Investigación de Operaciones
Árboles de Clasificación

Nicolás Rojas Morales
nicolas.rojasm@usm.cl

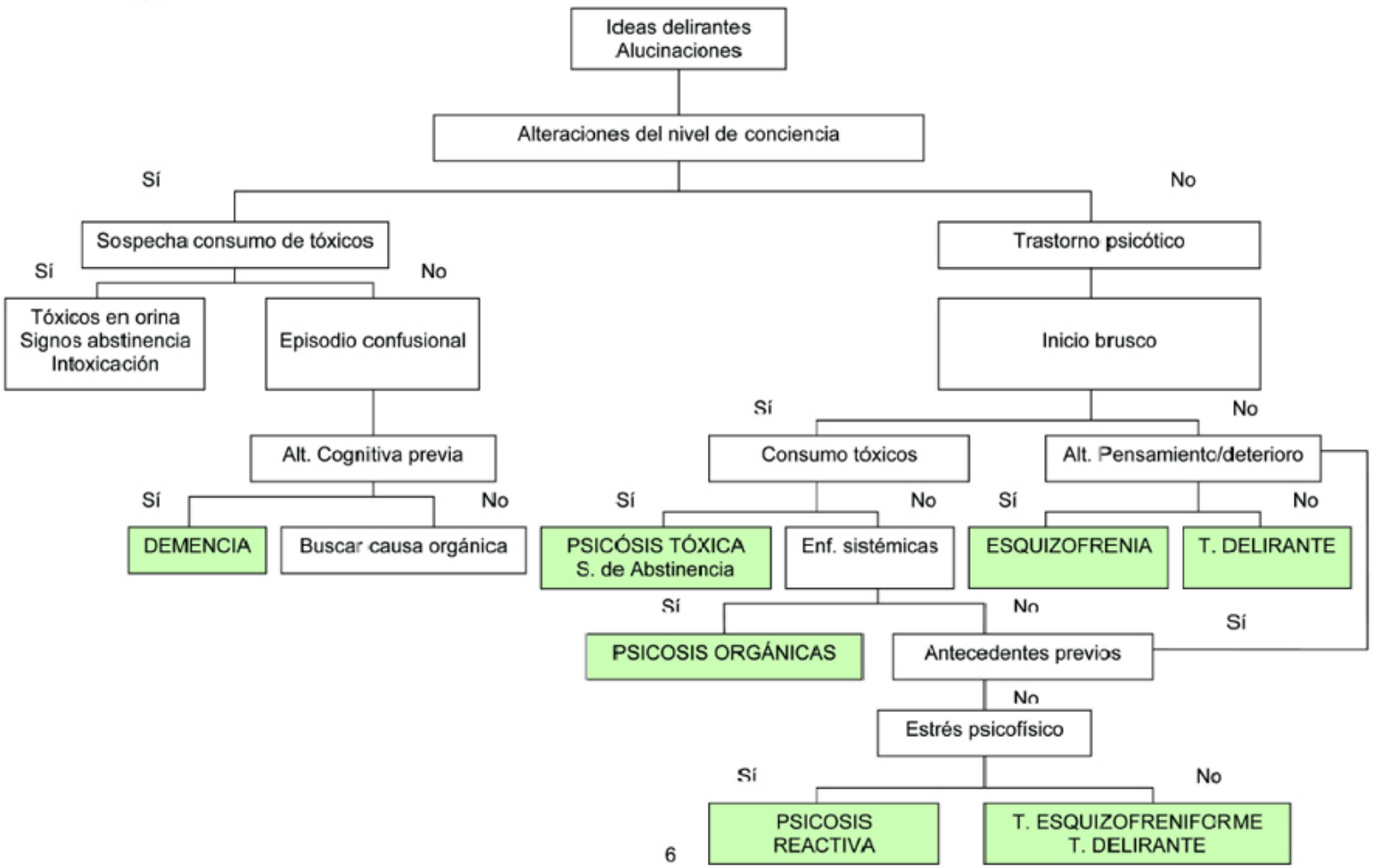
Departamento de Informática
Universidad Técnica Federico Santa María

Introducción

Introducción

Determinar una enfermedad

Tabla 2



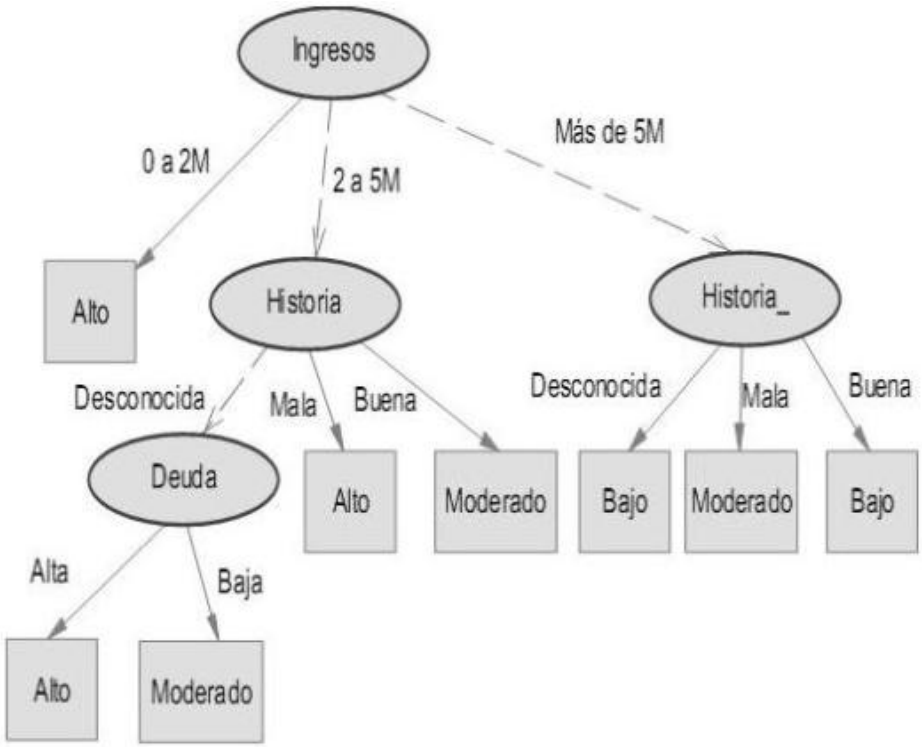
6

Introducción

Problema del Riesgo de Crédito (o Credit Scoring)

Introducción

Problema del Riesgo de Crédito (o Credit Scoring)



link: http://www.redalyc.org/html/750/75040605/

Introducción

Para determinar si un estudiante obtendrá una buena calificación en el próximo control de IO: ¿Qué variables podríamos determinar?

