Lecture 3 : Classification (2)

楊立偉教授

wyang@ntu.edu.tw

Decision Tree

Using Decision Tree for Classification

- 利用資料庫內每筆資料的已知欄位，預測目標欄位之值，並做為分類的依據
  - 可以將大量資料轉化成人類易於了解的知識樹
  - 常見應用：信用評等、消費行為預測、病症診斷

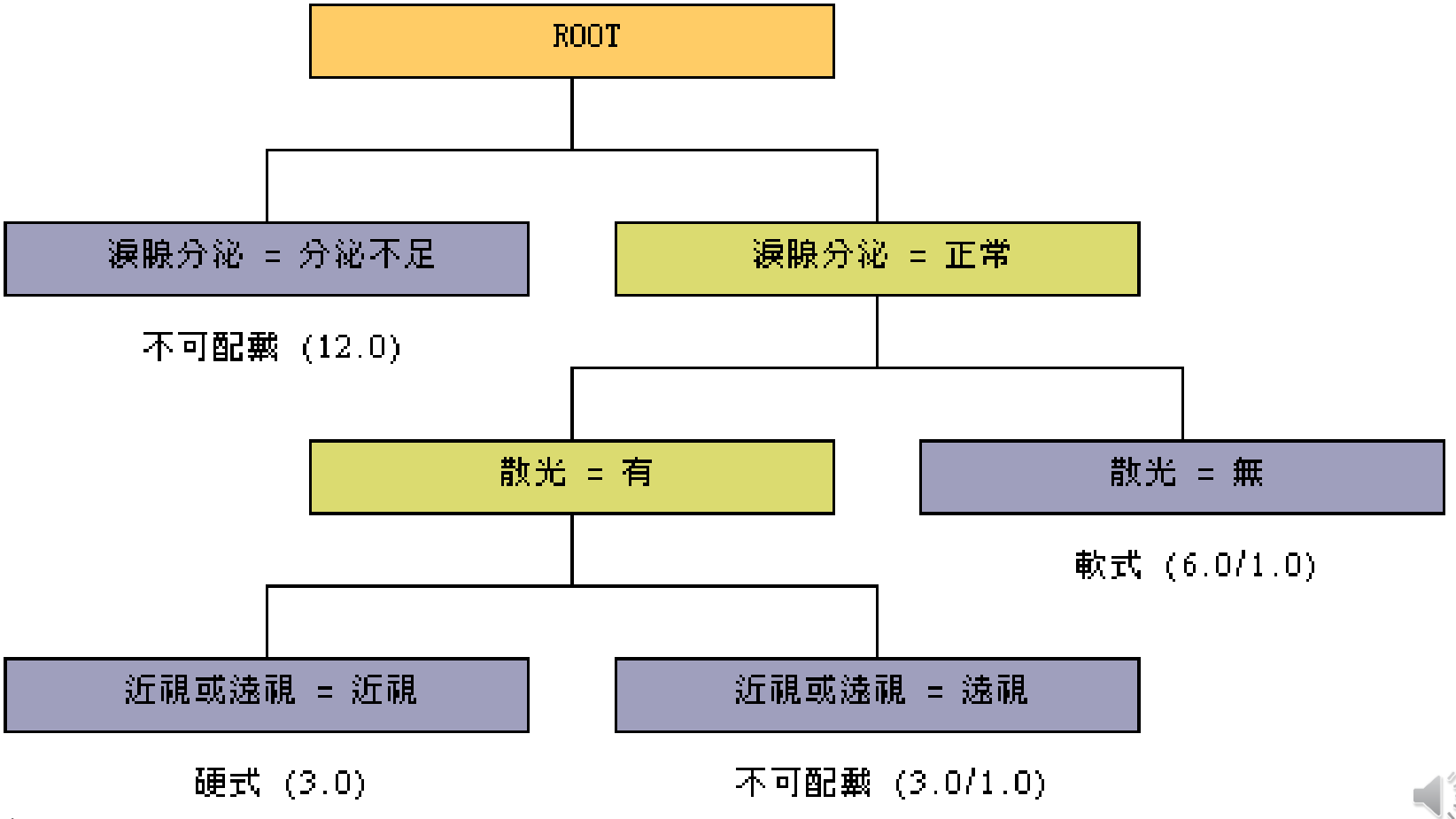
分類預測：眼科診所病例

年齡	近視或遠視	散光	淚腺分泌	隱形眼鏡處方
兒童	近視	無	分泌不足	不可配戴
兒童	近視	無	正常	軟式
兒童	近視	有	分泌不足	不可配戴
兒童	近視	有	正常	硬式
兒童	遠視	無	分泌不足	不可配戴
兒童	遠視	無	正常	軟式
兒童	遠視	有	分泌不足	不可配戴
兒童	遠視	有	正常	硬式
成人	近視	無	分泌不足	不可配戴
成人	近視	無	正常	軟式
成人	近視	有	分泌不足	不可配戴
成人	近視	有	正常	硬式
成人	遠視	無	分泌不足	不可配戴
成人	遠視	無	正常	軟式
成人	遠視	有	分泌不足	不可配戴
成人	遠視	有	正常	不可配戴
老年	近視	無	分泌不足	不可配戴
老年	近視	無	正常	不可配戴
老年	近視	有	分泌不足	不可配戴
老年	近視	有	正常	硬式
老年	遠視	無	分泌不足	不可配戴
老年	遠視	無	正常	軟式
老年	遠視	有	分泌不足	不可配戴
老年	遠視	有	正常	不可配戴

wyang@ntu.edu.tw

分類預測：眼科診所病例 (續)

