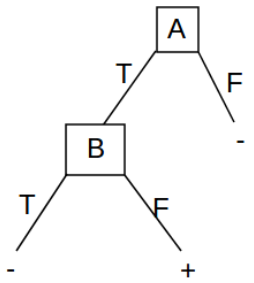
Упражнение 2

Денис Симеонов Михайлов
ФН: 25788

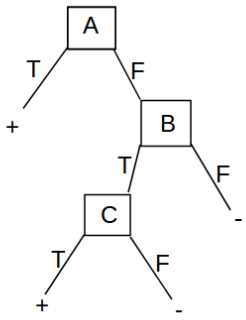
28 октомври 2017г.

1) Нарисувайте класификационни дървета, представящи следните Булеви функции:

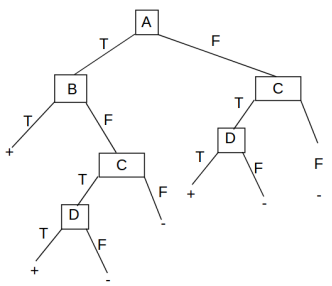
a) $A \wedge \neg B$



b) $A \vee [B \wedge C]$



c) $[A \wedge B] \vee [C \wedge D]$



2) Разгледайте следното множество от обучаващи примери: (T – истина; F – лъжа)

Пример	Класификация	A_1	A_2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

a) Каква е ентропията на това множество от обучаващи примери по отношение за целевата класификация?

Решение Ентропията се дефинира по следния начин:

$$Entropy(S) = -\sum_{i=1}^{c} p_i \cdot \log_2 p_i \tag{1}$$

В примерите имаме само два класа като три от примерите са положителни и три са отрицателни

$$Entropy(S) = -p_+ \cdot log_2 p_+ - p_- \cdot log_2 p_- \tag{2}$$

$$Entropy(S) = -0.5 \cdot (-1) - 0.5 \cdot (-1) = 1 \tag{3}$$

b) Каква е информационната печалба на атрибута A_2 по отношение на тези примери?

Решение Информационната печалба се дефинира по следния начин:

$$Gain(S, A) = Entropy(S) - \sum_{v \in ()} \frac{|S_v|}{|S|} \cdot Entropy(S_v) \tag{4}$$

Следователно за атрибута A_2 получаваме:

$$Gain(S, A_2) = Entropy(S) - \sum_{v \in \{T,F\}} \frac{|S_v|}{|S|} \cdot Entropy(S_v) \tag{5}$$

$$Gain(S, A_2) = 1 - \frac{|S_T|}{|S|} Entropy(S_T) - \frac{|S_F|}{|S|} \cdot Entropy(S_F) \tag{6}$$

В S_T попадат четири примера като два са положителни и два са отрицателни. В S_F попадат два примера като един е положителен и един е отрицателен. По този начин получаваме:

$$Entropy(S_T) = 1 \tag{7}$$

$$Entropy(S_F) = 1 \tag{8}$$

$$\Rightarrow Gain(S, A_2) = 1 - \frac{4}{6} \cdot 1 - \frac{2}{6} \cdot 1 = 1 - 1 = 0 \tag{9}$$

3)

а) Начертайте класификационното дърво, което ще бъде научено от ID3 алгоритъма след четири обучаващи примера на понятието *Харесва*, зададени в Таблица 1 от лекция 1:

Пример	Небе	Въздух	Влажност	Вятър	Вода	Прогноза	Харесва
1	Слънце	Топъл	Нормална	Силен	Топла	Същото	Да
2	Слънце	Топъл	Висока	Силен	Топла	Същото	Да
3	Дъжд	Студен	Висока	Силен	Топла	Промяна	Не
4	Слънце	Топъл	Висока	Силен	Студена	Промяна	Да

