Métodos de Análisis de Datos

Minería de Datos

Dra. Libertad Tansini

INTRODUCCIÓN A LA INGENIERÍA DE PRODUCCIÓN

Antecedentes

Grandes cantidades de información son producidos y almacenadas en la actualidad:

- Web Data
- E-comercio, tiendas de autoservicio, Bancos, e-banking
- Procesos industriales, sensores, Industria 4.0 (autooptimización, flexibilidad, sistemas dinámicos de producción - inteligencia sin intervención humana (o minimizarla))

La extracción de información se vuelve relevante, "oculta" y no evidente

Presión competitiva - proveer mejores servicios personalizados

Grandes computadoras bajan precios

Definición

La **Minería de Datos** es la extracción automática de información descriptiva o predictiva escondida desde bases de datos

La Minería de Datos estudia métodos y algoritmos que permiten la extracción automática de información sintetizada que permite caracterizar las relaciones escondidas

Las aplicaciones de la Minería de Datos se hacen sobre datos previamente recolectados. Los datos no cambian mientras están siendo analizados

Herramientas

Las Herramientas de la Minería de Datos:

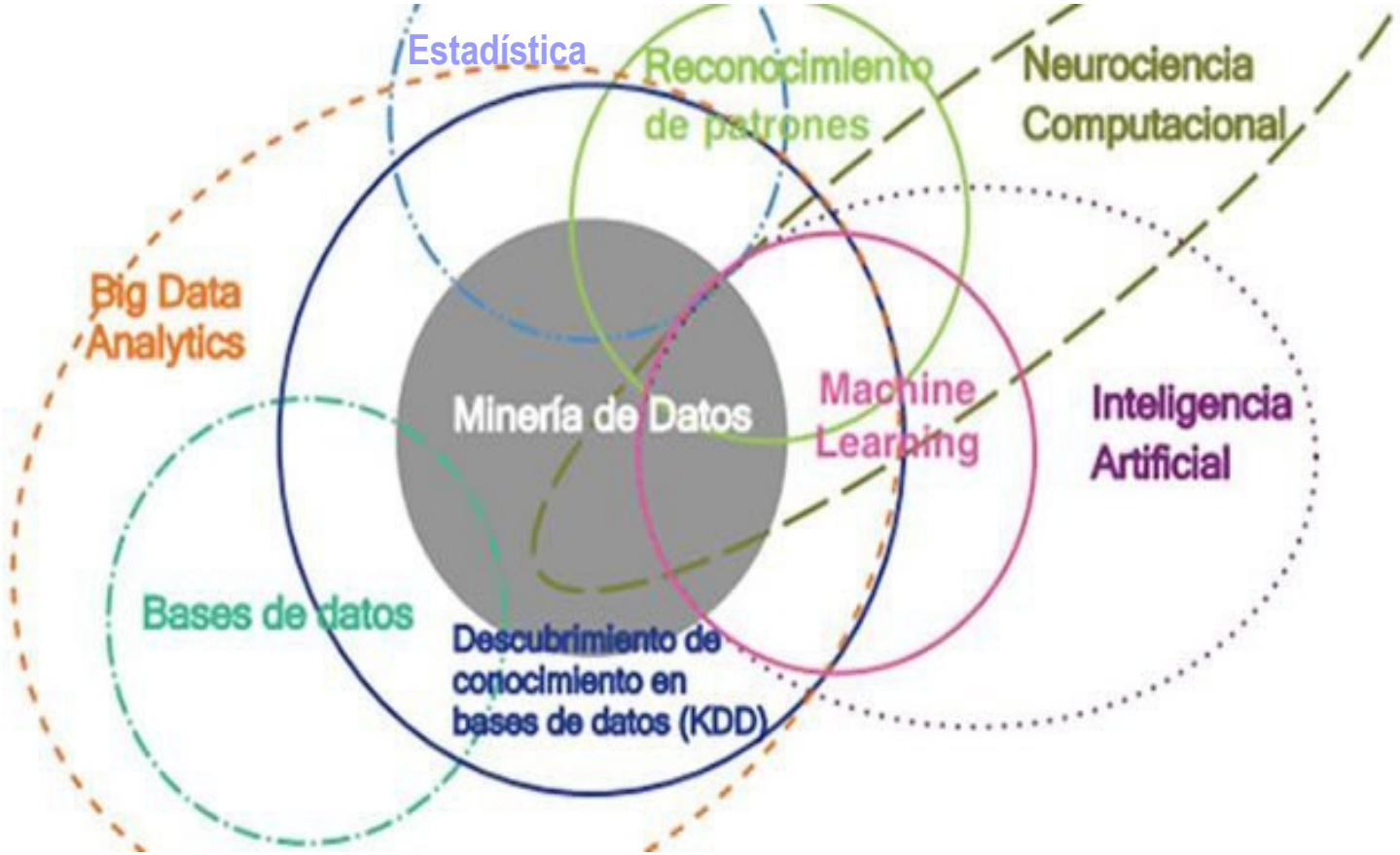
- Predicen tendencias futuras y comportamientos
- Pueden responder a preguntas que de otra forma llevaría demasiado tiempo resolver

Soportadas por tres tecnologías maduras:

- Colección masiva de datos
- Computadoras con multiprocesamiento
- Algoritmos de minería de datos

La automatización, provee herramientas típicas de soporte de decisión

Relación con otras disciplinas



Clasificación - Funcionalidad

Descriptiva

- **Agrupamiento** (clustering): clasificar individuos en grupos en base a sus características
- Por ejemplo, clasificar pacientes del hospital
  - MASSA, Fernando and ALVAREZ-VAZ, Ramón. Aplicación de técnicas de clustering para el desarrollo de perfiles epidemiológicos en estudios de salud. Saber Es, 2020, vol.12, n.2, http://www.scielo.org.ar/
  - Enfermedades no transmisibles (ENT), se usan variables categóricas para reflejar la presencia de determinadas enfermedades y comorbilidades o factores de riesgo (FR). Por ejemplo patologías orales (PO).

Clasificación - Funcionalidad

Descriptiva

- **Reglas de asociación**: conocer cómo se relacionan los datos o campos, encontrar correlaciones y co-ocurrencias entre conjuntos de datos
- Por ejemplo conocer en el hipermercado que un cliente que compra leche muy probablemente comprará también pan
- Servicios como Netflix y Spotify pueden utilizar las reglas de asociación para alimentar sus motores de recomendación de contenido

Clasificación - Funcionalidad

Descriptiva

- **Reglas de asociación**: conocer cómo se relacionan los datos o campos, encontrar correlaciones y co-ocurrencias entre conjuntos de datos
- Por ejemplo conocer en el hipermercado que un cliente que compra leche muy probablemente comprará también pan
- Servicios como Netflix y Spotify pueden utilizar las reglas de asociación para alimentar sus motores de recomendación de contenido

Clasificación - Funcionalidad

Predictiva

- En base a una **clasificación**: pronosticar si el cliente pagará o no pagará, o el tipo de dolencia de un paciente
- En base a una **Regresión**: calcular el tiempo previsible para corregir los errores de un desarrollo de software
- **Secuenciación** (Series de tiempo): predecir el valor de una variable en función del tiempo, por ejemplo la demanda de energía eléctrica
- **Redes Neuronales**, **SVM**, etc.

Algoritmos de Aprendizaje

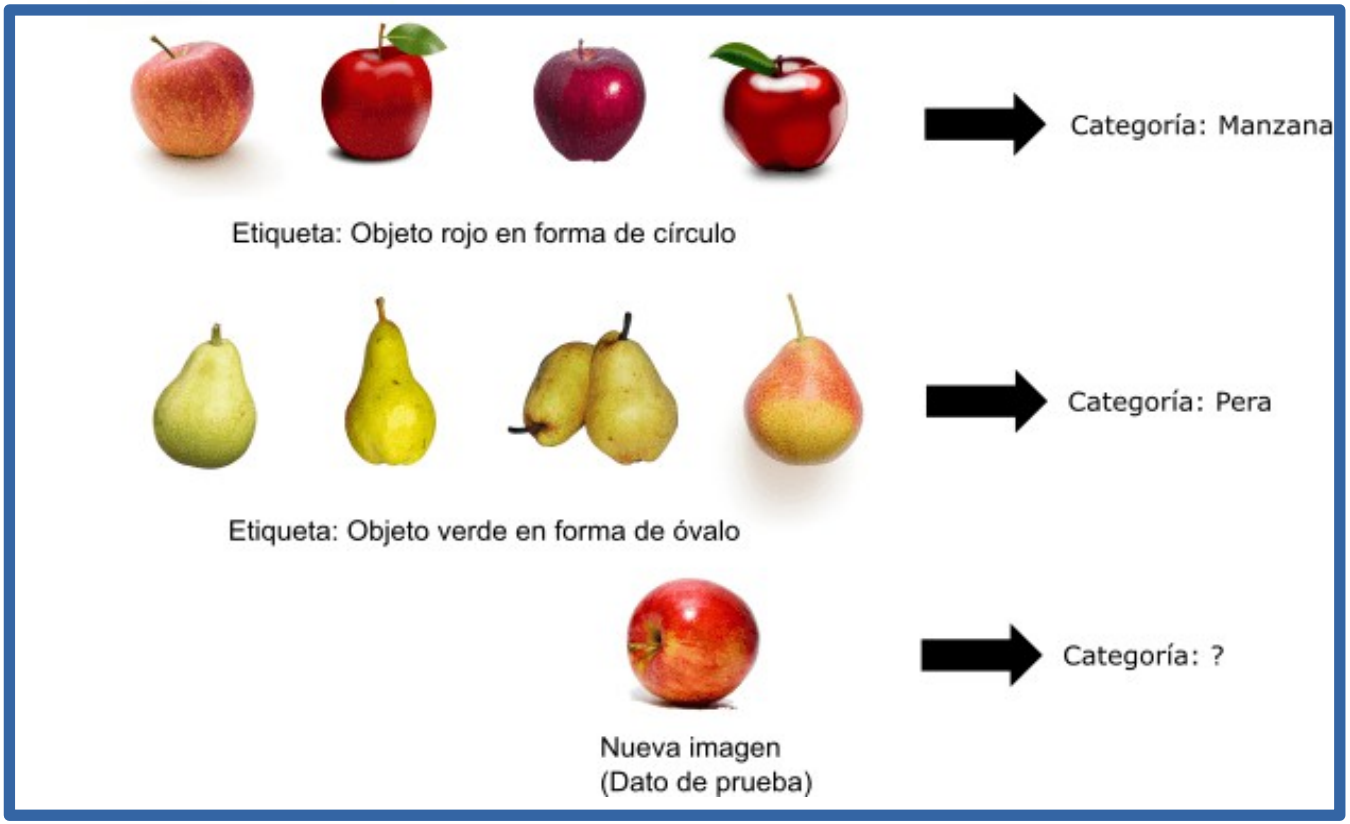
Supervisado y no supervisado

- Aprendizaje supervisado - parte de un conjunto de datos etiquetado previamente, se conoce el valor del atributo objetivo – clase o categoría

