School of Computing and Information Systems
The University of Melbourne
COMP30027 MACHINE LEARNING (Semester 1, 2019)

Tutorial exercises: Week 4

1. Consider the following 10 instances, given so-called "gold standard" labels (assuming a 3-class problem), and the output of four supervised machine learning models:

Instance	Gold	1	2	3	4
1	A	A	A or B	A	A
2	B	A	B or C	A	?
3	A	A	A	A	A
4	C	C	B or C	A	?
5	B	B	A or B or C	A	?
6	C	A	A or C	A	?
7	C	A	A or B or C	A	?
8	A	C	A or B	A	A
9	A	A	A	A	?
10	A	A	A or C	A	A

   (a) Where possible, calculate the **accuracy** and **error rate** of the four models.

   (b) Where possible, calculate the **precision** and **recall**, treating class A as the "positive" class. Do the same for the B and C classes, in turn, and then calculate the **macro-averaged precision and recall**.

2. What is the difference between evaluating using a **holdout** strategy and evaluating using a **cross–validation strategy**?

   (a) What are some reasons we would prefer one strategy over the other?

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	?

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

   (b) Classify the test instances using the method of 1-R.

   (c) Classify the test instances using the ID3 **Decision Tree** method:

      i. Using the **Information Gain** as a splitting criterion

      ii. Using the **Gain Ratio** as a splitting criterion

1. Consider the following 10 instances, given so-called "gold standard" labels (assuming a ~~3-class~~ problem), and the output of four supervised machine learning models:

Instance	Gold	1	2	3	4
1	A	A	A or B	A	A
2	B	A	B or C	A	?
3	A	A	A	A	A
4	C	C	B or C	A	?
5	B	B	A or B or C	A	?
6	C	A	A or C	A	?
7	C	A	A or B or C	A	?
8	A	C	A or B	A	A
9	A	A	A	A	?
10	A	A	A or C	A	A

(a) Where possible, calculate the **accuracy** and **error rate** of the four models.

(b) Where possible, calculate the **precision** and **recall**, treating class A as the "positive" class. Do the same for the B and C classes, in turn, and then calculate the **macro-averaged precision and recall**.

For binary class.

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

$1 - \text{Accuracy} = \text{error rate}$

Confusion matrix.

predict \ actual	I	U
I	TP	FN
U	FP	TN

(a) System 1, 3 normal :

Acc (Sys 1) = 60%

Acc (Sys 3) = 50%.

Acc (Sys 4) = 40%

For multi-class ! use macro - averaging

System 2 : TP = 10 , FP = 10

TN = 0

FN = 0

$$\text{Acc (Sys 2)} = \frac{10}{10+10} = 50\%$$

Instance	Gold	1	2	3	4
1	A	A	A or B	A	A
2	B	A	B or C	A	?
3	A	A	A	A	A
4	C	C	B or C	A	?
5	B	B	A or B or C	A	?
6	C	A	A or C	A	?
7	C	A	A or B or C	A	?
8	A	C	A or B	A	A
9	A	A	A	A	?
10	A	A	A or C	A	A

(b) $P = \frac{TP}{TP+FP} = \frac{\text{选对 A}}{\text{选所有 A}}$ — Precision (查准)

$R = \frac{TP}{TP+FN} = \frac{\text{选对 A}}{\text{实际有 A}}$ — Recall (查全)

Precision:

$P(sys\ 1; A) = \frac{4}{7}$ $\quad$ $P(sys\ 3; A) = \frac{5}{10}$

$P(sys\ 2; A) = \frac{5}{8}$ $\quad$ $P(sys\ 4; A) = \frac{4}{4}$.

$P(sys\ 1; B) = 1$ $\quad$ $P(sys\ 3; B) = \frac{0}{6}$

$P(sys\ 2; B) = \frac{2}{6}$ $\quad$ $P(sys\ 4; B) = 0$

Recall:

$R(sys\ 1; A) = \frac{4}{5}$ $\quad$ $R(sys\ 3; A) = 1$

$R(sys\ 2; A) = \frac{5}{8}$ $\quad$ $R(sys\ 4; A) = \frac{4}{5}$.

Macro-Average:

$$\text{macro-}P = \frac{1}{N} \sum_{k \in K} P_k$$

$$\text{macro-}R = \frac{1}{N} \sum_{k \in K} R_k$$

macro-P of system 3,4 can not calculated.

2. What is the difference between evaluating using a **holdout** strategy and evaluating using a **cross–validation strategy**?

   (a) What are some reasons we would prefer one strategy over the other?

In "hold-out" eval:

we partition the data into training set and testing set

build the model on the former

then evaluate on the later.

In "cross-validation" eval:

do the same, but a number of times, where each iteration uses one partition of the data as the test data and rest as training set.

(a) hold-hold is subject to some random variation. dependending on which instances are assigned to training set, which are testing set. This could mean our estimate result way-off.