Решение Нека началното множество да се казва S. Нека положителните примери са тези дни, които човекът харесва, а отрицателните да са тези, който човекът не харесва. В S има три положителни и един отрицателен пример като получаваме следната стойност за ентропията:

$$Entropy(S) = -\frac{1}{4} \cdot \log_2 \frac{1}{4} - \frac{3}{4} \cdot \log_2 \frac{3}{4} = 0.5 + 0.311 = 0.811 \tag{10}$$

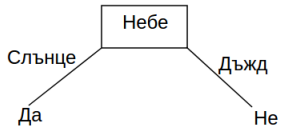
Така получаваме следните стойности за информационната печалба:

$$Gain(S, \text{ Небе}) = Entropy(S) - \frac{S_{\text{слънце}}}{S} \cdot Entropy(S_{\text{слънце}}) - \\ - \frac{S_{\text{дъжд}}}{S} \cdot Entropy(S_{\text{дъжд}}) \tag{11}$$

$$\Rightarrow Gain(S,\ \text{Небе}) = 0.811 - \frac{3}{4} \cdot 0 - \frac{1}{4} \cdot 0 = 0.811 \qquad (12)$$

Информационната печалба при разделяне по атрибута Небе е максимална, защото когато разделим по стойностите на Небе имаме множества, в които има само положителни или само отрицателни примери. Това означава, ентропиите на съответните множества ще са 0 и следователно информационната печалба ще е максимална - а именно равна на $Entropy(S)$. Това означава, че няма нужда да пресмятаме информационната печалба при разделяне по другите атрибути, защото тя ще е по-малка или равна на получената при разделяне по атрибута Небе.

Получава се следното класификационно дърво:



b) Добавете новия пример:

Пример	Небе	Въздух	Влажност	Вятър	Вода	Прогноза	Харесва
5	Слънце	Топъл	Нормална	Слаб	Топла	Същото	Не

Постройте ново дърво и укажете стойността на информационната печалба за всеки кандидат атрибут при всяка стъпка от построяването на дървото.

Решение С добавянето на новия пример получаваме и нова стойност за ентропията за множеството S:

$$Entropy(S) = -\frac{2}{5} \cdot \log_2 \frac{2}{5} - \frac{3}{5} \cdot \log_2 \frac{3}{5} = 0.529 + 0.442 = 0.971 \qquad (13)$$

Така получаваме и новата стойност за информационната печалба при разбиване по атрибута Небе:

$$Gain(S,\ \text{Небе}) = Entropy(S) - \frac{S_{\text{слънце}}}{S} \cdot Entropy(S_{\text{слънце}}) - \\ - \frac{S_{\text{дъжд}}}{S} \cdot Entropy(S_{\text{дъжд}}) \qquad (14)$$

За четири от петте примера имаме, че атрибута Небе е Слънце като три от тези четири имат стойност Да на атрибута Харесва и само

един има стойност Не. За един от петте примера имаме, че атрибута Небе приема стойност Дъжд и когато имаме само един пример в това множество ентропията е 0. Когато разбием S по атрибута Небе получаваме следните ентропии:

$$Entropy(S_{\text{слънце}}) = -\frac{1}{4} \cdot \log_2 \frac{1}{4} - \frac{3}{4} \cdot \log_2 \frac{3}{4} = 0.811 \tag{15}$$

$$Entropy(S_{\text{дъжд}}) = 0 \tag{16}$$

$$\Rightarrow Gain(S,\ \text{Небе}) = 0.971 - \frac{4}{5} \cdot 0.811 - \frac{1}{5} \cdot 0 = 0.322 \tag{17}$$

