分類模型的驗證指標

國立東華大學電機工程學系 楊哲旻

Outline

1. 混淆矩陣 與 準確度
2. 二元分類指標
3. PR Curve 震盪
4. 交叉驗證
5. 學習曲線

驗證指標　01. 混淆矩陣 與 準確度

混淆矩陣(Confusion Matrix)

在機器學習領域混淆矩陣，又稱為可能性表格或是錯誤矩陣。是用來評價算法或者說分類器的結果分析表。其每一列代表預測值，每一行代表的實際值。以下為二元分類為例：

Actual Class \ Predicted Class	0	1
0	TN	FP
1	FN	TP

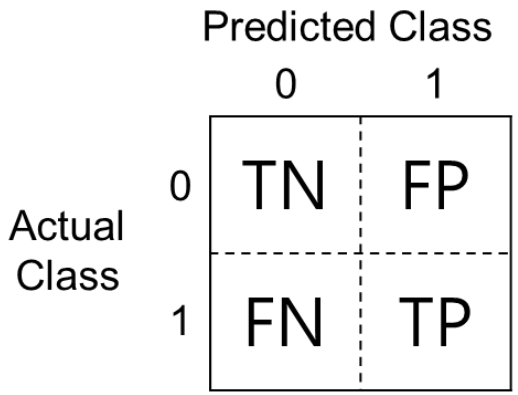
- 真陽性(True Positive, TP)：預測為有，實際上也有。
- 偽陽性(False Positive, FP)：預測為有，實際卻沒有。
- 真陰性(True Negative, TN)：預測為沒有，實際上也沒有。
- 偽陰性(False Negative, FN)：預測為沒有，實際卻有。

準確度(Accuracy)

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}$$

驗證指標　02. 二元分類指標

- 準確度(Accuracy)
- 真陰性率、特異度 (True Negative Rate、 Specificity)
- 真陽性率、敏感度、召回率(True Positive Rate、Sensitivity、Recall)
- 偽陽性率(False Positive Rate)
- 偽陰性率(False Negative Rate)
- 陽性預測值、精確度(Positive Predictive Value、Precision)
- 陰性預測值(Negative Predictive Value)
- 錯誤發現率(False Discovery Rate)
- 錯誤遺漏率(False Omission Rate)



$$Acc=\frac{TP+TN}{TP+TN+FP+FN}$$

$$TNR=\frac{TN}{TN+FP}$$

$$TPR=\frac{TP}{TP+FN}$$

$$FPR=\frac{FP}{FP+TN}$$

$$FNR=\frac{FN}{FN+TP}$$

$$PPV=\frac{TP}{TP+FP}$$

$$NPV=\frac{TN}{TN+FN}$$

$$FDR=\frac{FP}{FP+TP}$$

$$FOR=\frac{FN}{FN+TN}$$

驗證指標　02. 二元分類指標

- F_1 分數、平衡F分數 (F_1 score)
- F_β 分數 (F_β score)
- 盛行率(Prevalence Rate)
- 預兆得分(Threat score, TS)
- 平衡準確度(Balanced Accuracy, BA)
- 馬修斯相關係數(Matthews correlation coefficient, MCC)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{(\beta^2 \times Precision) + Recall}$$

$$Prevalence\ Rate = \frac{FN+TP}{TP+TN+FP+FN}$$

$$TS = \frac{TP}{TP+FP+FN}$$

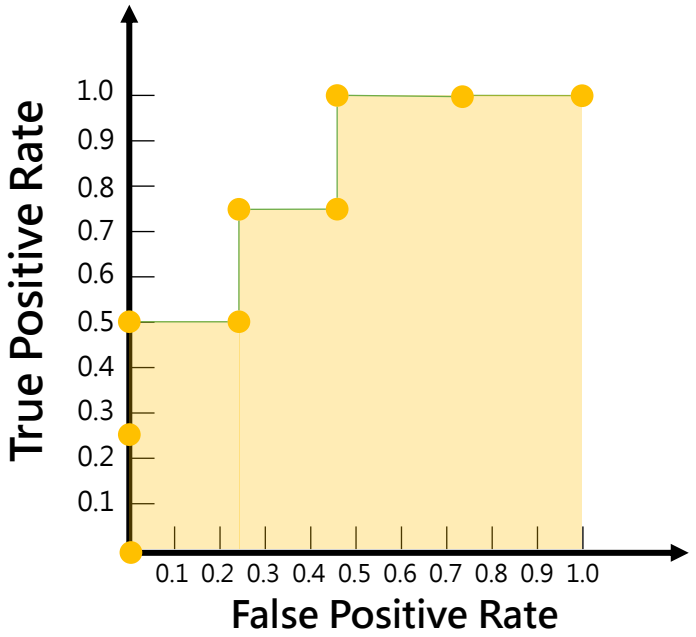
$$BA = \frac{TPR+TNR}{2}$$

		Predicted Class 0	Predicted Class 1
Actual Class	0	TN	FP
	1	FN	TP

驗證指標　02. 二元分類指標

- 接收者操作特徵曲線 (Receiver Operating Characteristic Curve, ROC)
- 曲線下面積 (Area under the curve, AUC)

