Facultad de Matemática, Astronomía y Física y Computación
Universidad Nacional de Córdoba

Ciencia de Datos

Práctico N°8: Árboles de Decisión

Problema 1: Al ingresar al bar el protagonista de este problema, encuentra que sus cinco amigos con quienes se reune los viernes al salir de trabajar, ya tienen sus bebidas sobre la mesa. María que tiene un cargo de gerente paga la ronda la mitad de las veces. Pablo paga la ronda la cuarta parte de las veces, mientras que Sara y Carlos, que son becarios, pagan indistintamente entre ambos la octava parte de las veces. Juan nunca sacó la billetera desde que se reunen los viernes.

a) ¿Qué fracción de las veces el protagonista paga la ronda?

b) Calcular la entropía de la distribución de probabilidad con la que cada uno paga la ronda. ¿Cuál es el número medio de preguntas (de respuesta binaria) que necesitan hacerse en promedio para saber quién paga la ronda?

c) Poco después de arribar al bar se suman dos antiguos amigos quienes no vivieron en la ciudad el último año. Ellos deciden que la próxima ronda la debe pagar el protagonista y conociendo que cursa esta materia, lo desafían a predecir que bebida tomará cada uno.
Con la información de las tres variables binarias: sexo, si es o no estudiante y si le gusta o no bailar, y recordando la elección de bebidas de la noche anterior, puede construirse la siguiente tabla:

Bebida	Sexo	Estudiante	Baile
cerveza	M	T	T
cerveza	M	F	T
vodka	M	F	F
vodka	M	F	F
vodka	F	T	T
vodka	F	F	F
vodka	F	T	T
vodka	F	T	T

Usando entropía de información entrenar un árbol de decisión a partir de la tabla anterior. Registrar todos los valores calculados para elegir las variables en cada nodo del árbol.

d) Proponer una poda posible del árbol y calcular la nueva *accuracy* resultante.

Problema 2: Estudiar la implementación de Árboles de Decisión para clasificación provistas por scikit-learn.

a) ¿Qué algoritmo de árboles de decisión está implementado en scikit-learn? ¿Qué nombre tiene el modelo? ¿Cuál es su principal limitación?

b) Identificar cuál opción del parámetro *criterion* se corresponde con la entropía de Shannon y cuál con Gini como medidas de impureza. ¿Qué parámetros controlan el tamaño del árbol y el *prunning*?

Problema 3: Entrenar un árbol de decisión usando el iris dataset (usar todos los datos y variables). Emplear entropía de Shannon y luego Gini.
a) Graficar los arboles resultantes usando la información disponible en la entrada de scikit-learn.
b) Identificar la variable utilizada como raíz y el valor de corte resultante en cada caso. Interpretar a partir del conocimineto adquirido sobre este dataset.
c) Utilizando la información disponible en Plot the decision surface of decision trees trained on the iris dataset, estudiar las fronteras de decisión bidimensionales generadas por el árbol de decisión aplicado sólo sobre las variables de sépalo del iris dataset.

Problema 4: Para evitar que un árbol se ajuste en exceso (overfitting), `DecisionTreeClassifier` proporciona los parámetros `min_samples_leaf` y `max_depth`. La poda usando una funcion de costo de complejidad proporciona otra opción para controlar el tamaño de un árbol. Esta técnica de poda está regulada por el parámetro `ccp_alpha`. Los valores más altos de `ccp_alpha` aumentan el número de nodos podados.
Con la información provista en Post pruning decision trees with cost complexity pruning, usar el Breast cancer Wisconsin dataset provisto por scikit-learn para:
a) graficar la impureza total de las hojas vs. `ccp_alpha`.
b) graficar el número de nodos y profundidad del árbolvs. `ccp_alpha`.
c) graficar la *accuracy de clasificación sobre los conjuntos de training y testing vs.* `ccp_alpha`.