Когато разбиваме по атрибута Въздух, получаваме същото разбиване на подмножества както при атрибута Небе:

$$S_{\text{слънце}} = S_{\text{топъл}} \tag{18}$$

$$S_{\text{дъжд}} = S_{\text{студен}} \tag{19}$$

Това означава, че ентропиите на подмножествата са равни и съответно и стойността на информационната печалба при разбиване по атрибута Въздух е същата като тази при разбиване по атрибута Небе:

$$Gain(S,\ \text{Въздух}) = Gain(S,\ \text{Небе}) \tag{20}$$

За информационната печалба при разбиване по атрибута Влажност получаваме:

$$Gain(S,\ \text{Влажност}) = Entropy(S) - \frac{S_{\text{нормална}}}{S} \cdot Entropy(S_{\text{нормална}}) - \\ - \frac{S_{\text{висока}}}{S} \cdot Entropy(S_{\text{висока}}) \tag{21}$$

В S имаме два примера като единия е положителен, а другият отрицателен.

$$Entropy(S_{\text{нормална}}) = 1 \tag{22}$$

В S имаме три примера като два са положителни, а другият отрицателен.

$$Entropy(S_{\text{висока}}) = -\frac{2}{3} \cdot log_2 \frac{2}{3} - \frac{1}{3} \cdot log_2 \frac{1}{3} = 0.39 + 0.53 = 0.92 \tag{23}$$

$$\Rightarrow Gain(S,\ \text{Влажност}) = 0.971 - \frac{2}{5} \cdot 1 - \frac{3}{5} \cdot 0.92 = 0.019 \tag{24}$$

Като следваме логиката за пресмятане от по-горе получаваме следната стойност за информационната печалба при разбиване по атрибута Вятър:

$$Gain(S,\ \text{Вятър}) = Entropy(S) - \frac{S_{\text{силен}}}{S} \cdot Entropy(S_{\text{силен}}) - \frac{S_{\text{слаб}}}{S} \cdot Entropy(S_{\text{слаб}}) \quad (25)$$

$$Entropy(S_{\text{силен}}) = -\frac{1}{4} \cdot \log_2 \frac{1}{4} - \frac{3}{4} \cdot \log_2 \frac{3}{4} = 0.811 \quad (26)$$

$$Entropy(S_{\text{слаб}}) = 0 \quad (27)$$

$$\Rightarrow Gain(S,\ \text{Вятър}) = 0.971 - \frac{4}{5} \cdot 0.811 - \frac{1}{5} \cdot 0 = 0.322 \quad (28)$$

Забелязва се, че когато разбиваме по този атрибут получаваме същата информационна печалба както при разбиването при атрибутите Небе и Въздух, въпреки че подмножествата са различни. Това се получава, защото всъщност състава на множествата е еднакъв. Т.е. и при трите атрибута се разбива на две множества като едното множество има само един елемент, а другото множество има четири като три от четирите примера са от една категория, а последния е от различна. Това означава, че ентропиите на множествата ще са еднакви и съответно и информационната печалба ще е същата. За информационната печалба при разбиване по атрибута Вода получаваме:

$$Gain(S,\ \text{Вода}) = Entropy(S) - \frac{S_{\text{топла}}}{S} \cdot Entropy(S_{\text{топла}}) - \frac{S_{\text{студена}}}{S} \cdot Entropy(S_{\text{студена}}) \quad (29)$$

$$Entropy(S_{\text{студена}}) = 0 \quad (30)$$

$$Entropy(S_{\text{топла}}) = -\frac{2}{4} \cdot \log_2 \frac{2}{4} - \frac{2}{4} \cdot \log_2 \frac{2}{4} = 1 \quad (31)$$

$$\Rightarrow Gain(S,\ \text{Вода}) = 0.971 - \frac{4}{5} \cdot 1 - \frac{1}{5} \cdot 0 = 0.171 \quad (32)$$

