CURSO DE INTELIGENCIA ARTIFICIAL MÁLAGA, 2 Junio de 2021

Introducción a la Inteligencia Artificial

Wolfram Rozas

Indice

Problemas de Ciencia de Datos

Problemas de Modelización Estadística/ Ciencia de Datos

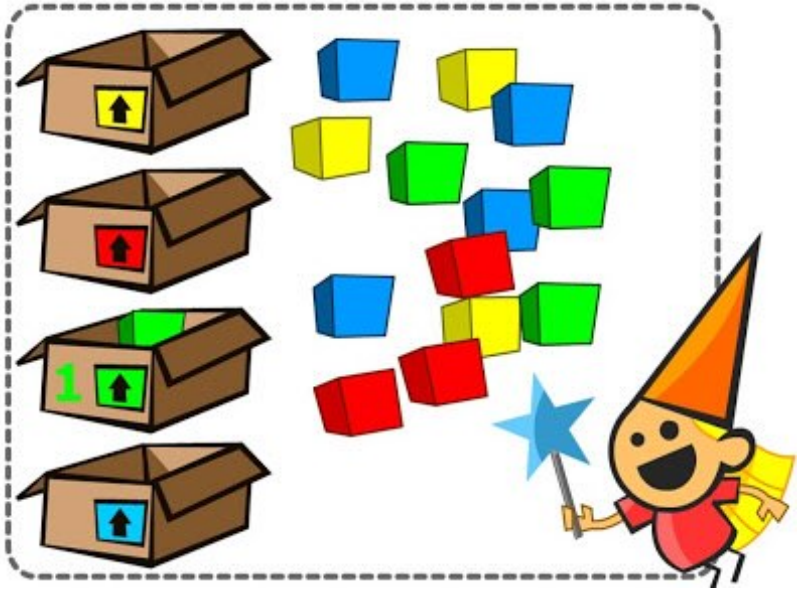
Supervisado Estimación	Supervisado Clasificación	No supervisado Clustering	No supervisado Asociación	Reducción Complejidad
Estimar (y es continua)	Clasificar (y es discreta)	Encontrar grupos homogéneos	Encontrar reglas afinidad	Menos atributos
Con objetivo (y error)	Con objetivo (y error)	Sin objetivo (ni error)	Muchos objetivos	Menos inputs
Regresión lineal, Redes Neuronales, SVM	Regresión Logística, Árboles de Decisión, KNN, SVM	K-Medias, Bietápico, Mapas de Kohonen, Detección de Anomalías	Reglas de Asociación, Patrones Secuenciales	Selección de Características, Componentes Principales, Análisis Factorial,

Estimación

- La Estimación o regresión especifica un modelo de aprendizaje supervisado que predice el valor de una variable continua
  - Estimar el valor futuro de un cliente (lifetime customer value)
  - Estimar la demanda de un flujo (electricidad, gas, agua, ...)
  - Predecir variables continuas como la renta