- 自動選擇最佳分支條件，產生決策樹

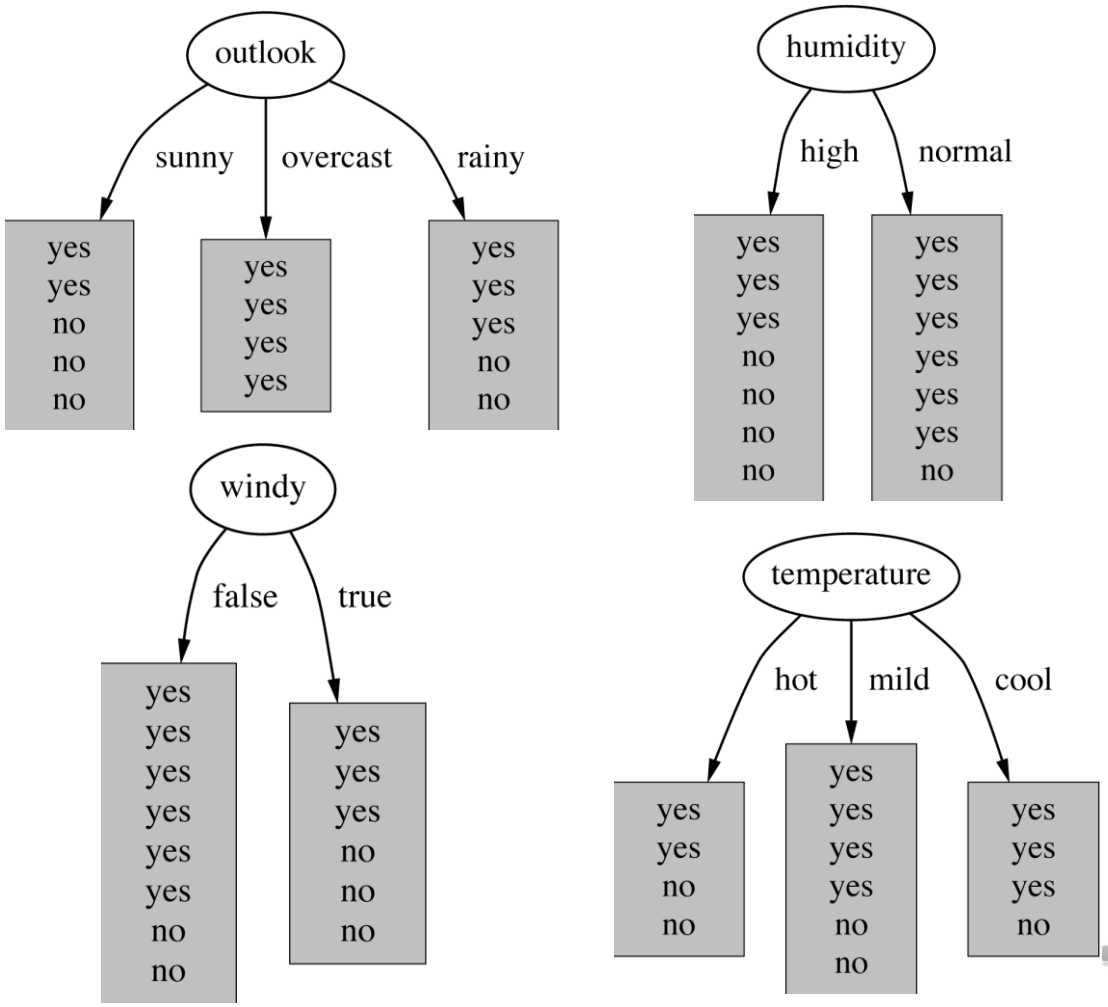


決策樹演算法 範例

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

- Weather Data: Play tennis or not ?

Which attribute to choose ?



