Artificial Intelligence III: Artificial Intelligence and Deep Learning

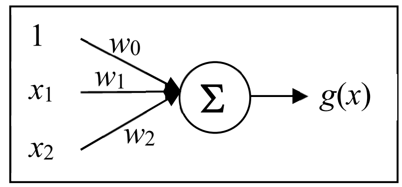
Ch04 – Classifiers

Tutorial Answer

1. For a two-class problem:

$$g(x) = w_0 + w_1x_1 + w_2x_2$$

where x_1 and x_2 is the feature 1 and 2 of a sample. If $g(x) > 0$, sample x belongs to class 1; otherwise, it is class 2. Given that w_0, w_1 and w_2 are 3.5, 5.6 and 2.5 respectively. What is the value of $g(x)$ and the class when



$g(x) = X^tW$
where $X = (1\ x_1\ x_2)^t$ and
$W = (w_0\ w_1\ w_2)^t$

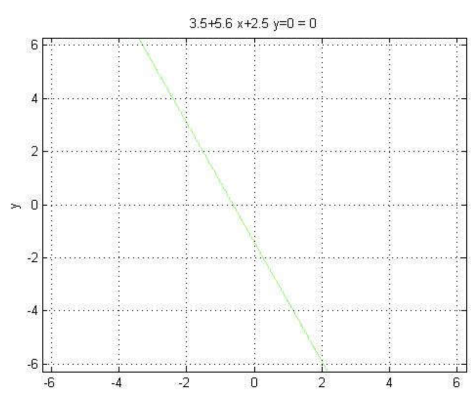
a) $x = (-3.0\ 2.1)$ b) $x = (1.5\ 3.6)$ c) $x = (-2.0\ 3.08)$

Answer:

a) Class 2

b) Class 1

c) On the decision boundary
(can be either class 1 and class 2)



2. Mary wants to solve a 4-class problem:

	x_1	x_2	ω
1	1	5	1
2	3	4	1
3	4	6	2
4	4	8	2

	x_1	x_2	ω
5	6	2	3
6	8	3	3
7	7	5	4
8	8	5	4

However, she only can use binary classifiers.

a) Can you suggest two methods to help Mary?

Answer:

1-against-all and 1-against-1

b) How many binary classifiers should be trained for each method? What are they? How do these methods work?

Answer:

1-against-all method needs 4 binary classifiers:

$g_{1vs234}(x)$, $g_{2vs134}(\mathrm{x})$, $g_{3vs124}(\mathrm{x})$, $g_{4vs123}(\mathrm{x})$

If $g_{1vs234}(x) > 0$, $g_{2vs134}(x) < 0$, $g_{3vs124}(x) < 0$, $g_{4vs123}(x) < 0$, x is class 1

1-against-1 method need 6 binary classifiers:

$g_{1vs2}(x)$, $g_{1vs3}(x)$, $g_{1vs4}(x)$,
$g_{2vs3}(x)$, $g_{2vs4}(x)$, $g_{3vs4}(x)$

If $g_{1vs2}(x)>0$, $g_{1vs3}(x)>0$, $g_{1vs4}(x)>0$, x is class 1

c) Please discuss the pros and cons of these two methods.

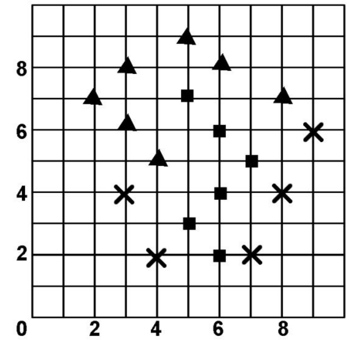
Answer:

1-against-All needs a smaller number of classifiers
1-against-1 has a smaller ambiguous region

3.

a) According to K-nn, which class the followings samples should belong to when K = 1?

i) (5, 5) ii) (7.5, 4) iii) (6, 7)



Answer:

i) Triangle
ii) cross
iii) Unknown (triangle or square)

b) If K = 3, what is the answer of part (a)?

Answer:

i) square
ii) square
iii) square

4. Bob wants to try a Decision Tree algorithm on a simple dataset listed as follows. The dataset has 8 instances each of which has 3 attributes and a class label. The attributes are abstracted as x1, x2, and x3 and the class label as Y. All of them are binary variables.

Sample	x_1	x_2	x_3	Y
1	0	0	0	0
2	0	0	1	0
3	0	1	0	1
4	0	1	1	1
5	1	0	1	1
6	1	0	1	1
7	1	1	0	0
8	1	1	0	0

a) What is the value of the Entropy of class labels, H(Y), in the dataset?

Answer:

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

b) What are the Information Gains of Y with respect to each of the three attributes, IG(Y,x_1), IG(Y, x_2) and IG(Y, x_3)?