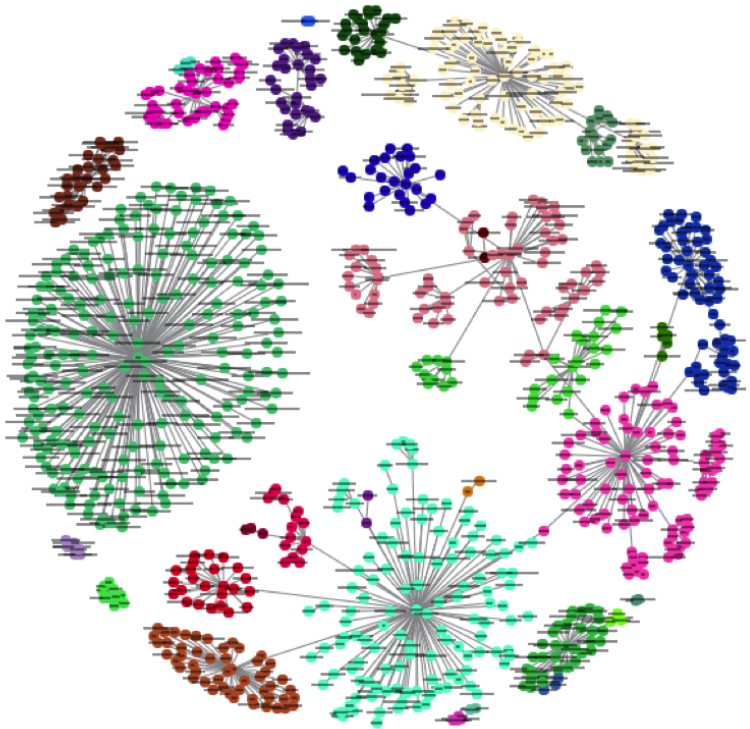
Clasificación

- La Clasificación es la asignación de registros del conjunto de datos a un número finito de clases de equivalencia con una probabilidad
- Es un modelo de aprendizaje supervisado donde la variable dependiente es categórica
  - Identificar el grado de pertenencia de un cliente a una clase
  - Determinar si el cliente aceptará la oferta comercial
  - Determinar si un cliente ha cometido fraude
  - Determinar si un cliente es de alto o bajo riesgo
  - Determinar si la máquina fallará en el próximo ciclo de producción



Clustering

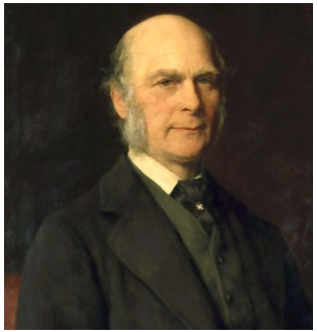
- El Clustering es un método de aprendizaje no supervisado que asocia elementos similares en grupos homogéneos:
  - Dividir un mercado en submercados
  - Identificar un subconjunto de interesados potenciales para un análisis más detallado
  - Confeccionar perfiles descriptivos de clientes
  - Agrupar los atributos de entrada que puedan contener conceptos similares
  - Crear “tablas de migración de estados” que muestren cómo se combina un cluster con otros
  - Identificar poblaciones homogéneas para mejorar la construcción de modelos



Algoritmia de Aprendizaje Supervisado

Técnicas de Estimación

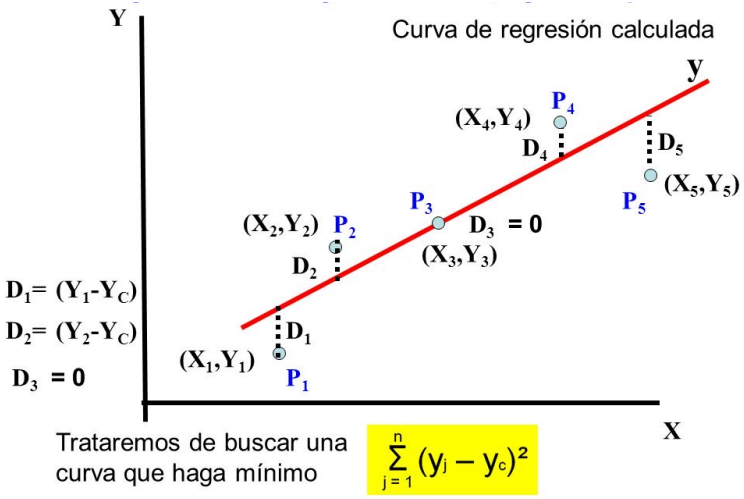
- Las técnicas de regresión suponen que la variable continua de predicción puede expresarse como una combinación lineal de las variables de entrada y un término de error, representando al resto de los factores



Sir Francis Galton

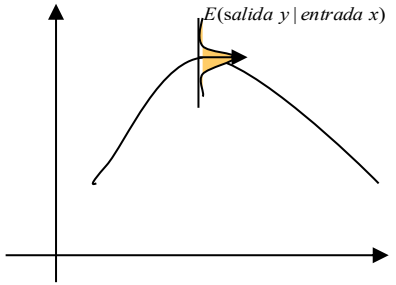
Galton acuñó el término regresión a hacia la media al observar que los hijos de los padres altos eran bajos, y los hijos de los padres bajos, eran altos

$$y = \beta_o + \beta_1 x_1 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$$



Ajuste por Mínimos Cuadrados Ordinarios

Adrien Marie Legendre



El término de error **ε** tiene una distribución predeterminada conocida como ruido blanco