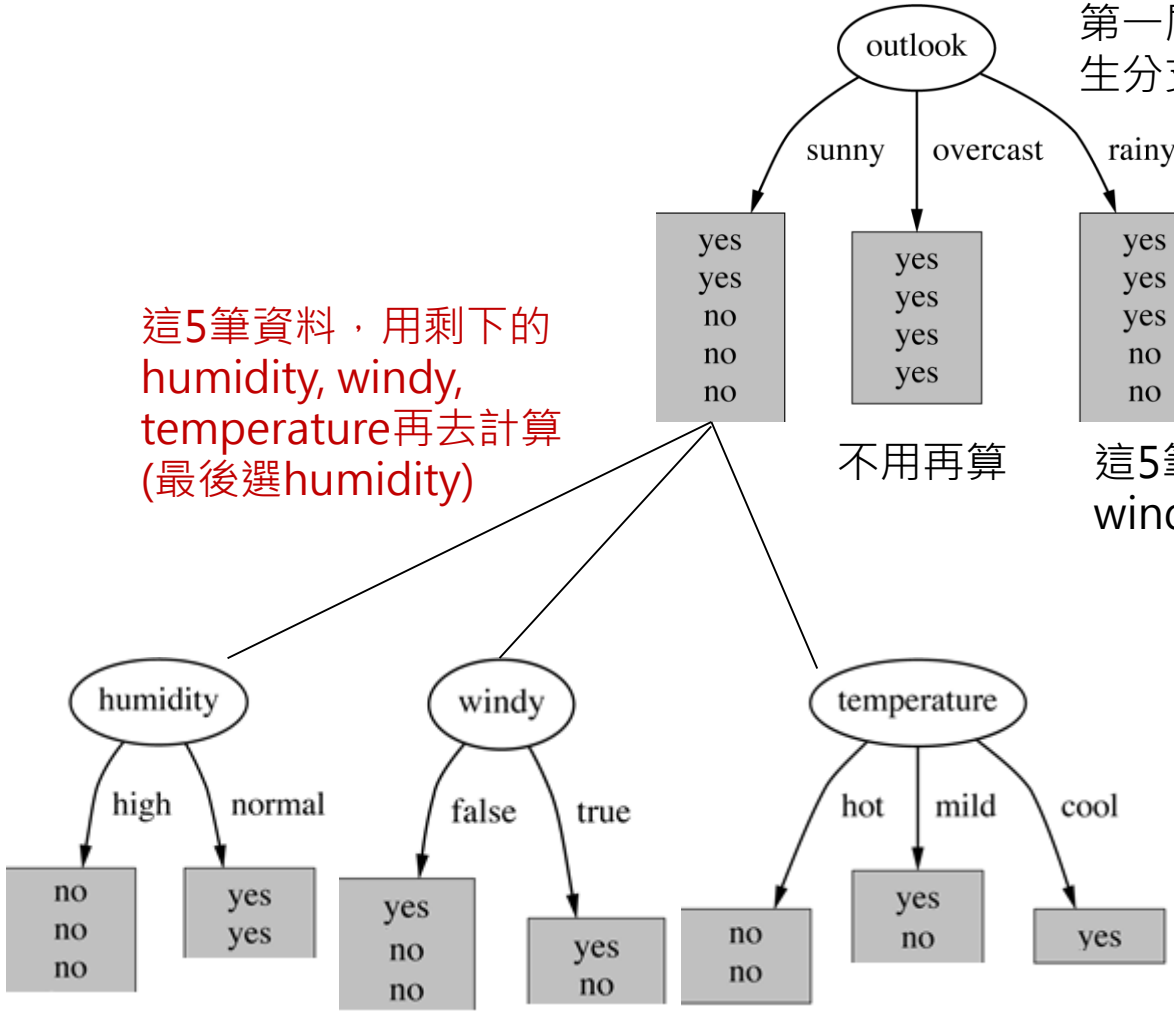
- Which attribute to choose ?
  - choose the attribute that produces the "purest" nodes
  - …and more informative
  - 常見演算法 Information gain (ID3, C4.5, C5)

