

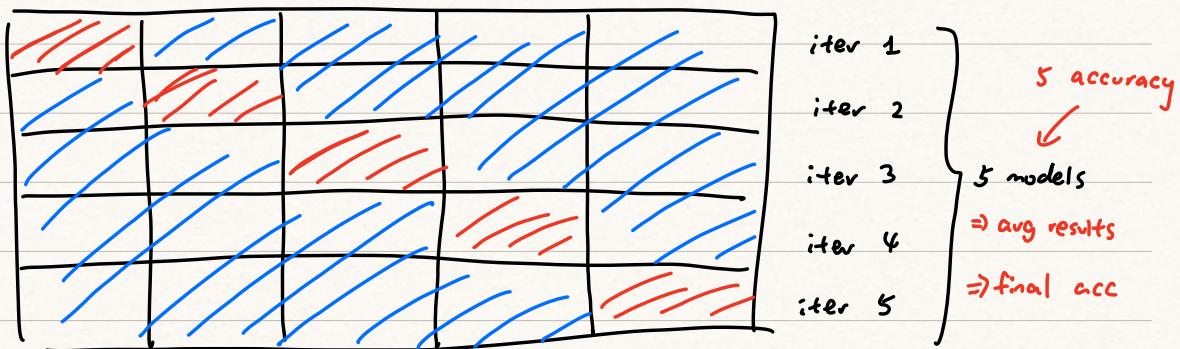
1. How is **holdout** evaluation different to **cross-validation** evaluation? What are some reasons we would prefer one strategy over the other?

Hold-out : fixed

		(eval)
		test
		40%
train		
60%		

subject to some random variation in allocation  
of instances (to train & test)

Cross-validation: K-fold ( $K=5$  partitions)



≡ for test      Each instance is used for testing.

≡ for train      (But CV takes longer)

more stable result! Computationally expensive!

2. A **confusion matrix** is a summary of the performance of a (supervised) classifier over a set of development ("test") data, by counting the various instances:

		Actual			
		a	b	c	d
Classified	a	10	2	3	1
	b	2	5	3	1
	c	1	3	7	1
	d	3	0	3	5

classified correctly

- (i). Calculate the **classification accuracy** of the system. Find the **error rate** for the system.

$$\text{Acc} = \frac{\# \text{ of correctly classified}}{\text{total } \# \text{ of instances}}$$

$$= \frac{10 + 5 + 7 + 5}{\text{Sum of the table}} = \frac{27}{50}$$

$$= 54\%$$

$$ER = 1 - \text{Acc} = 46\%$$

- (ii). Calculate the **precision**, **recall** and **F-score** (where  $\beta = 1$ ) for class  $d$ .

$$\text{Precision} = \frac{\# \text{ of instances "correctly" classified as } d}{\# \text{ of instances classified as } d} = \frac{TP}{TP+FP} = \frac{5}{5+6} = \frac{5}{11}$$

$$\text{Recall} = \frac{\# \text{ of instances "correctly" classified as } d}{\# \text{ of instances truly } d} = \frac{TP}{TP+FN} = \frac{5}{5+3} = \frac{5}{8}$$

		Actual			
		a	b	c	d
Classified	a	10	2	3	1
	b	2	5	3	1
	c	1	3	7	1
	d	3	0	3	5

$\beta = 1 : F_1 = \frac{2P \cdot R}{P + R} = \frac{2}{\frac{1}{P} + \frac{1}{R}}$  (harmonic mean)

$\text{FN} = 3$

$\text{FP} = 6$

$\text{TP} = 5$

$$= \frac{2 \cdot \frac{5}{11} \cdot \frac{5}{8}}{\frac{5}{11} + \frac{5}{8}} = 53\%$$

(iii). Why can't we do this for the whole system? How can we consider the whole system?

Why :

precision & recall: defined per-class (interesting class v.s. the rest)

How:

1. Calculate precision & recall for each class
2. Take average (e.g. Macro, Micro, Weighted Averaging)

↑  
emphasize small classes      ↑  
large classes

3. For the following dataset:

ID	Outl	Temp	Humi	Wind	PLAY					
						TRAINING INSTANCES				
A	s	h	h	f	n					
B	s	h	h	t	n					
C	o	h	h	f	y					
D	r	m	h	f	y					
E	r	c	n	f	y					
F	r	c	n	t	n					
TEST INSTANCES										
G	o	c	n	t	?					
H	s	m	h	f	?					

(i). Classify the test instances using the method of 0-R.

0-R : Majority class classifier

train: 3Y, 3N (choose either!)

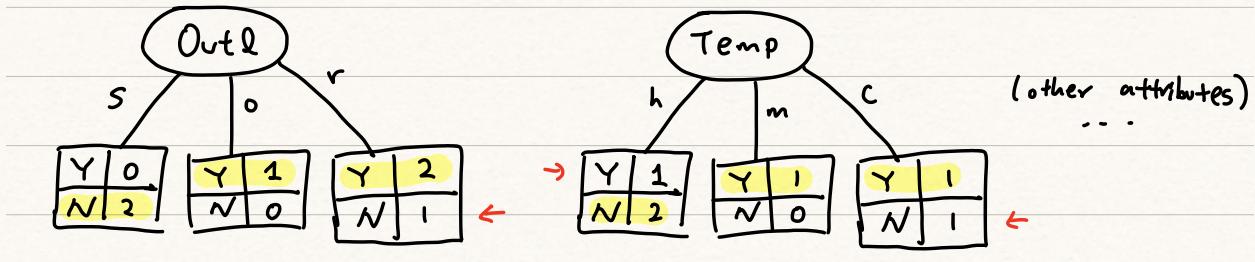
⇒ Let's choose "N"

