

Arboles de Decisión (V)

Carlos Hurtado L.

Depto de Ciencias de la
Computación, Universidad de
Chile

Poda basada en Principio de Descripción Mínima (PDM)

- Si tenemos muchos datos de entrenamiento y estos son representativos, los métodos de poda basados en reducción del error puede generar árboles grandes.
- Necesitamos un criterio de poda que considere compromiso entre tamaño del modelo y error.
- Por ejemplo, si lo que disminuye el error en una rama es muy poco en comparación con la complejidad de la rama podría ser conveniente podar la rama.

Principio de Descripción Mínima (PDM)

- La idea básica es que un modelo nos sirve para codificar los datos.
- El mejor modelo es aquel que minimiza el costo de codificar el modelo $C(M)$ más el costo de codificar los datos usando el modelo $C(D|M)$.
- $C(D|M)$ se puede ver como el error.
- El PDM integra error y complejidad del modelo en una medida única y en bits.

Costo de Descripción

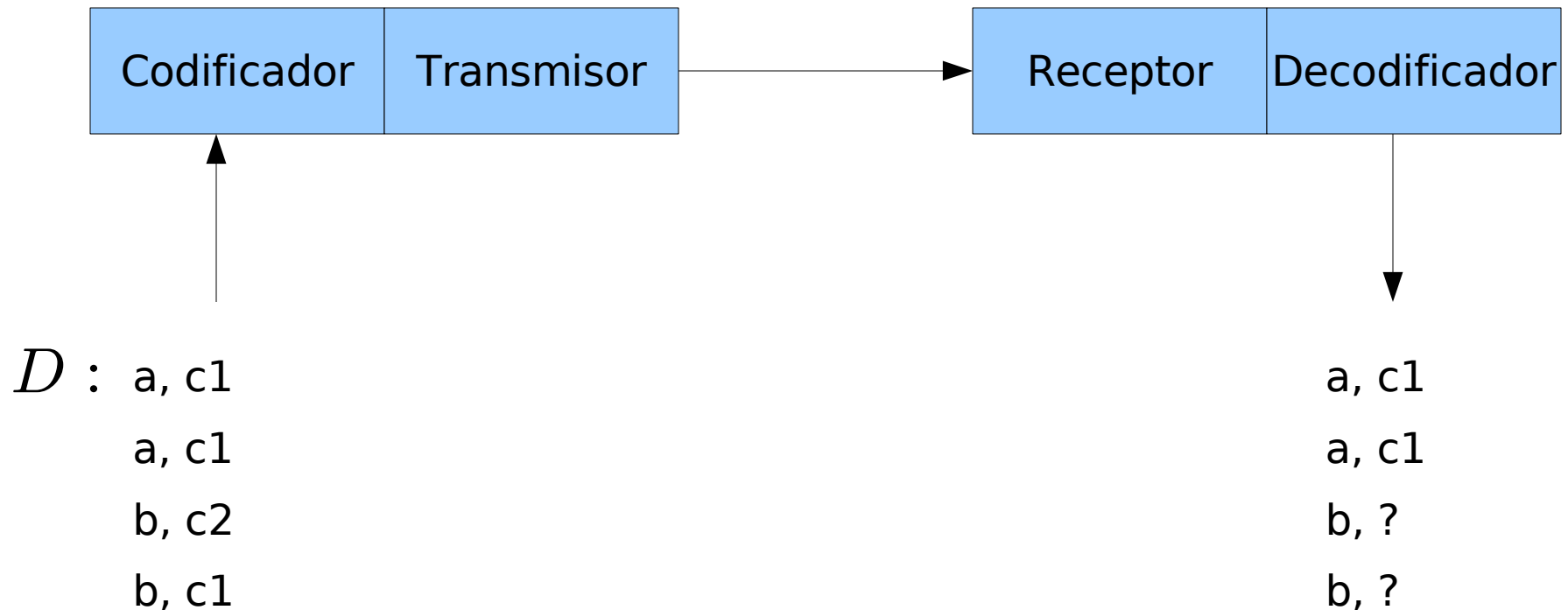
- Mide cantidad de información necesaria para codificar modelo + cantidad de información necesaria para codificar datos usando el modelo.

$$Costo(M, D) = Costo(M) + Costo(D|M)$$

Costo de Descripción

- $\text{Costo}(M)$:
 - Costo de codificación del árbol (como grafo) más costo de codificar los splits
- $\text{Cost}(D|M)$:
 - Enfoque simple: se estima como el costo de codificar los errores.
 - Enfoque avanzado: se estima usando entropía.