ID	Label	Predicted probability
1	0	0.1
2	0	0.3
3	1	0.4
4	0	0.6
5	1	0.65
6	0	0.7
7	1	0.85
8	1	0.9



Probability>0

0	4
0	4

Probability>0.1

1	3
0	4

Probability>0.3

2	2
0	4

Probability>0.4

2	2
1	3

Probability>0.6

3	1
1	3

Probability>0.65

3	1
2	2

Probability>0.7

4	0
2	2

Probability>0.85

4	0
3	1

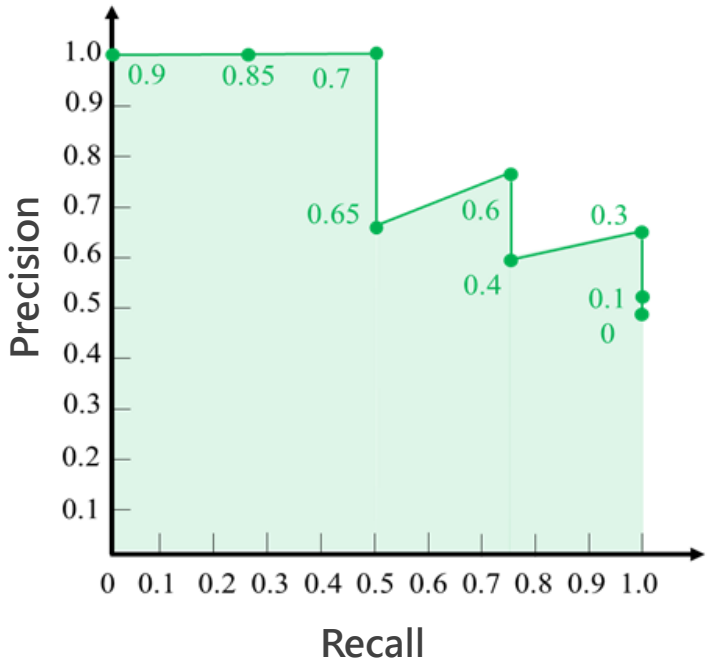
Probability>0.9

4	0
4	0

驗證指標　02. 二元分類指標

- 精確-召回曲線 (Precision-recall Curve, PRC)
- 曲線下面積 (Area under the curve, AUC)

ID	Label	Predicted probability
1	0	0.1
2	0	0.3
3	1	0.4
4	0	0.6
5	1	0.65
6	0	0.7
7	1	0.85
8	1	0.9