Soluciones:

Problema 1:

a)

nombre	*María*	*Pablo*	*Sara*	*Carlos*	*Juan*	*yo*
probabilidad	1/2	1/4	1/16	1/16	0	p

$$p = 1 - (1/2 + 1/4 + 1/16 + 1/16) = 1 - 7/8 = 1/8$$

b) $$S = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{2}{16}\log_2 16 + 0\log_2 0 + \frac{1}{8}\log_2 8 = \frac{1}{2} + \frac{2}{4} + \frac{8}{16} + 0 + \frac{3}{8} = \frac{15}{8} = 1{,}875$$

c)

• *S: 8 ejemplos*

C: Cerveza 2 $\rightarrow 1/4$

V: Vodka 6 $\rightarrow 3/4$

$H(S) = \frac{1}{4}\log_2 4 + \frac{3}{4}(\log_2 4 - \log_2 3) = 0{,}811$

Sexo:

M: 4 ejemplos $[2C, 2V]$, $S_M = 1$

F: 4 ejemplos $[0C, 4V]$, $S_F = 0$

$IG(S, Sexo) = 0{,}811 - (\frac{1}{2}\,1 + \frac{1}{2}\,0) = 0{,}311$

Baile:

T: 5 ejemplos $[2C, 3V]$, $S_T = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0{,}971$

F: 3 ejemplos $[0C, 3V]$, $S_F = 0$

$IG(S, Baile) = 0{,}811 - (\frac{5}{8}\,0{,}971 + \frac{3}{8}\,0) = 0{,}204$

Estudiante:

T: 4 ejemplos $[1C, 3V]$

F: 4 ejemplos $[1C, 3V]$

$S_T = S_F = -\frac{1}{4}\log_2\frac{1}{3} - \frac{3}{3}\log_2\frac{3}{4} = 0{,}0{,}811$

$IG(S, Estudiante) = 0{,}811 - \frac{2}{2}\,0{,}811 = 0$

• *En consecuencia se elije como primer nodo a* **Sexo**

F es hoja para vodka

$M \rightarrow S'$: *4 ejemplos*

C: 2 $\rightarrow 1/2$, *V: 2* $\rightarrow 1/2$, $H(S') = 1$

Baile:

T: 2 ejemplos $[2C, 0V]$

F: 2 ejemplos $[0C, 2V]$

$S_T = S_F = 0$

$IG(S, Baile) = 1 - \frac{2}{2}\,0 = 1$

Estudiante:

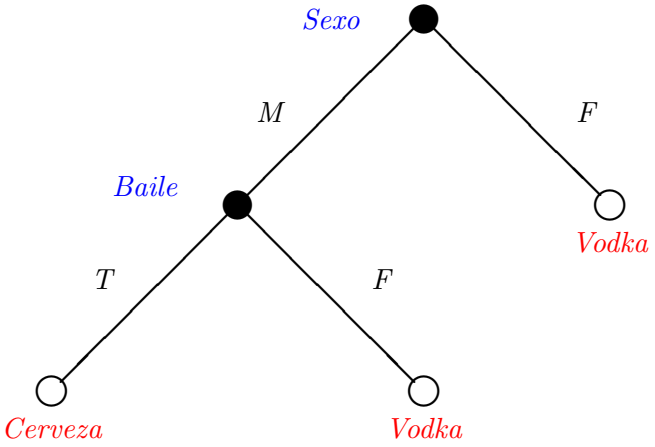
T: 1 ejemplo $[1C, 0V]$, $S_T = 0$

F: 3 ejemplos $[1C, 2V]$, $S_F = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0{,}918$

$IG(S', Estudiante) = 1 - (\frac{1}{4}\,0) + \frac{3}{4}\,0{,}918) = 0{,}312$

• *Luego se elije como segundo nodo a* **Baile**

Con sólo los nodos Sexo y Baile pueden clasificarse a la perfección todos los ejemplos, sin necesidad de utilizar el atributo Estudiante.



d) *La única rama a podar es la del nodo Baile. Esta poda bajaría la exactitud del clasificador sobre el conjunto de entrenamiento de* 1 *a* 0,75.