$$\varepsilon \approx iidN(0, \sigma_\varepsilon^2)$$

Algoritmia de Aprendizaje Supervisado

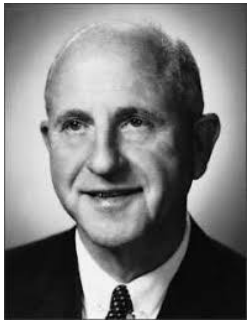
Técnicas de Clasificación

Regresión Logística

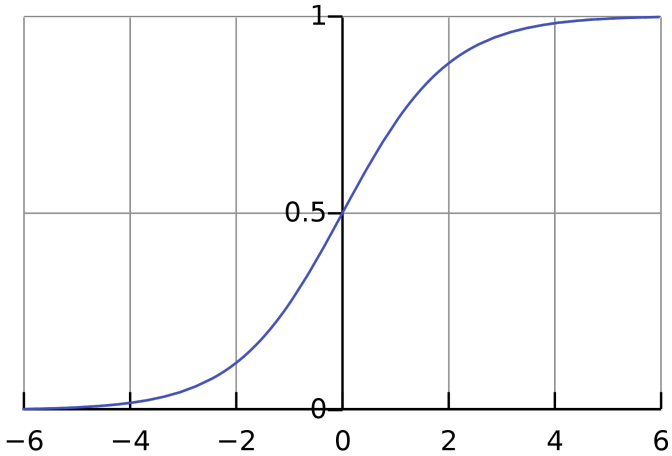
- La regresión logística busca la correlación de la probabilidad de una variable cualitativa binaria (asumiremos que puede tomar los valores reales "0" y "1") con una variable escalar x

- Aproxima la probabilidad de obtener "0" (no ocurre cierto suceso) o "1" (ocurre el suceso) con el valor de la variable explicativa x.

- En esas condiciones, la probabilidad estimada del suceso se aproximará mediante una función logística