Probability>0

0	4
0	4

Probability>0.1

1	3
0	4

Probability>0.3

2	2
0	4

Probability>0.4

2	2
1	3

Probability>0.6

3	1
1	3

Probability>0.65

3	1
2	2

Probability>0.7

4	0
2	2

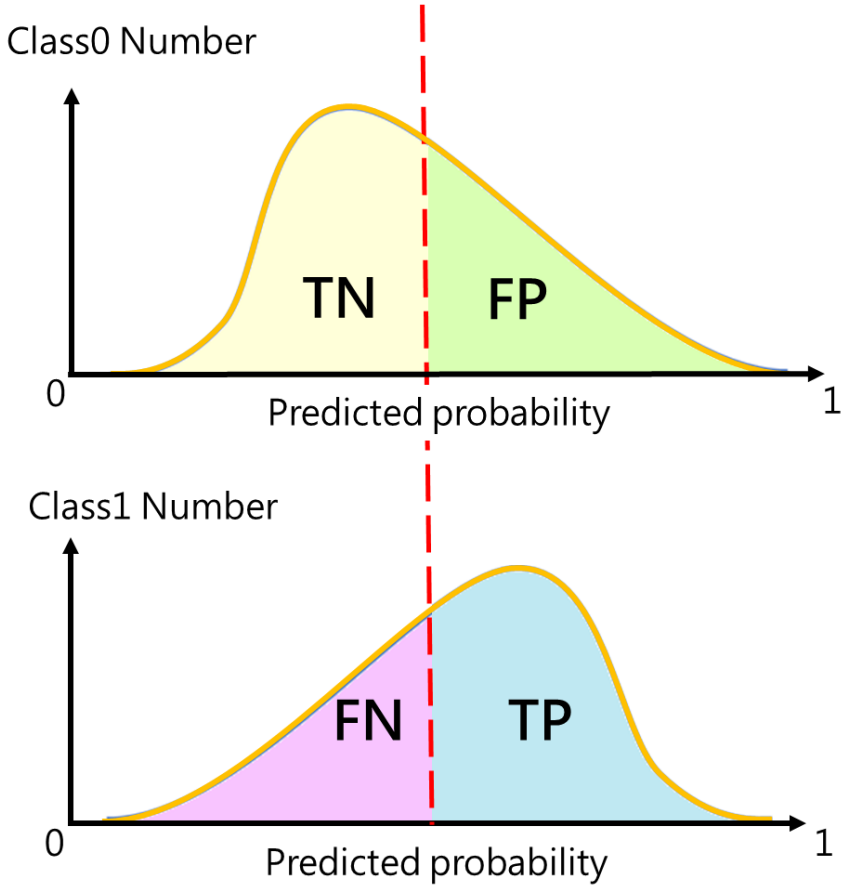
Probability>0.85

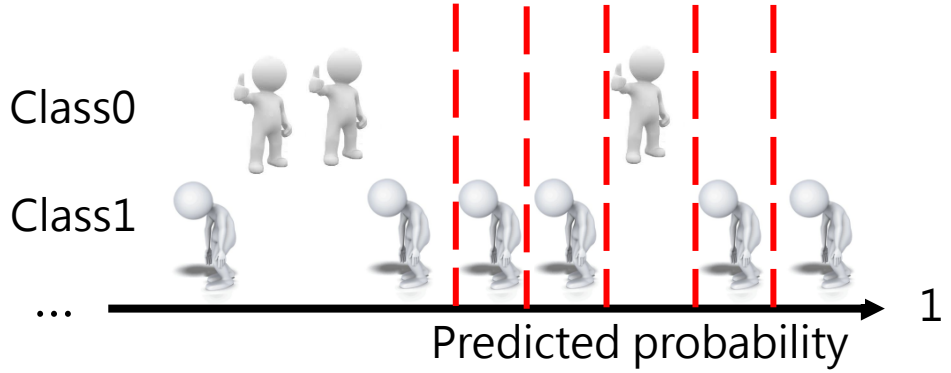
4	0
3	1

Probability>0.9

4	0
4	0

驗證指標　03. PR Curve 震盪