- Aprendizaje no supervisado parte de datos no etiquetados previamente

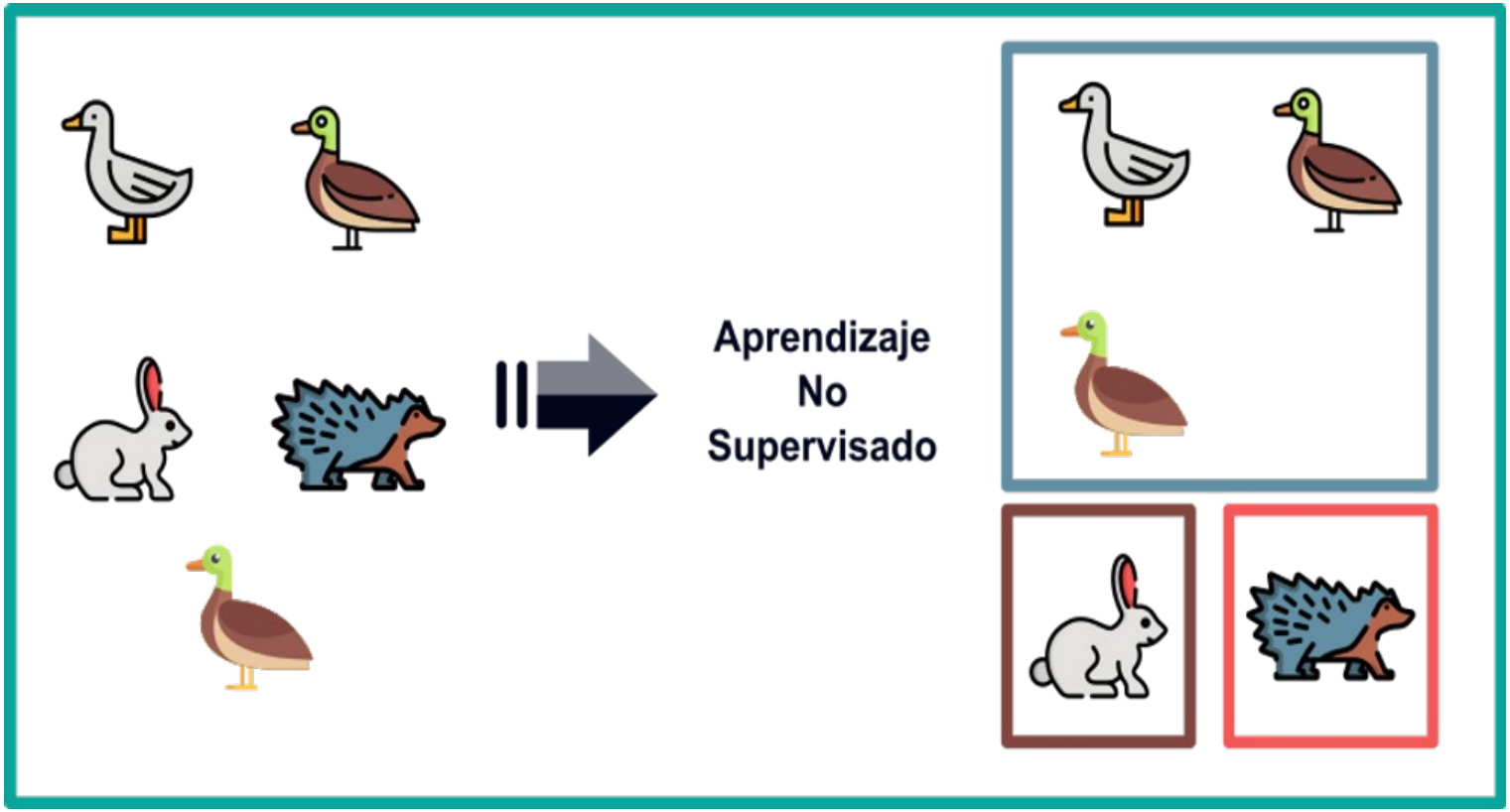
Algoritmos de Aprendizaje

Supervisado



Algoritmos de Aprendizaje

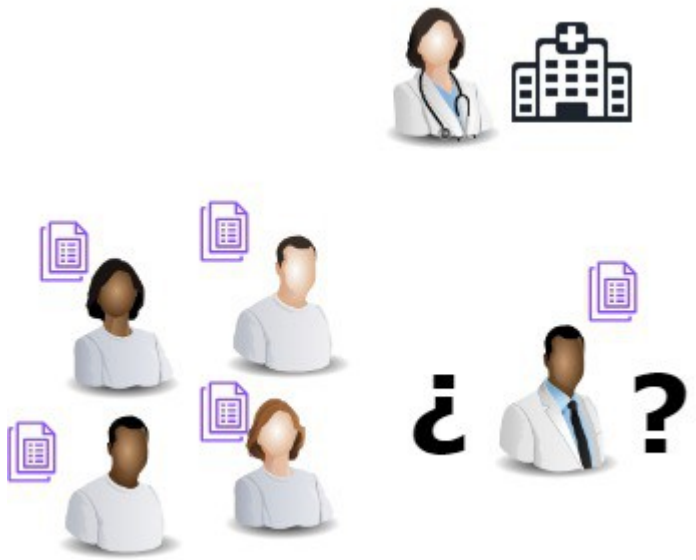
No supervisado



Algoritmos de Aprendizaje

Ejemplos de aprendizaje **supervisado** y **no supervisado**

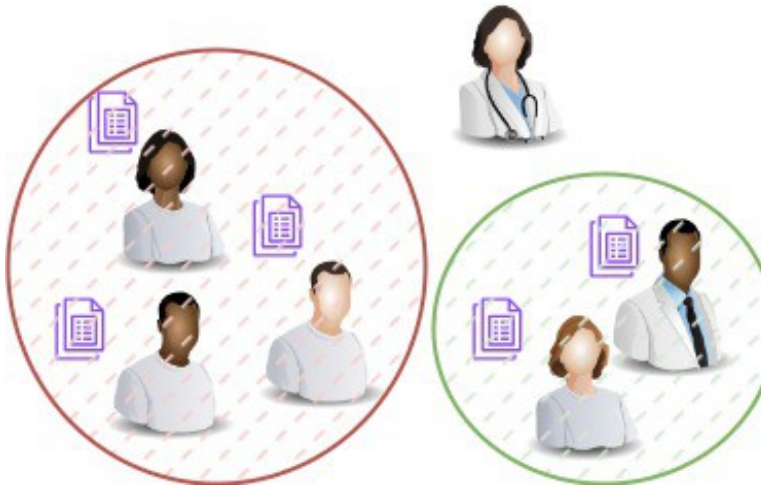
- **Supervisado** construcción de un modelo de reingresos en hospitalización, partiendo de un conjunto de datos previos de los que conocemos si el paciente reingresó o no (el atributo que nos indique la condición de reingreso en el conjunto de datos original sería la etiqueta) – otras variables (presión, temperatura, lesiones, etc.)



Algoritmos de Aprendizaje

Ejemplos de aprendizaje **supervisado y no supervisado**

- **No supervisado** objetivo de agrupar pacientes transplantados, sin un conocimiento previo de los grupos que queremos obtener; a partir de variables que permiten inferir estructuras no evidentes subyacentes en los datos (presión, glicemia, lesiones, etc.)



Algoritmos de AA

APRENDIZAJE AUTOMÁTICO CLÁSICO

los datos están clasificados en números o categorias

SUPERVISADO

predicen una categoría

CLASIFICACIÓN

clasifica pacientes por tipo (grupos de edad, sexo, otros)
Categorías discretas

predicen un número

REGRESIÓN

dividen pacientes por estatura o IMC
Categorías numéricas

los datos no estan clasificados de ninguna forma

NO SUPERVISADO

se agrupan por similitudes

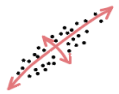
AGRUPAMIENTO

dividen pacientes por síndromes

se identifican dependencias ocultas

REDUCCIÓN DE DIMENSIONES

aprovecha las característi-cas para crear mejores modelos



se identifican secuencias

ASOCIACIÓN

agrupa pacientes de modos más convenientes

Algoritmos de Minería de Datos

Regresión

Definición

Modelo de regresión es aquel en el que las variables explicativas (x_{ip}) y la variable respuesta (y_i) son todas cuantitativas (numéricas)

$$y_i = r(x_{il}, \ldots, x_{ip}) + \varepsilon_i$$

- r es la función determinista que explica el comportamiento de y en función de las variables explicativas x_i
- ε_i es el error o la aleatoriedad en el comportamiento de cada individuo

Definición

El error:

$$E[\varepsilon_i] = E[y_i - r(x_{il}, \ldots, x_{ip})]$$

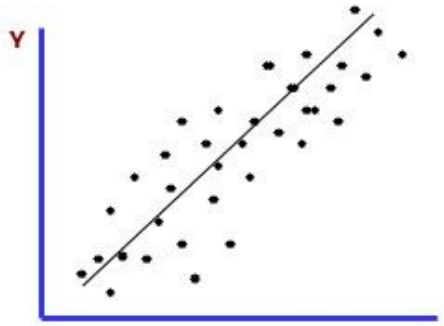
Buscamos una función tal que, en promedio, las desviaciones al cuadrado respecto de los puntos sean mínimas

Minimizamos el error cuadrático medio

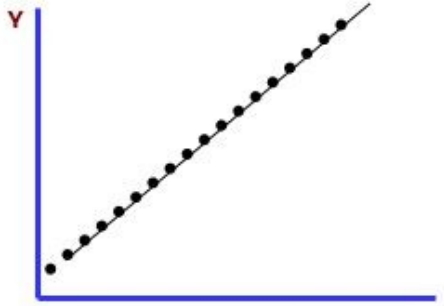
La función de regresión:

$$\min_r E[(y_i - r(x_{il}, \ldots, x_{ip}))^2]$$

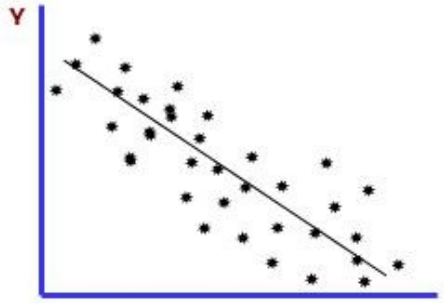
Relación entre variables cuantitativas



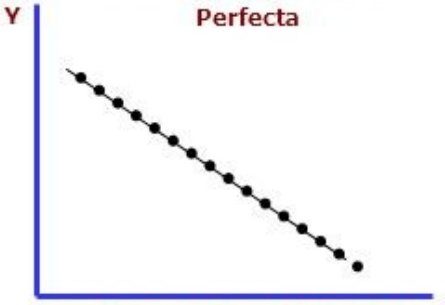
Relación Lineal Positiva



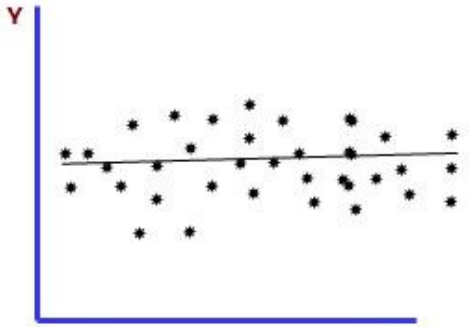
Relación Lineal Positiva Perfecta



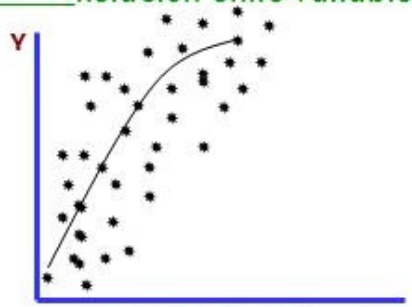
Relación Lineal Negativa



Relación Lineal Negativa Perfecta

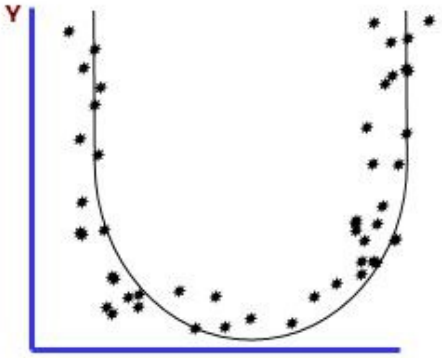


No exixte relación

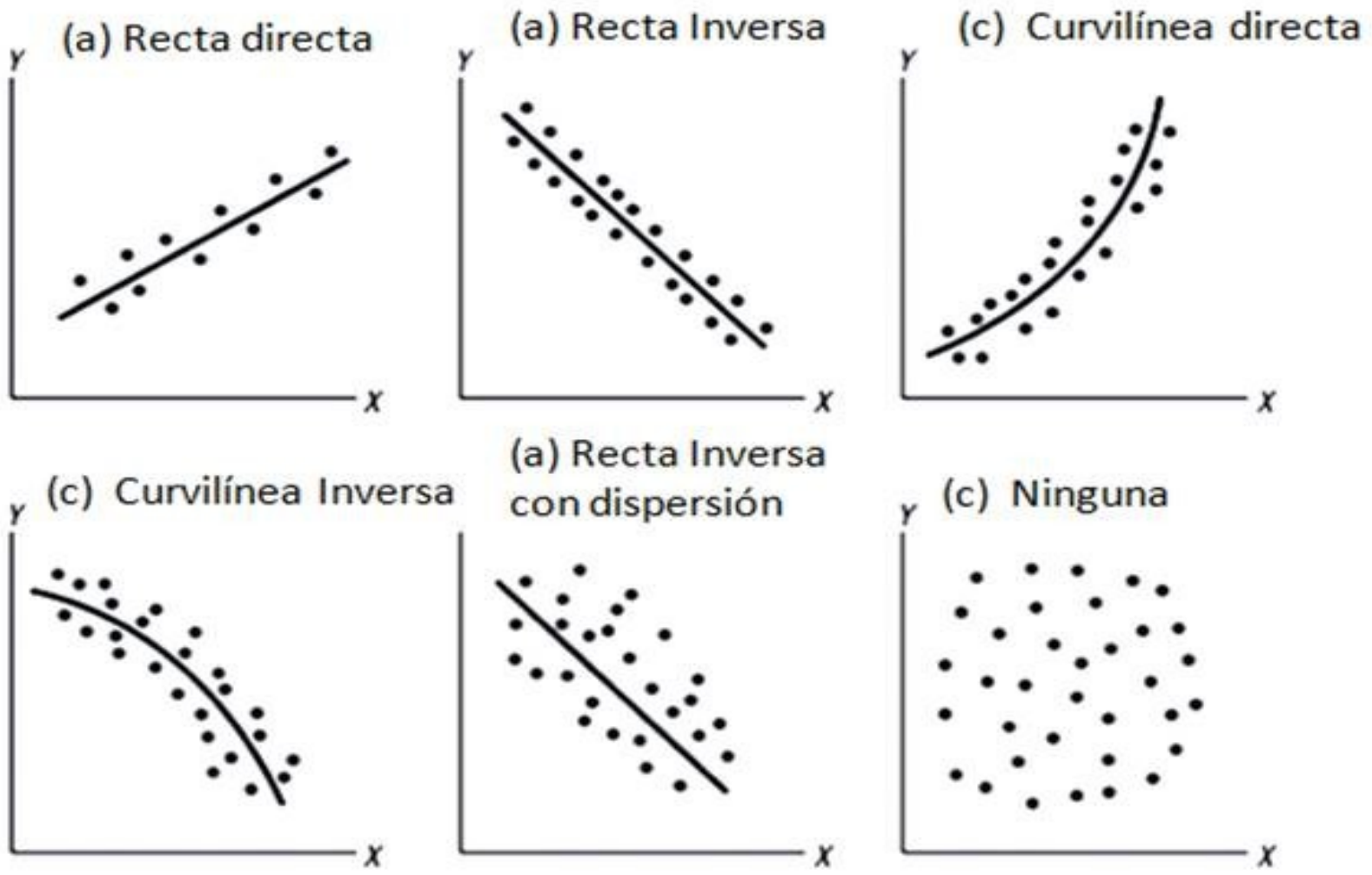


Relación Curvilinea Positiva

Relación Curvilinea Negativa



Relación Curvilinea Positiva


(a) Recta directa
Y
X
(a) Recta Inversa
Y
X
(c) Curvilínea directa
Y
X
(c) Curvilínea Inversa
Y
X
(a) Recta Inversa con dispersión
Y
X
(c) Ninguna
Y
X

