4105931 機器學習 (Machine Learning)

得分

姓名:_____

學號:_____

Midterm

總分 105 分

中文作答

1. (10%) 名詞解釋 a) Bootstrapping、b) Uniform Blending、c) Bagging、d) Deterministic noises and e) Stochastic noises in overfitting.

   (a) Bootstrapping:

   從原始具有 N 個 samples 的資料集 D 中，隨機抽取N′個 samples 組成新的資料集$\widetilde{D_t}$，注意原本的 N 個 samples 是可以被重複選取的，此過程即為 Bootstrapping。

   e.g. $D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$，Bootstrapping 取樣後$\widetilde{D_t} = \{(x_1, y_1), (x_1, y_1), (x_3, y_3)\}$。

   (b) Uniform Blending:

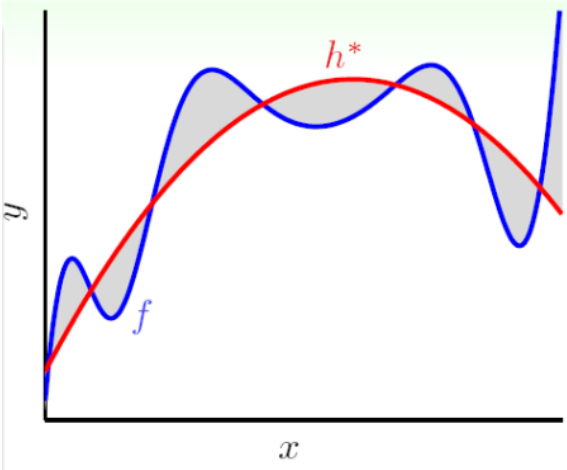
   給予每個小弱分類器g_t相同權重α，將它們 aggregate 起來，稱為 Uniform Blending。透過給予的統一權重的不同 e.g. 1, $\frac{1}{T}$，以及解決任務的不同，Uniform Blending 可以達到 voting、average 的效果。

   (c) Bagging:

   Bagging 即 Bootstrap Aggregation，其透過 Bootstrapping 特性取出不同的資料集並透過學習演算法A，進一步訓練出具有 diversity 的多個小弱分類器g_t，再將所有g_t以 Uniform Blending 方式 aggregation，得到 G。

   (d) Deterministic noises

   訓練集資料的理想函數 f 與理論上最佳模型h^*之間的差距，如下圖灰色部分。

   (e) Stochastic noises

   在模型訓練過程中，因收集到的訓練資料不夠具代表性或訓練資料在收集、採樣不佳導致的隨機雜訊。

2. (5%)請說明何謂機器學習?

   機器學習(Machine Learning = ML)是透過演算法將收集到的資料進行分類或預測模型訓練，當未來得到新的資料時，可以透過訓練出的模型進行預測 ，如果這些效能評估可以透過利用過往資料來提升的話，就叫機器學習 。