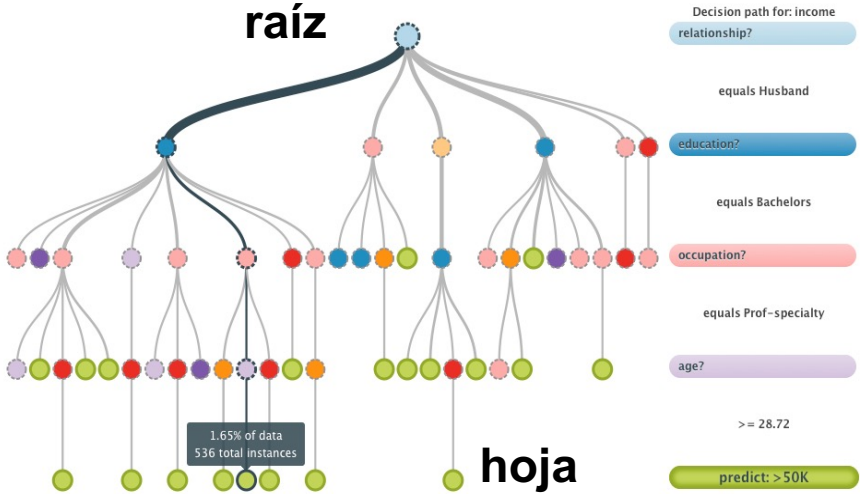
Joseph Berkson

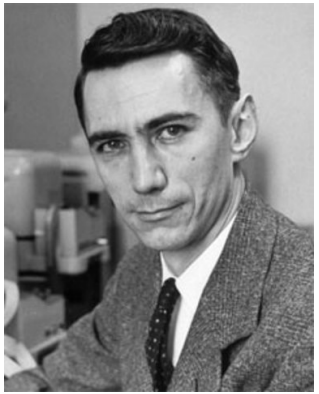


Función Logística

Arboles de Decisión

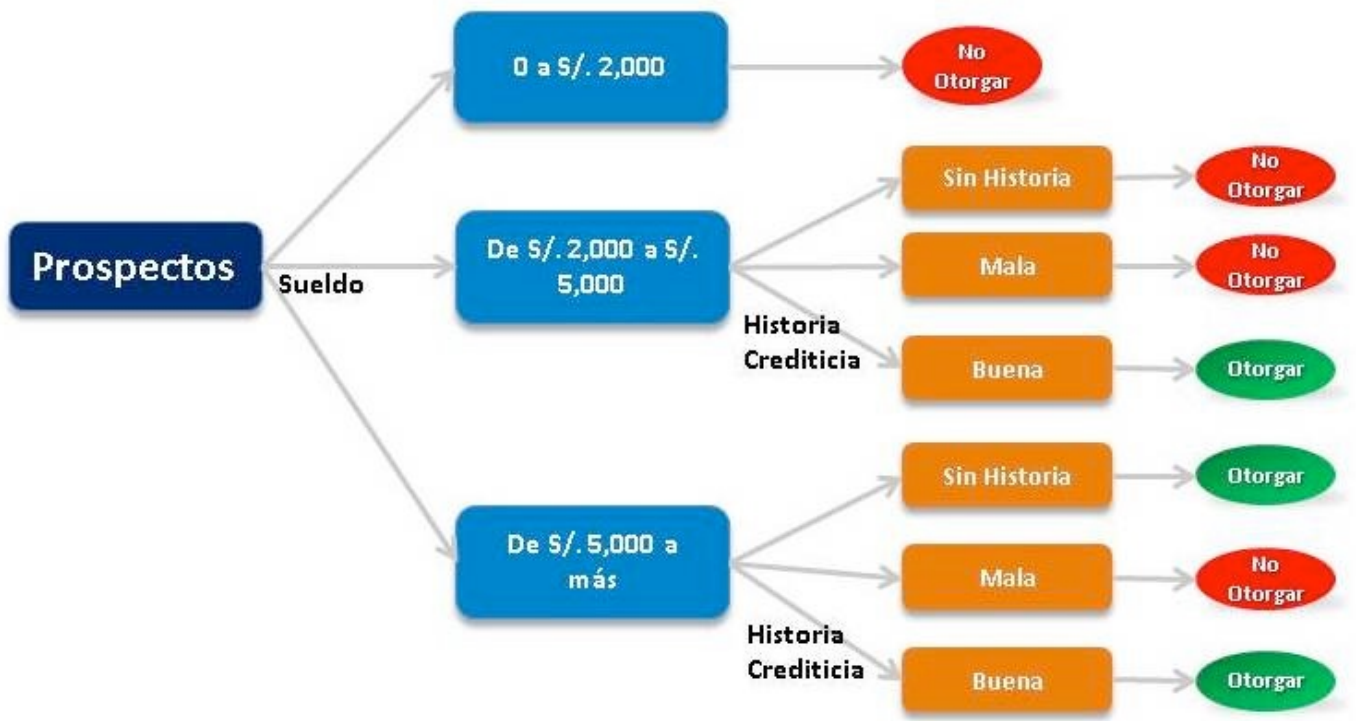
- Un Árbol de Decisión es un conjunto jerárquico de reglas. se construye de arriba hacia abajo desde un nodo raíz e implica la partición de los datos en subconjuntos que contienen instancias con valores similares (homogéneos)
- Fruto de la Teoría de la Decisión de Shannon, su objetivo es determinar estructuras (particiones) en los datos que no se deben al azar





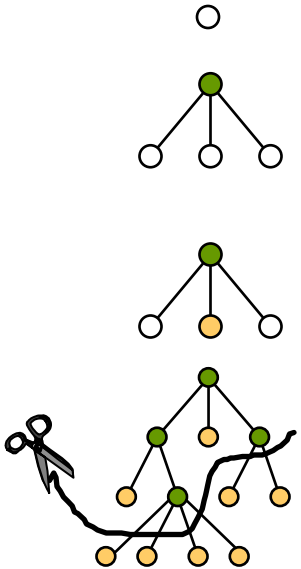
Claude Shannon

Se representan fácilmente por medio de reglas



Construcción de un Arbol de Decisión

- Los algoritmos que construyen un árbol de decisión comienzan con todo el juego de entrenamiento desde los nodos raíz, proceden iterando sobre los nodos que no tienen clases o reglas asignadas

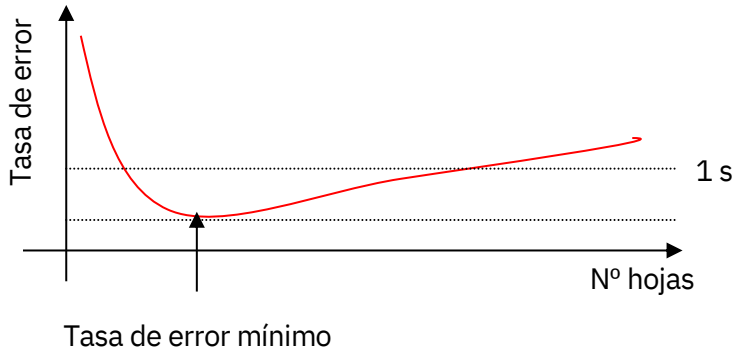


- Puntuar cada uno del conjunto de posibles divisiones de los datos en el nodo *t* utilizando una medida de ajuste. Asignar a la mejor división el grado de regla en *t*
- Crear nodos hijos para cada resultado de la regla y dividir los datos de acuerdo a las reglas
- Si todos los ejemplos de entrenamiento en el nodo *t* pertenecen a la misma clase o si no hay valor en seguir con la división, utilizar una regla simple como la mayoría en una votación para asignar la clase a la *hoja* t.

