Inducción de Árboles de Decisión

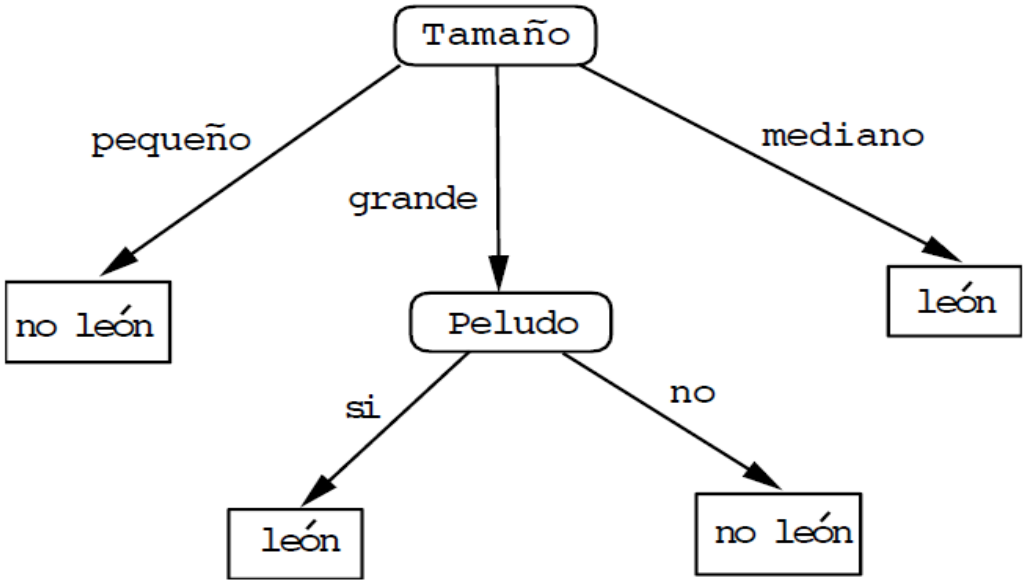
Autores: Eduardo Morales, Hugo Jair Escalante

Intructor: Alberto Reyes

Tareas de Aprendizaje

- **Clasificación**. Los datos son objetos caracterizados por atributos que pertenecen a diferentes clases (etiquetas discretas).
- La meta es inducir un modelo para poder predecir una clase dados los valores de los atributos.
- Se usan, por ejemplo, árboles de decisión, reglas, SVM, etc.

Árbol de Decisión



Método de aprendizaje:

- Ejemplos de entrenamiento y prueba
- Utilización de teoría de la información
- Incrementalmente por medio de “ventanas”

Tabla de Ejemplo

Ambiente	Temp.	Humedad	Viento	Clase
soleado	alta	alta	no	N
soleado	alta	alta	si	N
nublado	alta	alta	no	P
lluvioso	media	alta	no	P
lluvioso	baja	normal	no	P
lluvioso	baja	normal	si	N
nublado	baja	normal	si	P
soleado	media	alta	no	N
soleado	baja	normal	no	P
lluvioso	media	normal	no	P
soleado	media	normal	si	P
nublado	media	alta	si	P
nublado	alta	normal	no	P
lluvioso	media	alta	si	N

Inducción de Árboles de Decisión

Procedimiento de aprendizaje:

1. Junta una gran cantidad de ejemplos
2. Divídelos en dos conjuntos disjuntos: entrenamiento y prueba
3. Usa el algoritmo de aprendizaje para generar una hipótesis H
4. Mide el porcentaje de clasificación correcta de H en el conjunto de prueba
5. Repite los pasos 1 - 4 para diferentes tamaños de conjuntos seleccionados aleatoriamente