За информационната печалба при разбиване по атрибута Прогноза получаваме:

$$Gain(S,\ \text{Прогноза}) = Entropy(S) - \frac{S_{\text{същата}}}{S} \cdot Entropy(S_{\text{същата}}) - \frac{S_{\text{промяна}}}{S} \cdot Entropy(S_{\text{промяна}}) \quad (33)$$

$$Entropy(S_{\text{същата}}) = -\frac{2}{3} \cdot \log_2 \frac{2}{3} - \frac{1}{3} \cdot \log_2 \frac{1}{3} = 0.92 \tag{34}$$

$$Entropy(S_{\text{промяна}}) = -\frac{1}{2} \cdot \log_2 \frac{1}{2} - \frac{1}{2} \cdot \log_2 \frac{1}{2} = 1 \tag{35}$$

$$\Rightarrow Gain(S,\ \text{Прогноза}) = 0.971 - \frac{3}{5} \cdot 0.92 - \frac{2}{5} \cdot 1 = 0.019 \tag{36}$$

От получените стойности за информационната печалба имаме най-висока стойност от 0.322, която се постига при разбиването на един от атрибутите Небе, Въздух или Вятър. Нека си изберем и да разбием множеството по атрибута Небе. Така получаваме две множества: S_1, в което има примери 1, 2, 4 и 5 и S_2, в което има само пример 4. За множеството S_2 имаме ентропия 0, защото там има само един пример, затова там няма нужда да разбиваме повече, а можем да заключим, че ако стойността на небе е Дъжд, то човекът не харесва този ден.

Нека сега да видим по кой атрибут е най-добре да разбием множеството S_1. За целта отново трябва да изчислим информационните печалби. Първо нека сметнем стойността на ентропията за множеството S_1:

$$Entropy(S_1) = -\frac{3}{4} \cdot \log_2 \frac{3}{4} - \frac{1}{4} \cdot \log_2 \frac{1}{4} = 0.811 \tag{37}$$

Ако разбием по атрибута Въздух няма да получим никаква печалба, защото всички примери от множеството S_1 имат стойност за атрибута Топъл.

$$\Rightarrow Gain(S_1,\ \text{Въздух}) = 0 \tag{38}$$

За информационната печалба при разбиване по атрибута Влажност имаме:

$$\begin{aligned} Gain(S_1,\ \text{Влажност}) = \\ Entropy(S_1) - \frac{S_{1,\ \text{нормална}}}{S_1} \cdot Entropy(S_{1,\ \text{нормална}}) - \\ - \frac{S_{1,\ \text{висока}}}{S_1} \cdot Entropy(S_{1,\ \text{висока}}) \end{aligned} \tag{39}$$

$$Entropy(S_{1,\ \text{нормална}}) = -\frac{1}{2} \cdot \log_2 \frac{1}{2} - \frac{1}{2} \cdot \log_2 \frac{1}{2} = 1 \tag{40}$$

$$Entropy(S_{1,\ \text{висока}}) = 0 \tag{41}$$

$$\Rightarrow Gain(S,\ \text{Влажност}) = 0.811 - \frac{2}{4} \cdot 1 - \frac{2}{4} \cdot 0 = 0.311 \tag{42}$$

За информационната печалба при разбиване по атрибута Вятър имаме:

$$Gain(S_1,\ \text{Вятър}) = Entropy(S_1) - \frac{S_{1,\ \text{силен}}}{S_1} \cdot Entropy(S_{1,\ \text{силен}}) - \\ - \frac{S_{1,\ \text{слаб}}}{S_1} \cdot Entropy(S_{1,\ \text{слаб}}) \quad (43)$$

$$Entropy(S_{1,\ \text{силен}}) = 0 \quad (44)$$

$$Entropy(S_{1,\ \text{слаб}}) = 0 \quad (45)$$

$$\Rightarrow Gain(S,\ \text{Вятър}) = 0.811 \quad (46)$$

При разбиване по атрибута Вятър получаваме две подмножества с като в едното са само положителни, а в другото само отрицателни. Това означава, че ентропиите на съответните множества са 0, което от своя страна означава, че имаме максимална стойност на информационната печалба. Няма смисъл да разглеждаме информационната печалба за другите атрибути, защото тя няма как да е по-голяма от тази. Класификационното дърво има следния вид: