8. 機器學習應用 (佔分 10 分)

當我們要做機器學習應用時，常會選定一些屬性和一個類別來當資料收集的依據。例如我們現在要研究一公司內人員，會不會買筆記型電腦的情形時，我們可選定「**收入**」、「**工作經驗**」、「**主管**」、「**性別**」等四項來當屬性，選定「**買?**」來當類別。
根據這項目，我們去收集資料，假設我們得到下列 8 筆資料：

收入	工作經驗	主管	性別	買?
L	L	N	F	N
L	L	N	M	N
H	L	N	F	Y
M	M	N	F	Y
M	H	Y	F	Y
M	H	Y	M	N
H	H	Y	M	Y
L	M	N	F	N

從此資料我們可得知，「**收入**」和「**工作經驗**」這兩個屬性都有 **L**(low)、**H**(high)、**M**(middle)三種值，而「**主管**」這屬性有 **Y**(yes)、**N**(no)兩種值，而「**性別**」這屬性有 **M**(male)、**F**(female)兩種值。所有資料的「**買?**」這類別值則是 **Y** 或 **N** 兩者之一。

現在首先我們需要計算目標群組 T 之 Gini 值

$\text{Gini}(T) = 1 - \sum_{j=1}^{n} p_j^2$

其中 n 為 T 中類別種類數目(在上例中因只有 Y 或 N，所以是 2);
p_j 為類別值 j 在目標群組 T 中之分佈機率。

例如我們要根據上面所給的 8 筆資料來計算，$\text{Gini}(\text{收入=L}) = 1 - (0/3)^2-(3/3)^2 = 0$
因為符合收入=L 的資料為第 1,2,8 筆資料，此三筆資料之類別「買?」都是 N，所以 Y 的分佈機率為 0/3，N 的分佈機率為 3/3。

同理可得 $\text{Gini}(\text{收入=M})=1-(2/3)^2-(1/3)^2 = 0.44444444$
$\text{Gini}(\text{收入=H})=1-(2/2)^2-(0/2)^2 = 0$

接下來我們要計算某屬性 w 之 Gini-split 之值:

$\text{Gini-split}(w) = \sum_{i=1}^{m}(X_i / Y)\text{Gini}(T_i)$

其中 m 為屬性 w 中值種類之數目（在「收入」這屬性，有 L、H、M 三種值，所以 m=3）
X_i 為第 i 種屬性值的資料數目（如在收入=L 時有 3 筆資料，X_1=3）
Y 為全部資料筆數（在上例中 Y=8）
$\text{Gini}(T_i)$ 為第 i 種屬性值之 Gini 值

因此 $\text{Gini-split}(\text{收入}) = (3/8)*0+(3/8)*0.44444444+(2/8)*0 = 0.16666667$

同樣 Gini(工作經驗=L) = $1-(1/3)^2-(2/3)^2$ = 0.44444444
Gini(工作經驗=M) = $1-(1/2)^2-(1/2)^2$ = 0.5
Gini(工作經驗=H) = $1-(2/3)^2-(1/3)^2$ = 0.44444444
Gini-split(工作經驗) = (3/8)*0.44444444+(2/8)*0.5+(3/8)*0.44444444 = 0.45833333

Gini(主管=Y) = $1-(2/3)^2-(1/3)^2$ = 0.44444444
Gini(主管=N) = $1-(2/5)^2-(3/5)^2$= 0.48
Gini-split(主管) = (3/8)*0.44444444+(5/8)*0.48 = 0.46666667

Gini(性別=M) = $1-(1/3)^2-(2/3)^2$ = 0.44444444
Gini(性別=F) = $1-(3/5)^2-(2/5)^2$= 0.48
Gini-split(性別) = (3/8)*0.44444444+(5/8)*0.48 = 0.46666667

現我們要找出屬性中 Gini-split()最小者及其值，在上例中為收入，其值為 0.167(四捨五入到小數點後 3 位)

輸入說明：

第一列為資料筆數 Y，其中 1 < Y < 100000
第二列為屬性名稱，最後一個是類別名稱，各元素間以空白隔開，為避免中文編碼問題，本列皆改以英文代碼呈現。(收入→AAA、工作經驗→BBB、主管→CCC、性別→DDD、買?→buy?)
接下來為 Y 列資料，每筆資料的欄位以空白隔開。

輸出說明：

第一列印出各屬性中 Gini-split()最小者之名稱，若有相同最小值時印出最左邊之屬性名稱。
第二列印出其 Gini-split()值，四捨五入到小數點後 3 位

範例輸入：

```
8
AAA BBB CCC DDD buy?
L L N F N
L L N M N
H L N F Y
M M N F Y
M H Y F Y
M H Y M N
H H Y M Y
L M N F N
```

範例輸出：

```
AAA
0.167
```