3. (6%) 試說明機器學習的使用時機，須滿足那些條件？並依每一個條件舉實例，說明何種情形不該使用機器學習。

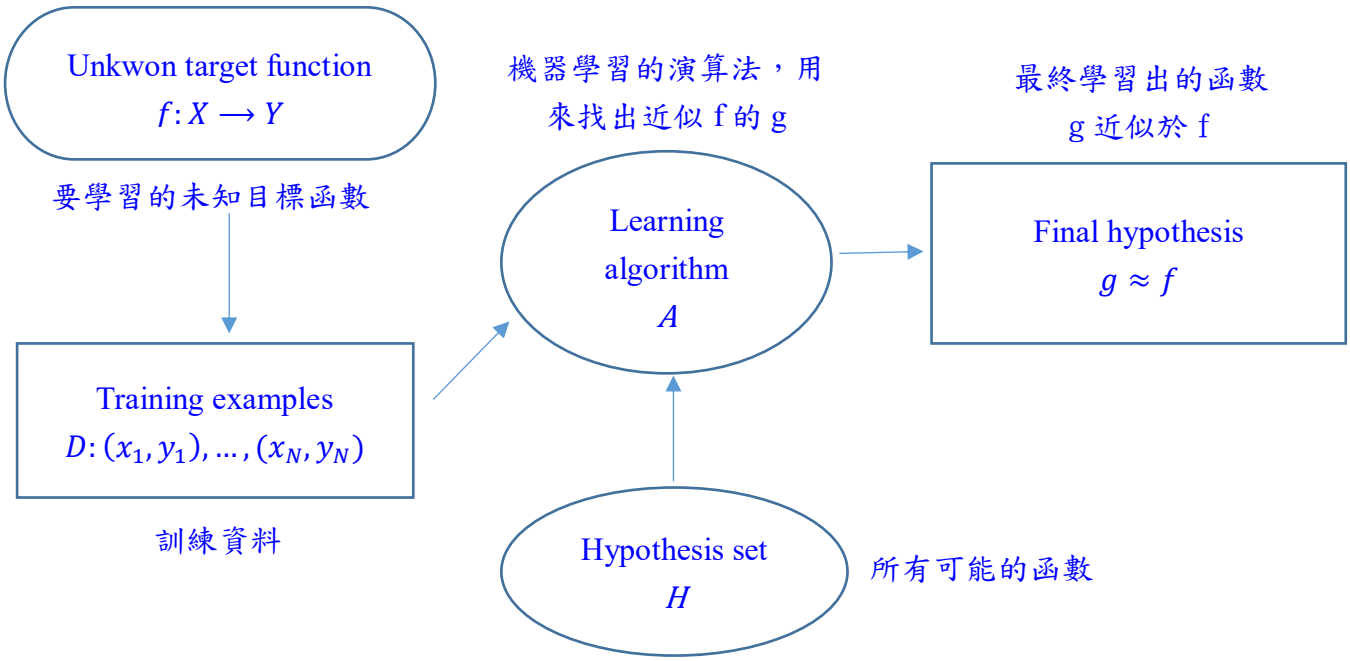
a. 訓練資料中存在某種規律是可以被學習的。
b. 要學習的技術不容易簡單的列出規則或沒有既存的演算法。
c. 存在可以代表這個要學習的規律的Data。

不適用的case (依上述三條件舉例):

a. 無法預測大樂透的開獎結果
b. 任何可用演算法解的問題皆不符合條件b. e.g.使用Dijkstra algorithm找最短路徑
c. 不存在可以預測世界末日何時發生的資料

4. a) (5%)機器學習的學習流程圖，包含五項主要的 component，試說明其意義。b) (5%)以 PLA 為例，假設要學習 PLA 的二元分類器，分出 2D 平面上分布的 100 個黑點跟白點，試說明這樣的問題如何對應到上述五項 component?

(a)



(b) Unkwon target function f:

能完美將所有黑、白點完全分類成功的二維平面直線或二維的線性二元分類器。

Training examples D:

100 個黑點及白點的二維座標與黑白的答案。

Learning algorithm A :

使用 $\text{sign}(W^T X_{n(t)}) \neq y_{n(t)}$，找出錯誤的點，再利用PLA 的方式更新權重。

$(W_{t+1} \leftarrow W_t + y_{n(t)} X_{n(t)})$

或者直接說 Perceptron Learning Algorithm (PLA)

Hypothesis set H :

2D 平面所有可能的直線 / 所有二維的線性二元分類器

Final hypothesis g:

修正完全部錯誤點後，從 Perceptron 回傳的 W，

或由演算法最終挑選出的分類器

5. (4%) 有鑑於理工科的男生都不太會穿搭，所以我做了下面這樣的事情。首先收集了 2023 年整年的紐約時裝周服飾照片，然後依照影像處理的作法，擷取了顏色、花紋、材質三種特徵，每一張影像將三種特徵串接成一種新的特徵表示法，然後分群成 100 群，接著我拿出我自己穿搭的照片，取出這種新的特徵表示法，跟 100 群每一群的群中心算距離，假設距離小於預設的門閥值 50，我就得一分，看看最終總得分的高低，來決定我的穿搭時不時尚。試以 Type of Learning 的角度，說明這個方法是屬於什麼樣的機器學習方法?