Actual Class \ Predicted Class	0	1
0	TN	FP
1	FN	TP

精準度=1/(1+0)=1

精準度=2/(2+0)=1

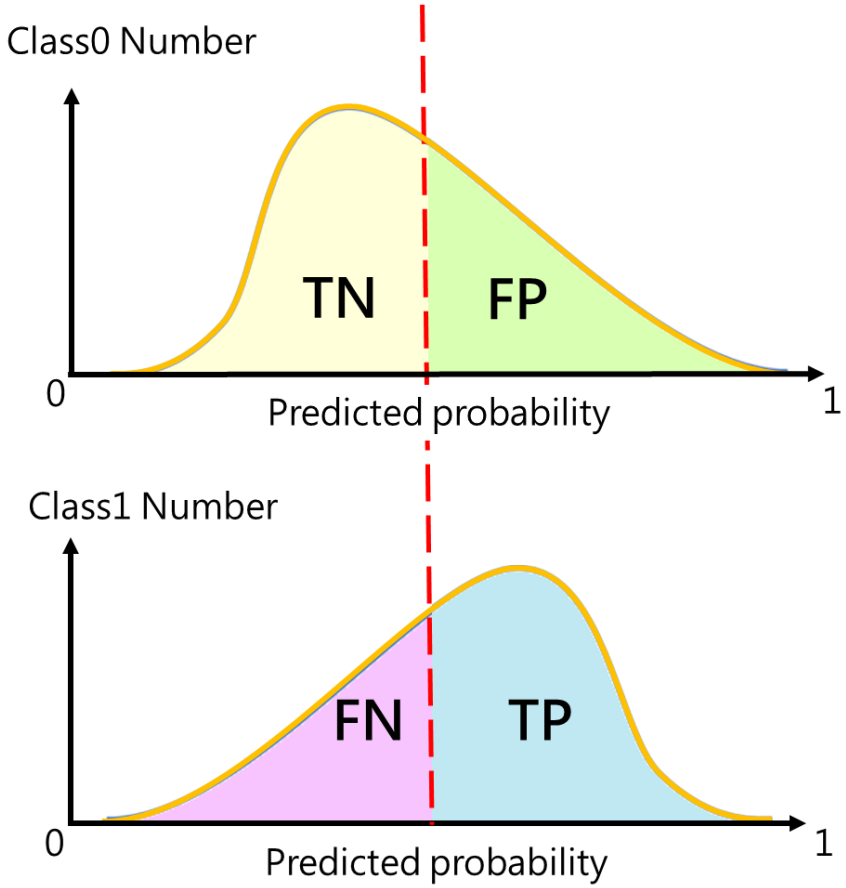
精準度=2/(2+1)=0.67

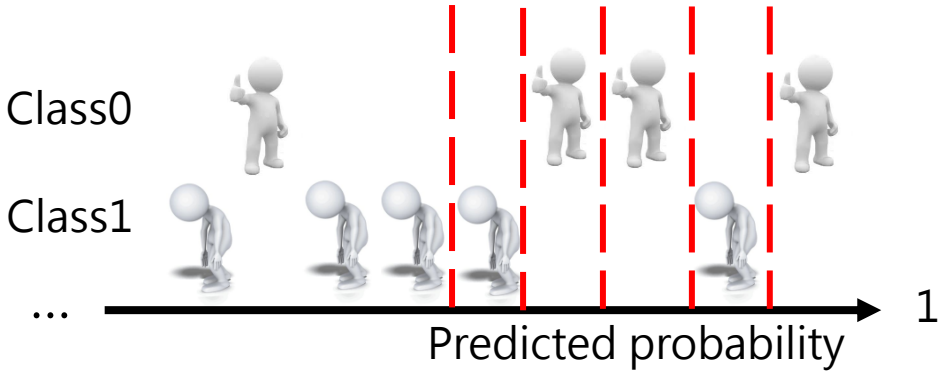
精準度=3/(3+1)=0.75

精準度=4/(4+1)=0.8

PR左半邊很平穩，因為隨切點右而左，FP數量少，TP數量高

驗證指標　03. PR Curve 震盪





切點機率為較高時，敏感度偏低，則為PR曲線的左半邊

		Predicted Class 0	Predicted Class 1
Actual Class	0	TN	FP