Inducción de Árboles de Decisión

- Idea: Probar primero el atributo “más importante”
- Este particiona los ejemplos y cada subconjunto es un nuevo problema con menos ejemplos y un atributo menos.
- Este proceso recursivo tiene 4 posibles resultados:
  1. Si existen ejemplos positivos y negativos, escoge el mejor atributo
  2. Si todos los ejemplos son positivos (o negativos), termina y regresa True (o False)
  3. No quedan ejemplos, regresa un default con base en la clasificación mayoritaria de su nodo padre
  4. No hay más atributos, pero seguimos con ejemplos positivos y negativos. Posible solución: Toma la clase mayoritaria

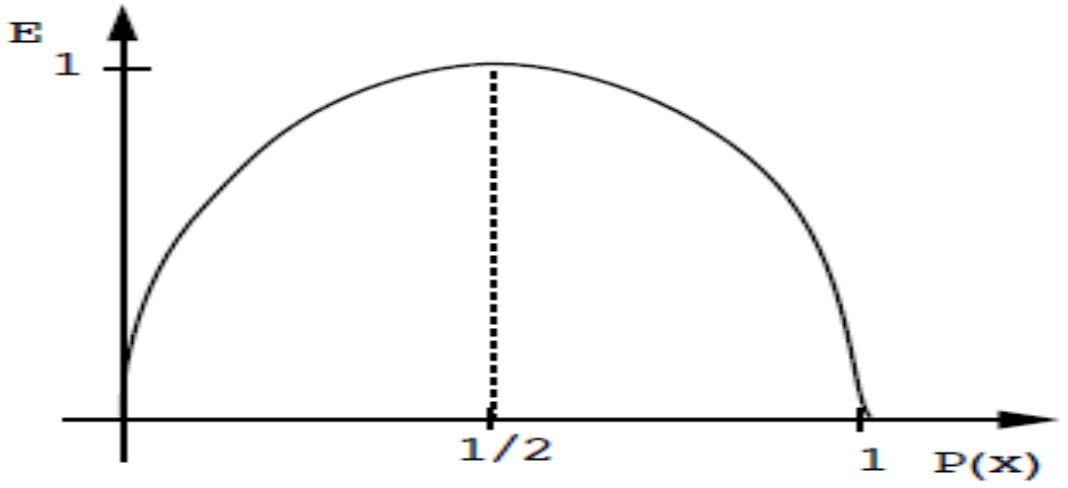
Entropía

Si se tienen v_i posibles respuestas con probabilidades $P(v_i)$, el **contenido de información** es:

$$I(P(v_1), \ldots, P(v_n)) = -\sum_{i=1}^{n} P(v_i) log_2 P(v_i)$$

Nos representa el contenido promedio de información para los diferentes eventos.

Función de Entropía



Evaluación de la ganancia de información

Para valores binarios (p y n) de la clase, la función de cantidad de información de valores de clase es:

$$I\,(p,n) = -p_1 \log_2 p_1 - n_1 \log_2 n_1$$

La entropía del atributo A es:

$$E(A) = \sum_{i=1}^{n} \frac{n_i + p_i}{n + p} I(n_i, p_i)$$

La ganancia del atributo A es:

$$\text{Ganancia}(A) = I(p,n) - E(A)$$

Ejemplo

Ambiente	Temp.	Humedad	Viento	Clase
soleado	alta	alta	no	N
soleado	alta	alta	si	N
nublado	alta	alta	no	P
lluvioso	media	alta	no	P
lluvioso	baja	normal	no	P
lluvioso	baja	normal	si	N
nublado	baja	normal	si	P
soleado	media	alta	no	N
soleado	baja	normal	no	P
lluvioso	media	normal	no	P
soleado	media	normal	si	P
nublado	media	alta	si	P
nublado	alta	normal	no	P
lluvioso	media	alta	si	N

Ejemplo

- Por ejemplo, si calculamos las ganancias para los atributos con los datos de la tabla de Golf (suponemos que: $0 \times log_2(0) = 0$):
  $I(9,5) = -\frac{9}{14} log_2(\frac{9}{14}) - \frac{5}{14} log_2(\frac{5}{14}) = 0.941$
- Para *Ambiente*:
  soleado: $p_1 = 2, n_1 = 3, I(p_1, n_1) = 0.971$
  nublado: $p_2 = 4, n_2 = 0, I(p_2, n_2) = 0$
  lluvioso: $p_3 = 3, n_3 = 2, I(p_3, n_2) = 0.971$
  Entropía(Ambiente) =
  $\frac{5}{14} I(p_1, n_1) + \frac{4}{14} I(p_2, n_2) + \frac{5}{14} I(p_3, n_3) = 0.694$

Ambiente	Temp.	Humedad	Viento	Clase
soleado	alta	alta	no	N
soleado	alta	alta	si	N
nublado	alta	alta	no	P
lluvioso	media	alta	no	P
lluvioso	baja	normal	no	P
lluvioso	baja	normal	si	N
nublado	baja	normal	si	P
soleado	media	alta	no	N
soleado	baja	normal	no	P
lluvioso	media	normal	no	P
soleado	media	normal	si	P
nublado	media	alta	si	P
nublado	alta	normal	no	P
lluvioso	media	alta	si	N

Ejemplo

- Para *Humedad*:
  alta: $p_1 = 3, n_1 = 4, I(p_1, n_1) = 0.985$
  normal: $p_2 = 6, n_2 = 1, I(p_2, n_2) = 0.592$
  Entropía(Humedad) = 0.798
- Para *Viento*:
  no: $p_1 = 6, n_1 = 2, I(p_1, n_1) = 0.811$
  si: $p_2 = 3, n_2 = 3, I(p_2, n_2) = 1.0$
  Entropía(Viento) = 0.892
- Para *Temperatura*, Entropía(Temperatura) = 0.9111

Ambiente	Temp.	Humedad	Viento	Clase
soleado	alta	alta	no	N
soleado	alta	alta	si	N
nublado	alta	alta	no	P
lluvioso	media	alta	no	P
lluvioso	baja	normal	no	P
lluvioso	baja	normal	si	N
nublado	baja	normal	si	P
soleado	media	alta	no	N
soleado	baja	normal	no	P
lluvioso	media	normal	no	P
soleado	media	normal	si	P
nublado	media	alta	si	P
nublado	alta	normal	no	P
lluvioso	media	alta	si	N

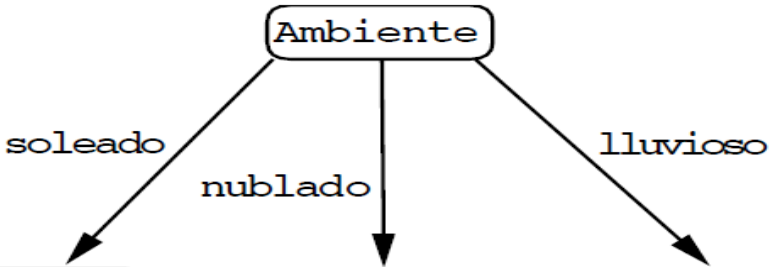
Ejemplo

Ganancia(Ambiente) = 0.941 - 0.694 = 0.246 (MAX)

Ganancia(Temperatura) = 0.940 – 0.9111= 0.029

Ganancia(Humedad) = 0.940 – 0.798 = 0.142

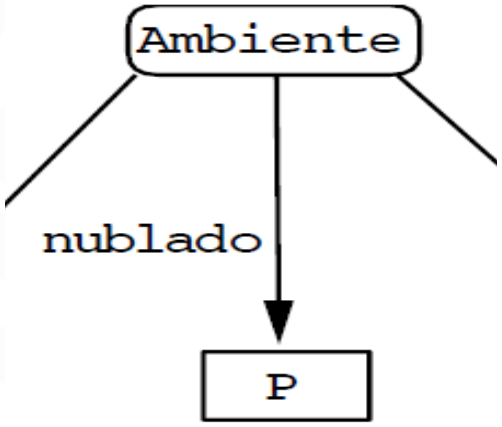
Ganancia(Viento) = 0.940 – 0.892 = 0.048



Por lo que se selecciona *Ambiente* como nodo raíz y procede a realizar el mismo proceso con los demás ejemplos de cada rama.

