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

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

En donde C es el conjunto de clases a las quepueden perteneces 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	Clasificacion
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 estan divididos exactamente a la mitad, es de esperarse que el valor de la entropía sea maximo, es decir, 1.

$E(S) = -p_+ log_2 p_+ - p_- log_2 p_- = -2plog_2 p = 1$

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

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

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:

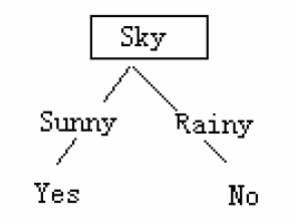

Figure 1: Árbol de decisión generado a partir de la tabla dada

   (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
      Entropia(X) = -3/5*log2(3/5)-2/5*log2(2/5) = 0.971
      G(X, cielo) = 0.971-4/5*(-3/4log2 (3/4)-(1/4)log2(1/4))-1/5*0 = 0.322
      G(X, tempAire) = 0.971-4/5*(-3/4log2 (3/4)-(1/4)log2(1/4))-1/5*0 = 0.322
      G(X, humedad) = 0.971-3/5*(-2/3log2 (2/3)-(1/3)log2(1/3))-2/5*1 = 0.02
      G(X, viento) = 0.971-4/5*(-3/4log2 (3/4)-(1/4)log2(1/4))-1/5*0 = 0.322
      G(X, tempAgua) = 0.971-4/5*(-2/4log2 (2/4)-(2/4)log2(2/4))-1/5*0 = 0.171
      G(X, pronostico) = 0.971-3/5*(-2/3log2 (2/3)-(1/3)log2(1/3))-2/5*1 = 0.02

         - El algorimto elige "cielo" como el atributo de testeo para la raiz.

      ii. Segundo
      Entropia(X) = -3/4*log2(3/4)-1/4*log2(1/4) = 0.8113
      G(X, tempAire) = 0

G(X, humedad) = 0.3113
G(X, viento) = 0.8113
G(X, tempAgua) = 0.1226
G(X, pronostico) = 0.1226

- El algorimto elige "viento".

Version final del arbol de decisión creado

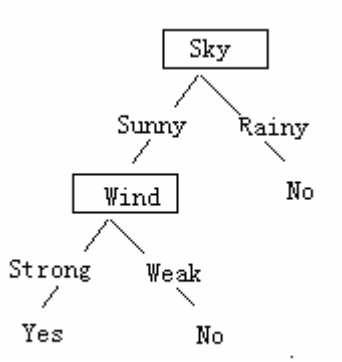

Figure 2: Nuevo árbol creado a partir de la tabla original con el nuevo ejemplo

3. Árbol generado para los lirios de *fisher*.

   (a) Resultado:

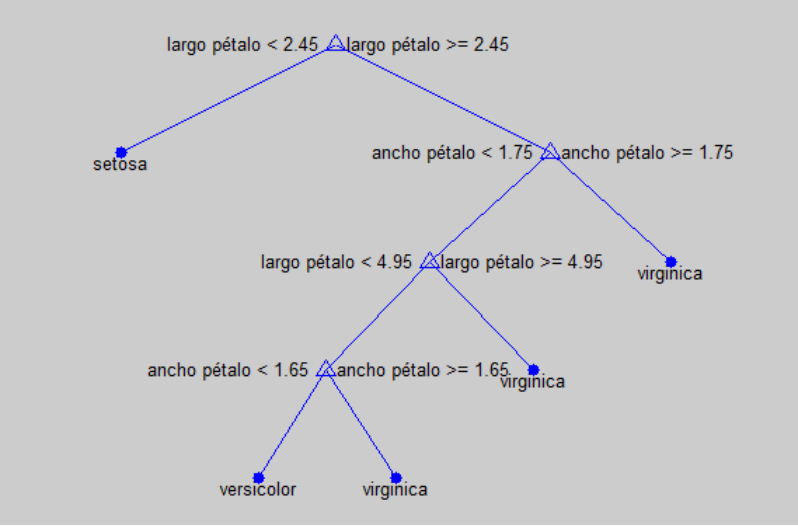

Figure 3: Ancho y largo de los pétalos.

(b) Resultado:

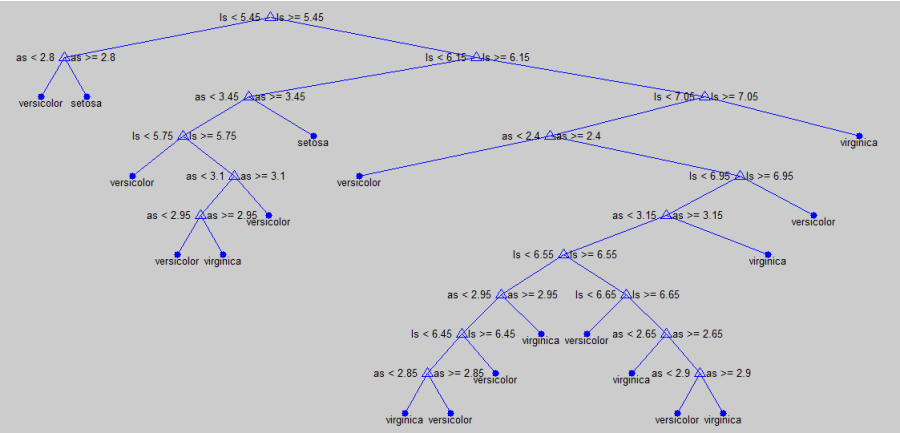

Figure 4: Ancho y largo de los sépalos

(c) Resultado:

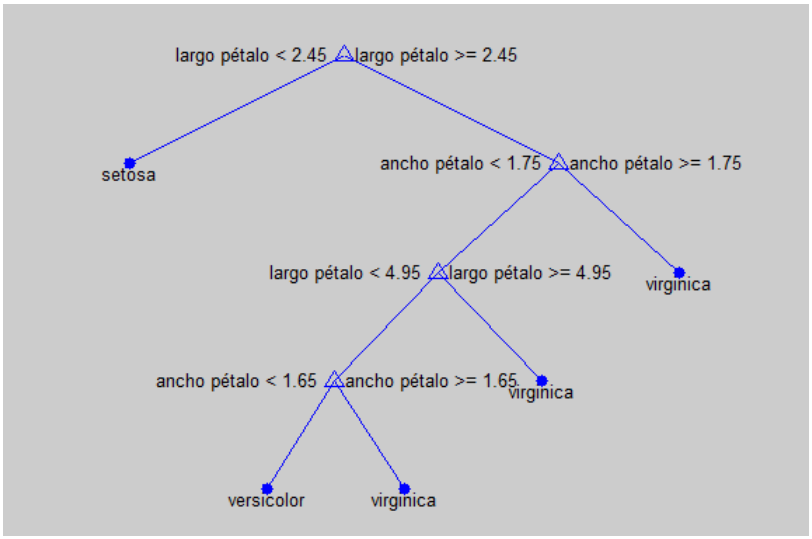

Figure 5: Considerando las cuatro variables

(d) Tabla de clasificacion según el grado de error

Punto	Mediciones	Proporcion correctamente clasificada
a	Ancho y largo del pétalo	0.98
b	Ancho y largo del sépalo	0.87
c	Ancho y largo del sépalo y pétalo	0.98

La clasificación que mayor porcentaje de error presenta con los que corresponden al largo y ancho del sépalo.

	Bayes	k-means	Arbol de decisión
Clasificación correcta	0.96	0.89	0.98

De todos los métodos de clasificacion utilizados, el que mejor error obtuvo es el de los árboles de decisión.