Ejemplo - Regresión

Anio es la variables explicativa

Ingresos la variable respuesta

Ingresos(Anio) = 326.47*Anio-654041.93

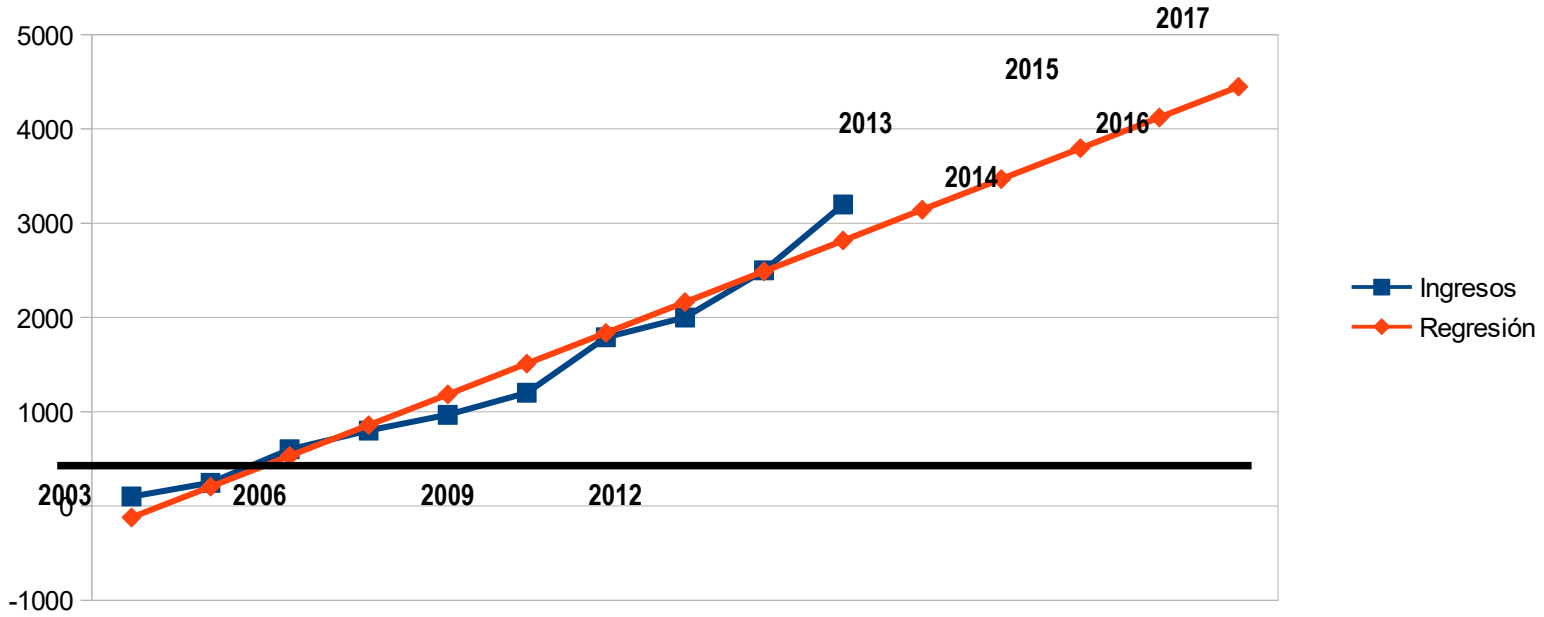
Ingresos(2013) = 326.47*2013-654041.93

= 3142,18

Anio	Ingresos
2003	100
2004	245
2005	600
2006	800
2007	967
2008	1200
2009	1789
2010	2000
2011	2500
2012	3198

Ejemplo - Regresión

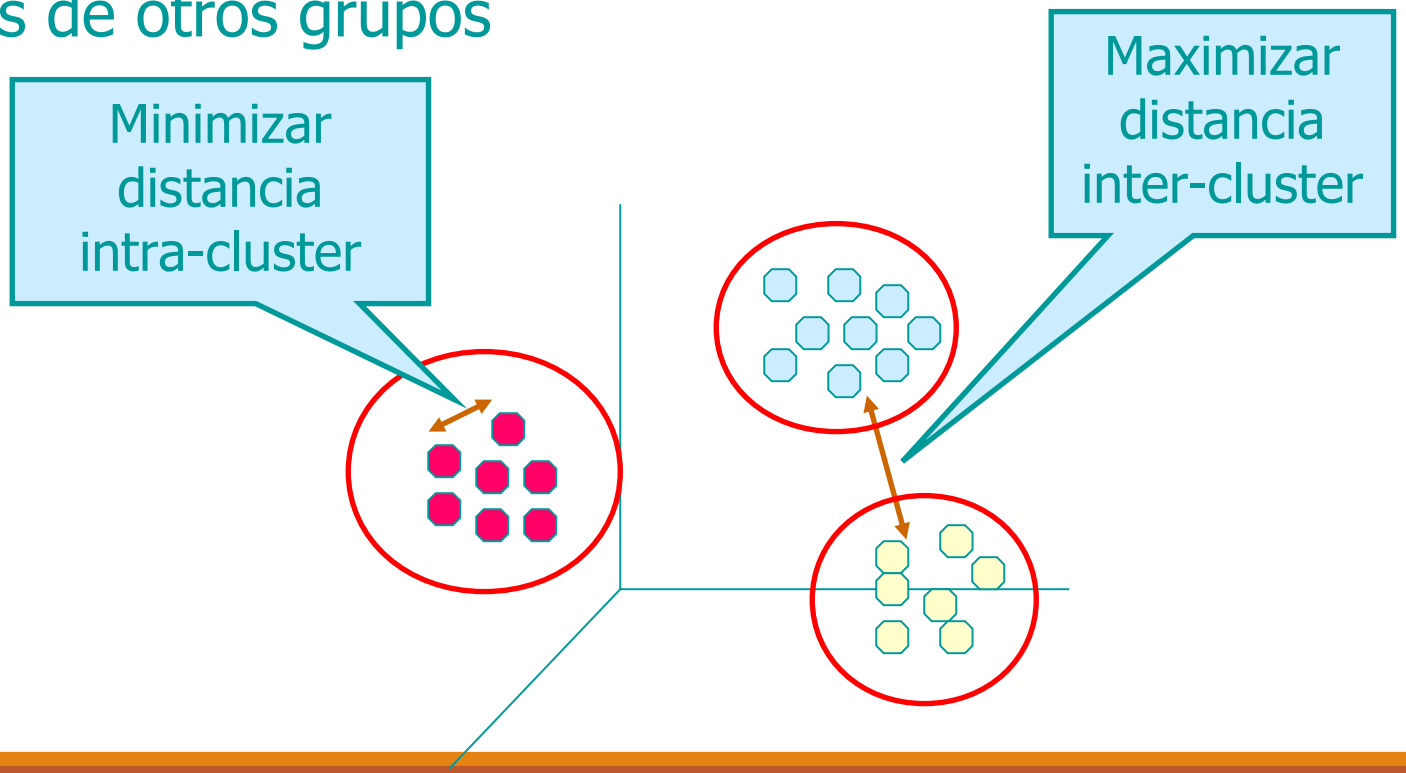
Anio es la variables explicativa, Ingresos es la variable respuesta



Clustering

Clustering

Encontrar agrupamientos de tal forma que los objetos de un grupo sean similares entre sí y diferentes de los objetos de otros grupos



Clustering

Medidas de Similitud

Usualmente, se expresan en términos de distancias:

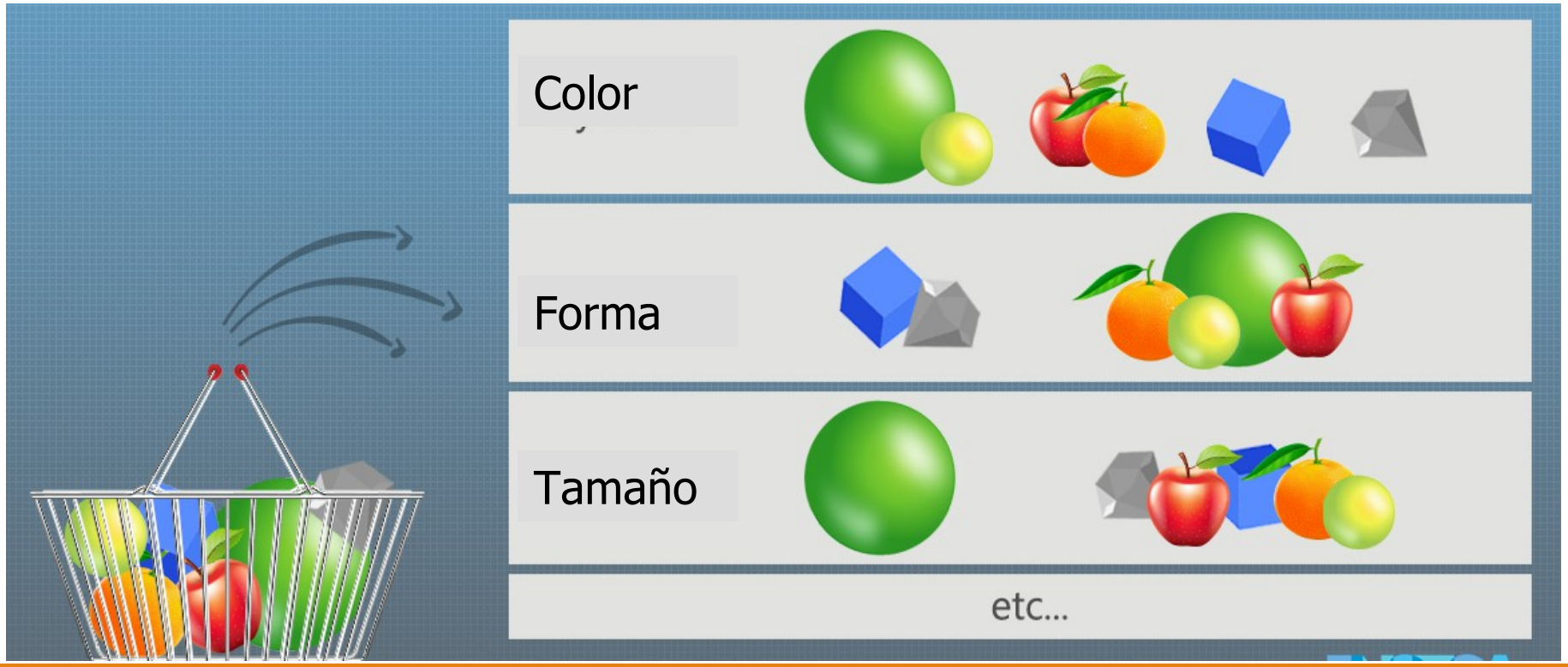
- **d(i,j) > d(i,k)** indica que el objeto i es más parecido a k que a j

La definición de la métrica de similitud/distancia dependerá del tipo de datos y de la interpretación semántica que se haga

En otras palabras, la similitud entre objetos es **subjetiva**

Clustering

Medidas de Similitud



Clustering

Métodos de Particionamiento

Métodos Jerárquicos

Basados en Densidad

Basados en Grillas

Basados en Modelos

Clustering

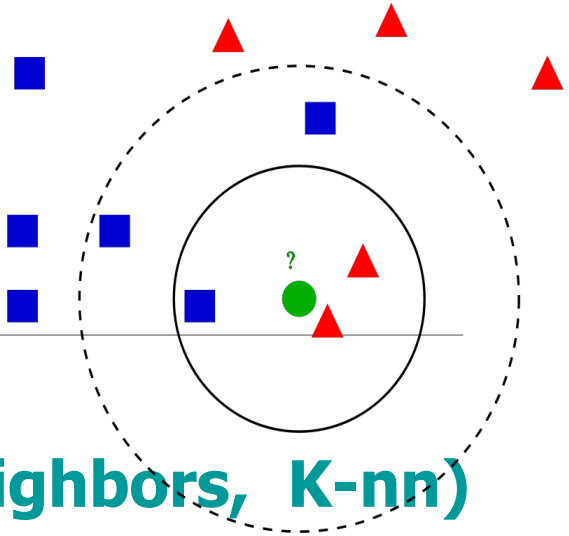
Métodos de Particionamiento

El conjunto de datos es particionado en un número pre-especificado de agrupamientos K

Iterativamente se va reasignando las observaciones a los agrupamientos hasta que algún criterio de parada (función a optimizar) se satisface (suma de cuadrados dentro de los clusters sea la más pequeña)

Ejemplos: K-nn, K-means, PAM, CLARA, SOM, Conglomerados basados en modelos de mezclas gausianas, Conglomerados difusos.

K-nn



K vecinos más cercanos (K-nearest neighbors, K-nn)

Algoritmo de Particionamiento en el que los elementos son vectores en un espacio multidimensional

Dado un conjunto de datos S y j número de clusters a formar

Se eligen elementos que serán los centros y una constante k

Cada punto en el espacio es asignado a la clase Cj, si esta es la clase más frecuente entre los k vecinos más cercanos.