Costo de Descripción de un Conjunto de Datos



Costo de Descripción de un Conjunto de Datos

- Para transmitir D debemos transmitir la distribución de las clases P_D y luego las clases de los datos.
- Esto, debido a que el decodificador que requiere el receptor se construye en base a P_D
- Luego, tenemos:

Costo de enviar las clases

$$\text{Costo}(D) = N_D \text{Entropy}(P_D) + \text{CostoCod}(P_D)$$

Costo de codificar la dist. de las Clases

Costo de Codificar la Distribución de las Clases

- Este costo es:

$$\text{CostoCod}(P_D) = \log \binom{N_D + m - 1}{m - 1}$$

donde m es el número de clases y N_D es el número de datos.

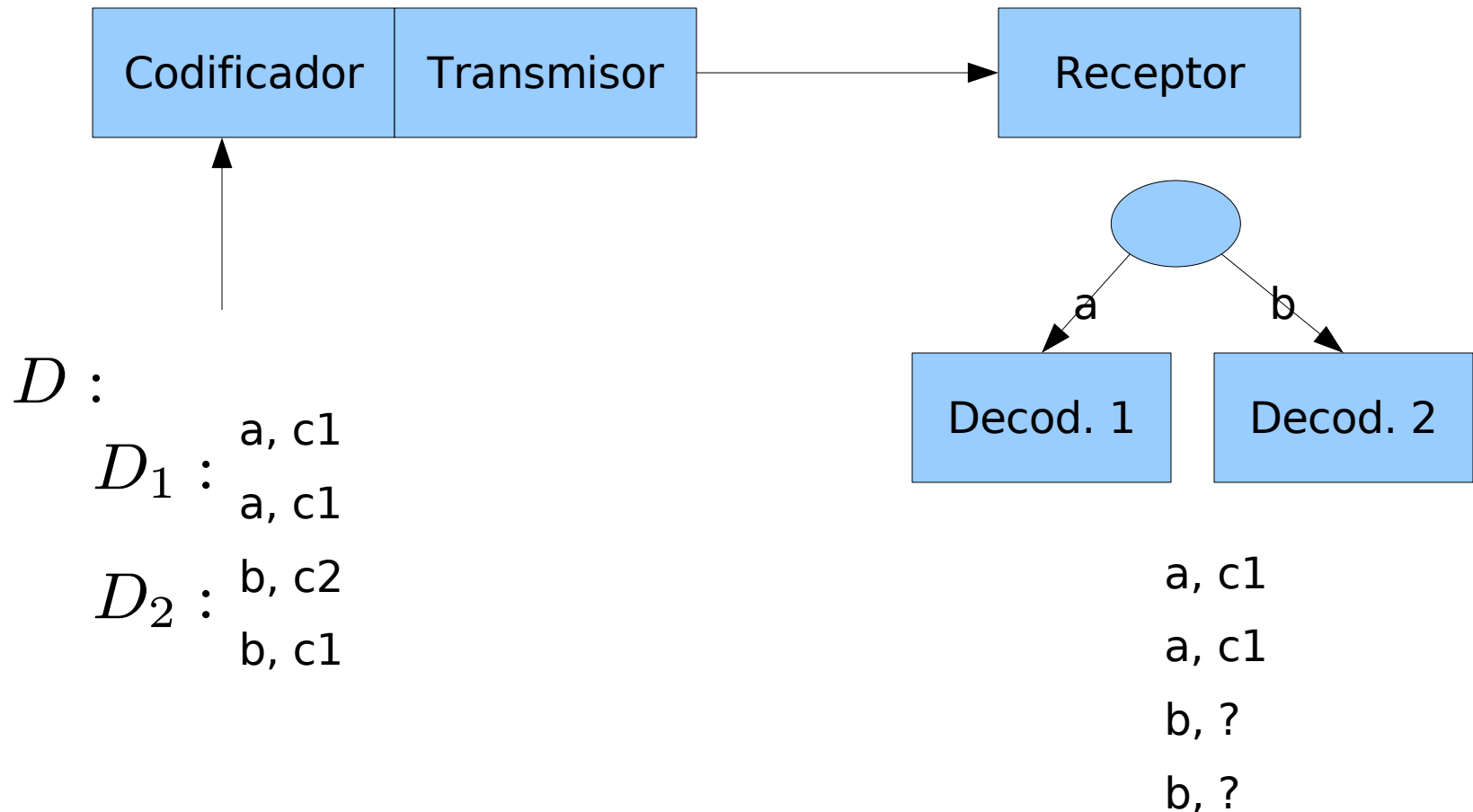
- Esto es equivalente al número de soluciones enteras positivas de la ecuación