Introducción

Para determinar si un estudiante obtendrá una buena calificación en el próximo control de IO: ¿Qué variables podríamos determinar?

- Nivel de motivación con la asignatura

Introducción

Para determinar si un estudiante obtendrá una buena calificación en el próximo control de IO: ¿Qué variables podríamos determinar?

- Nivel de motivación con la asignatura
- Veces que ha Tomado el Ramo (VTR)

Introducción

Para determinar si un estudiante obtendrá una buena calificación en el próximo control de IO: ¿Qué variables podríamos determinar?

- Nivel de motivación con la asignatura
- Veces que ha Tomado el Ramo (VTR)
- Promedio notas en asignaturas del "área"
- Prioridad académica

Introducción

Se traduce en:

$$\mathcal{C} = f(x_1, x_2, \ldots, x_n) + \epsilon \tag{1}$$

donde:

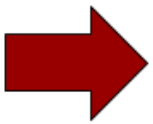
- $\mathcal{C}$ indica si el estudiante obtendrá una calificación Mala, Media, Buena o Excelente.
- x_1 es la Motivación por la Asignatura
- x_2 es el Promedio de Notas ...
- ϵ es el error de la predicción

Introducción

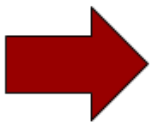
INPUT

Caso	Nombre	Apellido	Promedio	VTR	Motivación	Calificación
1	Juan	Torres	75	1	5	Excelente
2	Antonio	Tapia	55	2	4	Media
3	Raúl	Rivera	35	1	5	Buena
4	María	Reyes	68	1	3	Buena
5	Francisca	Rojas	90	2	5	Excelente
6	Marco	Piñera	25	3	1	Mala

(X)



CLASIFICADOR
f(...)



OUTPUT
Clase
(C)

Introducción

Objetivos:

- Estudiar y crear MODELOS que sean capaces de clasificar/predecir el valor de una variable
- Conocer una herramienta "básica" para predecir/clasificar datos

Introducción

Objetivos:

- Estudiar y crear MODELOS que sean capaces de clasificar/predecir el valor de una variable
- Conocer una herramienta "básica" para predecir/clasificar datos $\rightarrow$ Estudiar más a fondo herramientas del área

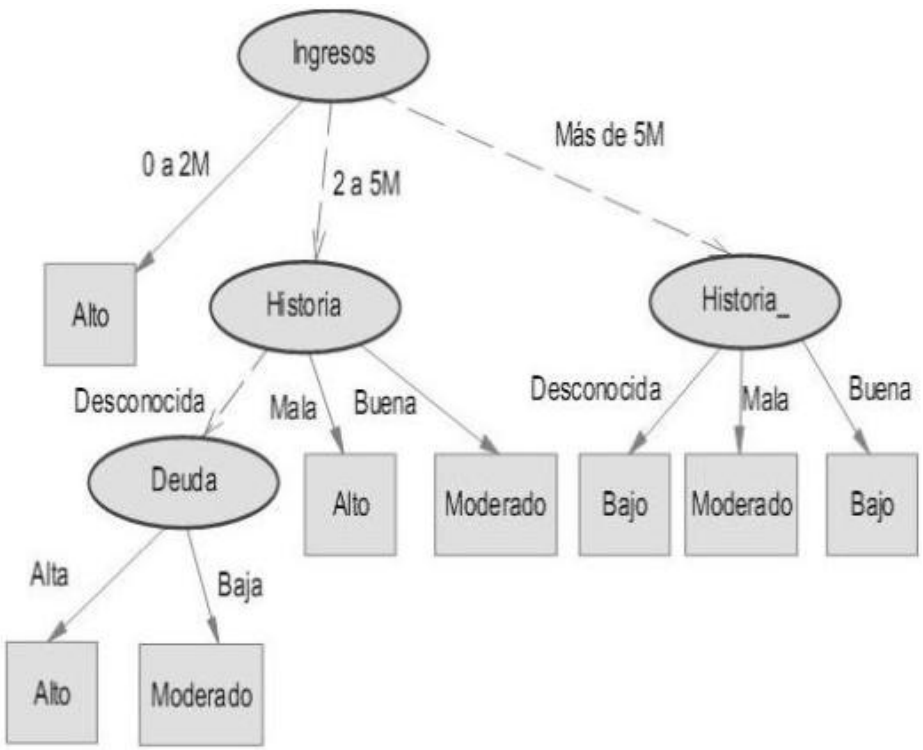
Conceptos Básicos

Conceptos Básicos
Árboles de Clasificación

Conceptos Básicos

Qué es un Árbol de Clasificación?

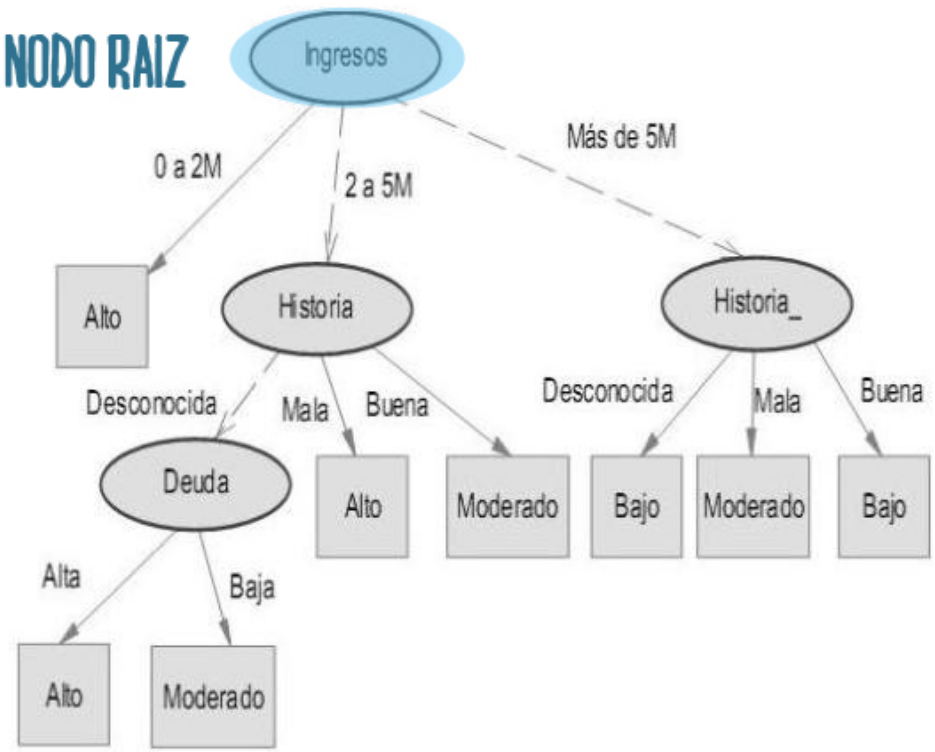
- Método de aprendizaje no paramétrico
- Son descriptivos → representación clara y simple
- Permite *predecir* la clase de un caso → Generalización deseada
- Aplicable en múltiples áreas: Data Mining, Medicina, Inteligencia Artificial, Economía, ...



