

計算思維與人工智慧

TA Class #06

RapidMiner3

主講者: 程至榮



政大
NATIONAL CHENGCHI UNIVERSITY



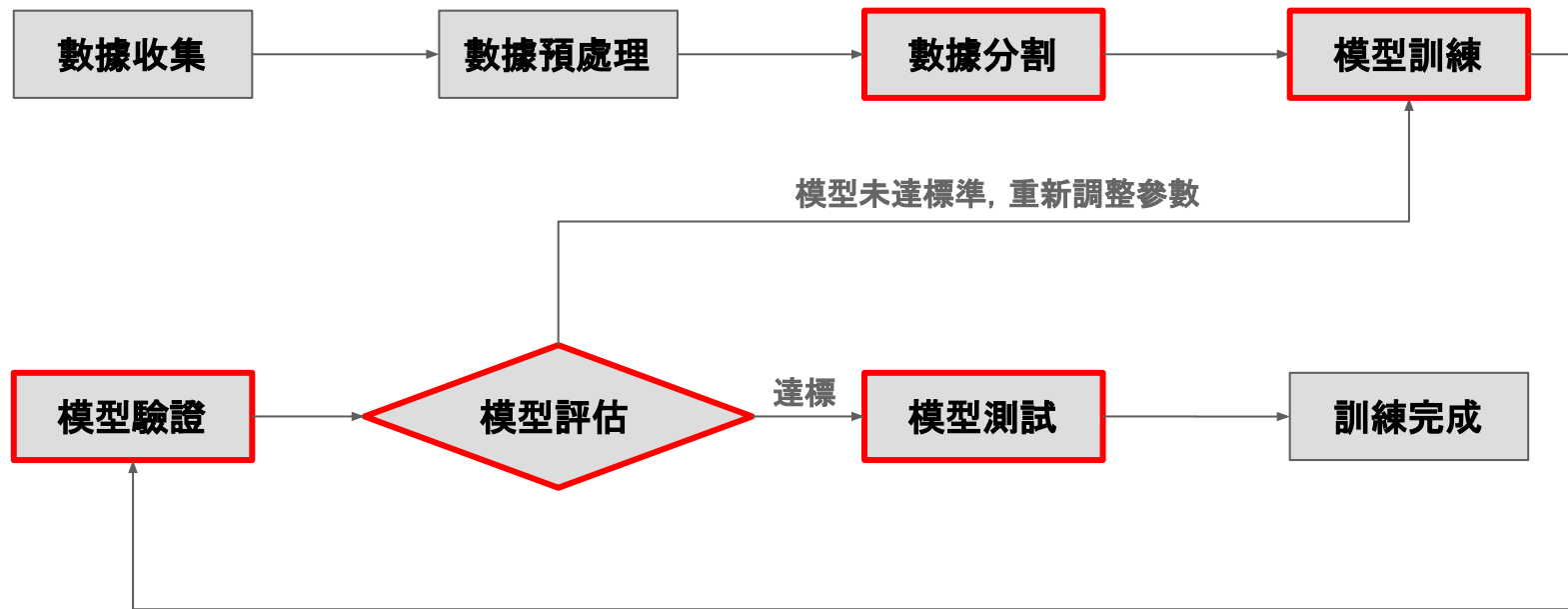
政大資訊科學系
Department of Computer Science, National Chengchi University

參考書目

大數據驅動商業決策:13 個 RapidMiner 商業預測操作實務

今日重點

模型訓練流程



常見的模型目標 & 今日重點

- The most common data science modeling tasks are these:

- Classification—Deciding if something belongs to one category or another
- Scoring—Predicting or estimating a numeric value, such as a price or probability
- Ranking—Learning to order items by preferences
- Clustering—Grouping items into most-similar groups
- Finding relations—Finding correlations or potential causes of effects seen in the data
- Characterization—Very general plotting and report generation from data



Evaluation (Classification)

Confusion matrix

Confusion matrix

- A good summary of classifier accuracy is the *confusion matrix*
 - which tabulates actual classifications against predicted ones

Confusion matrix

示例 [編輯]

如果已經訓練好了一個系統用來區分貓和狗，那混淆矩陣就可以概括算法的測試結果以便將來的檢查。假設一個**13**個動物的樣本，**8**隻貓和**5**隻狗，那混淆矩陣的結果可能如下表所示：

		預測的類別	
		貓	狗
實際的類別	貓	5	3
	狗	2	3

在這個混淆矩陣中，系統預測了**8**隻實際的貓，其中系統預測**3**隻是狗，而**5**隻狗中，則預測有**2**隻是貓。所有正確的預測都位於表格的對角線上（以粗體突出顯示），因此很容易從視覺上檢查表格中的預測錯誤，因為它們將由對角線之外的值表示。

<https://zh.wikipedia.org/zh-tw/%E6%B7%B7%E6%B7%86%E7%9F%A9%E9%98%B5>

Confusion matrix

- Definition (BadLoan => positive case)
 - True positive
 - False positive
 - True negative
 - False negative

pred			
Good.Loan	BadLoan	GoodLoan	
BadLoan	TP 41	FN 259	
GoodLoan	FP 13	TN 687	

Confusion matrix

Sources: [22][23][24][25][26][27][28][29][30] view · talk · edit

		Predicted condition			
Total population = P + N		Predicted Positive (PP)	Predicted Negative (PN)	Informedness, bookmaker informedness (BM) = TPR + TNR - 1	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
Prevalence $= \frac{P}{P + N}$		Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR-) $= \frac{FNR}{TNR}$
Accuracy (ACC) $= \frac{TP + TN}{P + N}$		False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) = $\frac{TN}{PN}$ $= 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$		F ₁ score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \frac{\sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}}{1}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

https://en.wikipedia.org/wiki/Confusion_matrix

Confusion matrix

Evaluation

		pred	
	Good.Loan	BadLoan	GoodLoan
	BadLoan	TP 41	FN 259
	GoodLoan	FP 13	TN 687

- Accuracy
 - # of items categorized correctly divided by #of items
 - $(TP+TN)/(TP+TN+FP+FN)$
- Precision
 - fraction of the items the classifier flags as being in the class actually are in the class = how often a positive indication turns out to be correct
 - $TP/(TP+FP)$
- Recall
 - what fraction of the things that are in the class are detected by the classifier
 - $TP/(TP+FN)$
- False positive rate = $FP/(FP+TN)$

Confusion matrix

Accuracy 在某些應用場景不實用 (舉例: Information Retrieval 關鍵字搜索)

以下的範例 (Retrieved = Positive) 其 Accuracy 將會趨近於 1, 但是 Retrieve 效果其實不好, 大部分和關鍵字相關的文件都沒有被撈回來 (82 筆)

	Retrieved	Not Retrieved
Relevant	18	82
Not Relevant	2	1,000,000,000

Confusion matrix

Precision 和 Recall 可以被人為操作, 讓數字很好看

- Recall = 1 (全猜 Positive)


	Retrieved	Not Retrieved
Relevant	18	0
Not Relevant	1,000,000,000	0

- 刻意降低 Recall 通常可以提高 Precision

Precision & Recall 有 Trade-off 的關係

- Picking thresholds other than 0.5 can allow the data scientist to trade *precision* for *recall*

```
(table(truth=spamTest$spam, prediction=spamTest$pred>0.9))  
(table(truth=spamTest$spam, prediction=spamTest$pred>0.5))  
(table(truth=spamTest$spam, prediction=spamTest$pred>0.1))
```

- 
- 假設模型預測的輸出結果為機率，只有當 "可能為 spam (垃圾信件) 的機率" 大於 0.9 or 0.5 or 0.1 時，模型才會把該樣本分類為 spam。
 - 機率門檻設定得愈低，則被預測為 Positive 的樣本會愈多、被預測為 Negative 的樣本會愈少
>> Precision 下降、Recall 上升。反之亦然
 - 所以我們通常會同時看 F1-Score

Confusion matrix

The F1 score

- Sørensen-Dice coefficient, Sørensen–Dice index, Sørensen index, Dice's coefficient
- Harmonic mean (調和平均數) of precision and recall
 - a useful combination of precision and recall.

$$\bullet \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2*precision*recall}{precision+recall} = \frac{2*TP}{2*TP+FP+FN}$$

Confusion matrix

Common classification performance measures

Table 5.5 Example classifier performance measures

Measure	Formula
Accuracy	$(TP+TN) / (TP+FP+TN+FN)$
Precision	$TP / (TP+FP)$
Recall	$TP / (TP+FN)$
Sensitivity	$TP / (TP+FN)$
Specificity	$TN / (TN+FP)$

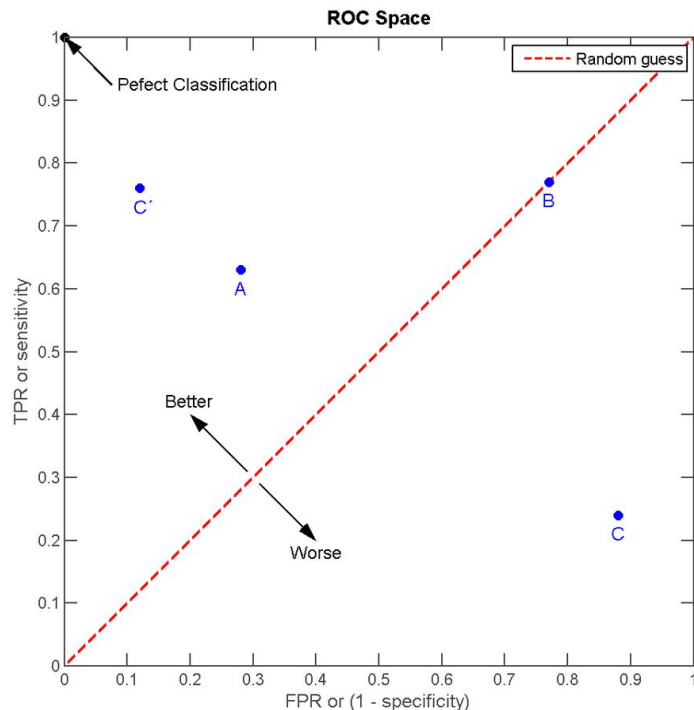
模型比較視覺化：ROC 曲線

- 當進行二元分類預測時，我們可以求得以下關係圖：

預測結果 與 實際結果的比較		實際結果		總數
		0	1	
預測結果	0	真陽性 (TP)	偽陽性 (FP)	P'
	1	偽陰性 (FN)	真陰性 (TN)	N'
總數		P	N	

- ROC 空間將偽陽性率 (FPR) 定義為 X 軸，真陽性率 (TPR) 定義為 Y 軸：

- $TPR = TP / (TP + FN)$
- $FPR = FP / (FP + TN)$



由 ROC_space.png: Indonderivative work: Kai walz (talk) - ROC_space.png, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=8326140>

模型比較視覺化：ROC 曲線

- ROC 空間將偽陽性率 (FPR) 定義為 X 軸，真陽性率

(TPR) 定義為 Y 軸：

- $TPR = TP / (TP + FN)$
- $FPR = FP / (FP + TN)$

真陽性率 (TPR, true positive rate)

又稱：命中率 (hit rate)、敏感度 (sensitivity)

$$TPR = TP / P = TP / (TP + FN)$$

偽陽性率 (FPR, false positive rate)

又稱：錯誤命中率，假警報率 (false alarm rate)

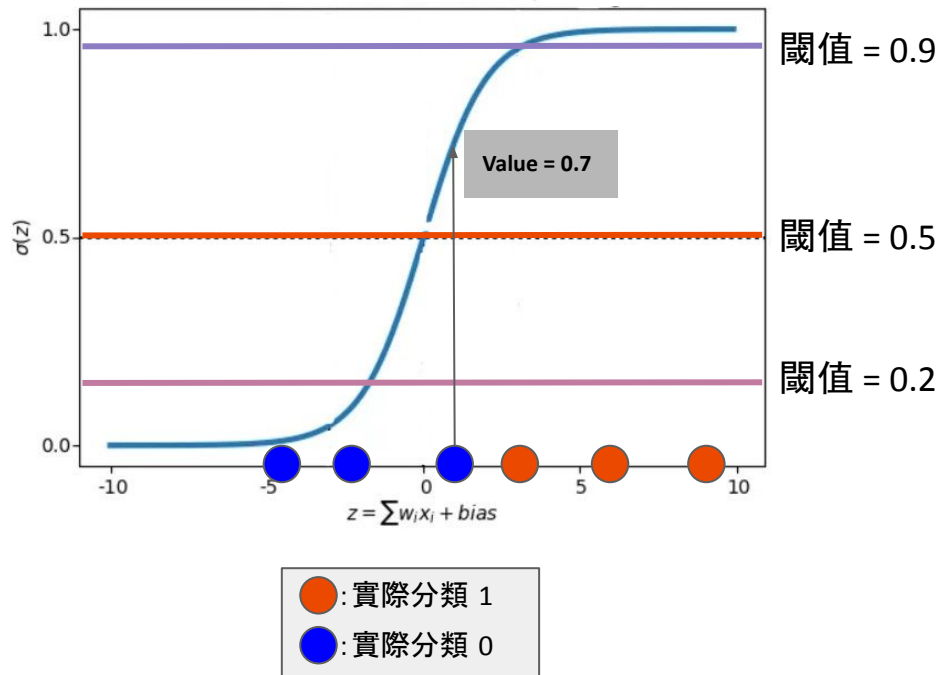
$$FPR = FP / N = FP / (FP + TN)$$

由 ROC_space.png: Indonderivative work: Kai walz (talk) - ROC_space.png,
CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=8326140>

製作者：TA 助教 - 林孝道

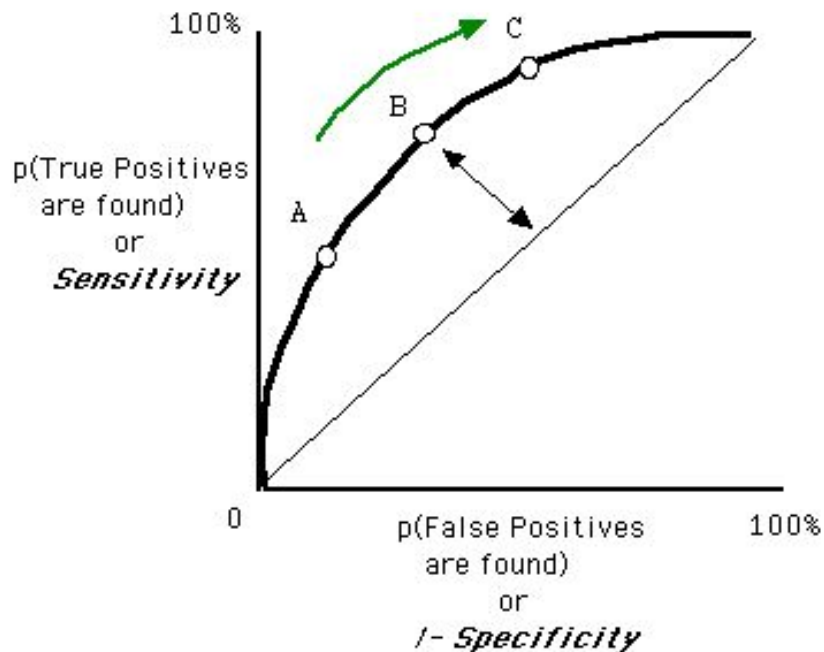
模型比較視覺化：ROC 曲線

- 進行分類時，我們會設置閾值(threshold)作為分類的分界。而 **設置不同的閾值會影響分類的結果**。
- 假設資料點 z 投影至 sigmoid function 後的值高於閾值分類為 1、反之則分類為 0。若此時閾值設定為 0.9，則六個資料點中左數第三個點可以被正確分類為 0 (投影值 0.7 < 0.9)。閾值設定為 0.5 或 0.2 都會讓第三個點分類錯誤



模型比較視覺化：ROC 曲線

- 因此當我們把所有的 threshold 都考慮進來的時候，我們就可以將 ROC 畫成一條曲線！
- **當曲線越靠近左上角時，代表該模型預測結果會越好！**因為這樣就有機會找到分類精確度更高的分界了！

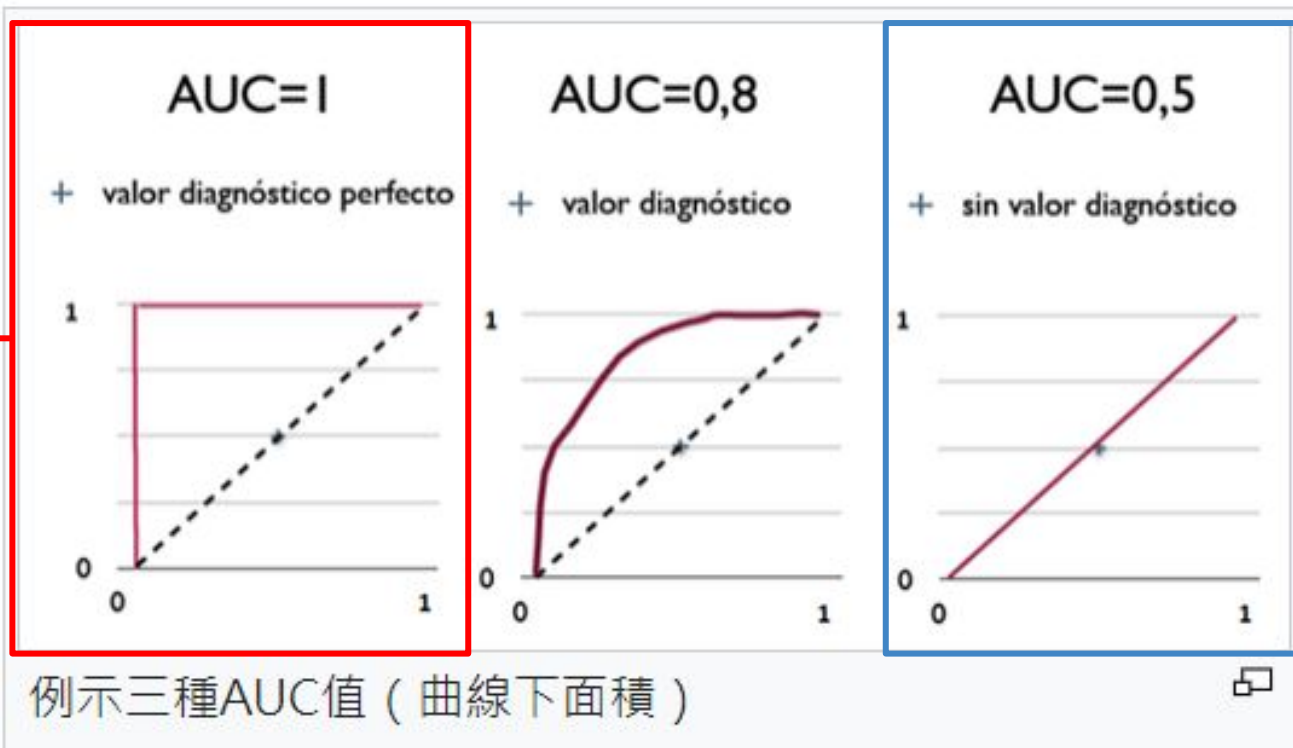


由 无法识别作者。根据版权声明推断作者为NekoJaNekoJa~commonswiki。 - 无法识别来源。根据版权声明推断为其自己的作品。 CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=407628>

ROC Curve & AUC Score

數字愈大愈好

最佳



無鑑別力

模型比較視覺化(Operator: Compare ROCs)

加入 Nominal to Numerical, 將自變數轉換成數值型態 (某些分類模型只接受數值型態的自變數)

Retrieve Titanic Training



Nominal to Numerical



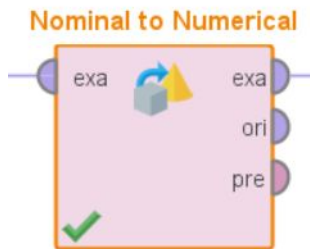
Compare ROCs




res
res
res

模型比較視覺化(Operator: Compare ROCs)


加入 Nominal to Numerical, 將自變數轉換成數值型態 (某些分類模型只接受數值型態的自變數)




Parameters ✕

 **Nominal to Numerical**

☐ create view ⓘ

attribute filter type  subset ⌵ ⓘ

attributes  Select Attributes... ⓘ

☐ invert selection ⓘ

☐ include special attributes ⓘ

coding type dummy coding ⌵ ⓘ

☐ use comparison groups ⓘ

unexpected value handling all 0 and warning ⌵ ⓘ


☐ use underscore in name ⓘ

Select Attributes: attributes ✕

Select Attributes: **attributes**
The attribute which should be chosen.