Output space y	Regression (輸出為一分數)
Different Data Label y	Unsupervised (使用 clustering)
Different Protocol f	Batch (訓練時只使用一批資料，無再追加)
Different Input Space x	Concrete feature: (擷取顏色、花紋、材質三種特徵，每一張影像將三種特徵串接成一種新的特徵表示法)

6. 如果我想訓練一個“預測碩班畢業會不會進台積電上班”的模型，試說明如何學習這樣的模型? 請依序回答: a) (2%)請從這個問題，**具體定義**模型的輸入資料、輸出資料。b) (2%)請說明如何收集與標記 a)中所提到的輸入輸出資料。c) (2%)請說明應該用什麼模型比較適合解這個問題。d) (4%)試說明如何訓練這樣的模型? 請從模型、損失函數、訓練、驗證到測試，說明完整步驟。

(a) Input :學歷、在校成績、side project 經驗、實習經驗…
Output :會/不會去台積電上班

(b) 學歷、在校成績可透過分級級距標記，side project 經驗、業界實習的有無可以用 0,1 表示法標記。上述資料收集、標記可由公司人事部門決定。

(c) 由於輸出結果為”會/不會”進台積電上班，故屬於二元分類問題，可用 perceptron、decision tree、SVM。

(d) 1. 收集過往錄取與不錄取者資料。
2. 將資料數值化。
3. 資料進行前處理(normalization 等)，將資料切分成訓練/測試集。
4. 決定使用的模型(decision tree)，決定損失函數為 classification error。
5. 整理資料格式並輸入模型訓練
6. 透過交叉驗證確認模型效果
7. 將測試資料輸入到訓練好的模型進行預測，比較預測結果與真實結果是否一致。

.

7. (5%) 你手上有一個包含 1000 筆資料的 dataset，其中訓練資料使用 800 筆、測試資料 200 筆，試說明如何利用 five-fold cross validation，訓練一個好的模型? 請詳細列出全部的步驟。

將訓練資料的800筆平分5等分，每個fold有160筆資料，選擇其中的4個fold作為訓練資料，剩餘的1個fold作為驗證資料來進行訓練，一共進行五次訓練，再將五次驗證的loss相加取平均得到平均誤差。

若結果不錯，將5個fold合併為一個訓練資料，重新進行訓練，訓練完畢後再由200筆測試資料作最後測試。

若結果不好，可以從重新調整模型超參數、降低模型複雜度、加入regularization (正則化)或選

擇另一種模型，再將修正後的模型以five-fold cross validation進行訓練，察看結果。
反覆執行上述做法直到five-fold cross validation輸出滿意的結果為止。最後再進行將5個fold合併為一個訓練資料，重新進行訓練，訓練完畢後再由200筆測試資料作最後測試。

8. a) (2%)試以一句話說明什麼是 Perceptron? b) (2%)試說明 Perceptron Learning Algorithm 中，尋找分類錯誤資料的方式為何? c) (2%)保證能讓 Perceptron Learning Algorithm 的停止條件是什麼? d) (2%)試比較 Perceptron Learning Algorithm 與 Pocket Algorithm **相同**與**相異**之處為何? e) (2%)試說明若資料維度是 d，為什麼 Perceptron Hypothesis h(x)的維度是 d+1?

(a) Perceptron 為模型中的單一節點，可以把它視為一個最基礎的模型，代表的是一個線性二元分類器。

(b) 找出錯誤點的方式: $\text{sign}(W^T X_{n(t)}) \neq y_{n(t)}$，由於 Perceptron 為線性二元分類器，若分類結果正確其輸出結果$W_t{}^T X_{n(t)}$和 Label:$y_{n(t)}$會同號，反之，若分類錯誤，兩者會異號，故可用前述公式找出錯誤的點。

(c) 訓練資料為線性可分割資料。

(d) 相同 ： 在線性可分中，都可以求到w的解；尋找錯誤樣本，以及更新w的方法是相同
相異 ： 在每一次的iteration中，Pocket需要確認Wt+1做完所有的資料後，整體結果有沒有比W好才更新，PLA每次只看一筆資料，不用算完所有的資料。因此Pocket比PLA慢，且Pocket演算法可以用在線性不可分

(e) $h(x) = \text{sign}\left(\left(\sum_{i=1}^{d} w_i x_i\right) - \text{threshold}\right)$也可化作