Conceptos Básicos

Qué es un Árbol de Clasificación?

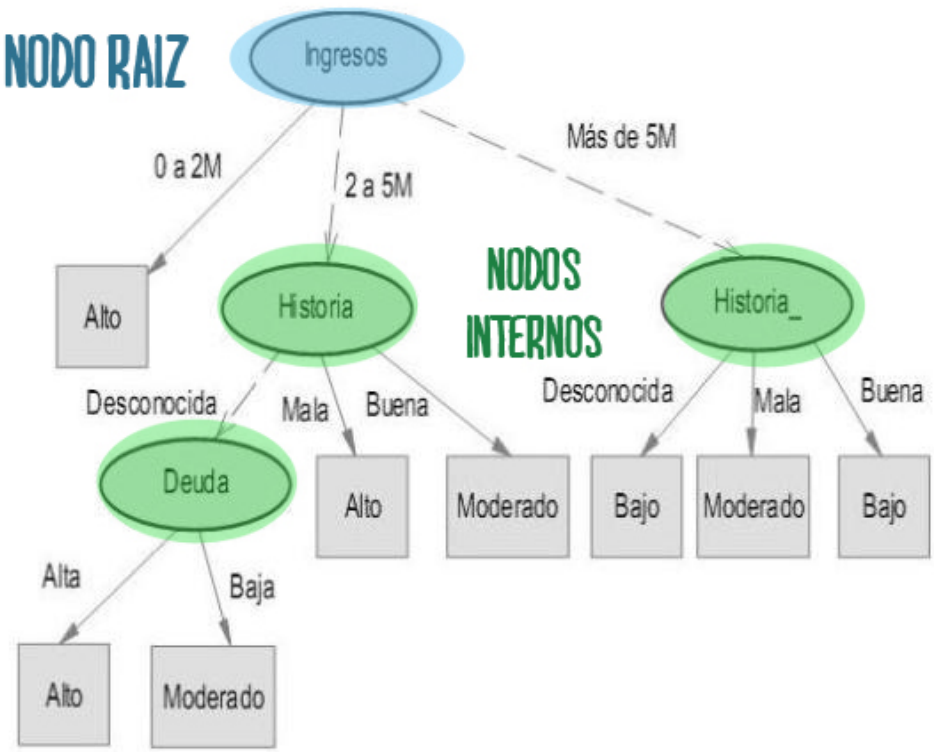
- Método de aprendizaje no paramétrico
- Son descriptivos → representación clara y simple
- Permite *predecir* la clase de un caso → Generalización deseada
- Aplicable en múltiples áreas: Data Mining, Medicina, Inteligencia Artificial, Economía, ...



Conceptos Básicos

Qué es un Árbol de Clasificación?

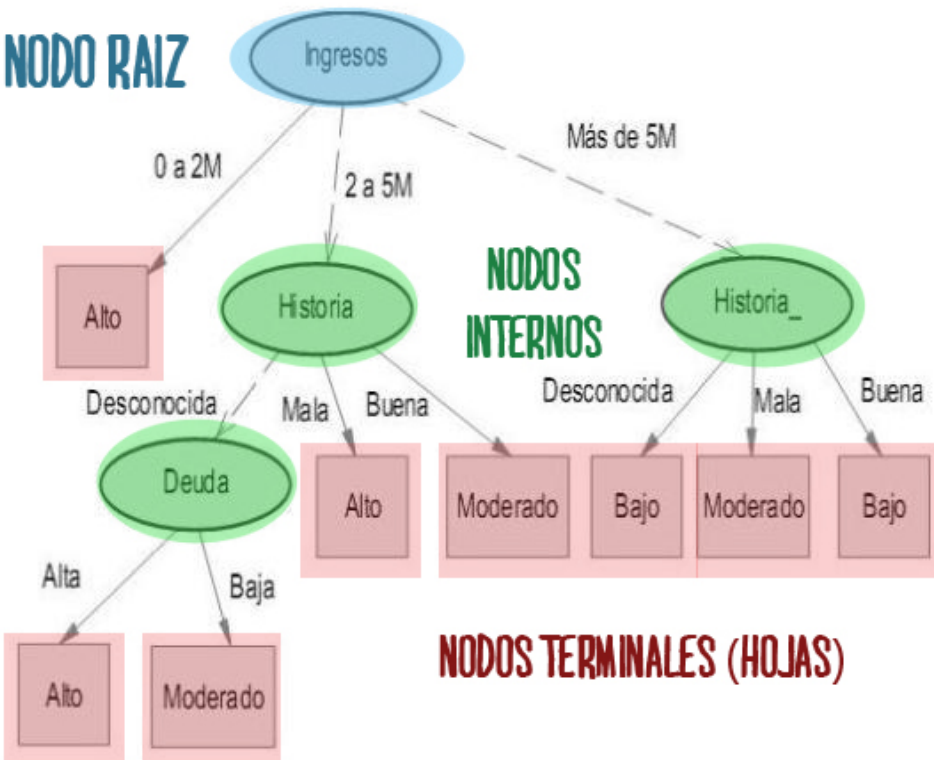
- Método de aprendizaje no paramétrico
- Son descriptivos → representación clara y simple
- Permite *predecir* la clase de un caso → Generalización deseada
- Aplicable en múltiples áreas: Data Mining, Medicina, Inteligencia Artificial, Economía, ...



Conceptos Básicos

Qué es un Árbol de Clasificación?