	1	FN	TP

精準度=0/(0+1)=0

精準度=1/(1+1)=0.5

精準度=1/(1+2)=0.33

精準度=1/(1+3)=0.25

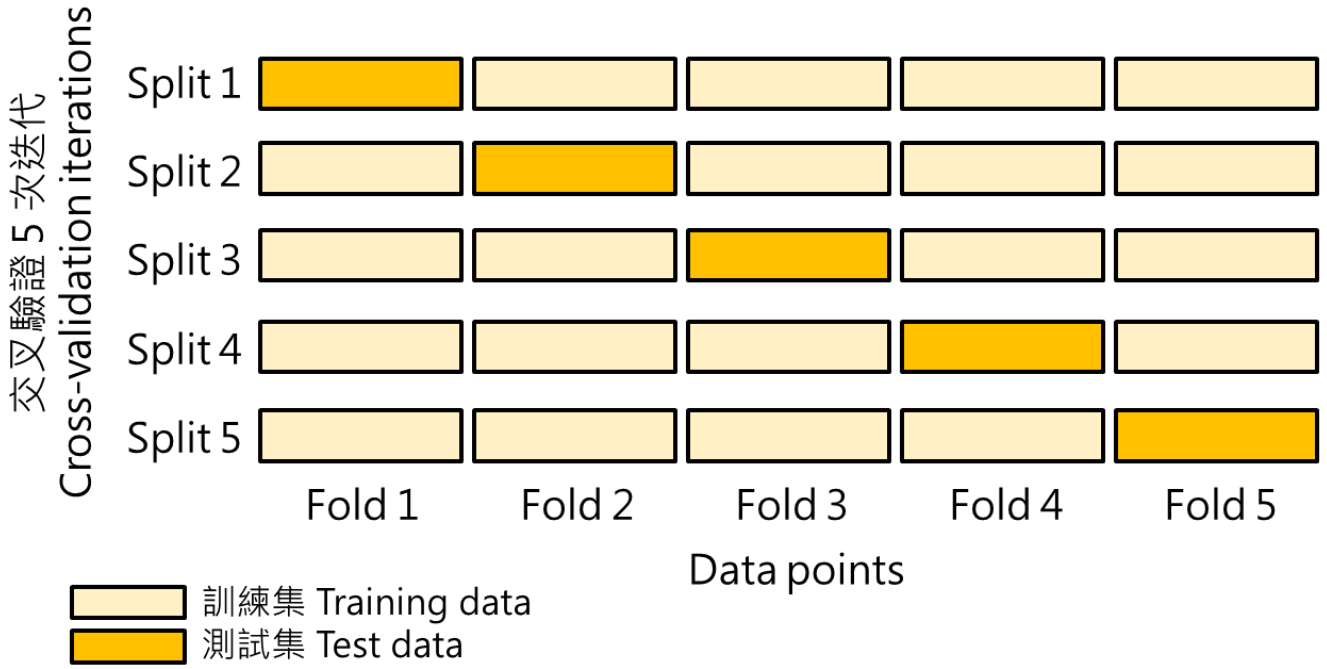
精準度=2/(2+3)=0.4

PR左半邊不穩定，因為隨切點右而左，FP穿插數量多，或TP數量不多

驗證指標　04. 交叉驗證

交叉驗證 (Cross-Validation)

資料被重複切割訓練出多個模型，分別計算他們的驗證指標後並求平均。其中k是資料被均分成k個相同數量，稱為格位(folds)



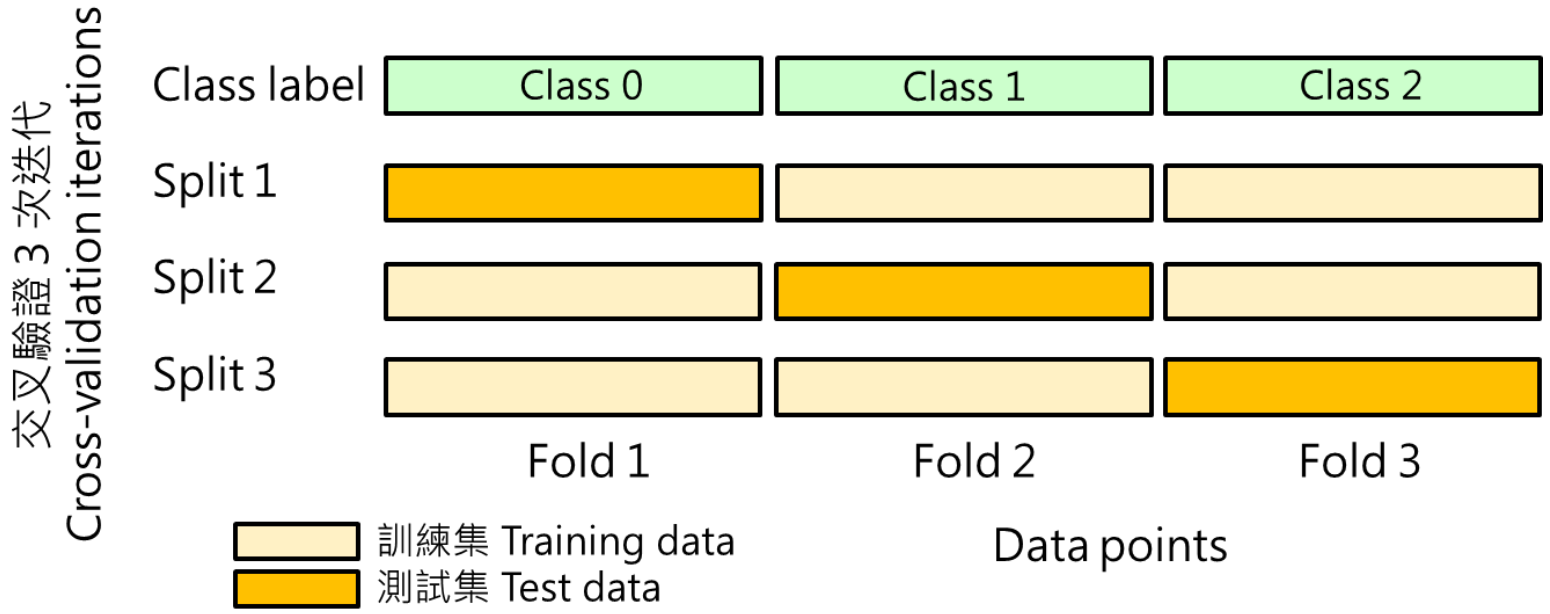
驗證指標　04. 交叉驗證

交叉驗證 (Cross-Validation)