- El árbol se completa cuando todos los nodos hayan sido asignados a clases o reglas. Para minimizar la posibilidad del *sobrejuste*, se “poda” el árbol. El “podado” es una forma de control de complejidad que mejora la generalización/estabilidad reduciendo el *overtraining*

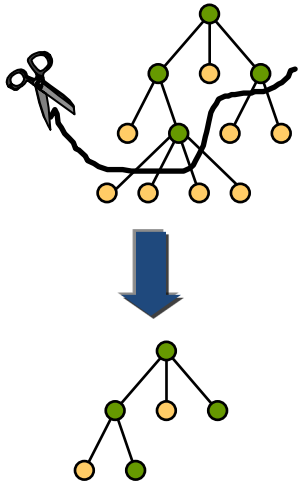
Podado (Pruning)

- Las algoritmos “podan” después de crecer el árbol
- Es un forma de control de complejidad
- Si el árbol es muy frondoso, puede memorizar todo el conjunto de datos
- La mayoría de los árboles más útiles apenas tienen unos niveles de profundidad
- El coste de complejidad del podado evalúa el valor aportado por una nueva subdivisión



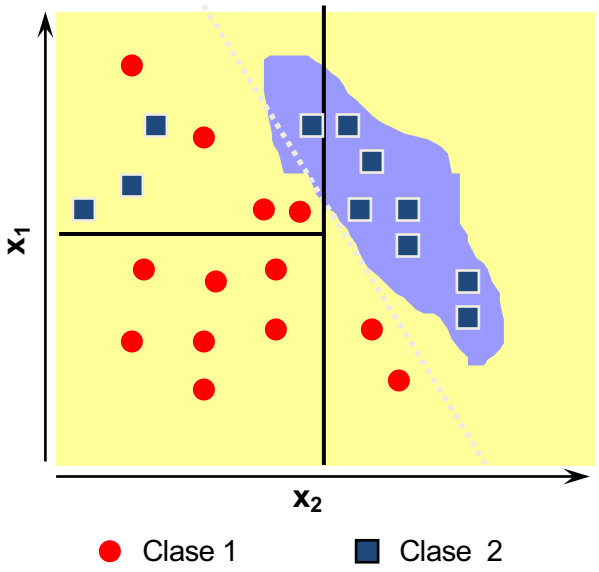
Coste complejidad podado

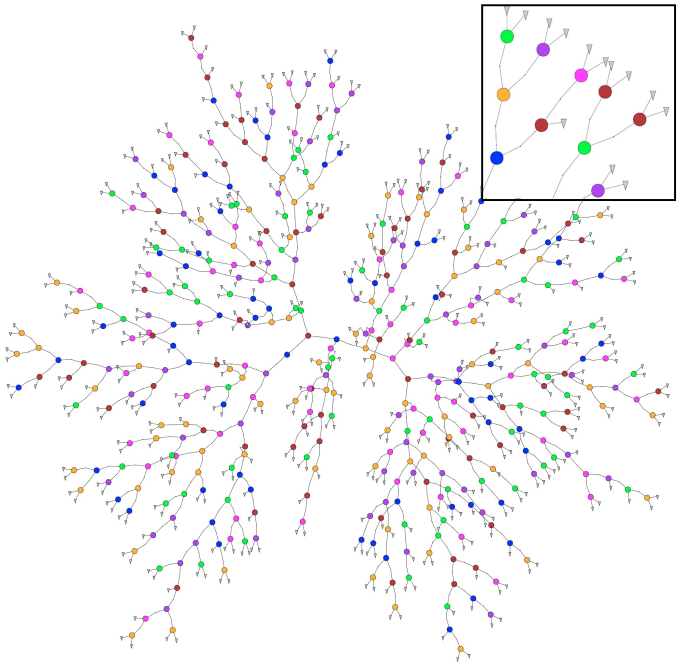
$$\alpha = \frac{Error_{arbolcompleto} - Error_{arbolpodado}}{\text{Tamaño podado - 1}}$$