Attributes


Search ✕

 Survived


Selected Attributes

Search ⊕ ✕

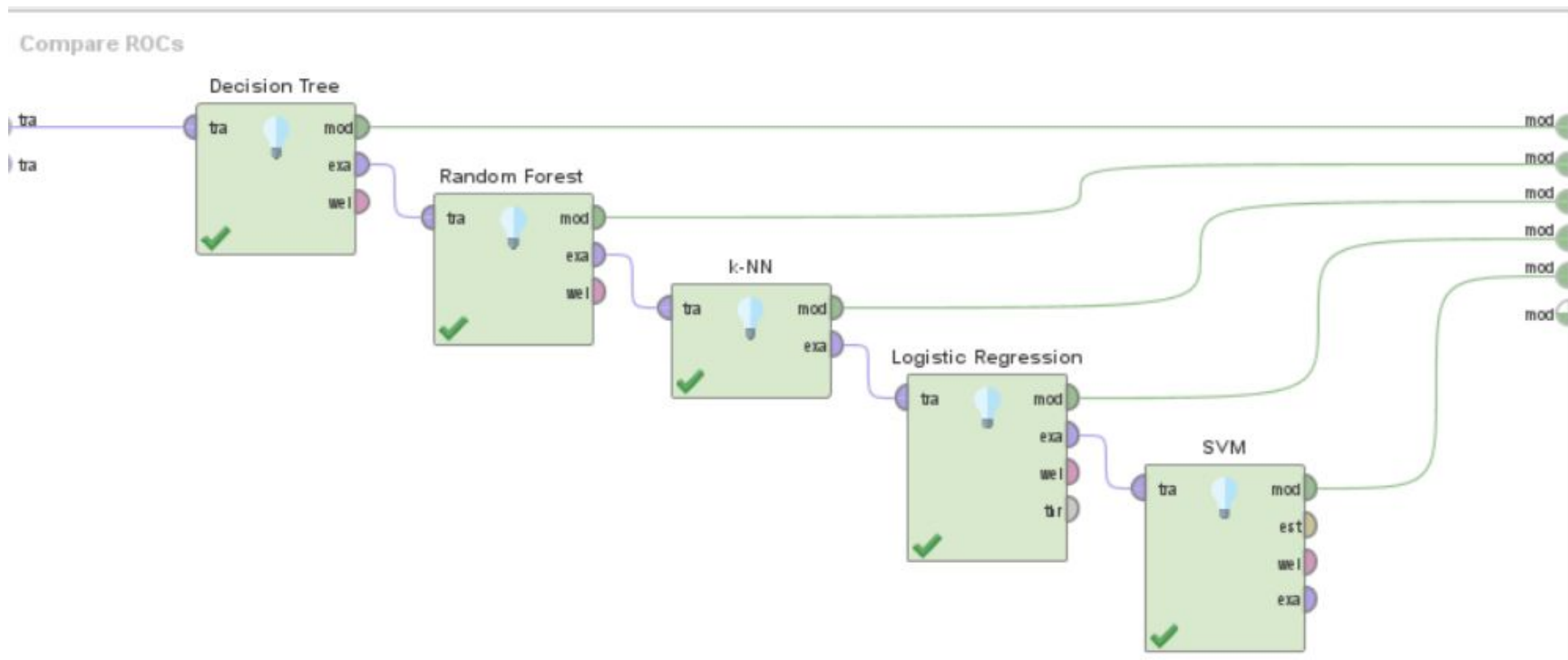
 Passenger Class

 Sex

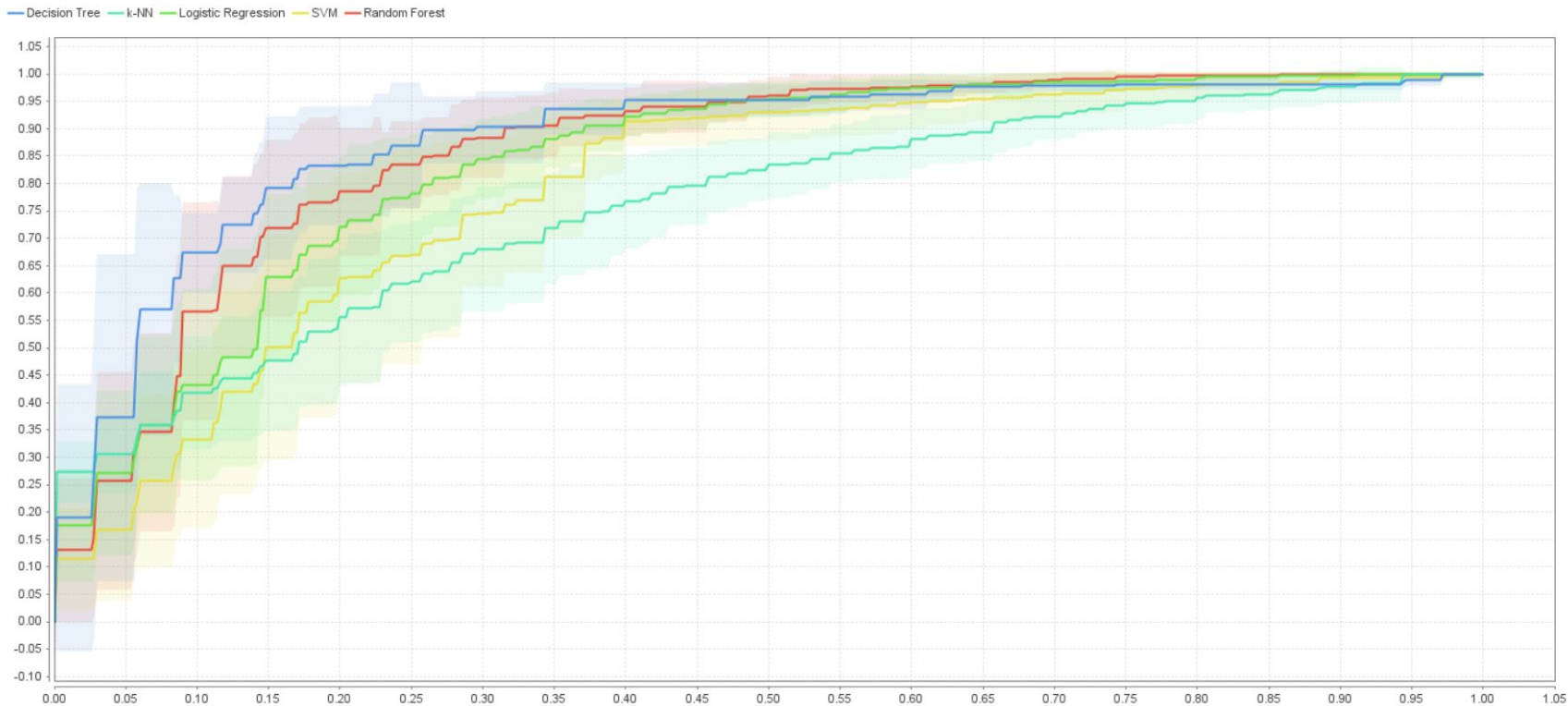
⬅ ➡

 Apply ✕ Cancel

模型比較視覺化：各種模型放入



模型比較視覺化：結果





Models (Classification)

常見分類模型

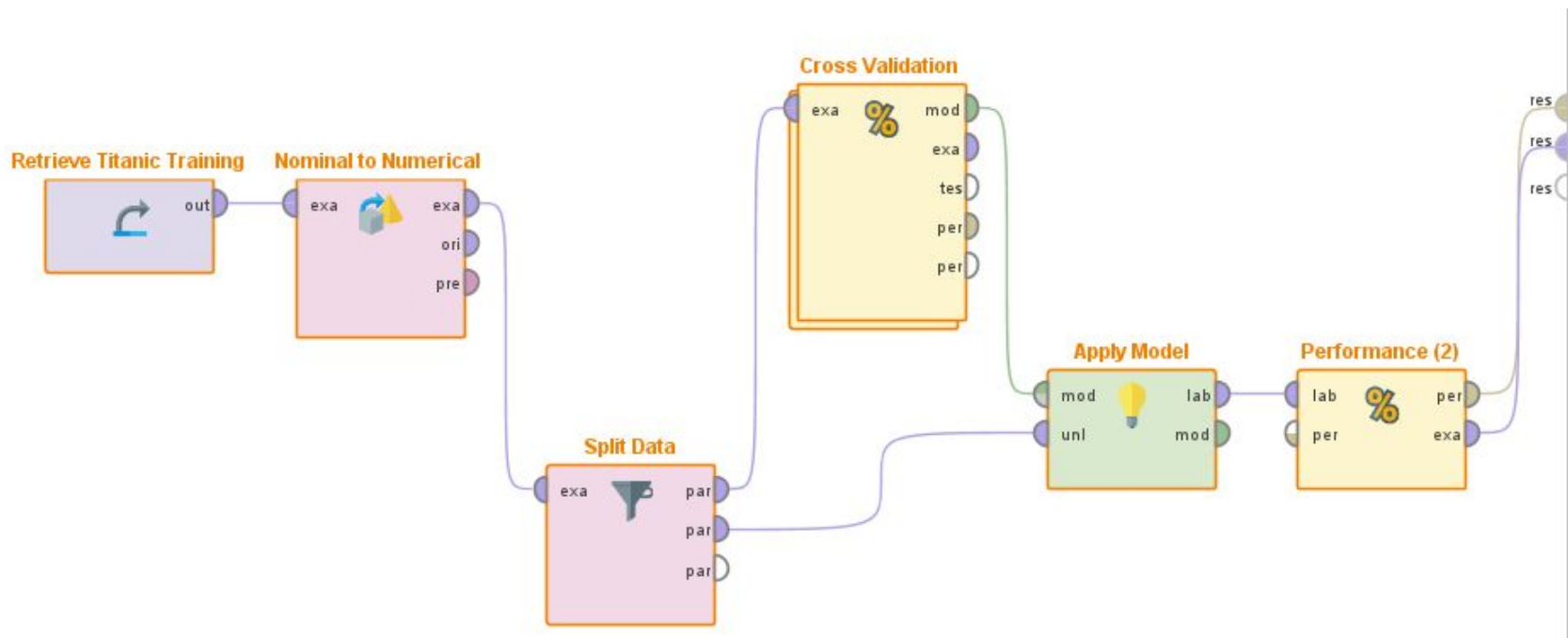


- Logistic Regression
- KNN
- 決策樹 (Decision Tree, Random Forest)
- Support Vector Machine
-

上課使用的範例流程檔



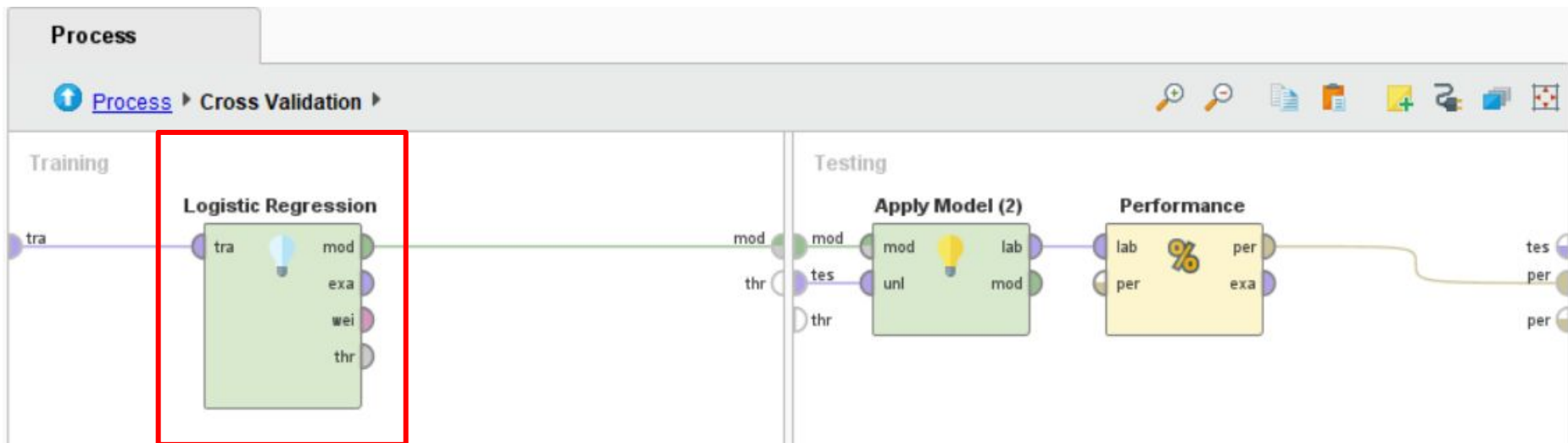
匯入RapidMiner3-Class-Example.rmp, 可以用於測試各種分類模型



上課使用的範例流程檔



點擊進入 Cross Validation 元件內部，可在紅色框起來的地方替換不同模型



Logistic Regression

還記得之前學過的線性迴歸嗎？

- Prediction

- $\hat{Y} = \hat{f}(X)$
- \hat{f} : our estimate for f
- \hat{Y} : resulting prediction for Y

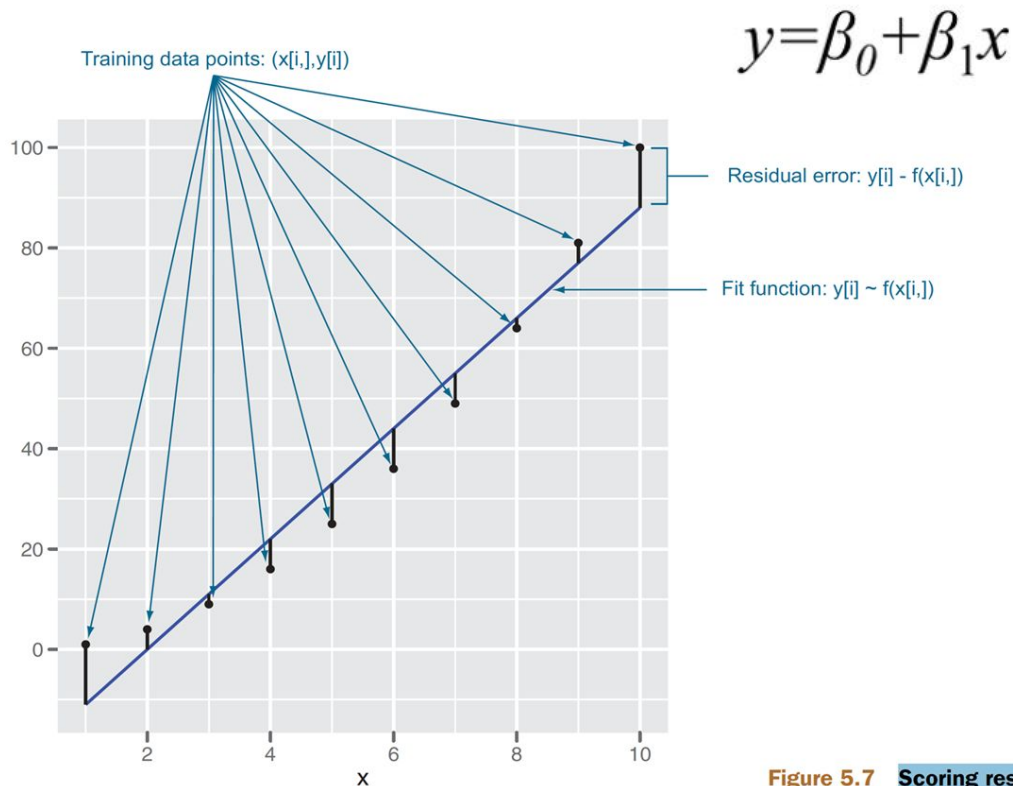
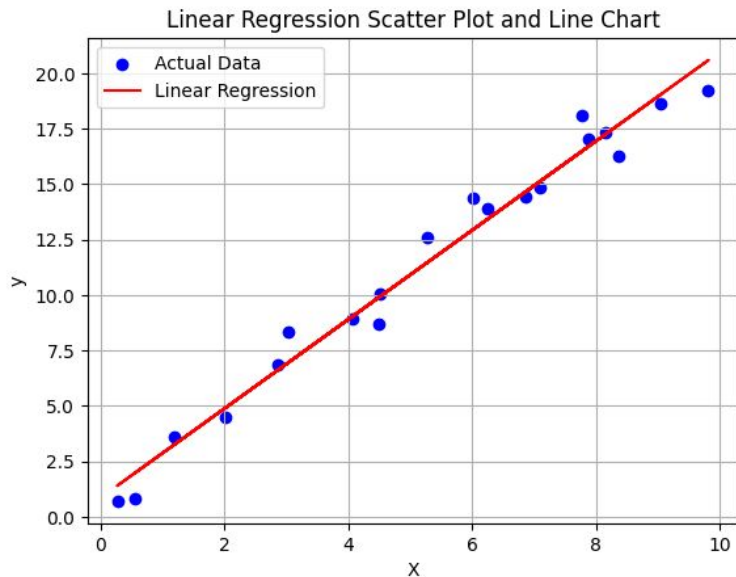


