

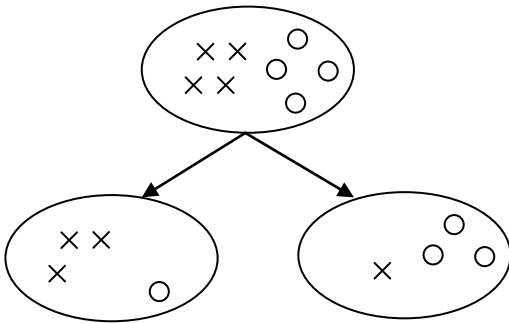
## INFS4203 / INFS7203 Data Mining

### Tutorial 3 – Classification I

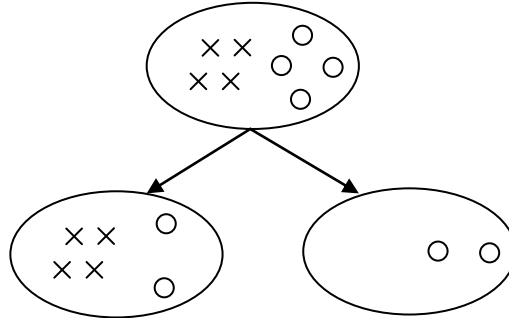
**Question**

Discuss how the following classifications can be evaluated. Use the following cases to illustrate the different evaluation methods. What conclusion you can make for each case?

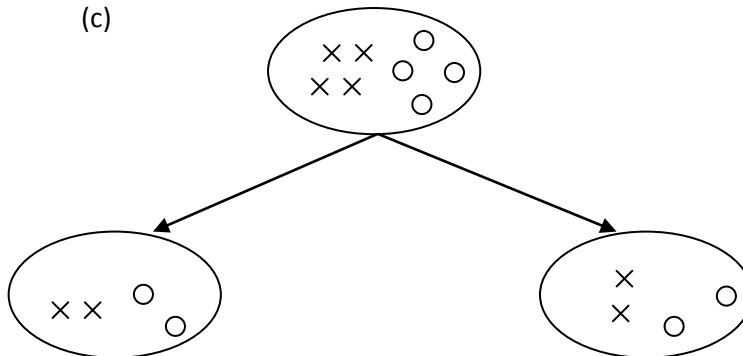
(a)



(b)



(c)



**Answer**

Apparently above classifiers are binary classifiers. We can assume “X” as Positive, and “O” as Negative in the above classifications. The above question can be rewritten in a tabular format, such as:

(a)

Attributes [omitted]	Class	Prediction
...	X	X
...	X	X
...	X	X
...	X	O
...	O	O
...	O	O
...	O	O
...	O	X

(b)

Attributes [omitted]	Class	Prediction
...	X	X
...	X	X
...	X	X
...	X	X
...	O	X
...	O	X
...	O	O
...	O	O

(c)

Attributes [omitted]	Class	Prediction
...	X	X
...	X	X
...	X	O
...	X	O
...	O	X
...	O	X
...	O	O
...	O	O

We use the following methods to evaluate the above classifiers: Accuracy, Precision/Recall, F1, and ROC.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision, } p = \frac{TP}{TP + FP}$$

ROC:

$$\text{Recall, } r = \frac{TP}{TP + FN}$$

$$fp \text{ rate} \approx \frac{\text{negatives incorrectly classified}}{\text{total negatives}}$$

$$\text{F - measure} = \frac{2rp}{r + p}$$

$$tp \text{ rate} \approx \frac{\text{Positives correctly classified}}{\text{Total positives}}$$

So the details calculations are as follows:

(a) Positives: TP = 3; FP = 1; Negatives: TN = 3; FN = 1

$$\text{Accuracy} = (3+3) / (3+1+3+1) = 6/8=0.75$$

$$\text{Precision } p = 3 / (3+1) = 0.75$$

$$\text{Recall } r = 3 / (3+1) = 0.75$$

$$\text{F-measure} = 2 \times 0.75 \times 0.75 / (0.75+0.75) = 0.75$$

$$\text{tp rate} = \frac{3}{4} = 0.75$$

$$\text{fp rate} = 1/4 = 0.25$$

(b) Positives: TP = 4; FP = 2; Negatives: TN = 2; FN = 0

$$\text{Accuracy} = (4+2) / (4+2+2+0) = 6/8=0.75$$

$$\text{Precision } p = 4 / (4+2) = 0.67$$

$$\text{Recall } r = 4 / (4+0) = 1.00$$

$$\text{F-measure} = 2 \times 0.67 \times 1.00 / (0.67 + 1.0) = 0.79$$

$$\text{tp rate} = 4/4 = 1.00$$

$$\text{fp rate} = 2/4 = 0.50$$

(c) Positives: TP =2; FP = 2; Negatives: TN = 2 ; FN = 2

$$\text{Accuracy} = (2+2) / (2+2+2+2) = 4/8 = 0.50$$

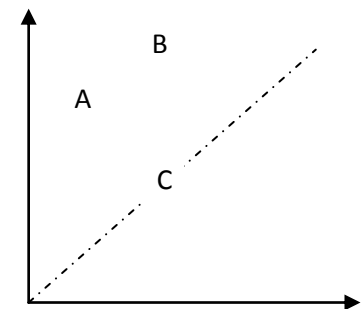
$$\text{Precision } p = 2 / (2+2) = 0.50$$

$$\text{Recall } r = 2 / (2+2) = 0.50$$

$$\text{F-measure} = 2 \times 0.50 \times 0.50 / (0.50 + 0.50) = 0.50$$

$$\text{tp rate} = 2 / 4 = 0.50$$

$$\text{fp rate} = 2 / 4 = 0.50$$



The ROC Graph of (a) , (b), and (c)

Discussions:

Case (a) (i.e., Point A in ROC Graph), makes less mistakes than Case (b) (i.e., A has a lower fp) but has a lower Recall (i.e., tp). Case (c) indicates a poor performance as if a random prediction. For an application that FP is not a problem but the Recall is important (like the cancer diagnoses for patients), Classifier (b) is preferred.