ig(outlook) = average(3/5, 4/4, 3/5) = 0.73

ig(humidity) = average(4/7, 6/7) = 0.71

ig(windy) = average(6/8, 3/6) = 0.63

ig(temperature) = average(2/4, 4/6, 3/4) = 0.64



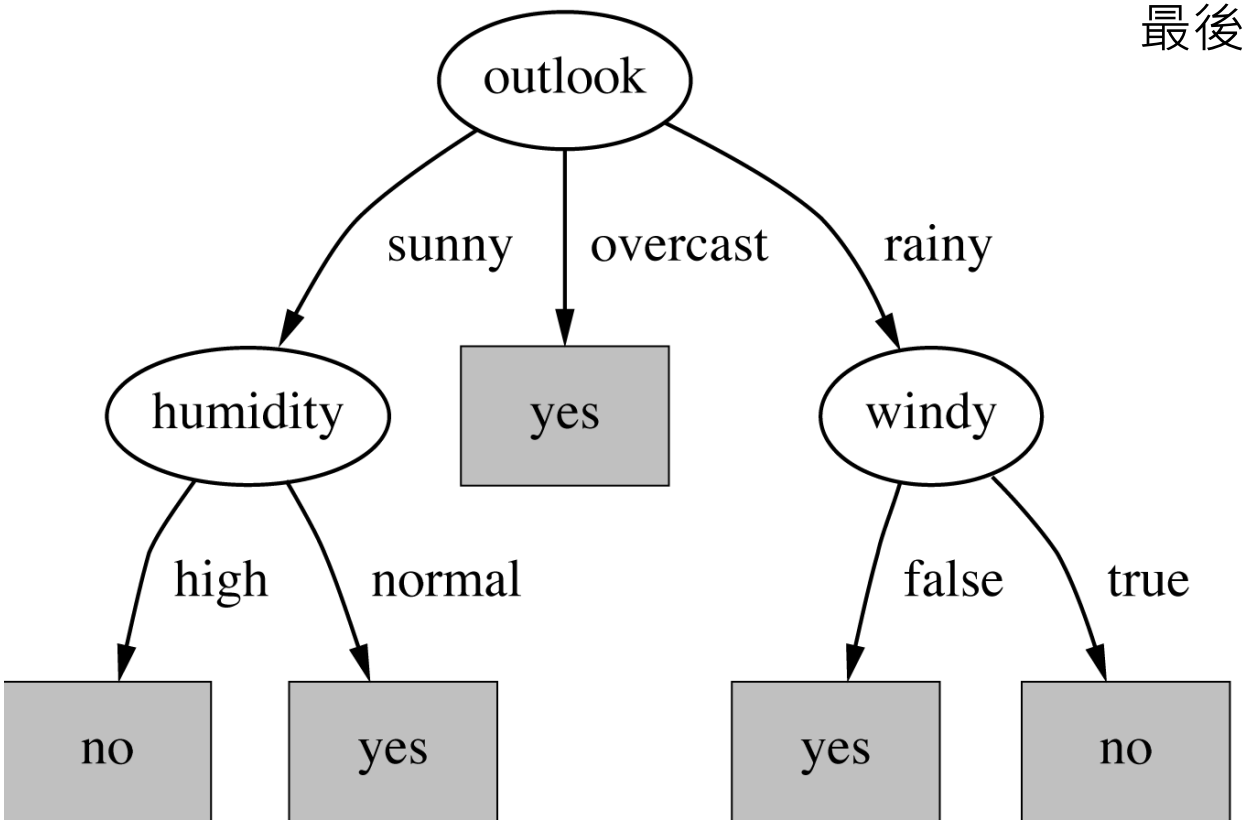
第一層選擇outlook，往下重複產生分支，直到結束或終止條件為止

這5筆資料，用剩下的humidity, windy, temperature再去計算 (最後選humidity)

不用再算

這5筆資料，同樣也用剩下的humidity, windy, temperature去算 (圖略)

最後的結果


outlook
sunny
overcast
rainy
humidity
yes
windy
high
normal
false
true
no
yes
yes
no

What does Decision Tree help ?

- 自動嘗試所有欄位排列組合
- 找出關鍵決策因素之優先順序
- 自動切割適當值
- 自動排除無關因素，並排除少數狀況 (pruning)
- 可用以解釋、協助決策或預測之用

Variation of Decision Trees

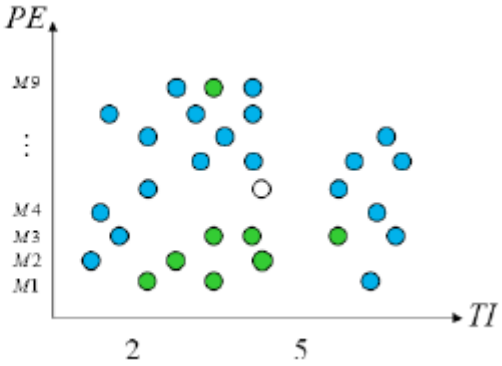
- Information gain, Entropy, Gini index (CART)
- Tree pruning
  - reduces the size of decision trees by removing sections of the tree that provide little power to classify instances.
  - reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.