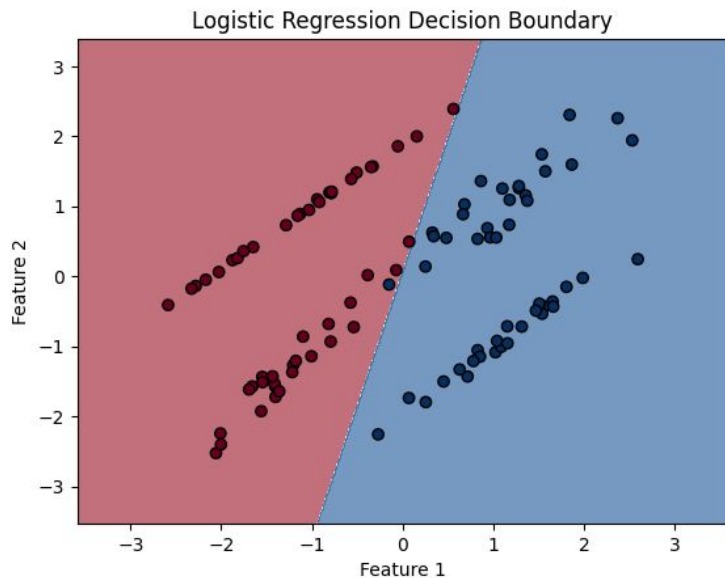
Figure 5.7 Scoring residuals

Logistic Regression

Logistic Regression 可以視為線性迴歸的一個變形，可用於分類任務



線性迴歸希望數據點都盡量符合紅線



Logistic Regression

Logistic Regression 會使用 Sigmoid 函數轉換數值, 讓最後的輸出結果為機率 (介於 0 和 1 之間); 接著根據閾值(Threshold) 做分類 (e.g. Threshold 為 0.5, 則 ≥ 0.5 為一類、 < 0.5 為一類)

Sigmoid function

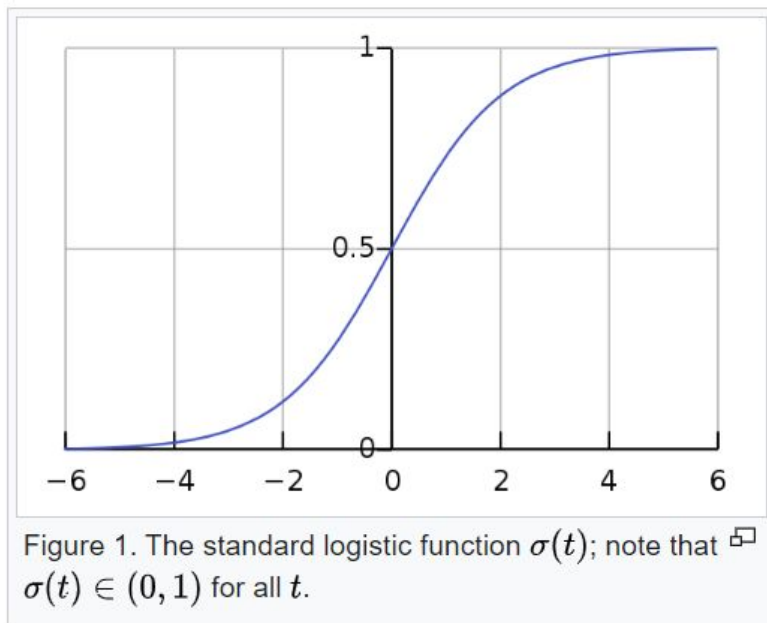
Article Talk

From Wikipedia, the free encyclopedia

A **sigmoid function** is a [mathematical function](#) having a characteristic "S"-shaped curve or **sigmoid curve**.

A common example of a sigmoid function is the [logistic function](#) shown in the first figure and defined by the formula:^[1]

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x} = 1 - \sigma(-x).$$



Logistic Regression

Linear Regression

$$f_{w,b}(x) = \sum_i w_i x_i + b$$

輸出值為任意數

Logistic Regression

$$f_{w,b}(x) = \sigma \left(\sum_i w_i x_i + b \right)$$

輸出值介於 0 和 1 之間

如何調整 FN 或 FP 的比率



以 Logistic Regression 為例，以下是原先範例的 PerformanceVector

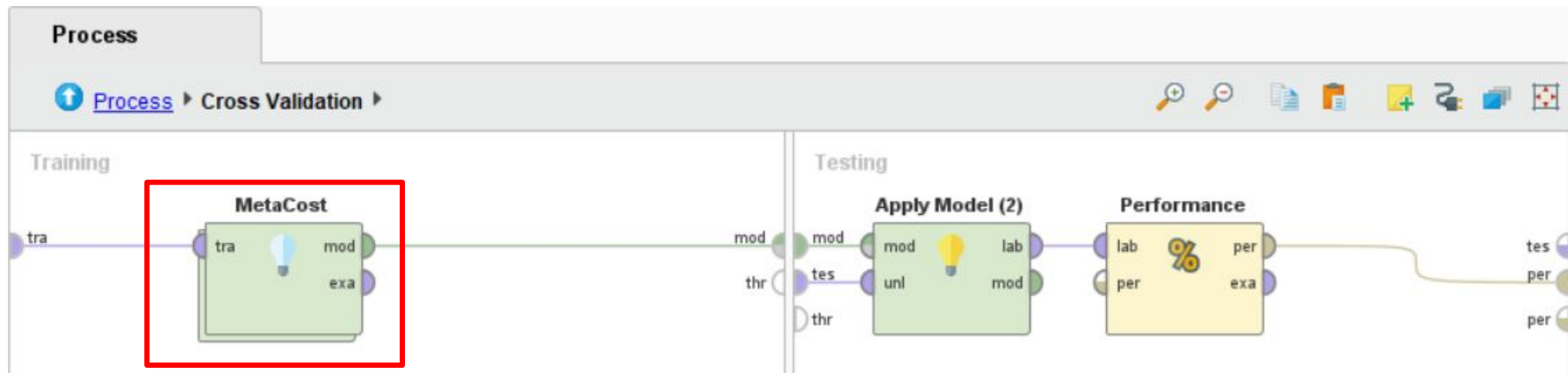
Q: 如果我要降低 False Negative (FN) 的比率，要怎麼做？

accuracy: 80.73%

	true Yes	true No	class precision
pred. Yes	78	26	75.00%
pred. No	27	144	84.21%
class recall	74.29%	84.71%	