Generalmente se usa la distancia euclidiana

Clustering

Métodos Jerárquicos

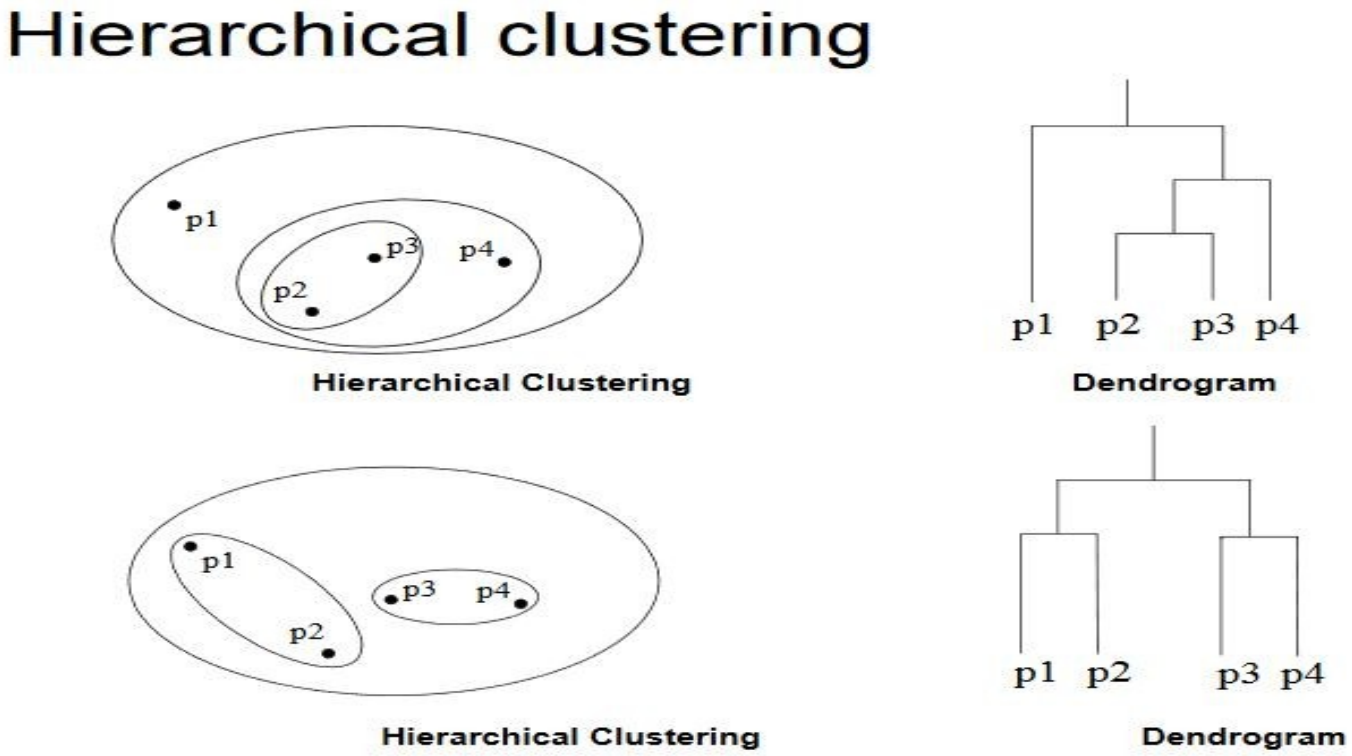
En estos algoritmos se generan sucesiones ordenadas (jerarquias) de agrupamientos

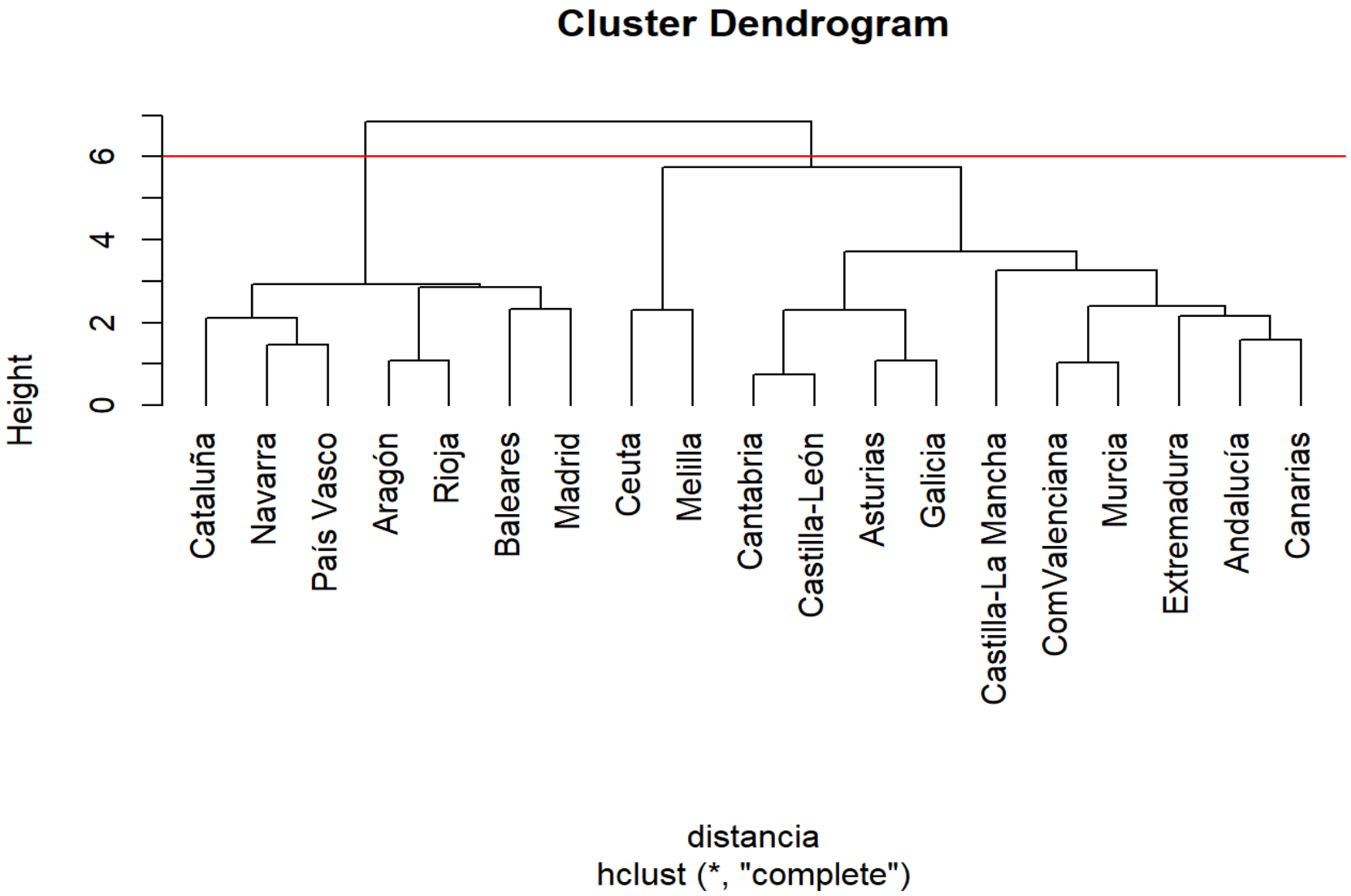
Pueden unir agrupamientos pequeños en mas grandes o dividir grandes clusters en otros mas pequeños

La estructura jerárquica es representada en forma de un árbol y es llamada Dendograma

Clustering

Métodos Jerárquicos




Cluster Dendrogram
Height
0
2
4
6
Cataluña
Navarra
País Vasco
Aragón
Rioja
Baleares
Madrid
Ceuta
Melilla
Cantabria
Castilla-León
Asturias
Galicia
Castilla-La Mancha
ComValenciana
Murcia
Extremadura
Andalucía
Canarias
distancia
hclust (*, "complete")

Clustering

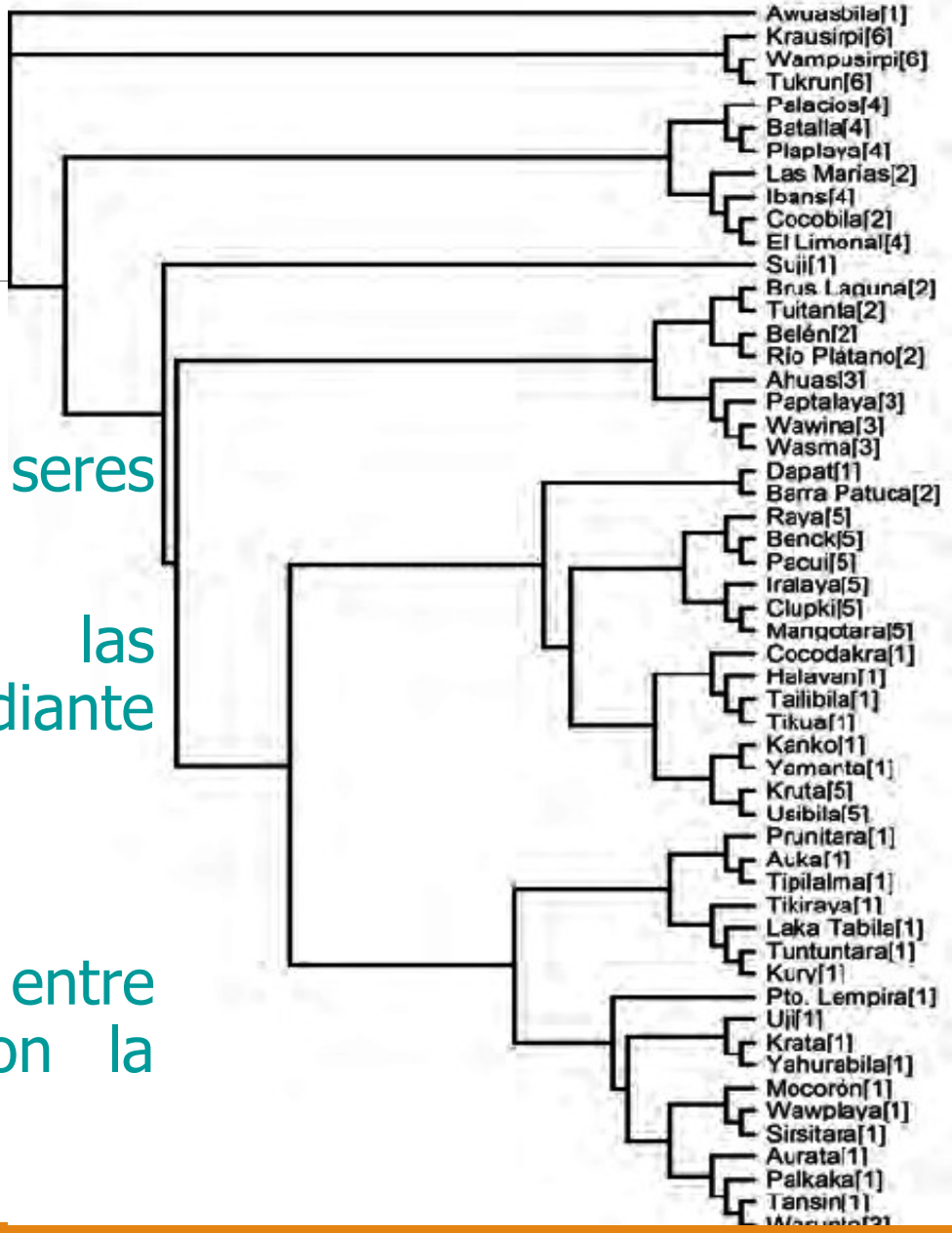
Dendograma

Diferencias fenotípicas entre seres humanos

Distancias utilizando las frecuencias de apellidos mediante el método de isonimia

54 comunidades

¿Diferencias genéticas entre poblaciones relacionadas con la distancia geográfica?



Clustering

Métodos Jerárquicos

Se dividen en dos tipos:

Algoritmos jerárquicos aglomerativos (bottom-up, inicialmente cada instancia es un cluster)

- AGNES (Agglomerative Nesting)

Algoritmos jerárquicos divisivos (top-down, inicialmente todas las instancias están en un solo cluster

- DIANA (DIvisive ANAlysis Clustering)

Reglas de Asociación

Definición

El objetivo de las reglas de asociación es encontrar asociaciones o correlaciones entre los elementos u objetos de bases de datos transaccionales, relacionales o data warehouse.

Las reglas de asociación tienen diversas aplicaciones como:

- Soporte para la toma de decisiones.
- Diagnóstico y predicción de alarmas en telecomunicaciones.
- Análisis de información de ventas

Problema

Dado un conjunto de transacciones, encontrar reglas que describen tendencias en los datos:

- Detectar cuándo la ocurrencia de un artículo está asociada a la ocurrencia de otros artículos en la misma transacción

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

Definiciones

El **Soporte** de un conjunto de items X en una base de datos D se define como la proporción de transacciones en la base de datos que contiene dicho conjunto de items. Ejemplo: Sop ({leche,pan})= 3/5=0,6

Confianza (o **eficiencia**), es la probabilidad condicional de que un item que contenga X también contenga Z. O la probabilidad de encontrar la parte derecha de una regla condicionada a que se encuentre también la parte izquierda.

Ejemplo: Conf({leche,pan}→{huevos})=
Sop({leche,pan}∩{huevos})/Sop({leche,pan})=
Sop({leche,pan,huevos})/Sop({leche,pan}) = **0,4 /0,6 =0,7**

Estrategias

¡¡Generar todas las reglas asociación para elegir las de alto soporte y confianza elevada es muy costoso!!

- Uso de técnicas de poda, reducción de transacciones y reducción de comparaciones
- Ejemplos: Algoritmos Apriori, DHP (Direct Hashing and Pruning) y AprioriTID

Árboles de Decisión

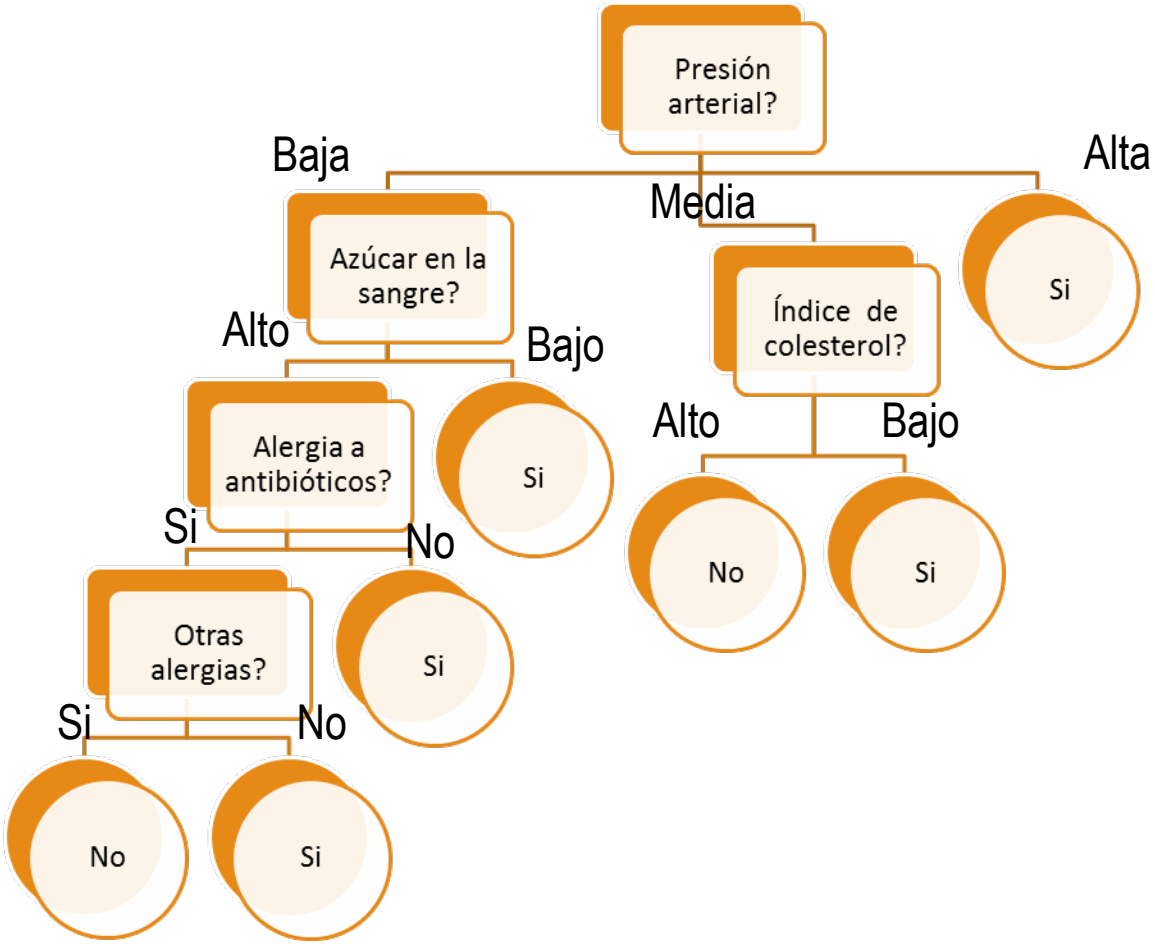
Árboles de Decisión

Un árbol de decisión es una forma gráfica y analítica de representar todos los eventos (sucesos) que pueden surgir a partir de una decisión asumida en cierto momento.