Otra representación de un árbol de decisión

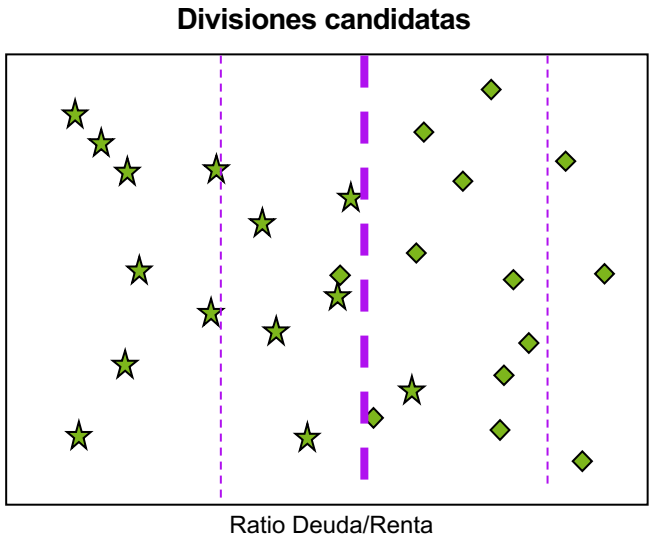
- Cada nodo no hoja divide un dominio e 2 regiones (nodos bidireccionales), cada división es perpendicular a un eje
- En un espacio multidimensional, las divisiones serían hiperplanos o cortes

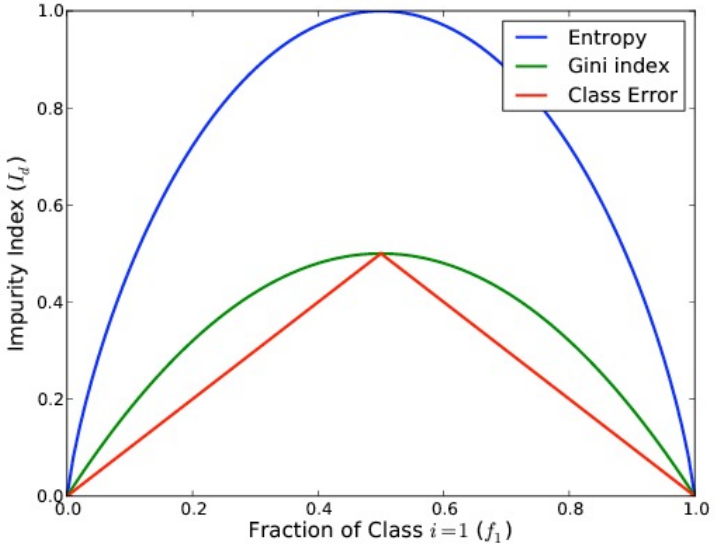




Pureza e Impureza

- La pureza determina la calidad de una clasificación
- Puede medirse con múltiples criterios: Gini, $\chi 2$, Entropía, etc.

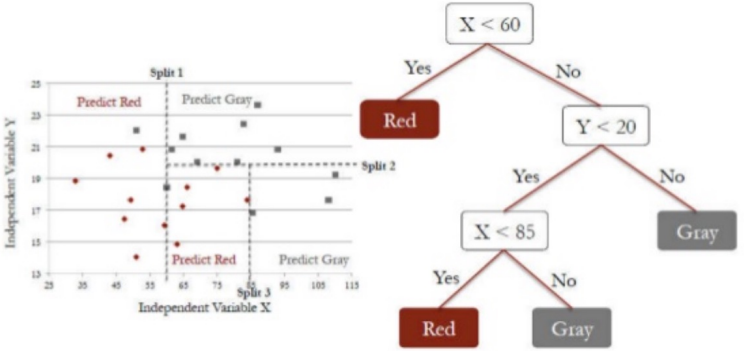




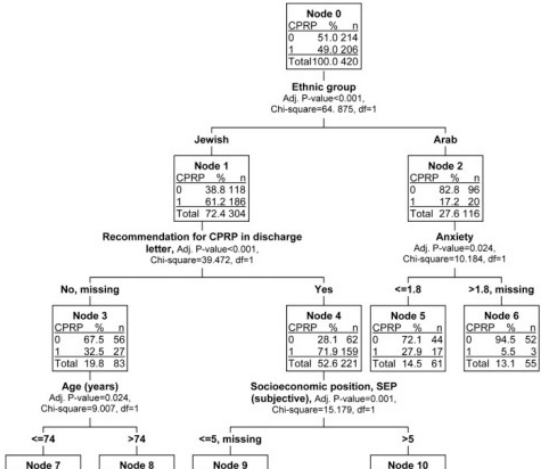
https://www.youtube.com/watch?v=2lEcfRuHFV4