- Método de aprendizaje no paramétrico
- Son descriptivos → representación clara y simple
- Permite *predecir* la clase de un caso → Generalización deseada
- Aplicable en múltiples áreas: Data Mining, Medicina, Inteligencia Artificial, Economía, ...

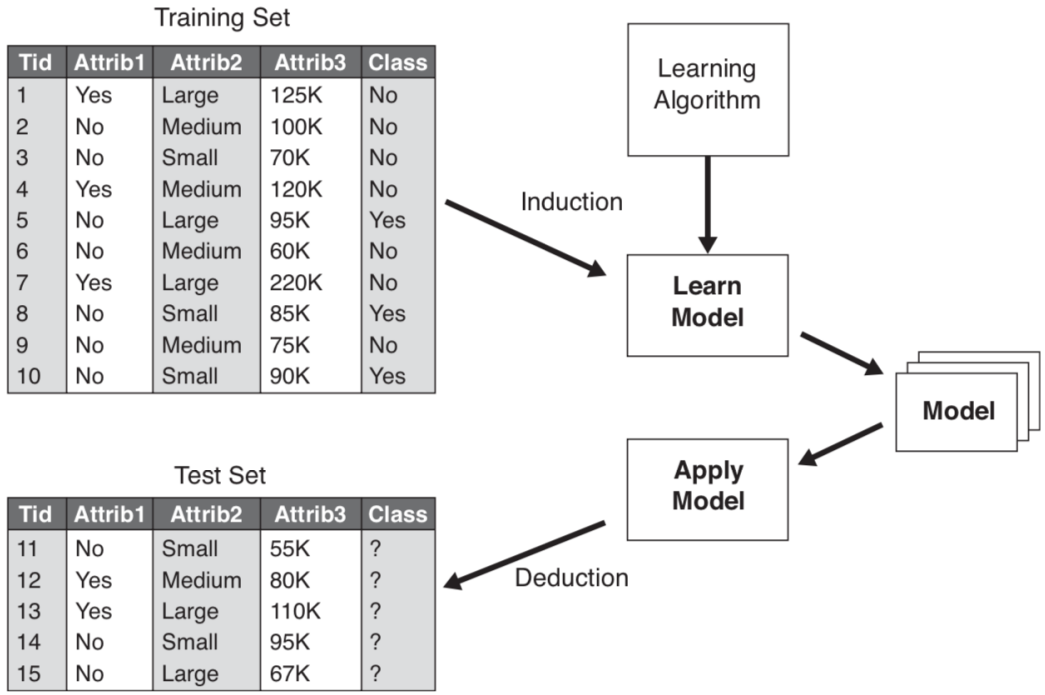


Conceptos Básicos

Conceptos

- Muestra de Datos: Conjunto de casos de dimensión conocida
  - Muestra podría contener ruido (ausencia de algún(os) valor(es)).
- Variables: Pueden ser contínuas, discretas o categóricas. En este escenario, cada variable puede ser Objetivo o Predictiva.
- Variable Objetivo (Target): Atributo discreto/categórico que deseamos predecir a partir de valores asignados a otras variables (Asignar o no un crédito)
- Variables son descritos de a pares (atributo, valor): (VTR,2).

Conceptos Básicos



Conceptos

- Muestra de Aprendizaje (training): Subconjunto de casos utilizados para construir un Árbol de Clasificación (se conocen las clases de cada caso)
- Muestra de Evaluación (testing): Subconjunto de casos utilizados para testear, donde la clase es desconocida (para el árbol) y debe ser determinada.

Conceptos Básicos

Conceptos

- Muestra de Aprendizaje (training): Subconjunto de casos utilizados para construir un Árbol de Clasificación (se conocen las clases de cada caso)
- Muestra de Evaluación (testing): Subconjunto de casos utilizados para testear, donde la clase es desconocida (para el árbol) y debe ser determinada.

¿Cuál es la idea?

Construir un árbol que nos permita clasificar *Individuos*, *Síntomas*, *Animales*, etc...

Conceptos Básicos

Conceptos

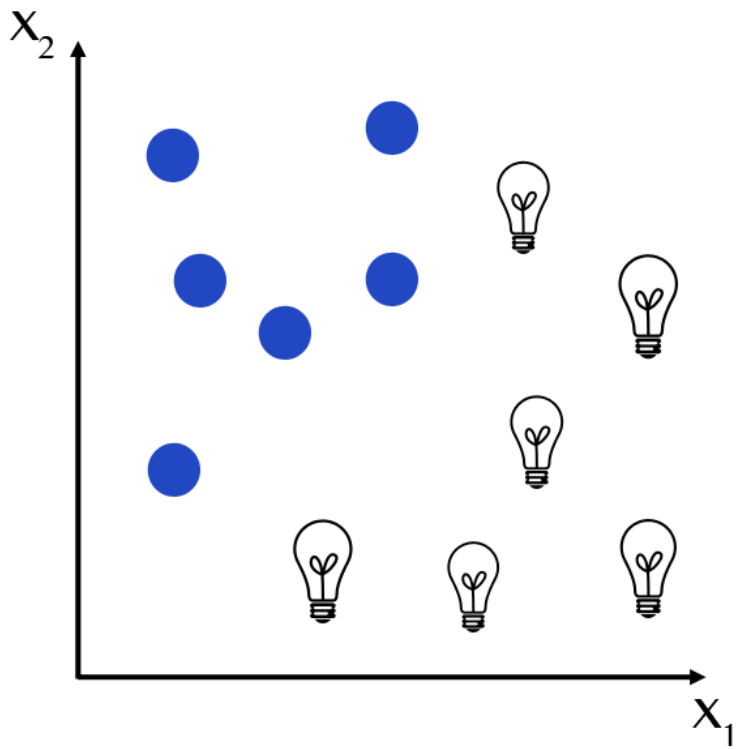
- Muestra de Aprendizaje (training): Subconjunto de casos utilizados para construir un Árbol de Clasificación (se conocen las clases de cada caso)
- Muestra de Evaluación (testing): Subconjunto de casos utilizados para testear, donde la clase es desconocida (para el árbol) y debe ser determinada.

¿Cuál es la idea?