$$X_1 + X_2 + \dots + X_m = N_D$$

Costo de Descripción de un Conjunto de datos

$$\textit{Costo}(D) = N_D \textit{Entropy}(D) + \log \binom{N_D + m - 1}{m - 1}$$

Costo de Descripción de un Conjunto de Datos + Split



Costo de descripción Conjunto de Datos + Split

$$\textit{Costo}(D, \textit{split}) = \textit{Costo}(\textit{split}) + \textit{Costo}(D|\textit{Split})$$

$$\textit{Costo}(D|\textit{split}) = \textit{Costo}(D_1) + \textit{Costo}(D_2)$$

Costo de Descripción Split

$$\text{Costo}(\textit{split}) = \log A + \log B$$

donde A es el número de variables y B es el número de posibles splits usando una variable.

Ejemplo: si tenemos 5 variables y consideramos splits simple, tenemos

$$\text{Costo}(\textit{split}) = \log 5 + \log 1$$

Costo de Descripción de un Conjunto de Datos + Arbol

$$\textit{Costo}(D, \textit{arbol}) = \textit{Costo}(\textit{arbol}) + \textit{Costo}(D|\textit{Arbol})$$

$$\textit{Costo}(D|\textit{arbol}) = \sum_i \textit{Costo}(D_i)$$

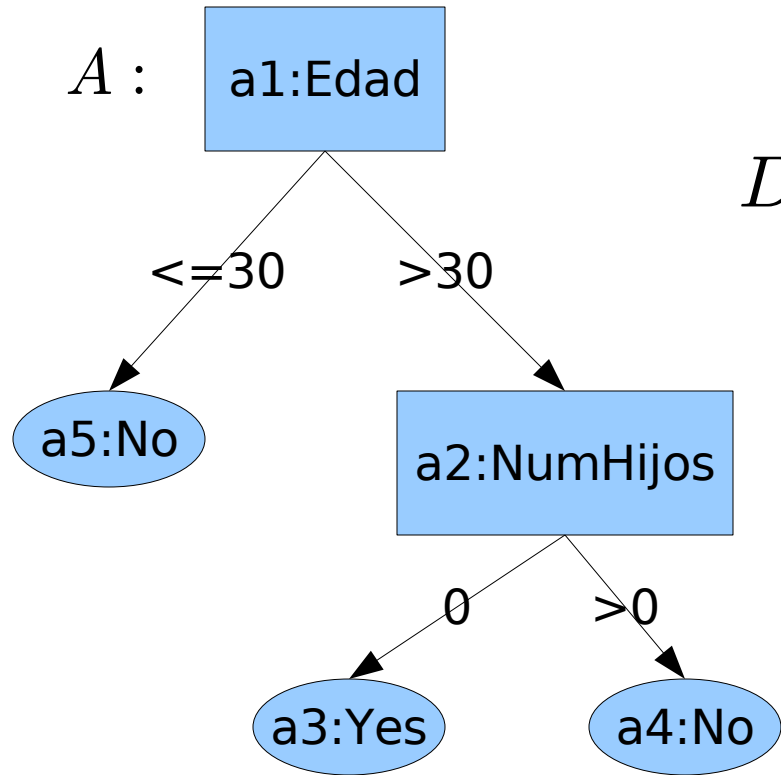
donde cada D_i son los datos de una hoja del árbol

