados que otros
- Análisis de trayectorias, un tipo particular de secuencias, como se mueven las aves?

REFERENCIAS

- Esta presentación fue hecha en base al material que acompaña al libro “Data Mining : Concepts and Techniques” de Han – Kamber
- Se basa en una presentación que hicimos junto con Gustavo Markel para la IEEE en el año 2006

REFERENCIAS

- http://www.kdnuggets.com/
- http://www.acm.org/sigkdd/
- http://www.computer.org/portal/site/transactions/tkde/content/index.jsp?pageID=tkde_home
- http://domino.research.ibm.com/comm/research.nsf/pages/r.kdd.html
- http://www.cs.waikato.ac.nz/~ml/weka/
- http://www.cs.umd.edu/users/nfa/dm_people_papers.html