Ayudan a tomar la decisión "más acertada", desde un punto de vista probabilístico, ante un abanico de posibles decisiones.

Permite desplegar visualmente un problema y organizar el las variables y decisiones que deben realizarse.

Ejemplo: Si administrar un fármaco o no



Algoritmos

Aprendizaje basado en tuplas de entrenamiento:

- **ID3** (Iterative Dichotomiser 3)
- **C4.5** (Sucesor de ID3)
- **ACR** (Árboles de Clasificación y Regresión) - CART
- **CHAID** (Detector automático de Chi-cuadrado de interacción). Realiza divisiones de múltiples niveles al calcular los árboles de clasificación
- **MARS**: Extiende los árboles de decisión para manejar mejor datos numéricos
- **Árboles de Inferencia Condicional**: utiliza pruebas no paramétricas como criterios de división, corregidos para múltiples pruebas para evitar el sobreajuste.

Redes Neuronales

Introducción ANN

Las Redes Neuronales Artificiales, ANN (Artificial Neural Networks) están inspiradas en las redes neuronales biológicas del cerebro humano

Están constituidas por elementos que se comportan de forma similar a la neurona biológica en sus funciones más comunes

Características

Las ANN presentan una serie de características propias del cerebro

Por ejemplo las ANN aprenden de la experiencia, generalizan de ejemplos previos a ejemplos nuevos

Aprender: adquirir el conocimiento de un elemento/tema por medio del estudio, ejercicio o experiencia

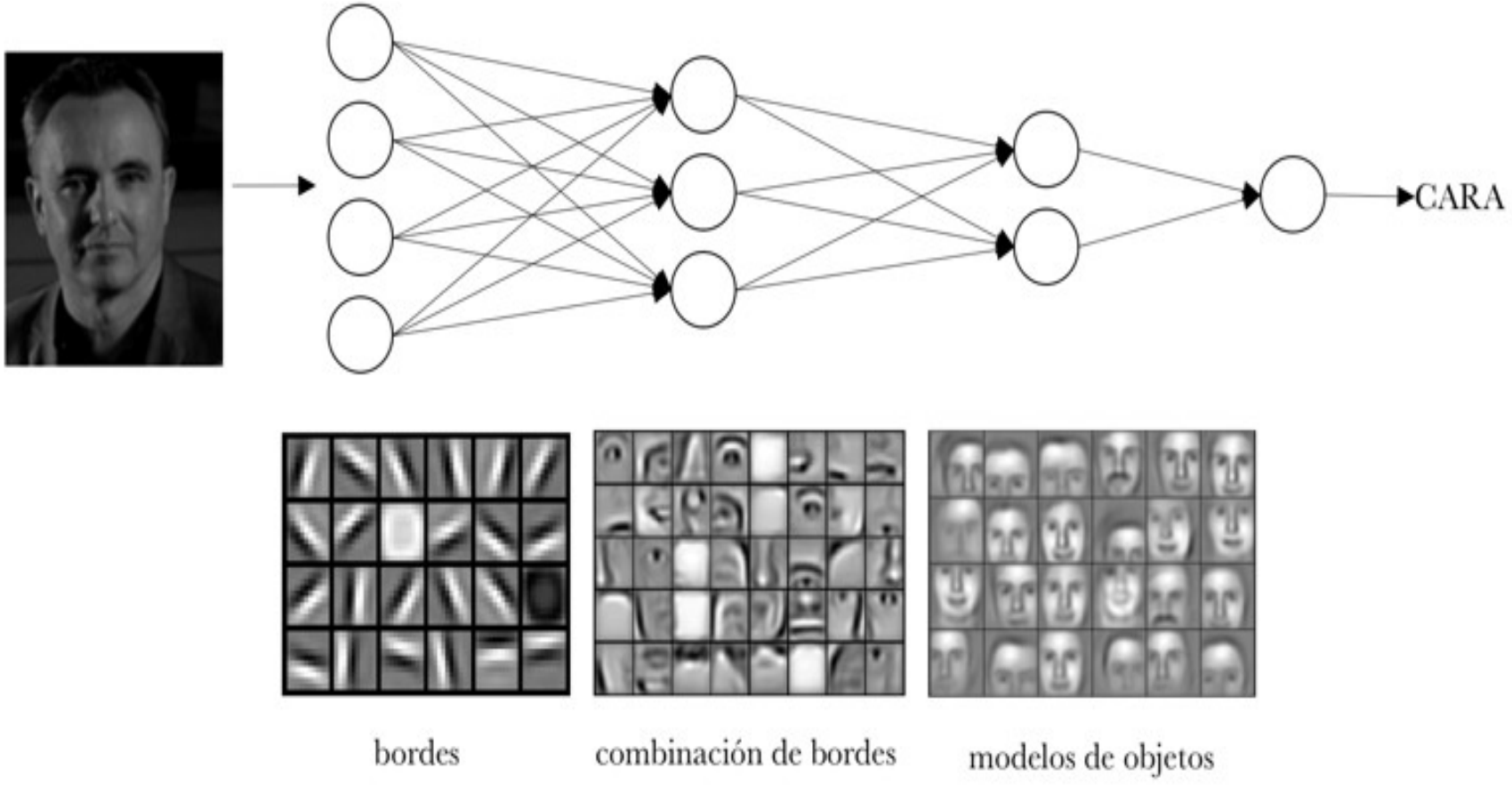
Las ANN pueden cambiar su comportamiento en función del entorno, se ajustan a un conjunto de entradas para producir salidas consistentes

Características

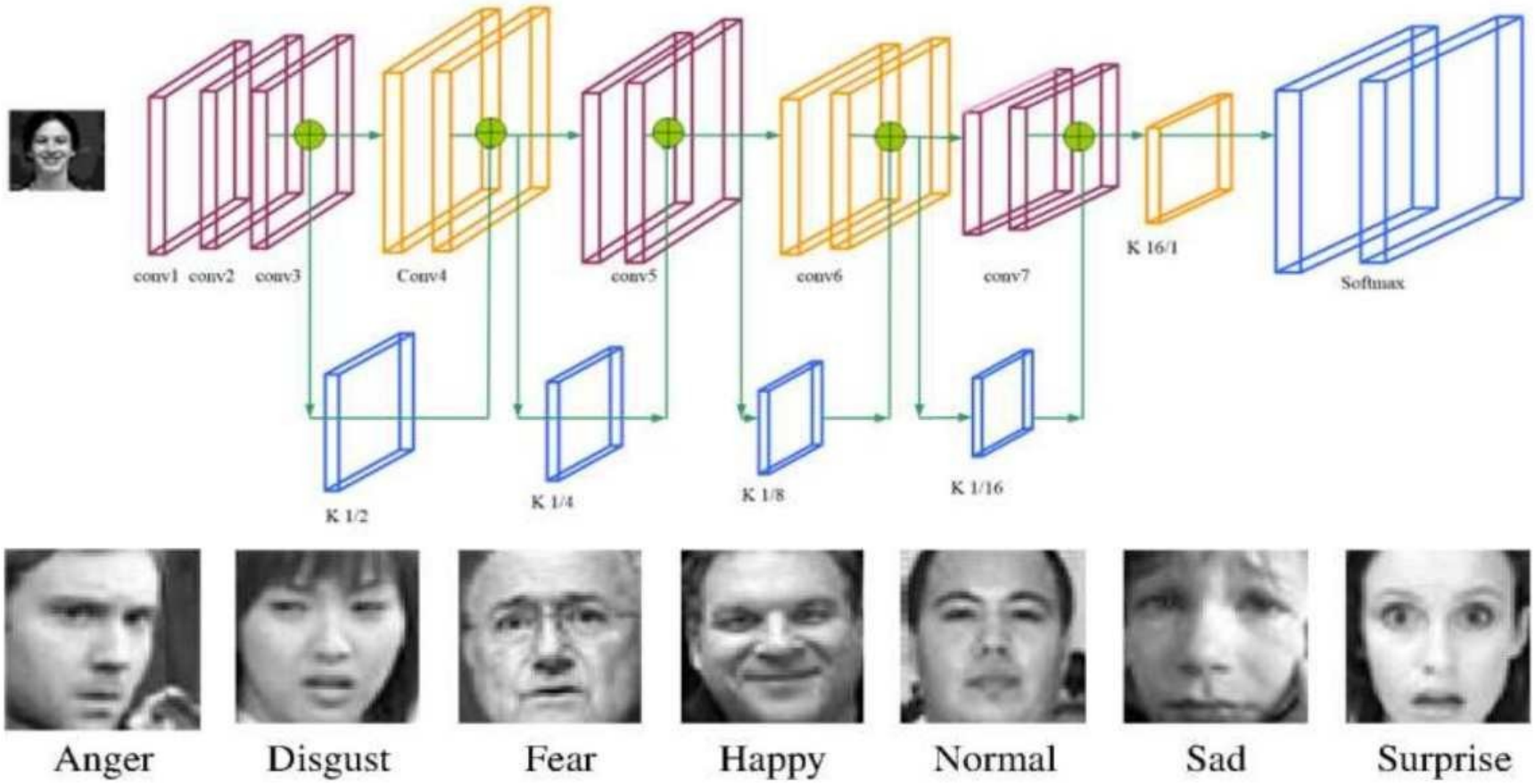
Generalizar: extender o ampliar un elemento/tema. Las ANN generalizan automáticamente debido a su propia estructura y naturaleza. Estas redes pueden ofrecer, dentro de un margen, respuestas correctas a entradas que presentan pequeñas variaciones debido a los efectos de ruido o distorsión.

Abstraer: aislar mentalmente o considerar por separado las cualidades de un objeto. Algunas ANN son capaces de abstraer la esencia de un conjunto de entradas que aparentemente no presentan aspectos comunes o relativos.

ANN



ANN


conv1
conv2
conv3
Conv4
conv5
conv6
conv7
K 16/1
Softmax
K 1/2
K 1/4
K 1/8
K 1/16
Anger
Disgust
Fear
Happy
Normal
Sad
Surprise

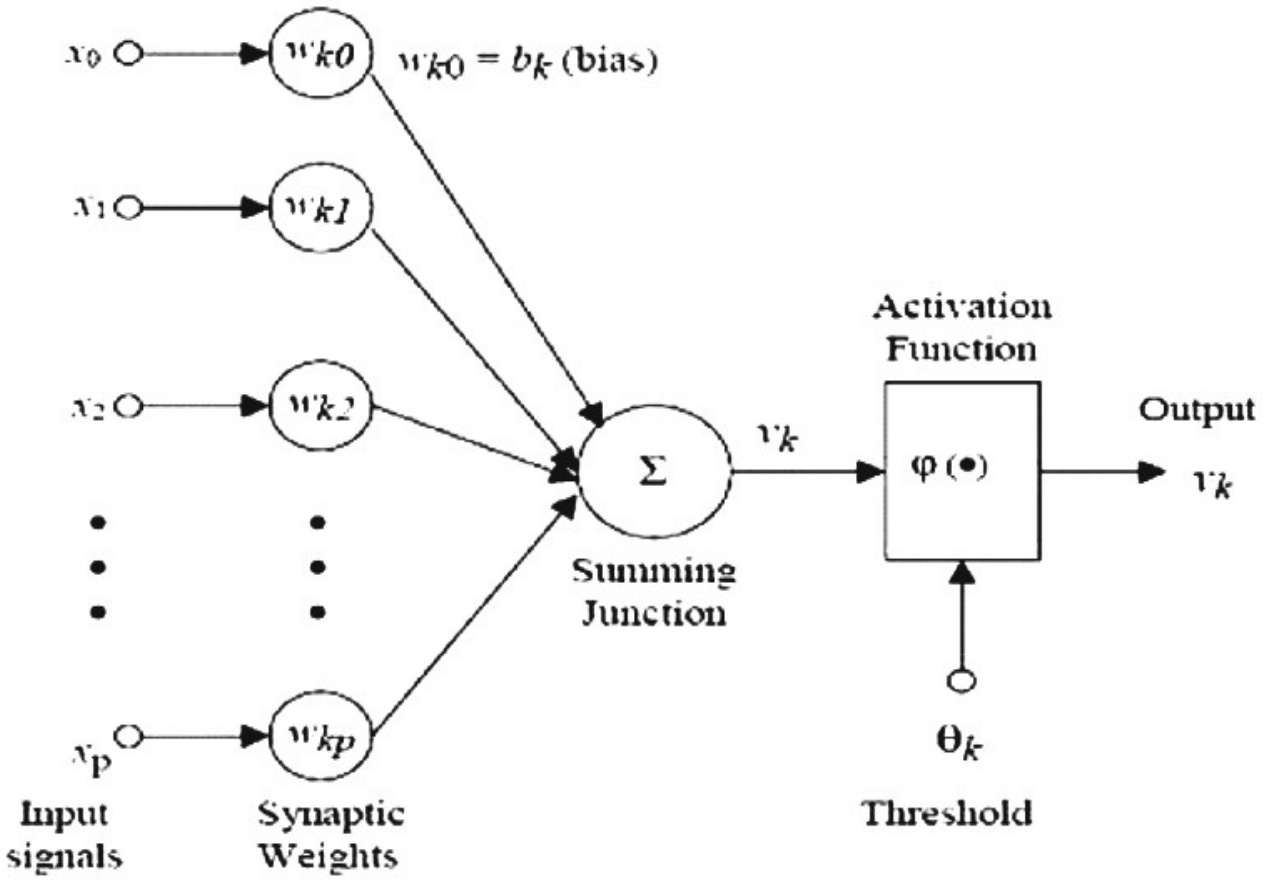
ANN - Estructura básica

En las Redes Neuronales Artificiales, ANN, la unidad análoga a la neurona biológica es el elemento procesador, PE (process element).