Answer:

$$\begin{aligned} H(Y \mid x_1) &= \frac{1}{2}\left(-\frac{1}{2}\log_2(\frac{1}{2})-\frac{1}{2}\log_2(\frac{1}{2})\right)+\frac{1}{2}\left(-\frac{1}{2}\log_2(\frac{1}{2})-\frac{1}{2}\log_2(\frac{1}{2})\right) \\ &= \left(-\frac{1}{2}\log_2(\frac{1}{2})-\frac{1}{2}\log_2(\frac{1}{2})\right) \\ &= 1 \end{aligned}$$

$$\begin{aligned} H(Y \mid x_2) &= \frac{1}{2}\left(-\frac{1}{2}\log_2(\frac{1}{2})-\frac{1}{2}\log_2(\frac{1}{2})\right)+\frac{1}{2}\left(-\frac{1}{2}\log_2(\frac{1}{2})-\frac{1}{2}\log_2(\frac{1}{2})\right) \\ &= \left(-\frac{1}{2}\log_2(\frac{1}{2})-\frac{1}{2}\log_2(\frac{1}{2})\right) \\ &= 1 \end{aligned}$$

$$\begin{aligned} H(Y \mid x_3) &= \frac{1}{2}\left(-\frac{1}{4}\log_2(\frac{1}{4})-\frac{3}{4}\log_2(\frac{3}{4})\right)+\frac{1}{2}\left(-\frac{1}{4}\log_2(\frac{1}{4})-\frac{3}{4}\log_2(\frac{3}{4})\right) \\ &= \left(-\frac{1}{4}\log_2(\frac{1}{4})-\frac{3}{4}\log_2(\frac{3}{4})\right) \\ &= 0.8113 \end{aligned}$$

$$IG(Y,x_1)=H(Y)-H(Y \mid x_1)=0$$
$$IG(Y,x_2)=H(Y)-H(Y \mid x_2)=0$$
$$IG(Y,x_3)=H(Y)-H(Y \mid x_3)=0.1887$$

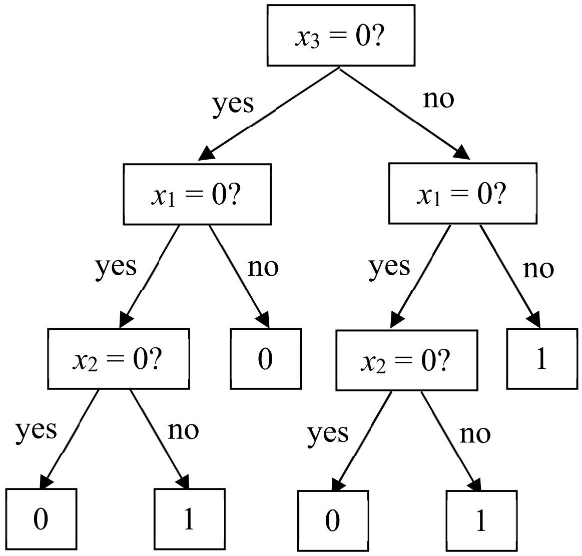
c) Which attribute should be chosen for the root according to the algorithm?

Answer:

Feature 3 should be used as it has the highest Information Gain among three features.

d) Build the rest of the tree according to the Information Gains.

Answer:



Sample	x_1	x_2	Y
1	0	0	0
3	0	1	1
7	1	1	0
8	1	1	0

Sample	x_1	x_2	Y
2	0	0	0
4	0	1	1
5	1	0	1
6	1	0	1

$$H(Y) = -\frac{3}{4}\log_2(\frac{3}{4}) - \frac{1}{4}\log_2(\frac{1}{4}) = 0.811$$

$$\begin{aligned} &H(Y \mid x_1) \\ &= \frac{1}{2}\left(-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})\right) \\ &+ \frac{1}{2}\left(-\frac{2}{2}\log_2(\frac{2}{2}) - \frac{0}{2}\log_2(\frac{0}{2})\right) \\ &= -\frac{1}{2}\log_2(\frac{1}{2}) = \frac{1}{2} \end{aligned}$$

$$\begin{aligned} &H(Y \mid x_2) \\ &= \frac{1}{4}\left(-\frac{1}{1}\log_2(\frac{1}{1}) - \frac{0}{1}\log_2(\frac{0}{1})\right) \\ &+ \frac{3}{4}\left(-\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3})\right) \\ &= 0.6887 \end{aligned}$$

$$H(Y) = -\frac{3}{4}\log_2(\frac{3}{4}) - \frac{1}{4}\log_2(\frac{1}{4}) = 0.811$$

$$\begin{aligned} &H(Y \mid x_1) \\ &= \frac{1}{2}\left(-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})\right) \\ &+ \frac{1}{2}\left(-\frac{2}{2}\log_2(\frac{2}{2}) - \frac{0}{2}\log_2(\frac{0}{2})\right) \\ &= -\frac{1}{2}\log_2(\frac{1}{2}) = \frac{1}{2} \end{aligned}$$

$$\begin{aligned} &H(Y \mid x_2) \\ &= \frac{1}{4}\left(-\frac{1}{1}\log_2(\frac{1}{1}) - \frac{0}{1}\log_2(\frac{0}{1})\right) \\ &+ \frac{3}{4}\left(-\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3})\right) \\ &= 0.6887 \end{aligned}$$

e) Bob wants to prune the tree as he thinks the tree is too complicated. Which node should be deleted and why?

Answer:
The either one of the nodes in middle layer [x1 = 0?] in the decision tree can be deleted. This is because removing one of these nodes only misclassifies one sample.