Arboles y Criterios de Pureza

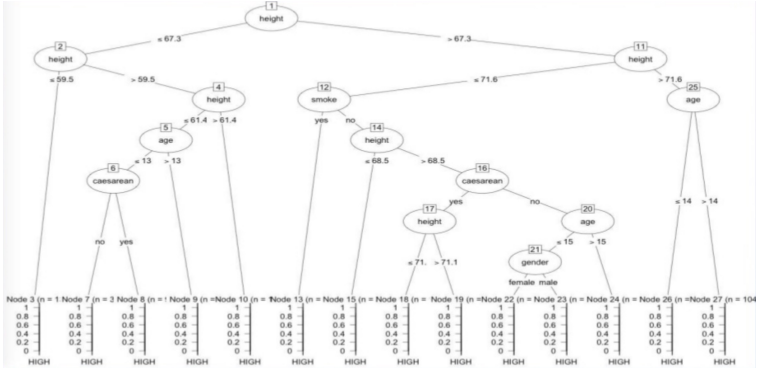
CART (Classification and Regression Trees): utiliza el índice (estadístico) de Gini. Splits dicotómicos



CHAID (Chi Square Automatic Interaction Detection): utiliza tests $\chi 2$



C5.0: emplea el criterio de entropía cruzada



Algoritmia de Aprendizaje No Supervisado
Técnicas de Clustering

Agrupamiento Homogéneo o Clustering

- La técnica de agrupamiento homogéneo o clustering es el proceso de dividir un conjunto de datos en grupos tales que los elementos pertenecientes al mismo grupo sean muy “parecidos”, y los que sean de grupos diferentes sean distintos

- La idea de similitud de los elementos, sin tener conocimiento previo de en qué son parecidos, es un concepto central en los algoritmos de clustering

Distancia Minkowski

$$d_{ij} = \left[\sum_{l=\text{todas variables}} \left| X_{il} - X_{jl} \right|^{2} \right]^{1/k}$$

Cliente*i*

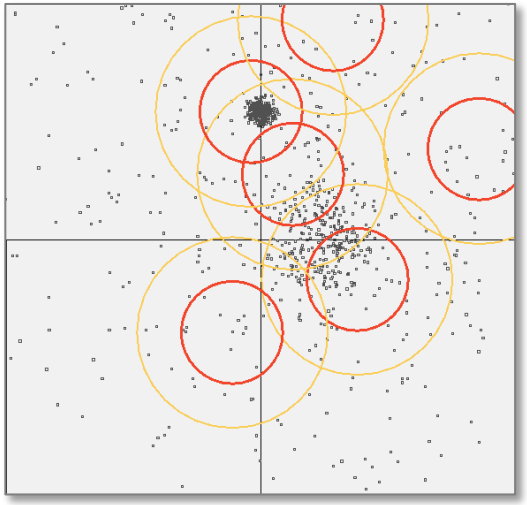
Cliente *j*

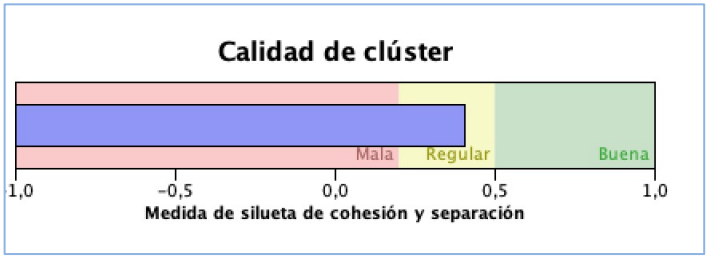
Variable *l* del Cliente *j*

- Euclídea
- Manhattan
- Euclídea Ponderada
- Pearson
- Mahalanobis
- etc…

¿Cómo saber cuándo se ha alcanzado un buen agrupamiento?

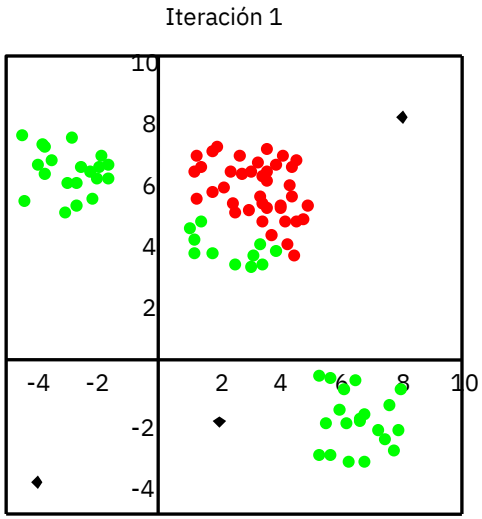
- Podemos medir la calidad del agrupamiento según el grado de separabilidad de los elementos o la compacidad de los grupo
- También podemos exportarlo a Excel y pintar las nubes de “colores” y analizar su densidad