如何調整 FN 或 FP 的比率

Ans: 使用 MetaCost 元件, 調高 FN 的模型懲罰



如何調整 FN 或 FP 的比率



Ans: 使用 MetaCost 元件, 調高 FN 的模型懲罰

Edit Parameter Matrix: cost matrix

Edit Parameter Matrix: **cost matrix**
The cost matrix in Matlab single line format

Cost Matrix	True Class 1	True Class 2
Predicted Class 1	0.0	1.0
Predicted Class 2	4.0	0.0

如何調整 FN 或 FP 的比率



PerformanceVector 的 FN 比率成功下降。

Q: 如果我要降低 False Positive (FP) 的比率, 要怎麼做?


accuracy: 73.09%

	true Yes	true No	class precision
pred. Yes	92	61	60.13%
pred. No	13	109	89.34%
class recall	87.62%	64.12%	

Kaggle 競賽

Smoker 預測

這次的作業將利用分類模型預測目標是否為抽菸者，預測目標為 smoking (0 沒抽, 1 有抽)

 Community Prediction Competition · Private

1121 W2_234 Computational Thinking - RapidMiner3

Binary Prediction of Smoker Status using - Classification Prediction. Please use the software "RapidMiner" exclusively for this competition.

Host Overview Data Discussion Leaderboard Rules Team Submissions ...

Overview

Binary Prediction of Smoker Status using Bio-Signals - Classification

The dataset used are from Kaggle Binary Prediction of Smoker Status using Bio-Signals

Playground Series - Season 3, Episode 24

<https://www.kaggle.com/competitions/playground-series-s3e24/>

Start

Set start date via the [launch checklist](#).

Close

a month to go



Competition Host

cheng-zhi-rong



Prizes & Awards

Kudos

Does not award Points or Medals

Participation

0 Competitors

0 Teams

0 Entries



Tags

Add Tags

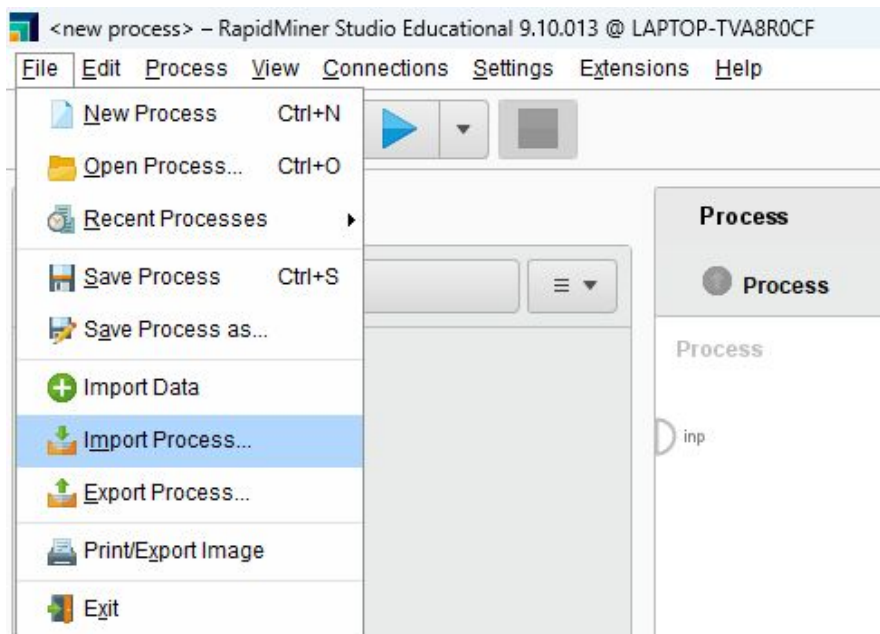
Smoker 預測 - 匯入資料

請從 Kaggle 下載 train.csv、test.csv >> 匯入 RapidMiner 並做以下設定 (記得勾選 Replace errors with missing values)

- **train.csv:** 使用 Change Role 將 id 改為 id、使用 Change Role 將 smoking 改為 label、使用 Change type 將 smoking 改為 binominal
- **test.csv:** 使用 Change Role 將 id 改為 id