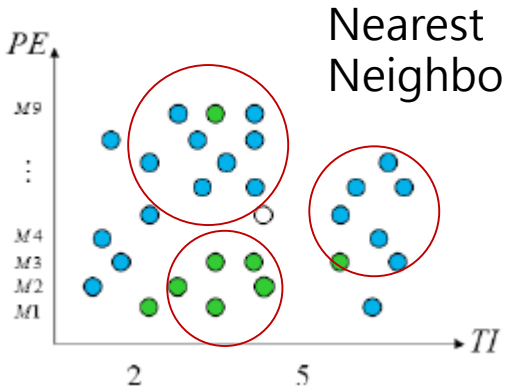
Ensemble Learning

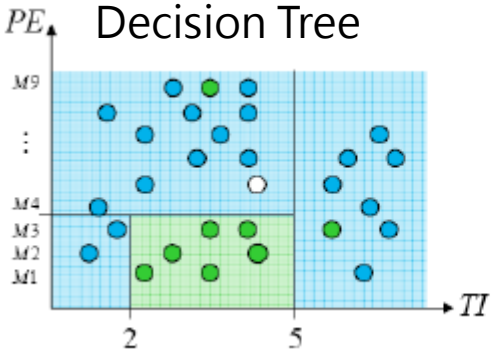
幾種分類器的概念比較

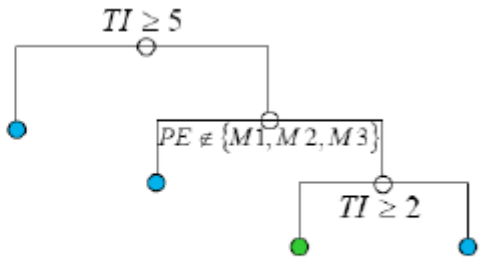
- Simple dataset with two predictors

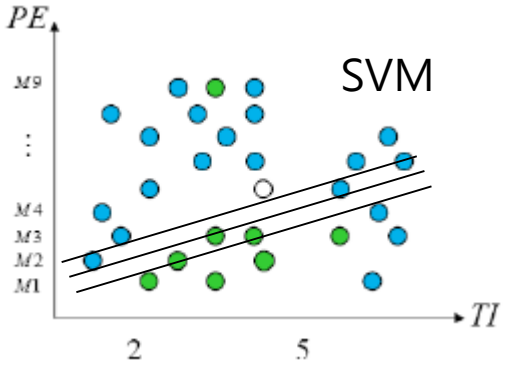
TI	PE	Response
1.0	M2	good
2.0	M1	bad
...	...	...
4.5	M5	?











- 組合算法 Bagging (Breiman, 1996)
  - 每種分類器的特性不同，可以使用一種或多種分類器，配合多回合的學習 (每一回合都從原訓練資料中抽取若干作為訓練資料)。最後，對於分類問題，採用投票方式作判別。

- 增強算法 Boosting (Shapire et al., 1998)
  - 在多回合的學習中，初始化時以等權重、有放回抽取樣方式進行訓練，接下來每次訓練要特別留意前一次分類失敗的訓練樣本，並賦以較大的權重進行抽樣。而在多分類器中，給予準確率較高者較大權重，反之較小。最終採用有權重的投票方式作判別。
    - 好像人在學習，訂正答案後，把上次做錯的於下次特別加強，經過如此針對性的練習，解題能力自然會上升
    - 與Bagging的差異，在於前者大部分採均勻抽取，而Boosting根據錯誤率來抽取。 Bagging和Boosting都可以提高分類的準確性，在多數研究中，Boosting又會比Bagging再高一些

Random Forest (Breiman, 2001)

- Construct the "forest" from decision trees
