COMP 472 Artificial Intelligence (Winter 2024)

Worksheet #3: Decision Trees & k-means Clustering

Decision Tree. Given the following training data:

	Features (X)				Output f(X)
Student	**'A' last year?**	**Black hair?**	**Works hard?**	**Drinks?**	**'A' this year?**
X1: Richard	Yes	Yes	No	Yes	**No**
X2: Alan	Yes	Yes	Yes	No	**Yes**
X3: Alison	No	No	Yes	No	**No**
X4: Jeff	No	Yes	No	Yes	**No**
X5: Gail	Yes	No	Yes	Yes	**Yes**
X6: Simon	No	Yes	Yes	Yes	**No**

Create a decision tree that decides if a student will get an 'A' this year, based on an input feature vector X. (Note: check that your tree would return the correct answer for all of the training data rows above.)

Your Decision Tree

Information Content. The *information content* (measured in bits) of an event x with $P(x) > 0$ is defined as:

$$-P(x) \cdot \log_2(P(x))$$

An *impossible event* ($P(x) = 0$) is defined as having an information content of 0. What's the information content of a *certain event* ($P(x) = 1$)? ....

Entropy. Using the definition of *Entropy* for a discrete random variable X with possible outcomes $x_1, x_2, \ldots, x_n$:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \cdot \log_2 p(x_i)$$

compute the entropy for the outcome of the color in the game of *Roulette*, where you have the numbers 1–36 (half red, half black) and the 0 with the color green:

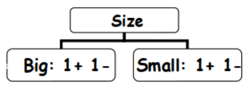
$H(X) =$ ....

Note: make sure you use $\log_2(x)$; if you have a calculator with $\log_{10}$ only, you can compute it using the formula $\log_2(x) = \log_{10}(x)/\log_{10}(2)$.

Information Gain. Compute the *Information Gain* (IG) for the following training data S when splitting using the "Size" attribute:

Size	Color	Shape	Output
Big	Red	Circle	+
Small	Red	Circle	+
Small	Red	Square	-
Big	Blue	Circle	-

$H(S) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1$



Note: by definition,
- Log 0 = -∞
- 0log0 is 0

$$\begin{aligned} &\text{gain}(S, A) \\ &= H(S) - H(S|A) \\ &= H(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \cdot H(S_v) \end{aligned}$$

$H(S|\text{Size}) =$ ....

$\text{gain}(\text{Size}) = H(S) - H(S|\text{Size}) =$ ....

F-Measure. Compute the *F-Measure*, which combines *precision* and *recall* into a single number, using $\beta = 1$ (called F_1-measure, P and R have an equal weight):

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}$$

For the systems s_2, s_3 from the previous lecture's worksheet:

1. $s_2 : P = 100\%, R = 60\% \Rightarrow F_1 =$ ....
2. $s_3 : P = 71\%, R = 100\% \Rightarrow F_1 =$ ....

k-Means Clustering. Here is a dataset with two attributes, to be grouped into two clusters. Compute the *distance* $d(\vec{p}, \vec{q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$ of each data point to the two initial centroids and assign each point to its closest cluster:

	Centroid	
	a1	a2
Cluster 1	1.0	1.0
Cluster 2	5.0	7.0

	a1	a2	Distance to C1	Distance to C2	Cluster
Data1	1.5	2.0			
Data2	3.0	4.0			
Data3	4.5	5.0			
Data4	3.5	4.5			