Aprendizaje automático

Departamento de Ingeniería en Informática
ITBA

Trabajo Práctico 5

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

Aprendizaje de árboles de decisiones

1. Considerar el siguiente conjunto de datos de entrenamiento:

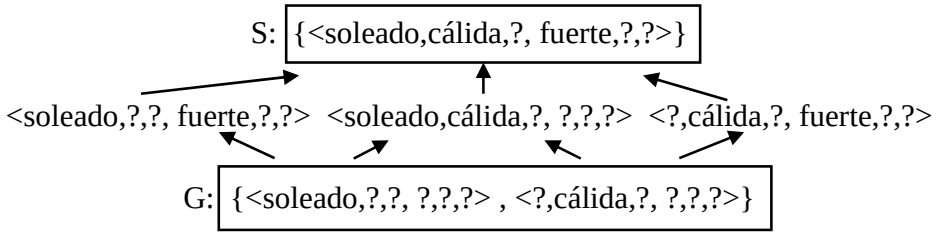
Instancia	clasificación	a_1	a_2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

   (a) ¿Cuál es la entropía de este conjunto de ejemplos de entrenamiento con respecto a la función objetivo **clasificación**?
   (b) ¿Cuál es la ganancia de información de a_2 respecto a los ejemplos de entrenamiento?

2. El algoritmo ID3 busca sólo hipótesis consistentes, mientras que al algoritmo de ELIMINACIÓN-DE-CANDIDATOS busca todas las hipótesis consistentes.
   Considerar la correspondencia entre los dos algoritmos.
   (a) Mostrar que el árbol de decisión que aprendería ID3 para la función objetivo **disfruta-deporte** para los siguientes ejemplos de entrenamiento

ejemplo	cielo	temperatura del aire	humedad	viento	temperatura del agua	pronóstico del tiempo	disfruta deporte
1	soleado	cálida	normal	fuerte	cálida	igual	si
2	soleado	cálida	alta	fuerte	cálida	igual	si
3	nublado	fría	alta	fuerte	cálida	cambiante	no
4	soleado	cálida	alta	fuerte	fría	cambiante	si

   (b) ¿Cuál es la relación entre el árbol de decisión aprendido y el espacio de versiones



que se aprendieron para este ejemplo?

(c) Agregar el siguiente ejemplo de entrenamiento y encontrar nuevamente el árbol de decisión.

ejemplo	cielo	temperatura del aire	humedad	viento	temperatura del agua	pronóstico del tiempo	disfruta deporte
5	soleado	cálida	normal	débil	cálida	igual	no

Mostrar el valor de la ganancia de información para cada atributo candidato en cada paso de crecimiento del árbol.

3. Clasificar, mediante un árbol de decisión, los “datos de los lirios Fisher” (ver ej. 5 tp2) en 3 grupos, tomando como medida de impureza de las divisiones el índice de Gini:
   (a) considerando el ancho y largo de los pétalos.
   (b) considerando el ancho y largo de los sépalos.
   (c) considerando las cuatro variables.
   (d) Calcular el porcentaje de datos correctamente clasificados para los ítems (a), (b) y (c) y comparar los resultados con los obtenidos en el ejercicio 5 del Tp2 y en el ejercicio 2 del Tp4.