

Aprendizaje Automático - Trabajo Práctico 5

Gonzalo Castiglione - 49138

June 22, 2012

Objetivo: Aprender a tomar decisiones basadas en un árbol de decisión

1 Aprendizaje de árboles de decisiones

1. La *entropía* para un conjunto de *ejemplos* S esta dada por la fórmula

$$E(S) = \sum_{i \in C} -p_i \log_2 p_i$$

En donde C es el conjunto de clases a las que pueden pertenecer dichos ejemplos y p_i es la probabilidad de que un ejemplo dado pertenezca a la clase i – *esima*.

- (a) Sea el siguiente conjunto de entrenamiento:

Instancia	a_1	a_2	Clasificación
1	T	T	+
2	T	T	+
3	T	F	-
4	F	F	+
5	F	T	-
6	F	T	-

$$p_- = p_+ = 0.5 = p$$

Dado que los patrones están divididos exactamente a la mitad, es de esperarse que el valor de la entropía sea máximo, es decir, 1.

$$E(S) = -p_+ \log_2 p_+ - p_- \log_2 p_- = -2p \log_2 p = 1$$

- (b) En base a la entropía, se define la *ganancia de información* como la disminución de la entropía que se produce al dividir un conjunto S de ejemplos según valores v_i de un atributo A . Es decir:

$$G(S, A) = E(S) - \sum_{v_i \in V} \frac{|S_{v_i}|}{|S|} E(S_{v_i})$$

donde V es el conjunto de valores que puede tomar el atributo A , y S_{v_i} , es el subconjunto de ejemplos de S cuyo atributo A tiene el valores v_i .

$$G(S, a_2) = 1 - \frac{4}{6} * 1 - \frac{2}{6} * 1 = 0$$

2. ELIMINACIÓN- DE-CANDIDATOS Vs ID3

(a) Arbol creado por el algoritmo:

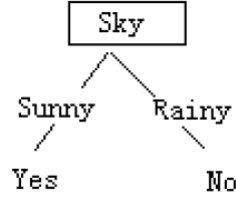


Figure 1: Árbol de decisión generado a partir de la tabla dada

(b) El espacio de versiones contiene todas las hipótesis consitentes con los ejemplos de entrenamiento, mientras que el árbol de decisión aprendido es una de las hipótesis consistentes con los ejemplos de entrenamiento.

(c) Resolución

i. Primero

$$\text{Entropia}(X) = -3/5 * \log_2(3/5) - 2/5 * \log_2(2/5) = 0.971$$

$$G(X, \text{cielo}) = 0.971 - 4/5 * (-3/4 \log_2(3/4) - (1/4) \log_2(1/4)) - 1/5 * 0 = 0.322$$

$$G(X, \text{tempAire}) = 0.971 - 4/5 * (-3/4 \log_2(3/4) - (1/4) \log_2(1/4)) - 1/5 * 0 = 0.322$$

$$G(X, \text{humedad}) = 0.971 - 3/5 * (-2/3 \log_2(2/3) - (1/3) \log_2(1/3)) - 2/5 * 1 = 0.02$$

$$G(X, \text{viento}) = 0.971 - 4/5 * (-3/4 \log_2(3/4) - (1/4) \log_2(1/4)) - 1/5 * 0 = 0.322$$

$$G(X, \text{tempAgua}) = 0.971 - 4/5 * (-2/4 \log_2(2/4) - (2/4) \log_2(2/4)) - 1/5 * 0 = 0.171$$

$$G(X, \text{pronostico}) = 0.971 - 3/5 * (-2/3 \log_2(2/3) - (1/3) \log_2(1/3)) - 2/5 * 1 = 0.02$$

- El algoritmo elige “cielo” como el atributo de testeo para la raíz.

ii. Segundo

$$\text{Entropia}(X) = -3/4 * \log_2(3/4) - 1/4 * \log_2(1/4) = 0.8113$$

$$G(X, \text{tempAire}) = 0$$

$G(X, \text{humedad}) = 0.3113$
 $G(X, \text{viento}) = 0.8113$
 $G(X, \text{tempAgua}) = 0.1226$
 $G(X, \text{pronostico}) = 0.1226$

- El algoritmo elige “viento”.

Version final del arbol de decisi3n creado

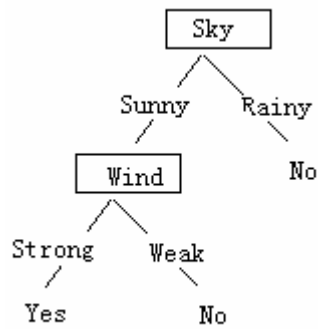


Figure 2: Nuevo 6rbol creado a partir de la tabla original con el nuevo ejemplo

3. 6rbol generado para los lirios de *fisher*.

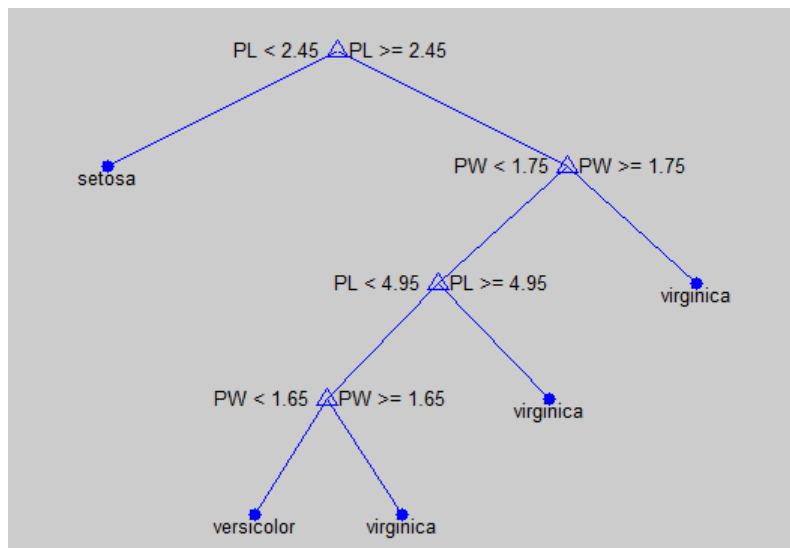


Figure 3: 6rbol creado a partir de las mediciones de los lirios de Fisher

if PL<2.45 then node 2

```
elseif PL>=2.45 then node 3
else setosa
    class = setosa
    if PW<1.75 then node 4
    elseif PW>=1.75 then node 5
    else versicolor
    if PL<4.95 then node 6
    elseif PL>=4.95 then node 7
    else versicolor 5

    class = virginica
    if PW<1.65 then node 8
    elseif PW>=1.65 then node 9
    else versicolor
    class = virginica
    class = versicolor
    class = virginica
```