Un elemento procesador tiene varias entradas y las combina, normalmente con una suma básica.

La suma de las entradas es modificada por una función de transferencia y el valor de la salida de esta función de transferencia se pasa a la salida del elemento procesador.

ANN - Estructura básica



ANN - Estructura básica

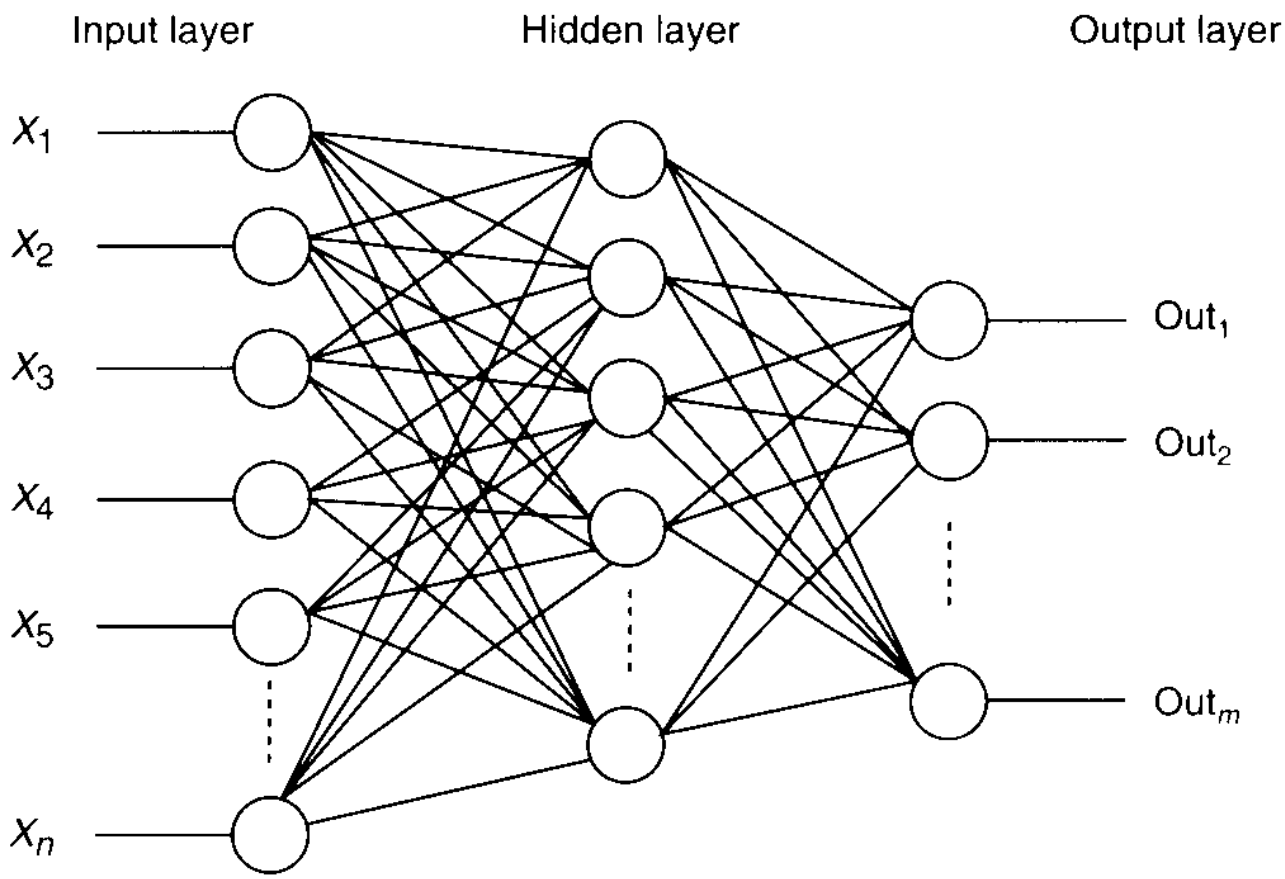
Una red neuronal consiste en un conjunto de unidades elementales PE conectadas de una forma concreta.

El interés de las ANN no reside solamente en el modelo del elemento PE sino en las formas en que se conectan estos elementos procesadores.

Generalmente los elementos PE están organizados en grupos llamados niveles o capas.

Una red típica consiste en una secuencia de capas con conexiones entre capas adyacentes consecutivas.

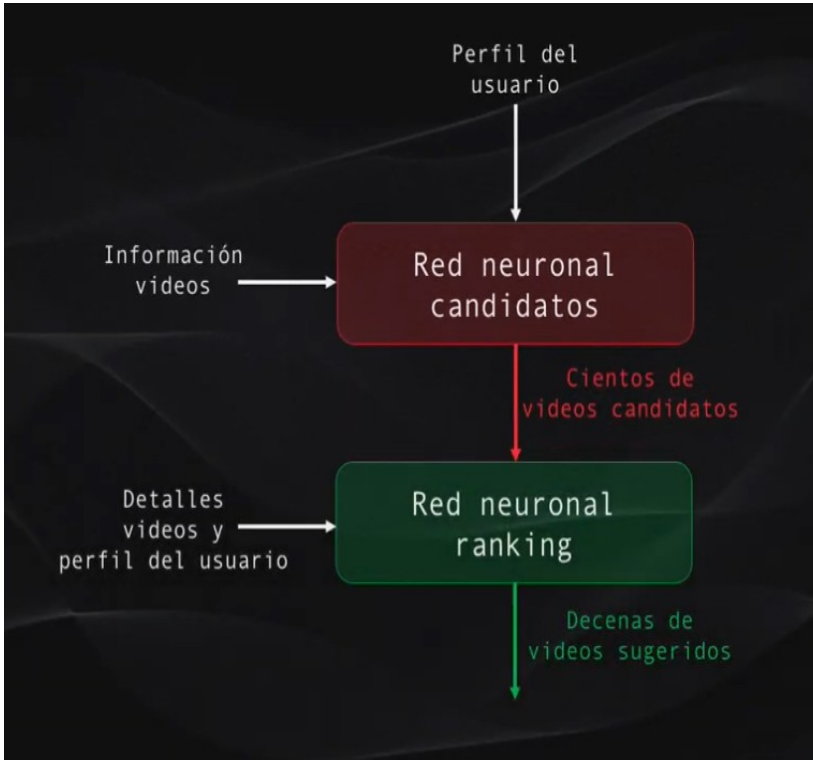
ANN - Estructura básica



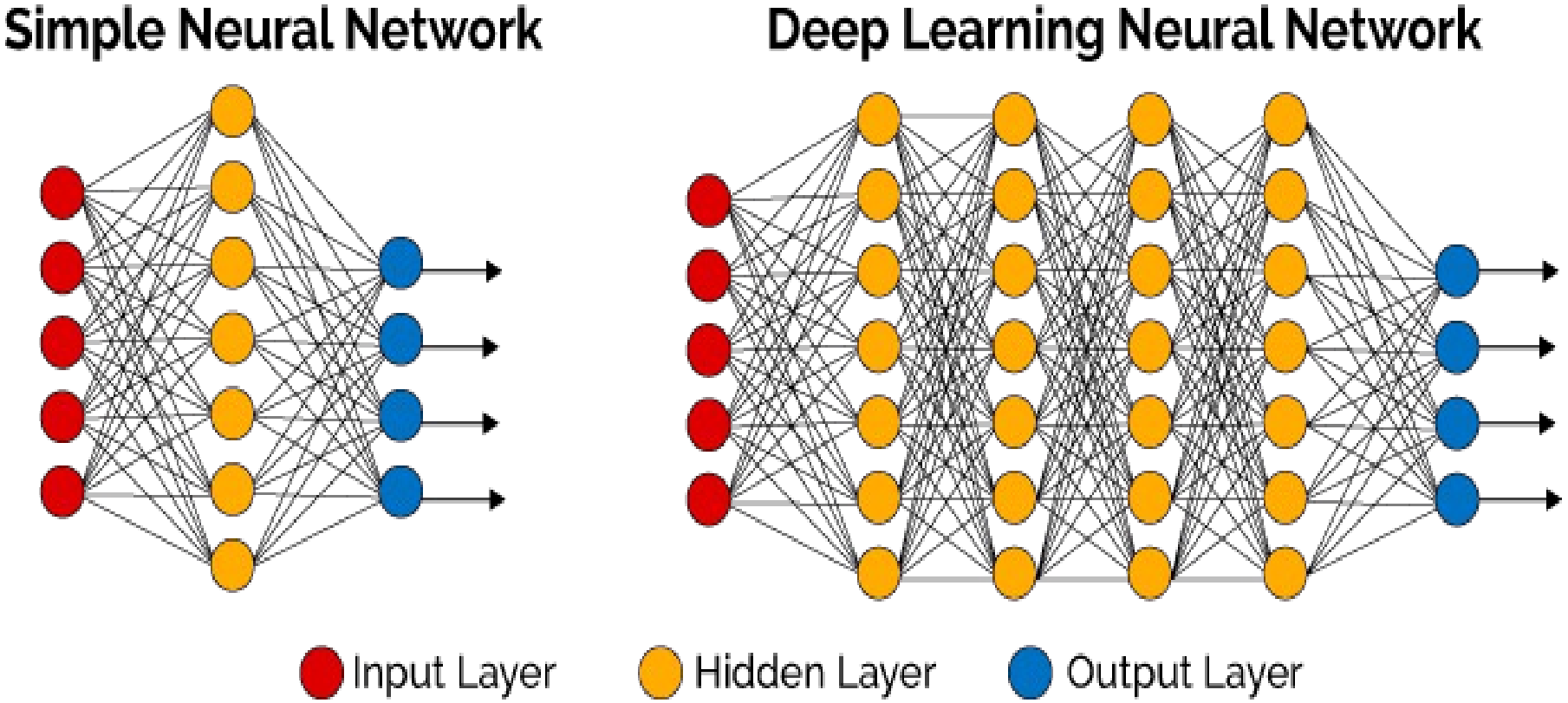
ANN - Aplicaciones

Algunas aplicaciones son: OCR, visión por computador, reconocimiento de voz, etc.

Youtube: una red neuronal consiste en generar una lista de posibles vídeos que se pueda recomendar a usuarios y otra red es capaz de establecer un ranking de los vídeos para seleccionar los más adecuados