Cross-validation solve this problem, averaging over a brunch of values - so one wired-partition won't affect result. But takes a long time. Since we need train a model for each partition.

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	?

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

(b) Classify the test instances using the method of 1-R.

(c) Classify the test instances using the ID3 **Decision Tree** method:

i. Using the **Information Gain** as a splitting criterion

ii. Using the **Gain Ratio** as a splitting criterion

What is 0-R ?

- Baseline classifier
- throw all attributes,
- predict each instance according to which label is most common in training set ("majority class")

What is 1-R ?

- Better Baseline classifier
- choose a single attribute which preferred.
- we will predict according to that attribute which label is most common in training set

(a) →

(b) How to choose a attribute?

Couting the errors made on the training instances

Let's choose "Outl" first.

Outl = s : ID A, B, Both label as N, no error

Outl = o : ID C. label as Y, no error

Outl = r : ID D, E, F. label Y/N. make one error (2-1)

"0-R"

total : 1 error for outl

Let's chose "Temp" then.

Temp = h : one error

Temp = m : on error

Temp = c : one error

total ; 2 errors for Temp

Assume "Outl" best. ID - G → Y . ID - H → N

(c) Classify the test instances using the ID3 **Decision Tree** method:

i. Using the **Information Gain** as a splitting criterion

ii. Using the **Gain Ratio** as a splitting criterion

For Information Gain (IG), at each level of decision tree.

we choose the attribute that has the largest difference (IG) between the entropy (H(R)) of the class distribution at the parent node and the average entropy which is MI (Mean information) accross the child nodes.

$$IG(\underbrace{A}_{\text{attribute}}|R) = \underbrace{H(R)}_{\text{class}} - \underbrace{\sum_{i\in A} P(A=i)H(A=i)}_{\text{MI}}$$

i at top level (root) of the tree.

$$H(R) = -\sum_{k\in R} P(R_k)\log(R_k) = -\left[\frac{3}{6}\log\frac{3}{6} + \frac{3}{6}\log\frac{3}{6}\right] = 1$$

unpredictable.

even distribution

we want daughters to have an uneven distribution.

- means we can select a class with more confidence.
- which means entropy will go down.

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	?

For example choose "Outl"

$H(A=s) = -[0\log_0 + 1\log 1] = 0$ ← this means at this branch we will choose N with confidence

$H(A=o) = -[1\log 1 + 0\log_0] = 0$

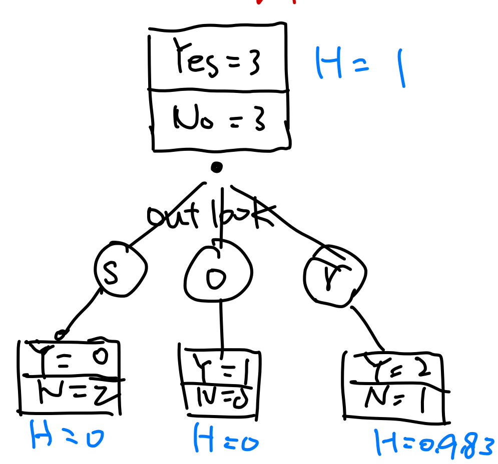
$H(A=r) = -\left[\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right] = 0.9183$

$MI(o) = \frac{2}{6}\cdot 0 + \frac{1}{6}\cdot 0 + \frac{3}{6}\cdot 0.9183 = 0.4592$