  - add an additional layer of randomness to bagging
    1. construct each tree using a different bootstrap sample of the data
    2. each node is split using the best among a subset of predictors randomly chosen (instead of all attributes)

Random Forest (Breiman, 2001)

- Comparison
  - counterintuitively but outperform very well
    - compared to discriminant analysis, support vector machines and neural networks
    - robust against overfitting
  - very user-friendly; only two parameters concerned
    - number of variables in the random subset at each node
    - number of trees in the forest

Ref: Classification and regression by randomForest (以R實作)

Random forest algorithm

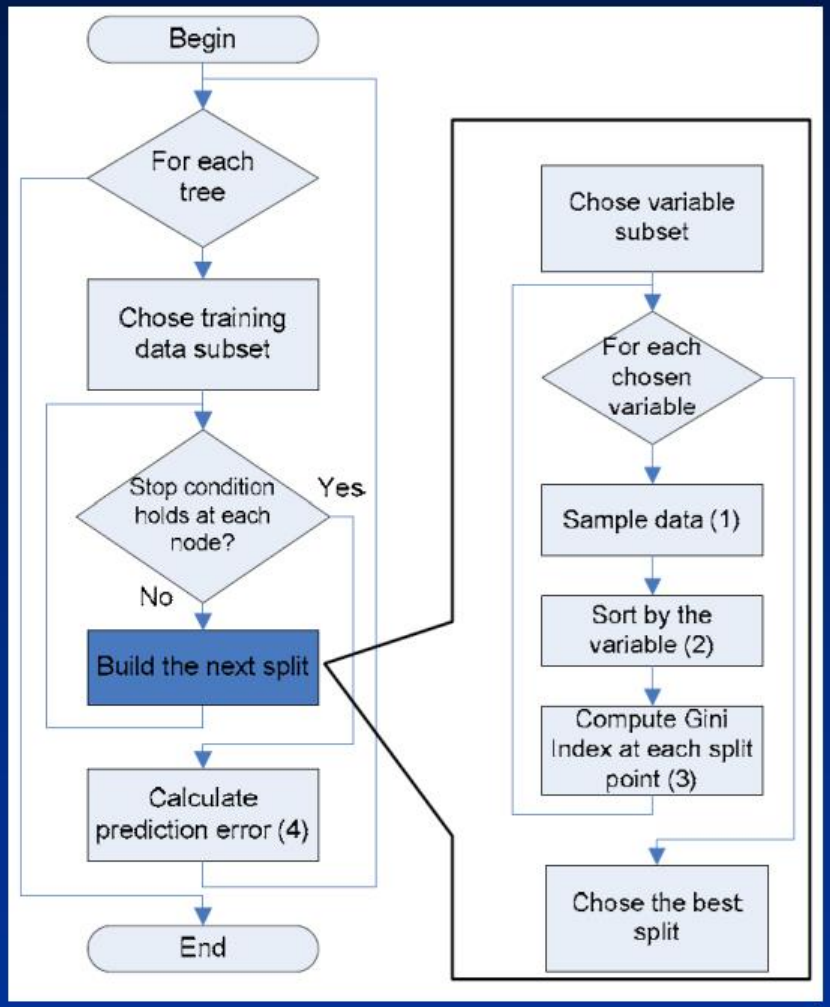
Let N_{trees} be the number of trees to build
for each of N_{trees} iterations

