Milan Zanussi

CSCI 4350/5350

Homework 7

Due: Thu. Nov. 8, 11:00 PM

1. (1 point) How many bits are required to encode an overall class membership of 12.5%? [hint: I(x) = -log_2(P(x))]

$12.5\% = \frac{1}{8}$. So $I(x) = I(\frac{1}{8}) = -\log_2(\frac{1}{8}) = \log_2(2^3) = \boxed{3}$

2. (1 point) What is the Shannon entropy of an overall class membership of 12.5%? [hint: H(x) = P(x)*I(x)]

$I(x) = 3$, $P(x) = \frac{1}{8}$. So $H(x) = P(x)I(x) = \boxed{\frac{3}{8}}$

3. (2 points) Given two classes, + and -, what is the total information for an overall class membership of 12.5% and 87.5%, respectively? [hint: I(x,y) = H(x)+H(y)]

$P(+) = \frac{1}{8}$, $P(-) = \frac{7}{8}$, $I(+) = 3$, $I(-) = -\log_2(\frac{7}{8}) = \log_2(\frac{8}{7}) = 3 - \log_2(7)$. So $I(+,-) = \frac{1}{8}(3) + \frac{7}{8}((3) - \log_2(7)) = \boxed{3 - \frac{7}{8}\log_2(7)}$

Use the data set below to answer all remaining questions:

size	shape	color	class
big	triangle	red	+
big	square	green	-
big	square	red	-
small	circle	green	-
big	triangle	green	+
small	square	red	+
small	square	green	+
small	circle	red	-

$P(+) = 0.5 = \frac{1}{2}$, $P(-) = \frac{1}{2}$

$P(big) = \frac{1}{2}$ $p(small) = \frac{1}{2}$

$P(+|big) = \frac{1}{2}$ $p(-|big) = \frac{1}{2}$

$P(-|small) = \frac{1}{2}$ $p(+|small) = \frac{1}{2}$

4. (2 points) What is I(+,-)?

$I(+,-) = H(+) + H(-) = p(+)I(+) + p(-)I(-) = \frac{1}{2}(1) + \frac{1}{2}(1) = \boxed{1}$

5. (2 points) What is P(big)?

$P(big) = \frac{1}{2}$

6. (2 points) What is the P(+|big)?

$P(+|big) = \frac{1}{2}$

7. (2 points) What is E(size)?
[hint: E(size)= - P(big)*[P(+|big)*log_2(P(+|big))+P(-|big)*log_2(P(-|big))] - P(small)*[P(+|small)*log_2(P(+|small))+P(-|small)*log_2(P(-|small))]

$E(Size) = -\frac{1}{2}\left[\frac{1}{2}\left(\log_2(\frac{1}{2})\right) + \frac{1}{2}\log_2(\frac{1}{2})\right] - \frac{1}{2}\left[\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{2}\log_2(\frac{1}{2})\right]$

$= -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = -\log_2(\frac{1}{2}) = \boxed{1}$

8. (2 points) What is gain(size)? [hint: gain(size) = I(+,-) - E(size)]

$gain(size) = I(+,-) - E(size) = 1 - 1 = \boxed{0}$

9. (2 points) Calculate gain(shape) and gain(color).

$gain(shape) = I(+,-) - E(shape) = \boxed{\frac{1}{2}}$ $\quad gain(color) = I(+,-) - E(color) = 0$

10. (1 point) Which is the best attribute to select for making the first split? (max gain)

Shape

11. (3 points) Draw a complete ID3 decision tree for the data (show your work for how you decided to make each split in the tree)

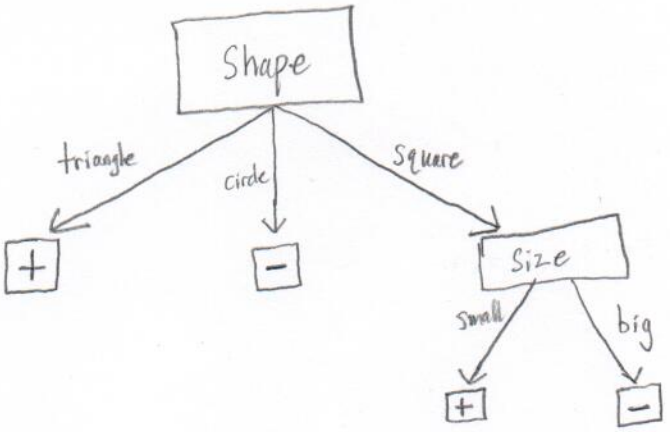
$P(triangle) = \frac{1}{4}$ $p(+|triangle) = 1$ $p(-|triangle) = 0$
$P(circle) = \frac{1}{4}$ $p(+|circle) = 0$ $p(-|circle) = 1$
$P(square) = \frac{1}{2}$ $p(+|square) = \frac{1}{2}$ $p(-|square) = \frac{1}{2}$

$\rightarrow E(shape) = -\frac{1}{4}\left[1\log_2(1) + 0\log_2(0)\right] - \frac{1}{4}\left[0\log_2(0) + 1\log_2(1)\right] - \frac{1}{2}\left[\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]$
$= -\frac{1}{2}\left[\log_2\left(\frac{1}{2}\right)\right] = \frac{1}{2}\log_2(2) = \frac{1}{2}$

$P(red) = \frac{1}{2}$ $p(+|red) = \frac{1}{2}$ $p(-|red) = \frac{1}{2}$
$P(green) = \frac{1}{2}$ $p(+|green) = \frac{1}{2}$ $p(-|green) = \frac{1}{2}$

$E(color) = 1$ (since all values are the same as the Size values)

This tree is complete.



Restricting Sample to Squares:

big	square	green	−
big	square	red	−
small	square	red	+
small	square	green	+

$P(+) = \frac{1}{2}$ $P(-) = \frac{1}{2}$ $I(+,-) = 1$

$\rightarrow E(size|shape) = -\frac{1}{2}\left[1\log_2(1) + 0\log_2(0)\right] - \frac{1}{2}\left[0\log_2(0) + 1\log_2(1)\right] = 0$

$P(big) = \frac{1}{2}$ $p(+|big) = 0$ $p(-|big) = 1$
$P(small) = \frac{1}{2}$ $p(+|small) = 1$ $p(-|small) = 0$

$\rightarrow E(color|shape) = -\frac{1}{2}\left[\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right] - \frac{1}{2}\left[\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right] = 1$

$P(red) = \frac{1}{2}$ $p(+|red) = \frac{1}{2}$ $p(-|red) = \frac{1}{2}$
$P(green) = \frac{1}{2}$ $p(+|red) = \frac{1}{2}$ $p(-|red) = \frac{1}{2}$

$gain(size|shape) = I(+,-|shape) - E(size|shape) = 1$
$gain(color|shape) = I(+,-|shape) - E(size|shape) = 0$

→ Size is chosen to branch from square case of shape