Construir un árbol que nos permita clasificar *Individuos*, *Síntomas*, *Animales*, etc...
¿Cómo lo construímos? . . . Realizamos particiones aplicando reglas (o preguntas)

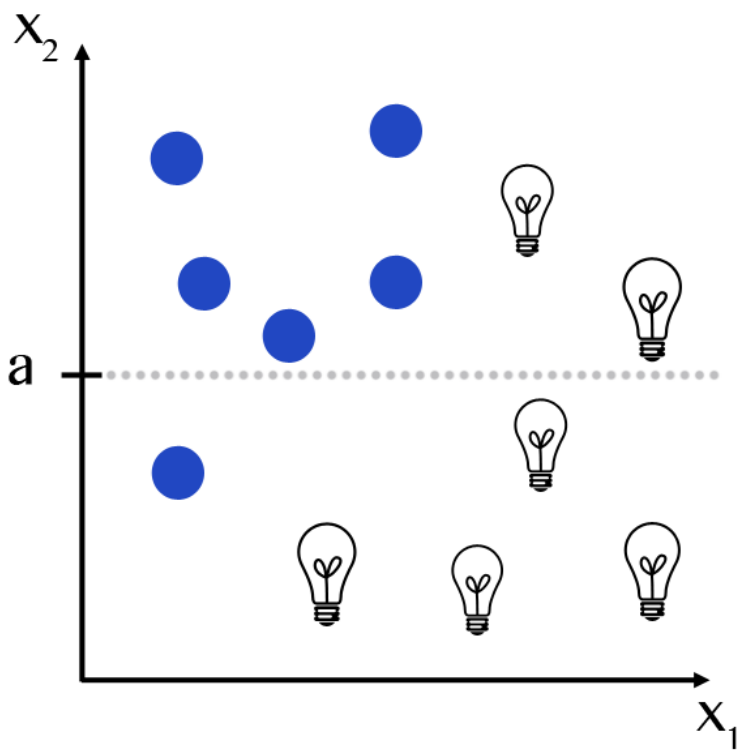
Conceptos Básicos

Construyendo un árbol pt.1



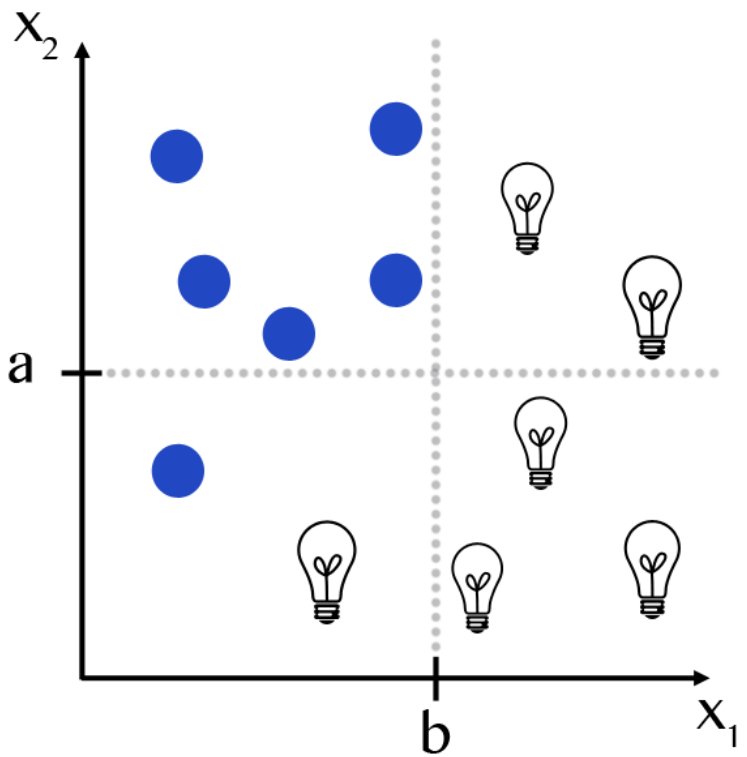
Conceptos Básicos

Construyendo un árbol pt.1



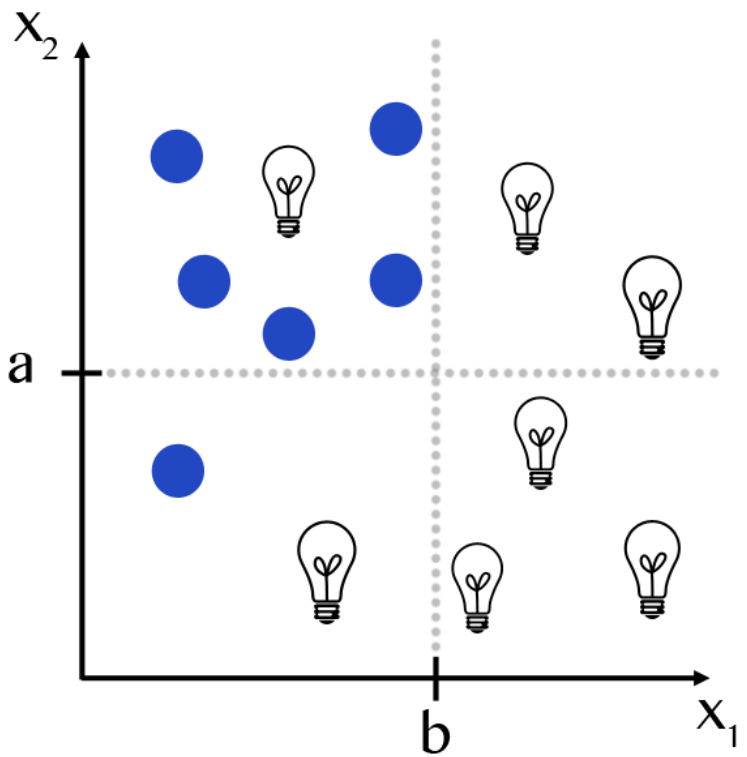
Conceptos Básicos

Construyendo un árbol pt.1



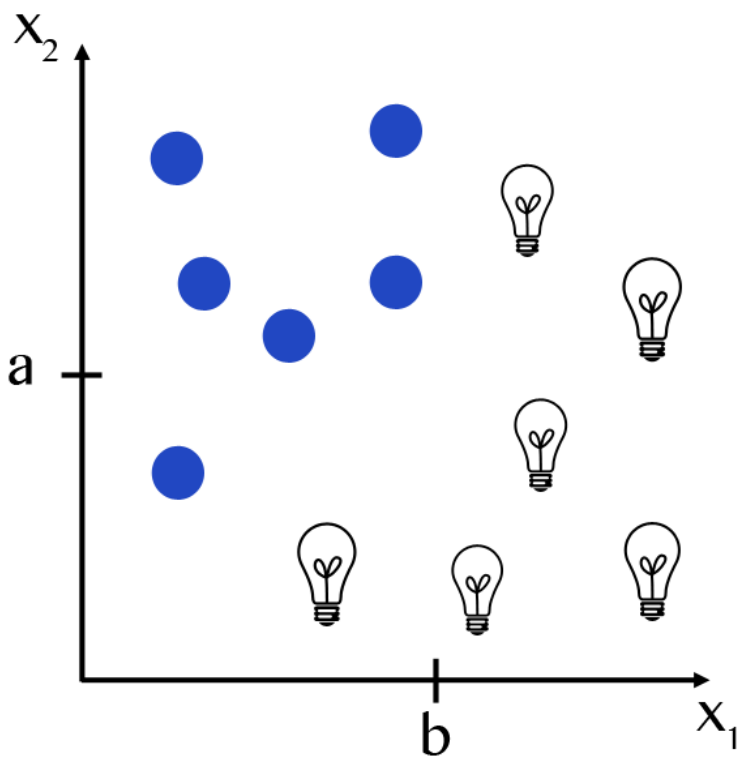
Conceptos Básicos

Construyendo un árbol pt.1



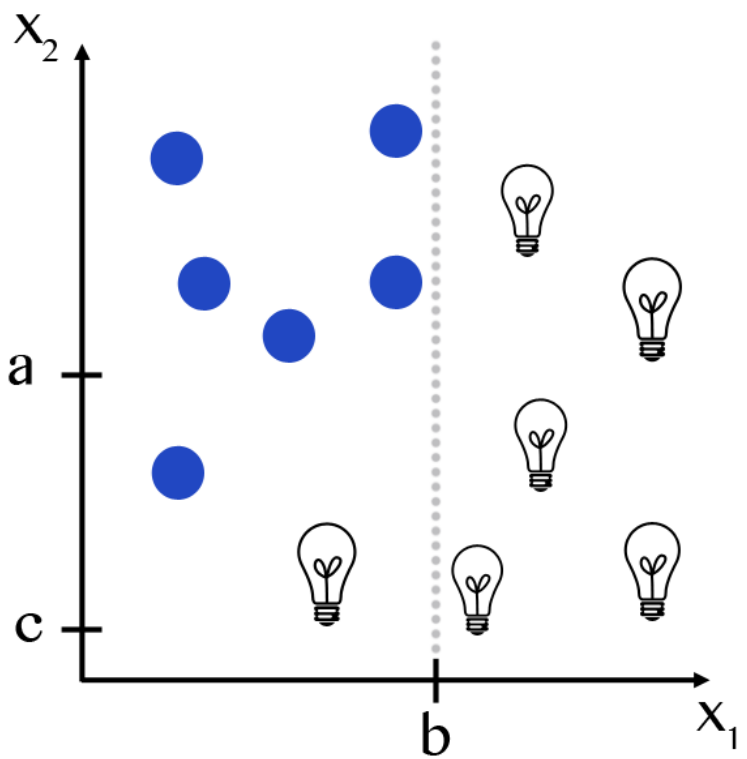
Conceptos Básicos

Construyendo un árbol pt.2



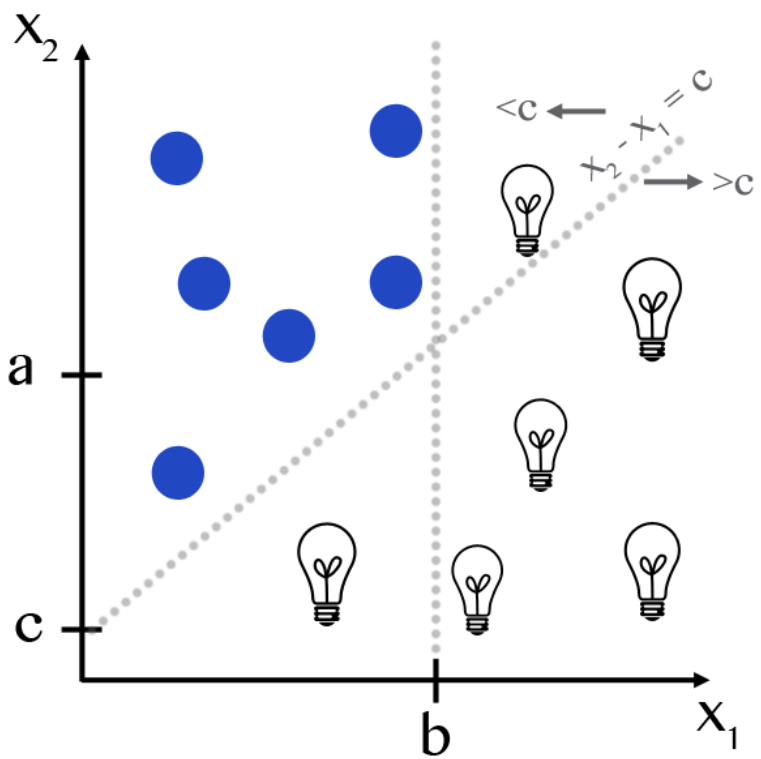
Conceptos Básicos

Construyendo un árbol pt.2



Conceptos Básicos

Construyendo un árbol pt.2



Construcción de Árboles

Construcción de Árboles

- Existe un número exponencial de árboles de clasificación
- Algunos son mejores que otros (bajo diferentes criterios)
- Encontrar el árbol óptimo es computacionalmente infactible (NP Completo).
- Variados algoritmos propuestos para la construcción de árboles de clasificación:
  - CART
  - ID3
  - C4 / C4.5

Construcción de Árboles

¿Cómo construir un árbol?

1. Comenzamos con un nodo raiz con todas las clases y casos
2. Para cada variable, determinar las reglas posibles para particionar los datos
3. Escoger la mejor regla y aplicarla
4. Realizar pasos 2 y 3 hasta cierto criterio de término

Construcción de Árboles

¿Cómo elegir la mejor partición?

- Utilizamos métricas para determinar cuál es la mejor regla para dividir los datos.
- Están definidas por la distribución de los datos, antes y después de aplicar una partición.
- Las métricas usualmente están basadas en el grado de Impureza en los nodos.

Construcción de Árboles

¿Cómo elegir la mejor partición?

