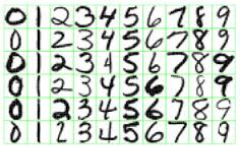
Clasificación Automática con árboles de decisión

Karim Pichara B.
Departamento de Ciencia de la Computación

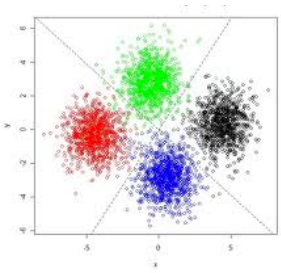
Clasificación Automática

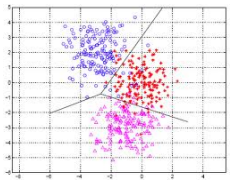
- Un problema de clasificación automática busca encontrar un sistema capaz de identificar para cada objeto la clase a la cual pertenece

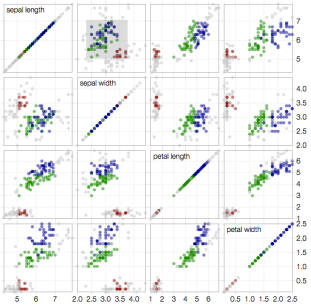


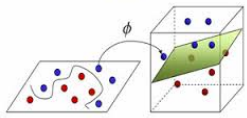
Clasificación Automática

- En el mundo de los datos





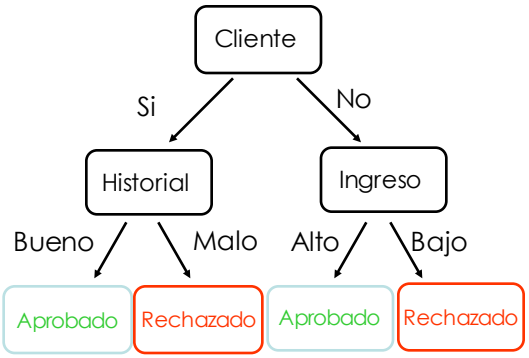




Árboles de Decisión

- Técnica de Clasificación
- Nodos internos del árbol representan atributos
- Cada nodo realiza un test basado en los valores del atributo al cual representa
- Links representan el resultado del test

Ej:



- Los nodos hoja representan el resultado de la clasificación
- Así para clasificar un registro se debe recorrer el árbol desde el nodo raíz a la hoja resultante
- El camino recorrido dependerá de los valores del registro
- Ej. ¿Cuál es la clasificación para el crédito de un cliente con buen historial y alto ingreso?

Consideraciones

- Los árboles de decisión son una técnica de aprendizaje supervisado
- Necesitan de un set de registros pre-clasificados
- Este set es conocido como set de entrenamiento

Consideraciones

- El set de entrenamiento permite ajustar el modelo, es decir:
  - Encontrar una estructura apropiada para el árbol, o en otras palabras:
  - Explorar el espacio de hipótesis buscando un buen clasificador

Cómo determinar el árbol

- Algoritmo básico
  - –Pertenecen todos los registros a la misma clase?
    - Si → Retornar un nodo hoja con la clase respectiva
  - –Tienen todos los registros el mismo valor para todos los atributos que determinan su clase
    - Si → Retornar un nodo hoja con la clase más común

Cómo determinar el árbol

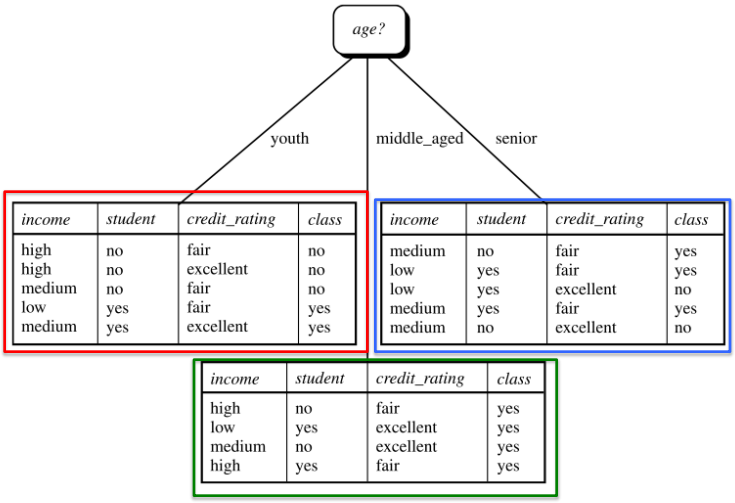
–De lo contrario:

1. Tomar el atributo que mejor separa los registros de las distintas clases
2. Usar ese atributo como nodo raíz
3. Dividir el set de entrenamiento de acuerdo a este atributo y para cada rama resultante continuar la construcción del árbol en forma recursiva

Al bajar por el árbol la base de datos se reduce

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Al bajar por el árbol la base de datos se reduce



income	*student*	*credit_rating*	*class*
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	*student*	*credit_rating*	*class*
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

Criterio de Separación

- ¿Cómo decidir cuál atributo es mejor para separar los registros?
  - –Cada atributo separa los datos en subgrupos
  - –Idealmente cada subgrupo debe ser lo más homogéneo posible
  - –Por tanto, se necesita una métrica de homogeneidad del subgrupo

Criterio de Separación

- Ejemplo:
  - –2 clases: +/-
  - –100 registros (50+ y 50-)
  - –A y B son dos atributos binarios
    - Registros con A=0: 48+, 2-
      Registros con A=1: 2+, 48-
    - Registros con B=0: 26+, 24-
      Registros con B=1: 24+, 26-

Criterio de Separación

- Separar usando A es mejor que separar usando B:
  - A produce una mejor separación de los ejemplos **+** y **–**
  - B produce una pobre separación de los ejemplos **+** y **-**

Entropía

- Entropía es una buena manera de medir homogeneidad
- La entropía mide el número de bits promedio que se necesita para codificar en forma óptima la clase de registros tomados al azar

- Breve paréntesis de teoría de la información:

Entropía (Cont..)

- Entropía:

$$H(S) = -\sum_{c_i} p_i \log_2 p_i$$

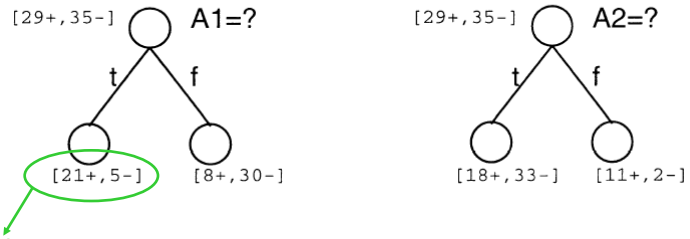
  - H(S) es la entropía del set S
  - c_i son las posibles clases
  - pi = fracción de registros de S que tienen la clase C_i

- Ejemplo de entropía:
  - 3 clases (A,B,C)
  - A ocurre en la mitad de los ejemplos
  - B y C ocurren en un ¼ de los ejemplos
  - Codificación óptima: A = 0, B = 10, C = 11
  - Entropía = número de bits promedio/registro = 1.5 bits

Ganancia de la información

La ganancia de información es la reducción esperada en entropía al separar según cierto atributo, digamos A:

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$



Proporciones en clase objetivo