Deep Learning



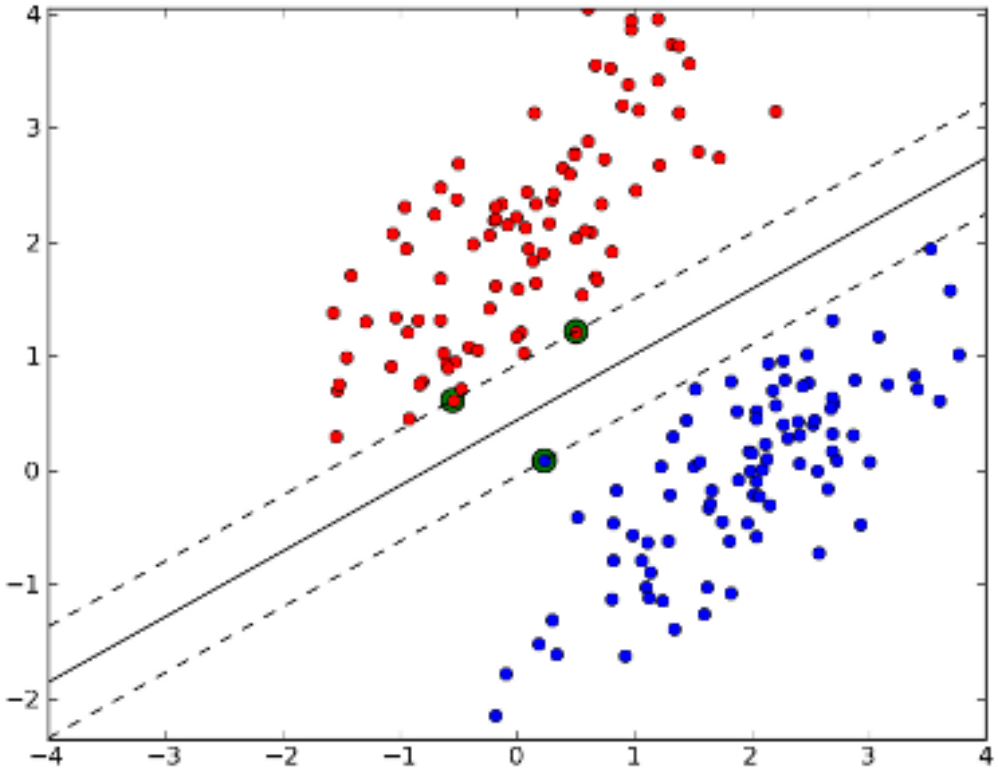
SVM

SVM - Definición

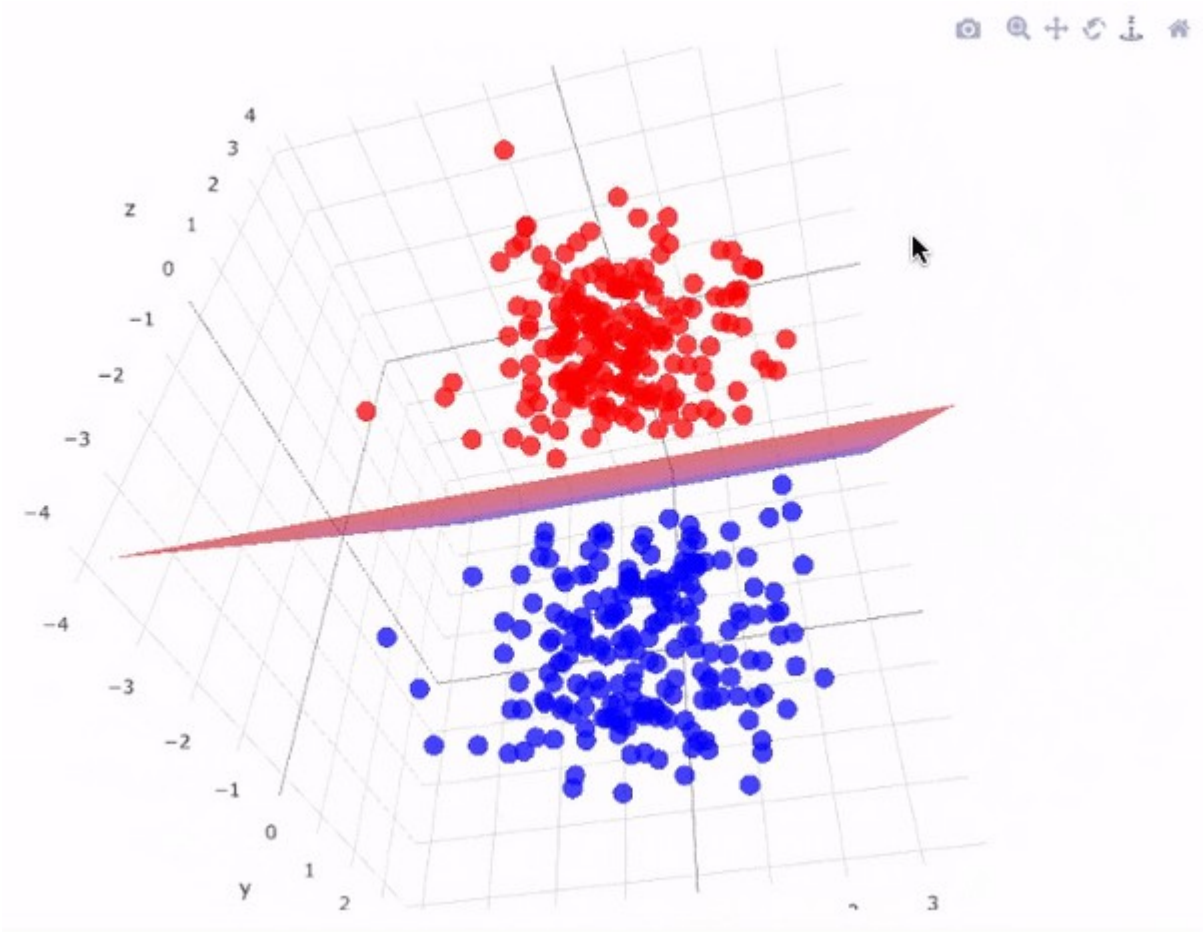
Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra.

SVM permite construir un modelo que representa a los puntos de muestra en el espacio, separando las clases en 2 subespacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre los 2 puntos, de las 2 clases, más cercanos al que se llama vector soporte.

Gráficamente 2D



Gráficamente 3D



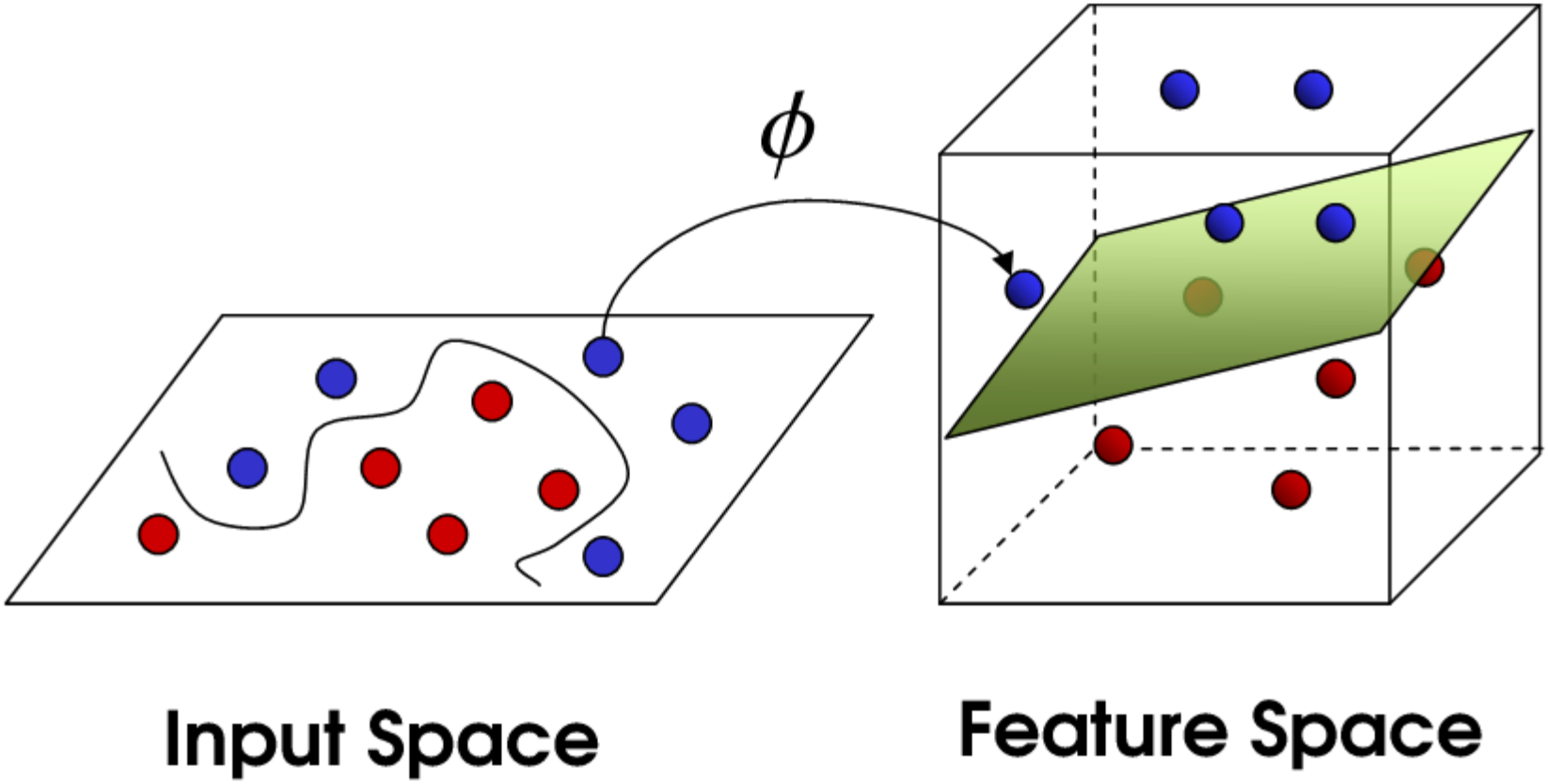
SVM

Las SVM son clasificadores derivados de la teoría de aprendizaje estadístico postulada por Vapnik y Chervonenkis.

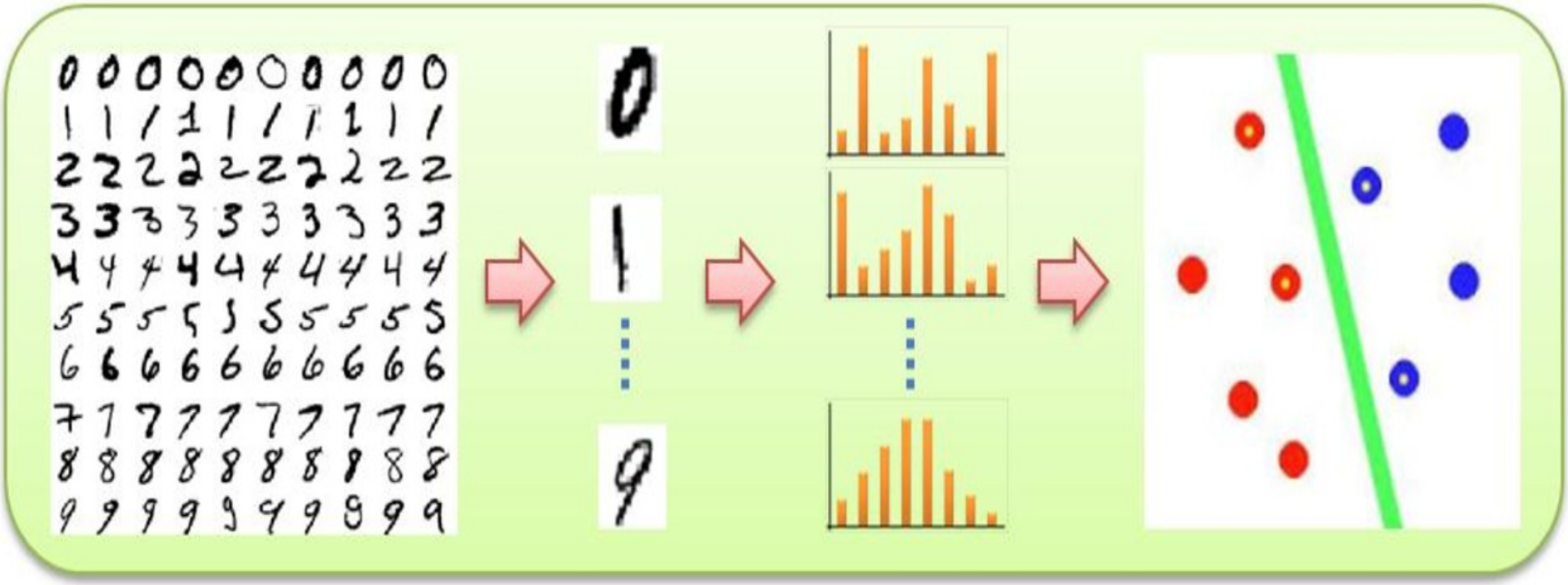
Las SVM fueron presentadas en 1992 y adquirieron fama cuando dieron resultados muy superiores a las redes neuronales en el reconocimiento de letra manuscrita, usando como entrada pixeles.

Transforma los datos a un espacio de dimensión muy alta a través de una función kernel -> se reformula el problema de tal forma que los datos se mapean implícitamente a otro espacio.

Linealmente separable



SVM - Aplicaciones



Evaluación y Uso de Modelos

Conjuntos de Datos

Para realizar los modelos hay 3 conjuntos de datos fundamentales:

- Conjunto de datos de entrenamiento: son los datos para entrenar los modelos
- Conjunto de datos de validación: para seleccionar el mejor de los modelos entrenados (por ejemplo, cambiando parámetros, clasificadores)
- Conjunto de datos de test: da el error real cometido con el modelo final seleccionado

Evaluación de Clasificación

Cada predicción puede ser uno de cuatro resultados, basado en cómo coincide con el valor real:

- Verdadero Positivo (TP): Predicho Verdadero y Verdadero en realidad
- Verdadero Negativo (TN): Predicho Falso y Falso en realidad
- Falso Positivo (FP): Predicción de verdadero y falso en la realidad
- Falso Negativo (FN): Predicción de falso y verdadero en la realidad

Evaluación de Clasificación

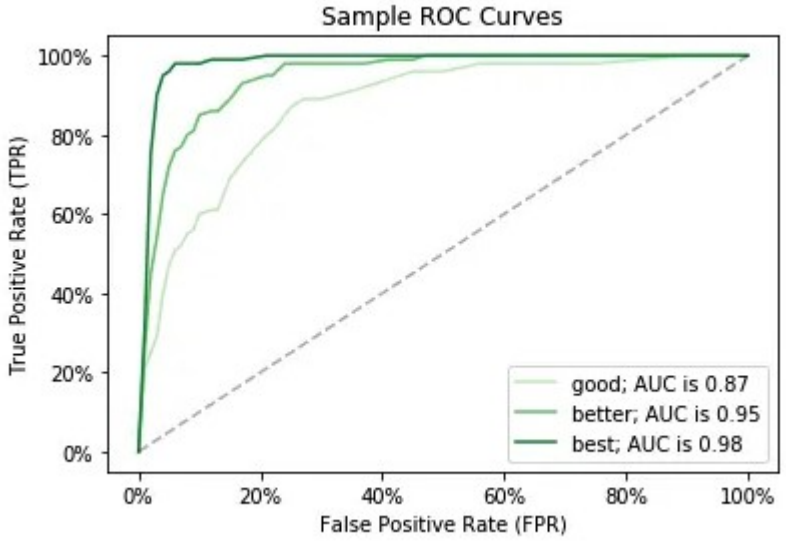
Cada predicción puede ser uno de cuatro resultados, basado en cómo coincide con el valor real: TP, TN, FP y FN.