$MI(x_1, x_2, \ldots x_m) = \sum_{i=1}^{M} P(x_i)H(x_i)$

$IG(O|R) = 1 - 0.4592 = 0.5408$

注：ID的IG最高，按键说选做root，因为each daughter is purely of a single class，但我们不用

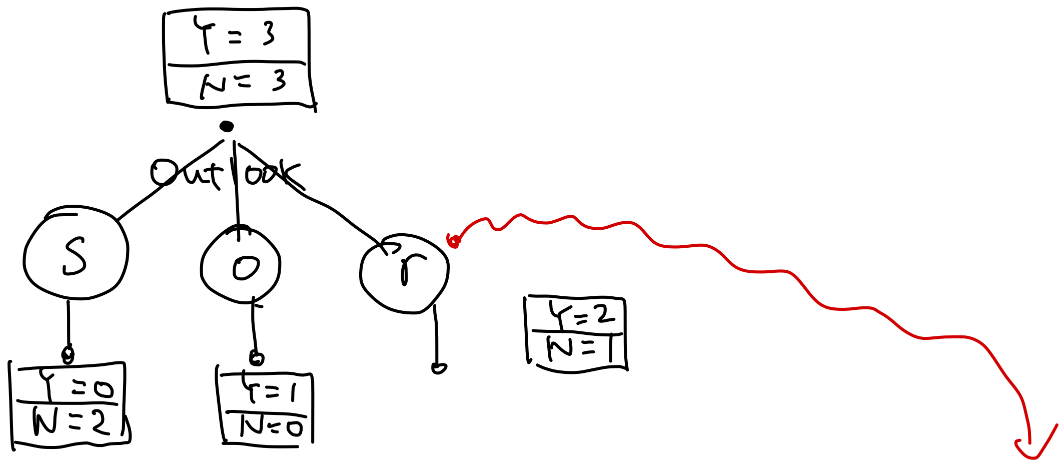


$MI = 0.4592$

把所有的IG的算出来，找出max IG 做为 root

	R	Outl			Temp			H		Wind		ID					
		s	o	r	h	m	c	h	n	T	F	A	B	C	D	E	F
Y	3	0	1	2	1	1	1	2	1	0	3	0	0	1	1	1	0
N	3	2	0	1	2	0	1	2	1	2	1	1	1	0	0	0	1
Total	6	2	1	3	3	1	2	4	2	2	4	1	1	1	1	1	1
$P(Y)$	$\frac{1}{2}$	0	1	$\frac{2}{3}$	$\frac{1}{3}$	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{3}{4}$	0	0	1	1	1	0
$P(N)$	$\frac{1}{2}$	1	0	$\frac{1}{3}$	$\frac{2}{3}$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{4}$	1	1	0	0	0	1
H	1	0	0	0.9183	0.9183	0	1	1	1	0	0.8112	0	0	0	0	0	0
MI		0.4592			0.7924			1		0.5408		0					
IG		0.5408			0.2076			0		0.4592		1					
SI		1.459			1.459			0.9183		0.9183		2.585					
GR		0.3707			0.1423			0		0.5001		0.3868					

Base on above. choose Outl as root



Now Root is Outlook.

$H(R) \neq 1 \Rightarrow H(R) = 0.9183$

If now split by Temp: $MI = \frac{1}{3} \times 0 + \frac{2}{3} \times 1 = 0.6667$

$IG = H(R) - MI = 0.9183 - 0.6667 = 0.2516$

- - - - - - - calculate each.

Final all daughters of r are pure.

$\longrightarrow$ $Outl = O \cup (Outl = r \cap wind = F) \Rightarrow Y$

final tree $Outl = S \cup (Outl = r \cap wind = T) \Rightarrow N$.

第一层 其余层

ii Gain Ratio.

Gain Ratio is Similar!

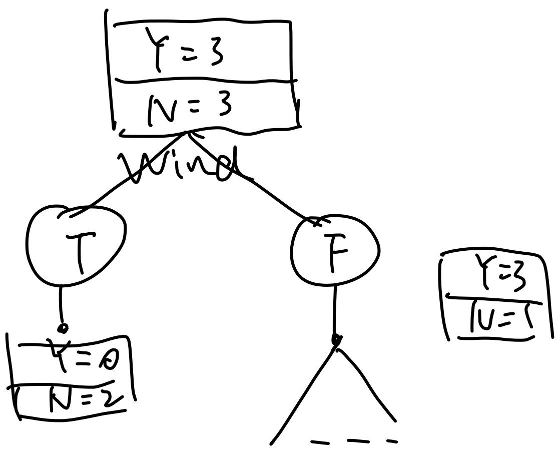
$$SI(A) = -\sum_{i \in A} P(A=i) \log P(A=i)$$

$\frac{\# \text{ of } i}{\# \text{ of instance}}$

$$GR(A) = \frac{IG(o)}{RG(A)}$$

	R	Outl			Temp			H		Wind		ID					
		s	o	r	h	m	c	h	n	T	F	A	B	C	D	E	F
Y	3	0	1	2	1	1	1	2	1	0	3	0	0	1	1	1	0
N	3	2	0	1	2	0	1	2	1	2	1	1	1	0	0	0	1
Total	6	2	1	3	3	1	2	4	2	2	4	1	1	1	1	1	1
$P(Y)$	$\frac{1}{2}$	0	1	$\frac{2}{3}$	$\frac{1}{3}$	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{3}{4}$	0	0	1	1	1	0
$P(N)$	$\frac{1}{2}$	1	0	$\frac{1}{3}$	$\frac{2}{3}$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{4}$	1	1	0	0	0	1
H	1	0	0	0.9183	0.9183	0	1	1	1	0	0.8112	0	0	0	0	0	0
MI		0.4592			0.7924			1		0.5408		0					
IG		0.5408			0.2076			0		0.4592		1					
SI		1.459			1.459			0.9183		0.9183		2.585					
GR		0.3707			0.1423			0		0.5001		0.3868					

GR of Wind is largest → Wind as Root



$H(\text{Wind} = F) = 0.8112$

- - - - Calculated

tree final : $\text{Wind} = F \cap (\text{Outl} = o \cup \text{Outl} = r) \Rightarrow Y$

$\text{Wind} = T \cup (\text{Wind} = F \cap \text{Outl} = s) \rightarrow N$