Ejemplo

- Para *Ambiente* tenemos tres subconjuntos: soleado $(2+, 3-)$, nublado $(4+, 0-)$, lluvioso $(3+, 2-)$. Para nublado, no tenemos que hacer nada, más que asignarle la clase *P*



Ambiente	Temp.	Humedad	Viento	Clase
soleado	alta	alta	no	N
soleado	alta	alta	si	N
nublado	alta	alta	no	P
lluvioso	media	alta	no	P
lluvioso	baja	normal	no	P
lluvioso	baja	normal	si	N
nublado	baja	normal	si	P
soleado	media	alta	no	N
soleado	baja	normal	no	P
lluvioso	media	normal	no	P
soleado	media	normal	si	P
nublado	media	alta	si	P
nublado	alta	normal	no	P
lluvioso	media	alta	si	N

Ejemplo

para *soleado* haríamos el mismo proceso:

Humedad|Ambiente=soleado

alta

I(0,3)=-0/3*log(0/3) -3/3*log (3/3)= 0

normal

I(2,0)=-2/2*log(2/2) – 0 = 0

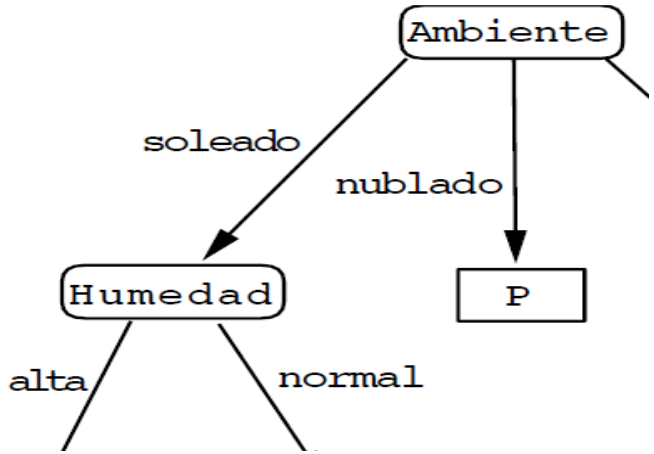
E(Humedad)= 3/5(0) + 2/5(0)=0

Ganancia(Humedad) = 0.97 - [(3/5)0 + (2/5)0] = 0.97 (MAX)

Ganancia(Temperatura) = 0.97 - [(2/5)0 + (2/5)1 + (1/5)0] = 0.570

Ganancia(Viento) = 0.97 - [(2/5)1 + (3/5)0.918] = 0.019

Ambiente	Temp.	Humedad	Viento	Clase
soleado	alta	alta	no	N
soleado	alta	alta	si	N
soleado	media	alta	no	N
soleado	baja	normal	no	P
soleado	media	normal	si	P



Uso del Árbol de Decisión

- Con el árbol construido, podemos preguntar si está bien jugar el sábado en la mañana con ambiente soleado, temperatura alta, humedad alta y con viento, a lo cual el árbol me responde que no
- ID3 sigue una estrategia hill-climbing, sin backtracking
- Tiende a preferir construir árboles pequeños con atributos con ganancia de información alta cerca de la raíz

INEEL
INSTITUTO NACIONAL DE ELECTRICIDAD Y ENERGÍAS LIMPIAS

INEEL CECSE
CENTRO ESPECIALIZADO DE CAPACITACIÓN PARA EL SECTOR ENERGÍA

GRACIAS

Alberto Reyes
areyes@ineel.mx