Tomamos la decisión basado en la impureza en los nodos:

1. Sean K las clases existentes en los casos (i.e, Círculos y Ampolletas)
2. Sea **t** un nodo y $N(k|t)$ es la cantidad de casos del nodo **t** que pertenecen a la clase **k**
3. Sea $p(k|t)$ la proporción de casos del nodo **t** que pertenecen a la clase **k**:

$$p(k|t) = \frac{N(k|t)}{N(t)} \tag{2}$$

con N(t) la cantidad total de casos en el nodo **t** y $\sum_{k=1}^{N} p(k|t) = 1$.

Nota: En algunos casos, $p(k|t)$ puede ser abreviado como p_k, obviando el nodo t

Construcción de Árboles

¿Cómo elegir la mejor partición?

Sea $\mathcal{I}(t)$ la función de impureza:

$$\mathcal{I}(t) = f(p(1|t), p(2|t), \ldots, p(k|t)) \tag{3}$$

- La impureza es máxima cuando todas las clases están igualmente representadas en t
- La impureza es mínima cuando en t sólo hay casos de una sola clase (máxima homogeneidad)

Construcción de Árboles

Podemos utilizar las siguientes funciones de impureza:
Entropía de Shannon:

$$\mathcal{I}(t) = -\sum_{k=1}^{N} p(k|t) * \log_2(p(k|t)) \tag{4}$$

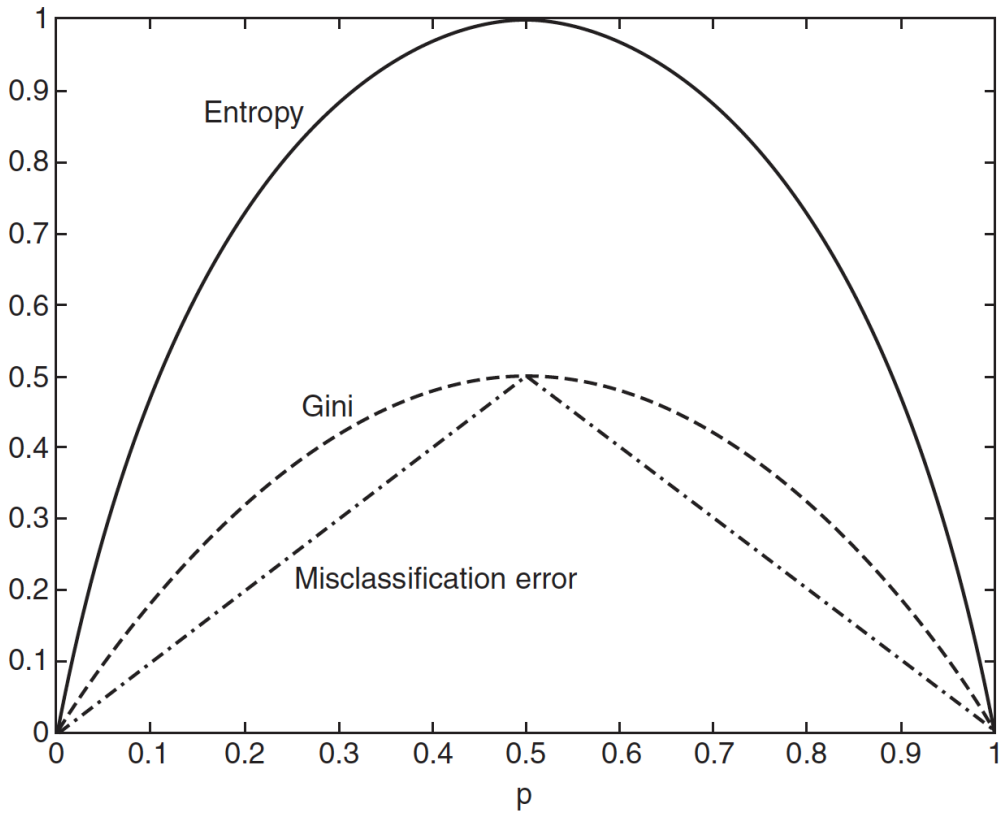
Índice de Gini:

$$\mathcal{I}(t) = 1 - \sum_{k=1}^{N} [p(k|t)]^2 \tag{5}$$

Misclassificacion Error:

$$\mathcal{I}(t) = 1 - max_k[p(k|t)] \tag{6}$$

Construcción de Árboles



Comparación entre medidas de Impureza para un problema de clasificación binaria

Construcción de Árboles

Analizamos cada partición posible para escoger la mejor opción.
Evaluamos el efecto de aplicar una regla mediante la Bondad de la partición:

$$\Delta\mathcal{I}(t,s) = \mathcal{I}(t) - \sum_{j=1}^{J}[p_j * \mathcal{I}(j)] \tag{7}$$

- t el nodo padre
- s la partición/condición evaluada
- J los nodos hijos de t
- p_j la proporcion de elementos en el nodo hijo j

Elegir aquella partición que maximiza $\Delta\mathcal{I}(t,s)$

Construcción de Árboles

¿Cuándo dejamos de aplicar particiones?

Decidimos si hacer un nodo terminal o aplicar una nueva división. Algunos criterios simples son:

- Porcentaje de la clase dominante:

$$k(t) = k, ssi : p(k|t) = max_{j=1...K}\{p(j|t)\} \tag{8}$$

  - En caso de empate, se elige de manera aleatoria
- Disminución de la impureza, menor que un cierto parámetro determinado:
  - Parámetro con valor bajo genera muchas particiones
  - Parámetro con valor alto genera pocas particiones

Construcción de Árboles

Construir un árbol con la siguiente muestra de datos

Caso	Clima	Temperatura	Humedad	Viento	Jugar
1	Soleado	Caluroso	Alta	Débil	No
2	Soleado	Caluroso	Alta	Fuerte	No
3	Nublado	Caluroso	Alta	Débil	Si
4	Lluvioso	Medio	Alta	Débil	Si
5	Lluvioso	Frío	Normal	Débil	Si
6	Lluvioso	Frío	Normal	Fuerte	No
7	Nublado	Frío	Normal	Fuerte	Si
8	Soleado	Medio	Alta	Débil	No
9	Soleado	Frío	Normal	Débil	Si
10	Lluvioso	Medio	Normal	Débil	Si
11	Soleado	Medio	Normal	Fuerte	Si
12	Nublado	Medio	Alta	Fuerte	Si
13	Nublado	Caluroso	Normal	Débil	Si
14	Lluvioso	Medio	Alta	Fuerte	No