問題：

1. 增加計算成本，訓練k個模型，複雜度高於一般的k倍

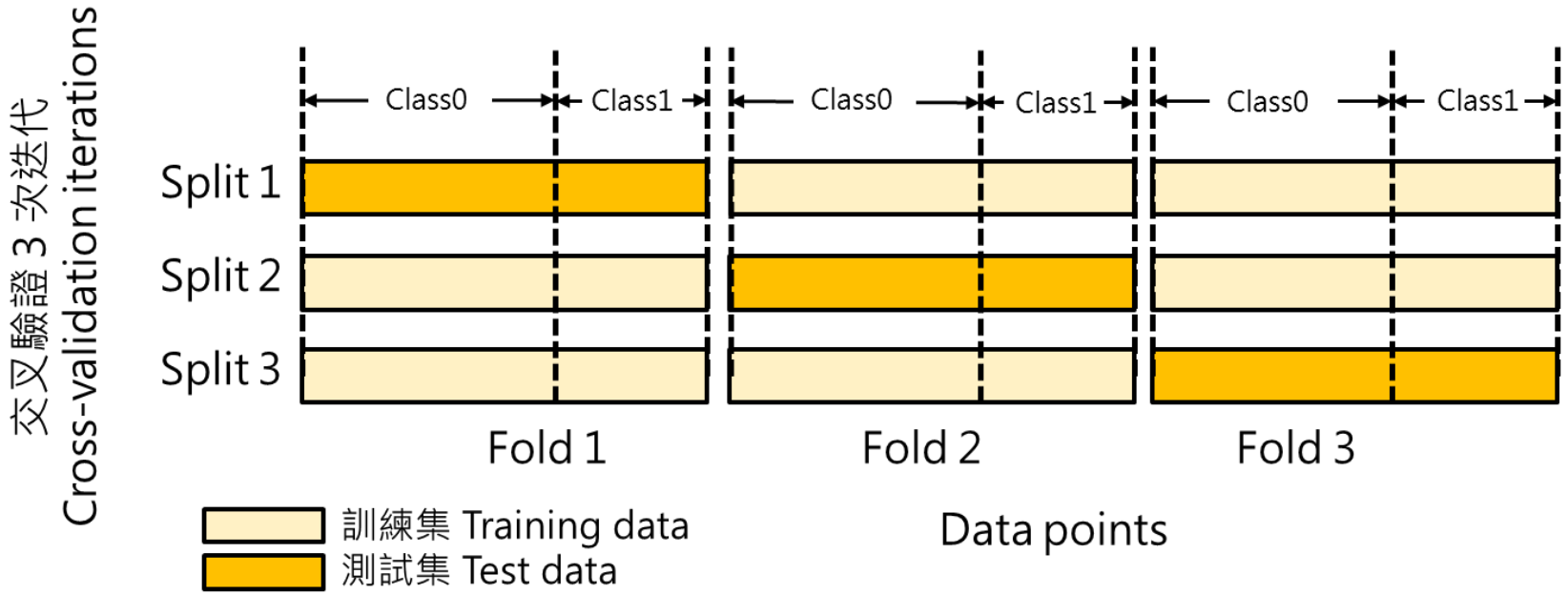
2. 若資料其類別份散如下圖，這樣切割 k 格位方法是非常糟糕的



驗證指標　04. 交叉驗證

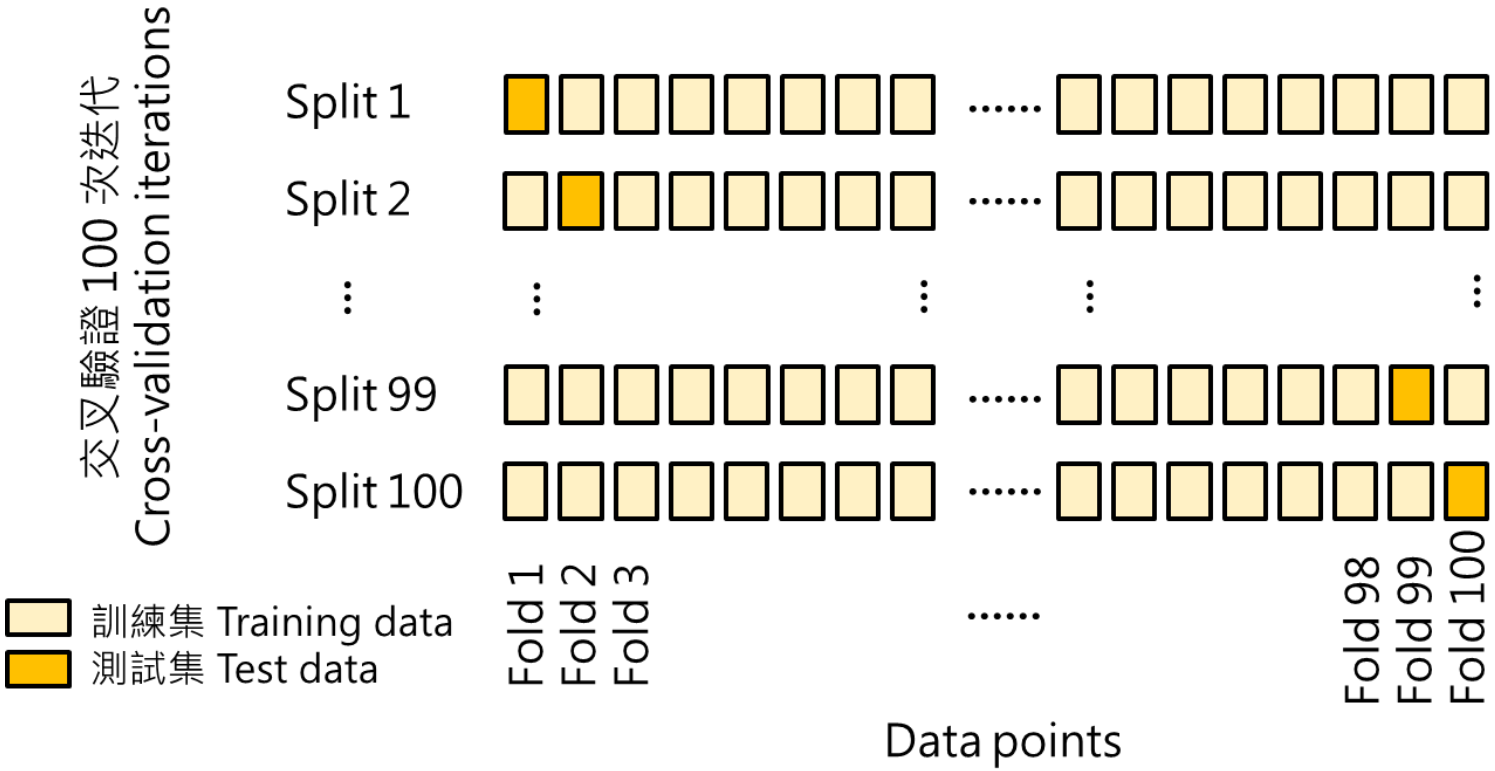