K-means: El concepto básico

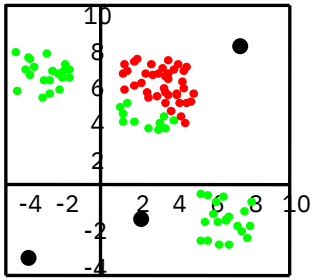
- Este método pide un número predefinido de clusters, e iterativamente asigna registros a clusters ajustando los centroides hasta que ningún refinamiento pueda mejorar el modelo.



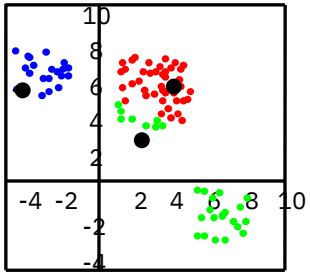
1. Elegir aleatoriamente la posición “media”
2. Agrupar los datos según los valores más recientes de las medias; *cada media posee un conjunto de datos*
3. Mover las medias hacia el centroide del grupo seleccionado

John A. Hartigan

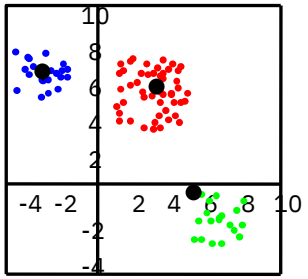
Iteración 1



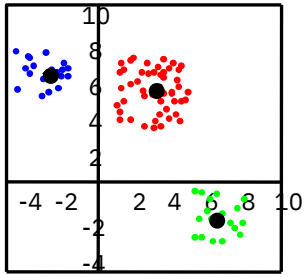
Iteración 2



Iteración 3



Iteración 4



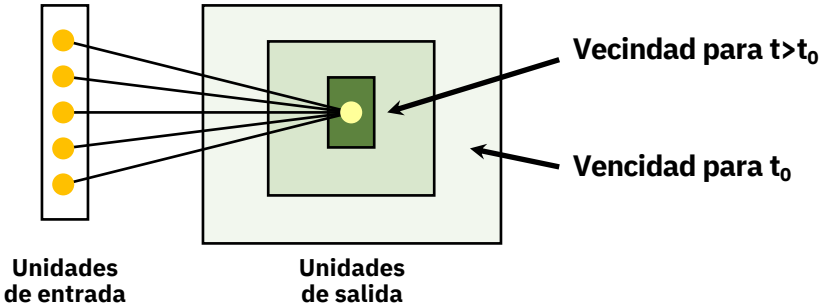
- Trabaja mal con “cúmulos” de clústers
- No es robusto con los outliers

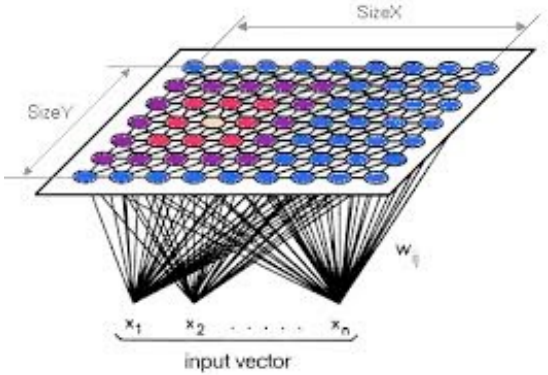
- En las dos últimas iteraciones, la variación de la posición de las medias es muy pequeña
- En esta iteración para el algoritmo.

Mapas Autoorganizados: Redes de Kohonen

- El entrenamiento de una red neuronal de Teuvo Kohonen, está basado en el aprendizaje competitivo
- Cuando la red está entrenada, los registro sque son similares se colocarán juntos en el mapa de salida, mientras que los registros distintos estarán separados
- El ganador de cualquier competición es la neurona más cercana al vector de entrada. Está neurona es la que recibe mayor activación
- Todas las neuronas están conectadas con todas las salidas, por eso es un mapa que acaba autoorganizándose. No existen neuronas ocultas. Se suele representar en un gráfico bidimensional con coordenadas cartesianas
- No precisa de un número de clúster predefinido

Teuvo Kohonen





Práctica: Descripción de un modelo Predictivo

Descripción del Dataset e Identificación de Clústers

¿Preguntas?

Wolfram Rozas