Construcción de Árboles

Nodo Raíz - Primera Partición

Caso	Clima	Temperatura	Humedad	Viento	Jugar
1	Soleado	Caluroso	Alta	Débil	No
2	Soleado	Caluroso	Alta	Fuerte	No
3	Nublado	Caluroso	Alta	Débil	Si
4	Lluvioso	Medio	Alta	Débil	Si
5	Lluvioso	Frío	Normal	Débil	Si
6	Lluvioso	Frío	Normal	Fuerte	No
7	Nublado	Frío	Normal	Fuerte	Si
8	Soleado	Medio	Alta	Débil	No
9	Soleado	Frío	Normal	Débil	Si
10	Lluvioso	Medio	Normal	Débil	Si
11	Soleado	Medio	Normal	Fuerte	Si
12	Nublado	Medio	Alta	Fuerte	Si
13	Nublado	Caluroso	Normal	Débil	Si
14	Lluvioso	Medio	Alta	Fuerte	No

Construcción de Árboles

Nodo Raíz - Segunda Partición

Caso	Clima	Temperatura	Humedad	Viento	Jugar
1	Soleado	Caluroso	Alta	Débil	No
2	Soleado	Caluroso	Alta	Fuerte	No
3	Nublado	Caluroso	Alta	Débil	Si
4	Lluvioso	Medio	Alta	Débil	Si
5	Lluvioso	Frío	Normal	Débil	Si
6	Lluvioso	Frío	Normal	Fuerte	No
7	Nublado	Frío	Normal	Fuerte	Si
8	Soleado	Medio	Alta	Débil	No
9	Soleado	Frío	Normal	Débil	Si
10	Lluvioso	Medio	Normal	Débil	Si
11	Soleado	Medio	Normal	Fuerte	Si
12	Nublado	Medio	Alta	Fuerte	Si
13	Nublado	Caluroso	Normal	Débil	Si
14	Lluvioso	Medio	Alta	Fuerte	No

Construcción de Árboles

Caso	Clima	Temperatura	Humedad	Viento	Jugar
1	Soleado	Caluroso	Alta	Débil	No
2	Soleado	Caluroso	Alta	Fuerte	No
3	Nublado	Caluroso	Alta	Débil	Si
4	Lluvioso	Medio	Alta	Débil	Si
5	Lluvioso	Frío	Normal	Débil	Si
6	Lluvioso	Frío	Normal	Fuerte	No
7	Nublado	Frío	Normal	Fuerte	Si
8	Soleado	Medio	Alta	Débil	No
9	Soleado	Frío	Normal	Débil	Si
10	Lluvioso	Medio	Normal	Débil	Si
11	Soleado	Medio	Normal	Fuerte	Si
12	Nublado	Medio	Alta	Fuerte	Si
13	Nublado	Caluroso	Normal	Débil	Si
14	Lluvioso	Medio	Alta	Fuerte	No

Construcción de Árboles

Construir un árbol con la siguiente muestra de datos

Caso	Clima	Temperatura	Humedad	Viento	Jugar
1	Soleado	Caluroso	Alta	Débil	No
2	Soleado	Caluroso	Alta	Fuerte	No
3	Nublado	Caluroso	Alta	Débil	Si
4	Lluvioso	Medio	Alta	Débil	Si
5	Lluvioso	Frío	Normal	Débil	Si
6	Lluvioso	Frío	Normal	Fuerte	No
7	Nublado	Frío	Normal	Fuerte	Si
8	Soleado	Medio	Alta	Débil	No
9	Soleado	Frío	Normal	Débil	Si
10	Lluvioso	Medio	Normal	Débil	Si
11	Soleado	Medio	Normal	Fuerte	Si
12	Nublado	Medio	Alta	Fuerte	Si
13	Nublado	Caluroso	Normal	Débil	Si
14	Lluvioso	Medio	Alta	Fuerte	No

Construcción de Árboles

Caso	Clima	Temperatura	Humedad	Viento	Jugar
1	Soleado	Caluroso	Alta	Débil	No
2	Soleado	Caluroso	Alta	Fuerte	No
8	Soleado	Medio	Alta	Débil	No
9	Soleado	Frío	Normal	Débil	Si
11	Soleado	Medio	Normal	Fuerte	Si

Construcción de Árboles

Caso	Clima	Temperatura	Humedad	Viento	Jugar
4	Lluvioso	Medio	Alta	Débil	Si
5	Lluvioso	Frío	Normal	Débil	Si
6	Lluvioso	Frío	Normal	Fuerte	No
10	Lluvioso	Medio	Normal	Débil	Si
14	Lluvioso	Medio	Alta	Fuerte	No

Construcción de Árboles

Errores de Clasificación

Puede que no todos los elementos sean bien clasificados con el árbol construido.

- Árbol demasiado específico y complejo (por ejemplo: Un caso por hoja) $\rightarrow$ Overfitting
- Árbol demasiado simple (por ejemplo: Sólo dos niveles) $\rightarrow$ Underfitting

Construcción de Árboles

Errores de Clasificación

Puede que no todos los elementos sean bien clasificados con el árbol construido.

- Árbol demasiado específico y complejo (por ejemplo: Un caso por hoja) $\rightarrow$ Overfitting
- Árbol demasiado simple (por ejemplo: Sólo dos niveles) $\rightarrow$ Underfitting

¿Qué hacer para evitar casos de generalización? $\rightarrow$ **Poda**

Construcción de Árboles

Generalización

Cómo enfrentar el problema de generalización:

- Pre-poda: Detener el crecimiento del árbol en su construcción
- Post-poda:
  - Permitir que el árbol crezca libremente: Hasta encontrar un nodo puro o Hasta que el nodo tenga pocos elementos (cantidad a definir).
  - Podar: Selección del mejor árbol podado

Construcción de Árboles

Error

Sea T un árbol y $\bar{T}$ los nodos terminales. $R(T)$ es el estimador del error

$$R(T) = \sum_{t \in \bar{T}} R(t) = \sum_{t \in \bar{T}} r(t)p(t) \tag{9}$$

donde

- $r(t) = 1 - \text{máx}_k\{p(k|t)\}$
- $p(t)$: proporción de casos en el nodo t (nodo padre)

Apuntes útiles:

- https://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf
- http://bit.ly/arbol_io2
- http://www.redalyc.org/html/750/75040605/