分層交叉驗證 (Stratified Cross-Validation)

分層交叉驗證更能可靠的歸納效能評估模型，目前論文撰寫的k格交叉驗證方法均是指分層交叉驗證



驗證指標　04. 交叉驗證

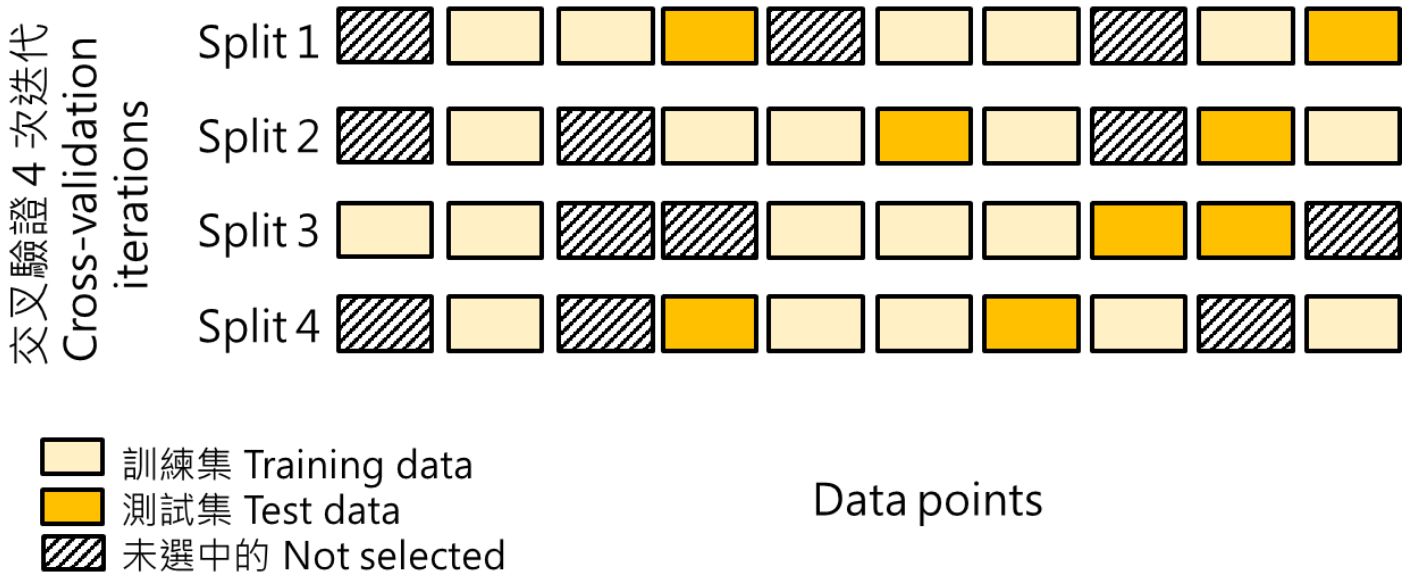
留一法交叉驗證
(Leave-One-Out Cross-Validation, LOO CV)