1. Select a new bootstrap sample from training set
2. Grow an un-pruned tree on this bootstrap.
3. At each internal node, randomly select m_{try} predictors and determine the best split using only these predictors.
4. Do not perform cost complexity pruning. Save tree as is, along side those built thus far.

Output overall prediction as the average response (regression) or majority vote (classification) from all individually trained trees

Random forest algorithm (flow chart)

For computer scientists:



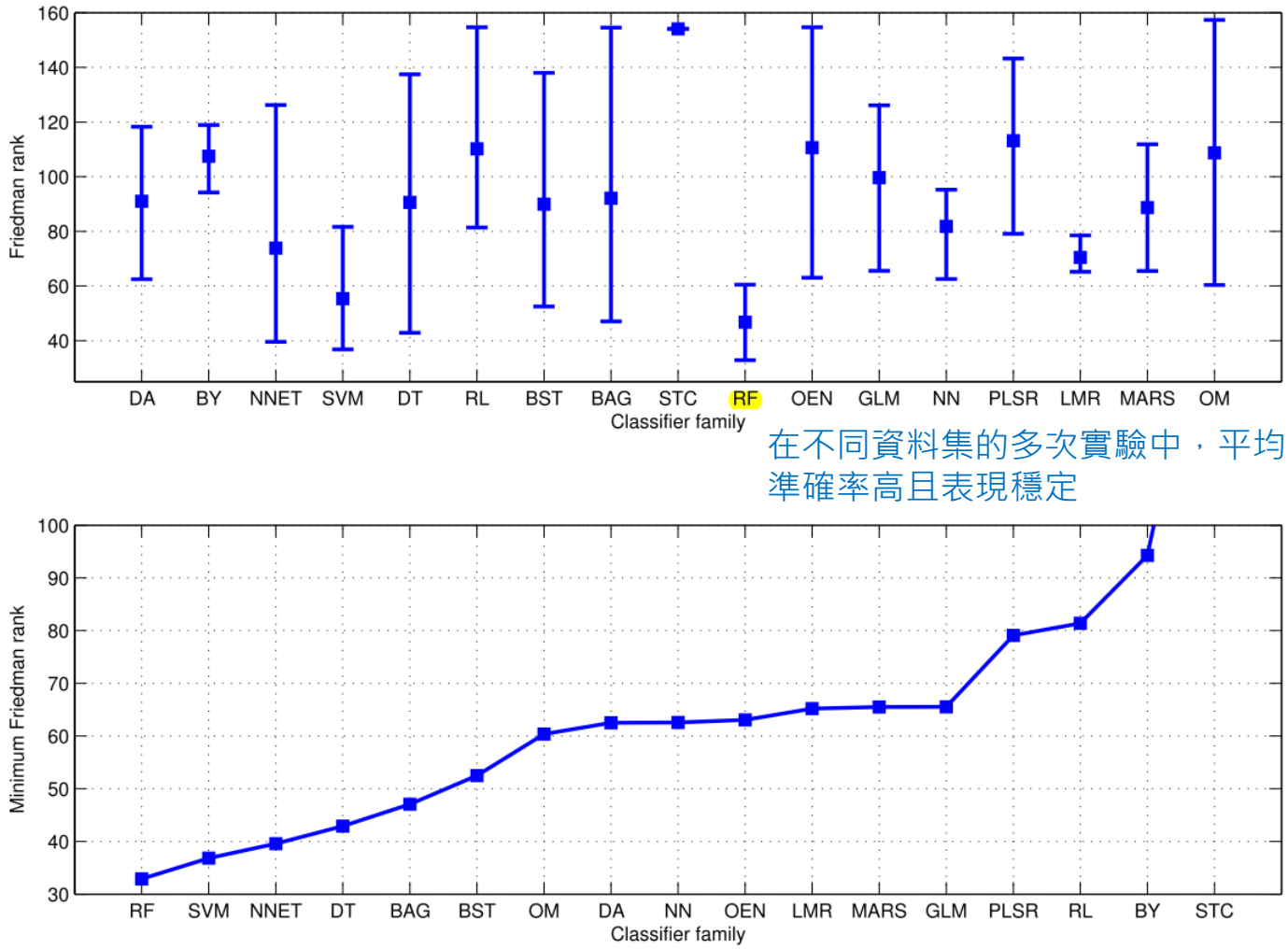
Comparison (Fernández-Delgado, Manuel, et al, 2014)