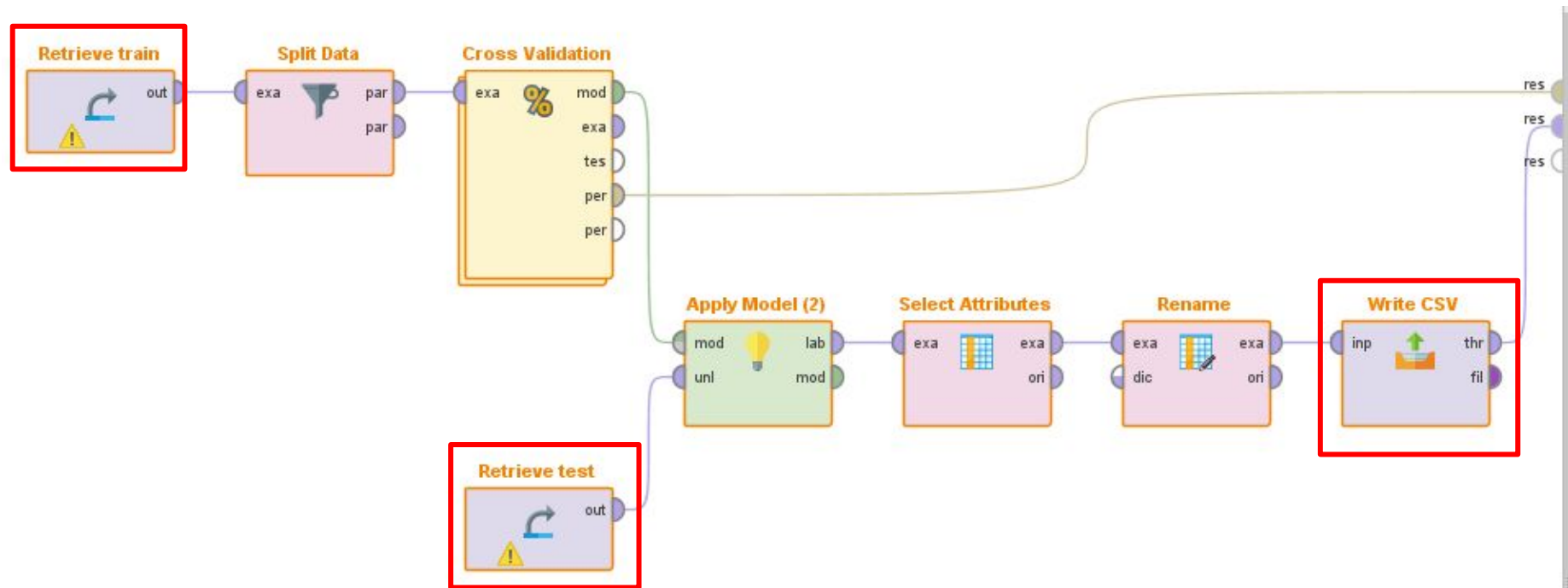
Smoker 預測 - 匯入資料

請從 Moodle 下載 RapidMiner3_Smoker_Status.rmp 並匯入 RapidMiner



Smoker 預測 - 設計流程

記得按照之前教學的方式更改檔案路徑、CSV 輸出位置



注意事項



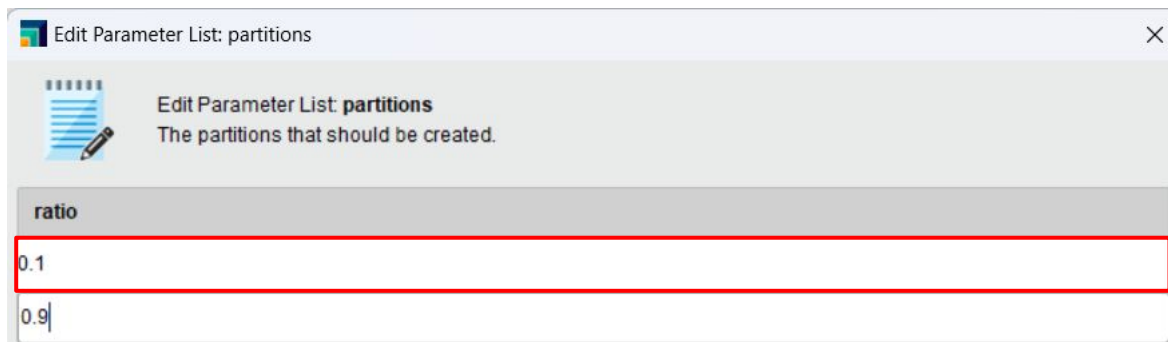
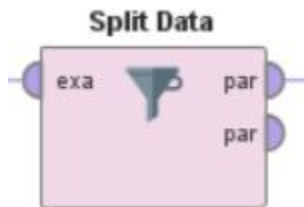
這次的 training dataset 大小為 9 mb 左右，照道理來講並不是很大的一筆資料，但是使用 RapidMiner 訓練模型卻需要花大量時間。所以這次的作業範例沒有使用 Optimize Parameters，且另外使用 Split data 將訓練資料再切小一點，避免大家做作業的時候等待太久。

期末報告盡量也不要找太大的資料集，找 kb 等級的資料比較安全。

注意事項



1. 以下分類模型不建議在這次的作業中使用，會計算很久
 - Support Vector Machine
2. 如果發現訓練時間還是太長的話，可以把 Split Data 第一格的 ratio 再調小一點 (下面範例表示只用 0.1 的 training data 訓練模型)



Smoker 預測 - 上傳格式

預測目標為 smoking (值為 Integer, 1 或 0)

sample_submission.csv (477.78 kB)



Detail Compact Column

2 of 2 columns ▾

About this file



This file does not have a description yet.

id	# smoking
111k	0
159k	0
111479	0
111480	0
111481	0
111482	0

Data Explorer

13.97 MB

sample_submission.csv

test.csv

train.csv

Evaluation- AUC Score

上傳分數愈高愈好

Evaluation



Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target.

Higher score is better

Smoker 預測 - Baseline

以下是今天範例的 Public 和 Private score, 要拿到 "加分題" 的同學你的 **Private score** 必須優於 Baseline

✓ Sandbox Submissions

Upload a Submission CSV and make sure it produces the expected score. These submissions are private unless tagged as a Benchmark, which appears on the Leaderboard.

Create sandbox submission

Submission and Description

Private Score ⓘ

Public Score ⓘ

Benchmark ⓘ



Smoker_Baseline.csv

Complete · 16s ago

0.75273

0.74709



更改 Kaggle Team Name

請注意 Team name 務必改成以下格式 學號-系級-名字

Overview Data Code Discussion Leaderboard Rules Team

Submissions

Submit Predictions

...

Your Team

Everyone that competes in a Competition does so as a team - even if you're competing by yourself. [Learn more.](#)

General

TEAM NAME

111753151-資碩一-程至榮

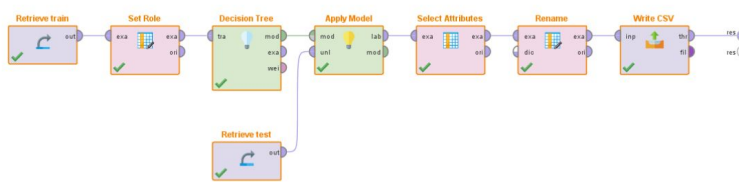
This name will appear on your team's leaderboard position.

基本題 4 分

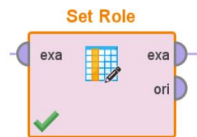
1. 不限定模型，成功上傳 kaggle(即 Leaderboard 有你的名字，且**格式正確**),
2. **Public Score** 優於 Baseline
3. 在 Moodle **上傳至少 1 頁 PDF**，說明你的設計流程、參數設定，並附上截圖(請參考簡報前面的教學是怎麼做的)
4. **上傳你的 Process file** (檔名: 學號_RapidMiner3.rmp, 範例: 111753151_RapidMiner3.rmp)

基本設計流程

以下為基本的設計流程，請透過 RapidMiner 的搜尋功能找出以下元件並排好。接著會說明參數設定，沒有特別講就是不用改設定。



參數設定



Parameters	
Set Role	
attribute name	Survived
target role	label
set additional roles	Edit List (0)...


加分題 1 分

1. 結算後在 **Leaderboard** 排名前 50% 且 **Private Score** 優於 **Baseline** 的同學可以得到額外的分數。

Public

Private

The private leaderboard is calculated with approximately 70% of the test data.

#	△	Team	Members	Score	Entries	Last
		Smoker_Baseline.csv		0.75273		

透過上課教過的方法 + 上網查詢、調整資料前處理的方法 or 模型參數。

接下來的作業都沒有標準答案, 請大家盡可能的去嘗試!

作業注意事項

- 為了公平起見，且大家的**期末專題海報**預計會與 micro:bit 或 RapidMiner 有關，此作業**限定使用 RapidMiner 產生的 Submission 參賽**，請確認繳交的 Process file 可以產生正確的 Submission 檔案
- **請不要抄襲、或是直接拿別人的 Submission 檔案上傳**，助教會隨機抽查是否有排名分數與 Process file 不一致的問題。

Reference

- [大數據驅動商業決策: 13 個 RapidMiner 商業預測操作實務](#)
- [RapidMiner 人工智慧機器學習軟體](#)
- [Data Science course by professor Jia-Ming Chang](#)
- [基礎統計名詞介紹網頁](#)
- [2021 iThome 鐵人賽 - 全民瘋 AI 系列 2.0](#)
- [Hung-yi Lee 機器學習](#)

Tools

- [ZoomIt - Sysinternals - Microsoft Learn](#)