驗證指標　04. 交叉驗證

隨機分割交叉驗證 (Shuffle-split Cross-Validation)

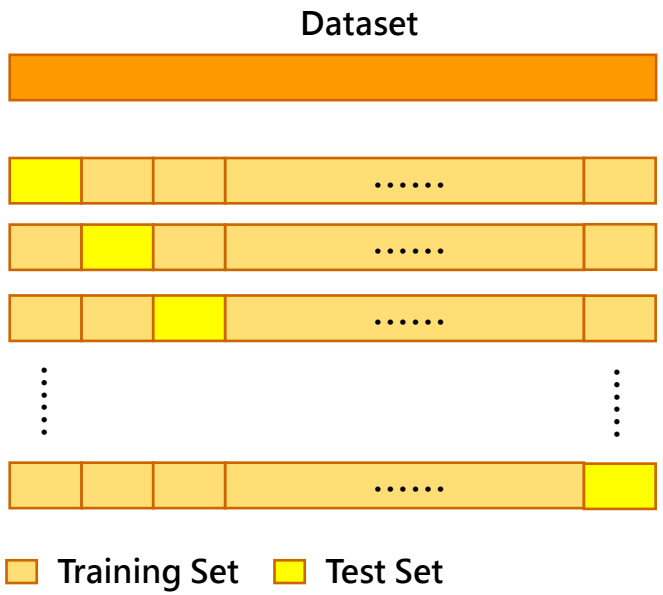
允許控制訓練集與測試集的大小以及迭代次數



驗證指標　04. 交叉驗證

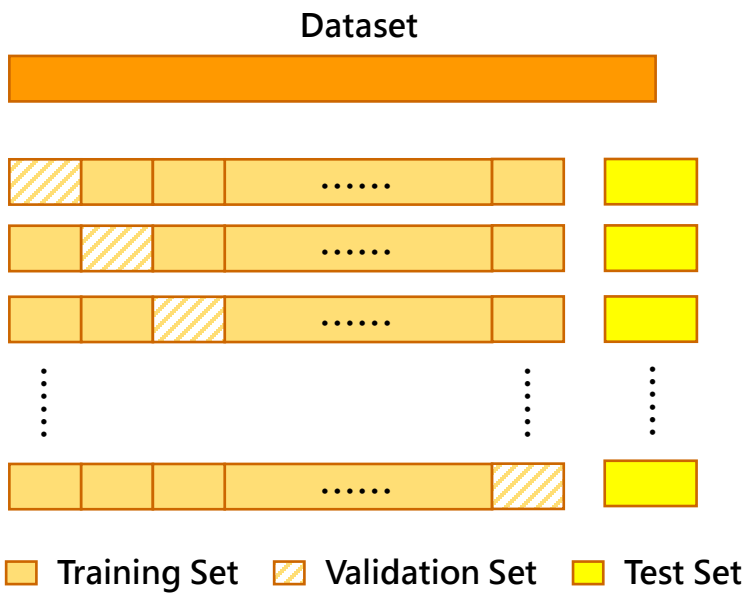
k格交叉驗證 (k-folds Cross-Validation)

- 訓練集與測試集



- 訓練集、驗證集與測試集

  測試集不能參與**權重的訓練**與**超參數的優化**



驗證指標　05. 學習曲線

學習曲線 (Learning Curve)

學習曲線為性能與經驗的曲線圖。性能是模型的錯誤率或準確性等，而經驗可能是用於**學習的訓練數量**或**用於優化模型的超參數(迭代數量)**。其曲線目的為：比較模型、選擇模型參數、調整優化以提高收斂性以及確定訓練的數據量