⇒ Instances: G → N

H → N

(ii). Classify the test instances using the method of **1-R**. (for H assume  $Outl = s$ )

**1 - R:** Classify instances based on one attribute



$$ER = \frac{1}{6}$$

$$ER = \frac{2}{6}$$

Choose attribute with lowest ER : Assume "Outl" is best

$\Rightarrow$  Instances : G:  $Outl = o \rightarrow Y$

H:  $Outl = s \rightarrow N$

ID	Outl	Temp	Humi	Wind	PLAY
TRAINING INSTANCES					
A	s	h	h	F	N
B	s	h	h	T	N
C	o	h	h	F	Y
D	r	m	h	F	Y
E	r	c	n	F	Y
F	r	c	n	T	N
TEST INSTANCES					
G	o	c	n	T	?
H	s	m	h	F	?

4. Given the above dataset, we wished to perform feature selection on this dataset, where the class is PLAY:

- (i). Which of Humi and Wind has the greatest *Pointwise Mutual Information* for the class Y? What about N?

PMI :

$$\text{PMI}(a, c) = \log_2 \frac{P(a, c)}{P(a) P(c)} \Rightarrow \text{find } a \text{ with largest PMI}$$

$\underbrace{\quad}_{\sim 1} \Rightarrow \text{Independent}$

$\ll 1 \Rightarrow \text{Negatively correlated}$

$\cancel{\gg 1} \Rightarrow \text{occur together (positively correlated, if one occur, the other likely to occur)}$

For  $\text{Humi} \& \text{Y}$  : (let  $\text{Humi}=h$ )

$$\text{PMI}(\text{Humi}=h, \text{Play}=Y) = \log_2 \frac{\frac{2}{6}}{\frac{4}{6} \cdot \frac{3}{6}} = \log_2 1 = 0 \text{ (uncorrelated)}$$

For  $\text{Wind} \& \text{Y}$  : (Let  $\text{Wind}=T$ )

$$\text{PMI}(\text{Wind}=T, \text{Play}=Y) = \log_2 \frac{\frac{0}{6}}{\frac{2}{6} \cdot \frac{3}{6}} = \log_2 0 = -\infty \text{ (neg correlated)}$$

Neg class (N) :

Humi : still uncorrelated

Wind : pos correlated

(per-class)

- (ii). Which of the attributes has the greatest *Mutual Information* for the class, as a whole?

$$MI(A, C) = \underbrace{\sum_{i \in \{a, \bar{a}\}} \sum_{j \in \{c, \bar{c}\}}}_{\text{weighted - avg}} P(i, j) \log_2 \frac{P(i, j)}{P(i) P(j)}$$

↑ 4 terms

PMI

$$\Rightarrow MI(A, C) = \sum_{i \in A} \sum_{j \in \{c, \bar{c}\}} P(i, j) \log_2 \frac{P(i, j)}{P(i) P(j)}$$

↑ all values

MI: Combine PMIs of the attribute occurring together with each class, as well as not occurring.

For Outl & Play:

$$\begin{aligned}
 MI(\text{Outl}, \text{Play}) &= P(s, Y) \log_2 \frac{P(s, Y)}{P(s) P(Y)} + P(o, Y) \log_2 \frac{P(o, Y)}{P(o) P(Y)} \\
 MI(\text{Outl}) &\quad + P(r, Y) \log_2 \frac{P(r, Y)}{P(r) P(Y)} + P(s, N) \log_2 \frac{P(s, N)}{P(s) P(N)} \\
 &\quad + P(o, N) \log_2 \frac{P(o, N)}{P(o) P(N)} + P(r, N) \log_2 \frac{P(r, N)}{P(r) P(N)} \\
 &= \frac{o}{6} \log_2 \frac{\frac{o}{6}}{\frac{3}{6} \cdot \frac{3}{6}} + \frac{1}{6} \log_2 \frac{\frac{1}{6}}{\frac{1}{6} \cdot \frac{3}{6}} + \dots \\
 &= 0.541
 \end{aligned}$$

For Temp & Play:

$$\begin{aligned}
 MI(\text{Temp}, \text{Play}) &= P(h, Y) \log_2 \frac{P(h, Y)}{P(h) P(Y)} + P(h, N) \log_2 \frac{P(h, N)}{P(h) P(N)} \\
 MI(\text{Temp}) &\quad + P(m, Y) \log_2 \frac{P(m, Y)}{P(m) P(Y)} + P(m, N) \log_2 \frac{P(m, N)}{P(m) P(N)} \\
 &\quad + P(c, Y) \log_2 \frac{P(c, Y)}{P(c) P(Y)} + P(c, N) \log_2 \frac{P(c, N)}{P(c) P(N)} \\
 &= 0.110
 \end{aligned}$$

For Humi & Play : MI = 0

For Wind & Play : MI = 0.459

} exercise

⇒ Outl is the best attribute (MI)

{ PMI : attribute value & a class value ( $C$  or  $\bar{C}$ )  
MI : an attribute (consider all values) & a class ( $C$  &  $\bar{C}$ )