Costo de descripción de un Arbol

- Podemos codificar un árbol de la sgte forma:
 - Lista con la estructura del árbol:
 - Esta es una lista que contiene 1 bits por nodo
 - Supone que se recorre el árbol en pre-orden.
 - El bit de cada nodo dice si es un nodo interior o es hoja.
 - Lista con los códigos de los splits
 - Codificamos cada split
 - La correspondencia de un split a un nodo está dada por el orden de las dos listas.
- Luego tenemos:

$$Costo(arbol) = \sum_i Costo(split_i) + \log P$$

Ejemplo: Dado el arbol A y datos D



$D :$

Nodo	NumYes	Num No
a3	20	5
a4	4	30
a5	4	30

Costo de Descripción Arbol A

$$\text{Costo}(A) = \text{Costo}(\text{split}_1) + \text{Costo}(\text{split}_2) + \log 5$$

Suponemos splits binarios para edad máxima 90 y
núm hijos máximo 10, respectivamente. Suponemos dos
variables.

$$\text{Costo}(\text{split}_1) = \log 2 + \log 90 = 7.49$$

$$\text{Costo}(\text{split}_2) = \log 2 + \log 10 = 4.32$$

Luego:

$$\text{Costo}(A) = 2.32 + 7.49 + 4.32 = 14.13$$

Costo de Descripción Datos dado el Arbol

Nodo	NumYes	Num No
a3	20	5
a4	4	30
a5	4	30

$$Costo(D|A) = Costo(D_3) + Costo(D_4) + Costo(D_5)$$

$$Costo(D_3) = 25 Entropy\left(\left(\frac{5}{25}, \frac{20}{25}\right)\right) + \log \binom{25 + 2 - 1}{2 - 1}$$

$$Costo(D_4) = 34 Entropy\left(\left(\frac{4}{34}, \frac{30}{34}\right)\right) + \log \binom{34 + 2 - 1}{2 - 1}$$

$$Costo(D_5) = 34 Entropy\left(\left(\frac{4}{34}, \frac{30}{34}\right)\right) + \log \binom{34 + 2 - 1}{2 - 1}$$

Costo de Descripción Datos dado el Arbol

Nodo	NumYes	Num No
a3	20	5
a4	4	30
a5	4	30

$$\text{Costo}(D_3) = 250.72 + \log 26 = 22.7$$

$$\text{Costo}(D_4) = 340.52 + \log 35 = 22.81$$

$$\text{Costo}(D_5) = 340.52 + \log 35 = 22.81$$

$$\text{Costo}(D|A) = 22.7 + 22.81 + 22.81 = 68.32$$

Costo de Descripción Datos + Arbol

$$\textit{Costo}(D, A) = \textit{Costo}(A) + \textit{Costo}(D|A)$$

$$\textit{Costo}(D, A) = 14.13 + 68.32 = 82.45$$

¿Conviene podar el Nodo a1 y reemplazarlo por una hoja?

$$\text{Costo}(D, A) = 14.13 + 68.32 = 82.45$$

NumYes	Num No
20	5
4	30
4	30
28	65

$$\text{Costo}(D) = 93 \text{Entropy}\left(\left(\frac{28}{93}, \frac{65}{93}\right)\right) + \log \binom{93 + 2 - 1}{2 - 1}$$

$$\text{Costo}(D) = 82.08 + 6.55 = 88.63$$

Algoritmo de Poda Basado en PDM

- Recorrer el árbol desde las hojas hacia arriba verificando si podar o no (similar a algoritmo de poda basado en reducción del error)
- En un paso del recorrido, sea t el nodo que visitamos, D los datos asociados a t y A el subárbol bajo t (después de podar los nodos bajo t)
- Reemplazamos t por una hoja si:

$$Costo(D) \leq Costo(D, A)$$

Algoritmo de Poda Basado en PDM

- Al calcular $C(D, A)$, podemos usar la siguiente propiedad:

$$Costo(D, A) = Costo(split) + \frac{|D_i|}{|D|} \sum_i Costo(D_i, A_i)$$

- donde los A_i son los subárboles bajo los hijos de t.