En base a esto se definen las medidas:

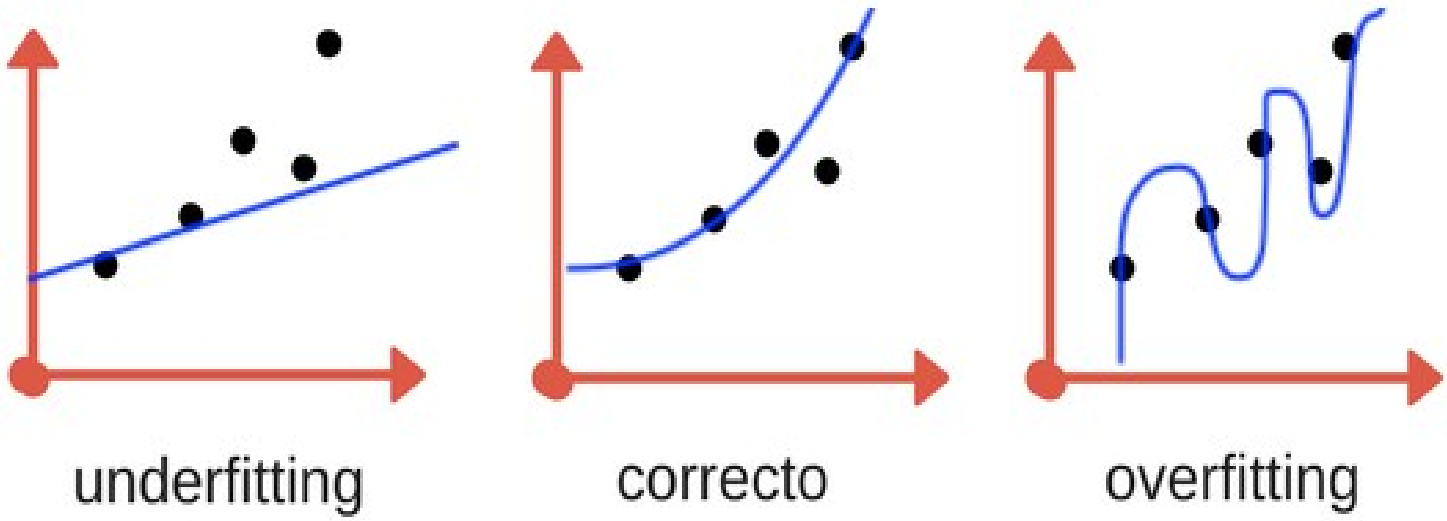
- Exactitud = (TP+TN)/total
- Tasa de clasificación errónea = (FP+FN)/total
- P=Precisión=TP/predicciones sí
- R=Tasa positiva verdadera (Exhaustividad) = TP/Si reales
- F = (P*R)/(P+R)

Evaluación de Clasificación

Medir el área bajo la curva ROC, tasa positiva verdadera (sensibilidad) frente a la tasa de falsos positivos (1-especificidad), se obtiene la curva de Característica Operativa del Receptor (ROC), muestra el equilibrio entre la tasa de verdaderos positivos y la tasa falsos positivos:



Ajuste de los modelos



Contacto

Departamento de Investigación Operativa,
Instituto de Computación,
Facultad de Ingeniería, UDELAR

Dra. Libertad Tansini
libertad@fing.edu.uy

Metodologías

Para la Minería de Datos

Comparación de Metodologías

Las tres metodologías dominantes para el proceso de la minería de datos son: KDD, CRISP-DM y SEMMA.

Azevedo, A. and Santos, M. F., KDD, SEMMA and CRISP-DM: a parallel overview. In Proceedings of the IADIS European Conference on Data Mining 2008, pp 182-185. Archived January 9, 2013.

SAS Enterprise Miner website, 2012.

Shearer C., El modelo CRISP-DM: el nuevo plan para la minería de datos, almacenamiento de los datos, 2000.

Metodología KDD

KDD es una metodología propuesta por Fayyad et al. en 1996, que propone 5 fases: selección, preprocesamiento, transformación, minería de datos y evaluación e implantación. Es un proceso iterativo e interactivo.

"From Data Mining to Knowledge Discovery in Databases", Usama Fayyad, Gregory Piatetsky-Shapiro, *From Data Mining to Knowledge Discovery in Databases*, American Association for Artificial Intelligence, 1996.

KDD

KDD (Knowledge Discovery from Databases) es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia, comprensibles a partir de los datos.

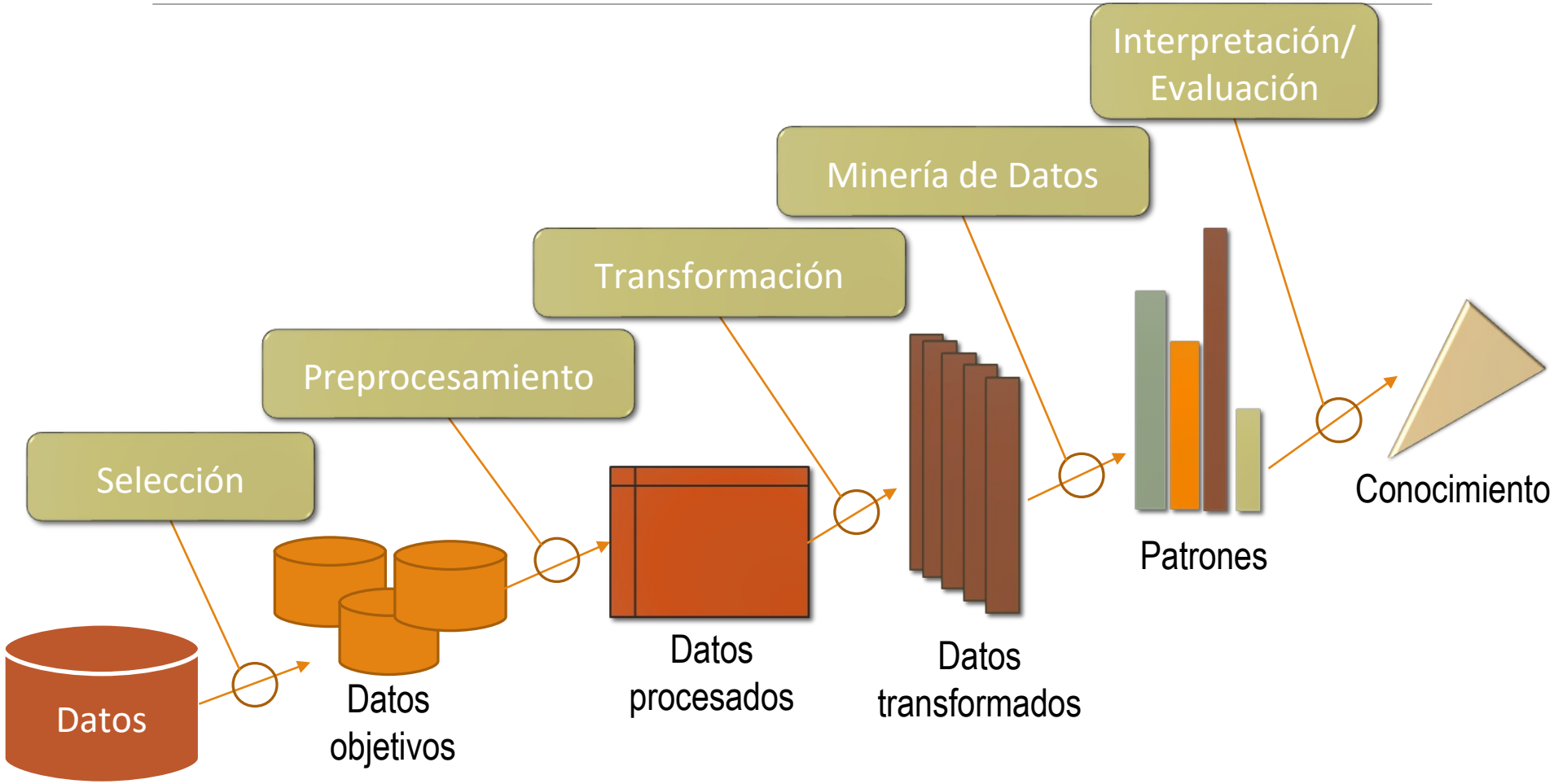
El **objetivo** fundamental del KDD (Knowledge Discovery from Databases), es encontrar conocimiento útil, válido, relevante y nuevo sobre una determinada actividad mediante algoritmos, dadas las crecientes órdenes de magnitud en los datos

KDD

KDD (Knowledge Discovery from Databases) es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia, comprensibles a partir de los datos.

El **objetivo** fundamental del KDD (Knowledge Discovery from Databases), es encontrar conocimiento útil, válido, relevante y nuevo sobre una determinada actividad mediante algoritmos, dadas las crecientes órdenes de magnitud en los datos

Etapas de KDD



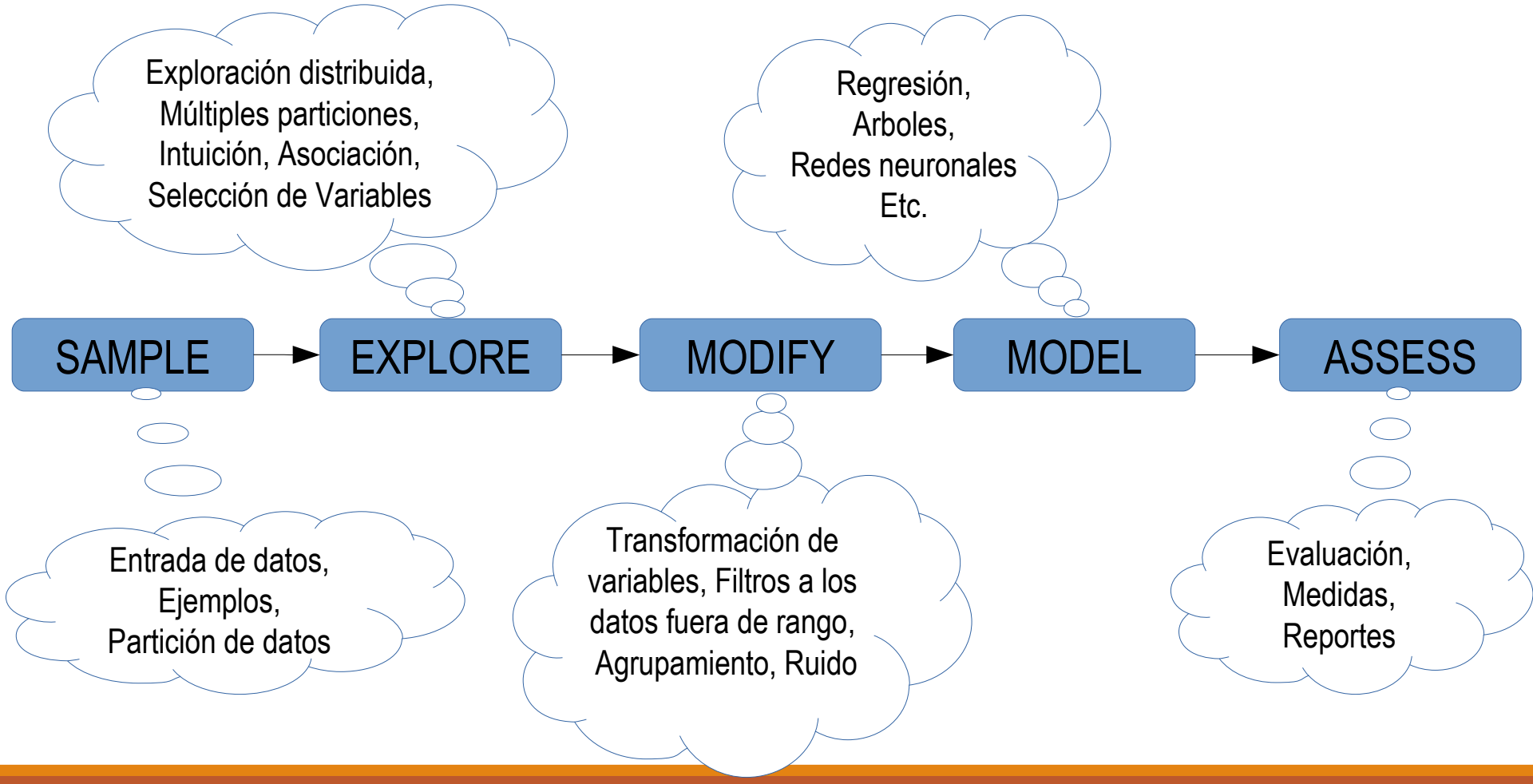
SEMMA

SEMMA es el acrónimo a las cinco fases: (Sample, Explore, Modify, Model, Assess).

La metodología es propuesta por SAS Institute Inc, la define como: "… proceso de selección, exploración y modelamiento de grandes cantidades de datos para descubrir patrones de negocios desconocidos…".

http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html

SEMMA


Exploración distribuida,
Múltiples particiones,
Intuición, Asociación,
Selección de Variables
Regresión,
Arboles,
Redes neuronales
Etc.
SAMPLE
EXPLORE
MODIFY
MODEL
ASSESS
Entrada de datos,
Ejemplos,
Partición de datos
Transformación de variables, Filtros a los datos fuera de rango, Agrupamiento, Ruido
Evaluación,
Medidas,
Reportes

CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISPDM). (1996 - 1997)

Iniciativa financiada por la Comunidad Europea se unió para desarrollar una plataforma para Minería de Datos. Dirigido por cinco empresas: SPSS, Teradata, Daimler AG, NCR y Ohra

Objetivos:

- Fomentar la interoperabilidad de las herramientas a través de todo el proceso de minería de datos.
- Eliminar la experiencia misteriosa y costosa de las tareas simples de minería de datos.

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Comparación KDD, SEMMA y CRISP-DM

KDD	SEMMA	CRISP-DM
Pre KDD	----	Comprensión del negocio
Selección	Muestra	Comprensión de los datos
Preprocesamiento	Exploración	
Transformación	Modificación	Preparación
Minería de datos	Modelo	Modelado
Interpretación/Evaluación	Evaluación	Evaluación
Post KDD	----	Despliegue

Selección, Limpieza y Transformación de Datos

Limpieza y selección de datos

Limpieza y selección de datos

Transformación de Atributos

- Discretización de Atributos
- Numerización de Atributos
- Normalización de Atributos
- Reducción de Dimensionalidad
- Métodos de Selección de Características
- Aumento de Dimensionalidad