- "Do we need hundreds of classifiers to solve real world classification problems?"
  - evaluate 179 classifiers arising from 17 families
    - discriminant analysis, Bayesian, neural networks, support vector machines, decision trees, rule-based classifiers, boosting, bagging, stacking, random forests and other ensembles, generalized linear models, nearest-neighbors, partial least squares and principal component regression, logistic and multinomial regression, multiple adaptive regression splines and other methods)
    - implemented in Weka, R (+caret package), C and Matlab
    - 121 data sets from UCI

- #1 Random Forest
  - R (+ caret package), 94.1% accuracy
- #2 SVM
  - C + LibSVM with Gaussian kernel, 92.3% accuracy

– Top 20

Rank	Acc.	κ	Classifier
32.9	82.0	63.5	parRF_t (RF)
33.1	**82.3**	**63.6**	rf_t (RF)
36.8	81.8	62.2	svm_C (SVM)
38.0	81.2	60.1	svmPoly_t (SVM)
39.4	81.9	62.5	rforest_R (RF)
39.6	82.0	62.0	elm_kernel_m (NNET)
40.3	81.4	61.1	svmRadialCost_t (SVM)
42.5	81.0	60.0	svmRadial_t (SVM)
42.9	80.6	61.0	C5.0_t (BST)
44.1	79.4	60.5	avNNet_t (NNET)
45.5	79.5	61.0	nnet_t (NNET)
47.0	78.7	59.4	pcaNNet_t (NNET)
47.1	80.8	53.0	BG_LibSVM_w (BAG)
47.3	80.3	62.0	mlp_t (NNET)
47.6	80.6	60.0	RotationForest_w (RF)
50.1	80.9	61.6	RRF_t (RF)
51.6	80.7	61.4	RRFglobal_t (RF)
52.5	80.6	58.0	MAB_LibSVM_w (BST)
52.6	79.9	56.9	LibSVM_w (SVM)
57.6	79.1	59.3	adaboost_R (BST)



在不同資料集的多次實驗中，平均準確率高且表現穩定

Figure 6: Friedman rank interval for the classifiers of each family (upper panel) and minimum rank (by ascending order) for each family (lower panel).

Discussions

Exercise

- 思考決策樹與迴歸的特性
  - 目標值可能是一個不連續、非線性的分布
  - 造成任何一個變數與目標值的相關性都不高
  - 但是可能多個變數合成之後相關性就提高
  - 若再加上條件判斷 (conditional) 則相關性更高