$h(x) = \text{sign}\left(\left(\sum_{i=1}^{d} w_i x_i\right) + (-\text{threshold}) \cdot (+1)\right)$接著將$(-\text{threshold})$視為$w_0$、$(+1)$視為$x_0$，代回公式後可以得到$h(x) = \text{sign}(\sum_{i=0}^{d} w_i x_i)$，因此 Perceptron Hypothesis h(x)的維度是 d+1。

9. (7 %)假設 dataset X 有五筆資料 x1, x2,⋯,x5，其資料維度為 2，每一筆資料的答案為 y1,y2,⋯,y5，a) (2%)試寫出要預測 y 的 regression model 公式解。b) (3%)試寫出公式解中每一個矩陣或向量的維度。c) (2%)假設 dataset 每一筆資料維度為 10，利用這個 dataset 求出的線性回歸模型，共會有多少的參數?

(a) 目標:找到W_{LIN}使得$\frac{2}{N}(X^T X w - X^T y) = \nabla E_{in}(w) = 0$

- $X^T X$ 是可逆的(invertible)時:
  - $W_{LIN} = (X^T X)^{-1} X^T y$
  - 通常情況因為$N \gg d + 1$
- $X^T X$ 是奇異的(singular)時:
  - 多個最佳解
  - 其中一解: $W_{LIN} = X^{\dagger} y$

(b) 資料維度是 2，再加上常數項，維度變為 3
$X: \mathbf{5 \times 3}$, $X^T: \mathbf{3 \times 5}$, $(X^T X)^{-1}: \mathbf{3 \times 3}$, $y: \mathbf{5 \times 1}$, $W_{LIN} = (X^T X)^{-1} X^T y: \mathbf{3 \times 1}$

(c) X_n維度:10, $x^T w + b = y$, 共會有$10 + 1 = \mathbf{11個參數}$

10. a) (4%)試說明導致過度擬合發生的四個原因? b) (3%)從資料、模型、訓練流程三個面向，說明有哪些實際作法可以降低過度擬合發生的?

(a) Data size 太小、stochastic noise 太大、deterministic noise 太大、excessive power 太大

(b) 資料:增加資料集大小、資料清洗、資料裁切(Data Cleaning/Pruning)
模型:加入 regularization (正則化)，降低模型複雜度(或挑選較簡單的模型)
訓練:交叉驗證

11. (5%)試以 Bias & Variance 的角度，說明為什麼將模型 uniform blending 的結果會比各別模型的平均表現來的好?

$G(x)=\frac{1}{T}\sum_{t=1}^{T} g_t(x)$

$$\begin{aligned} avg((g_t(x)-f(x))^2) &= avg(g_t^2-2g_tf+f^2) \\ &= avg(g_t^2)-2Gf+f^2 \\ &= avg(g_t^2)-G^2+(G-f)^2 \\ &= avg(g_t^2)-2G^2+G^2+(G-f)^2 \\ &= avg(g_t^2-2g_tG+G^2)+(G-f)^2 \\ &= avg((g_t-G)^2)+(G-f)^2 \end{aligned}$$

$$avg(E_{out}(g_t)) = \underbrace{avg(\varepsilon(g_t-G)^2)}_{\text{variance}} + \underbrace{E_{out}(G)}_{\text{bias}} \geq 0 + E_{out}(G)$$

uniform blending:透過降低 variance 以達到穩定的性能。
上述公式中，$E_{out}(G)$代表 uniform blending 的結果、$avg(E_{out}(g_t))$為各別模型的平均表現，因此可知各別模型的平均表現的 error 會大於等於 uniform blending 的結果。

12. Adaptive Boosting 演算法中，a) (6%)試說明哪些項目是需要學習的?請列出這些項目、以及它們學習的方式。b) (4%)試說明在這個演算法中，資料權重的調整是依據什麼概念? 不需要列公式、做法，只需要解釋調整的精神。

(a) 1) 資料的權重:

資料權重$u_n^{(t)}$的 re-weighting 是希望

$$(\text{total } u_n^{(t+1)} \text{ of incorrect}) = (\text{total } u_n^{(t+1)} \text{ of correct})$$

更新$u_n^{(t)}$時會以 multiply incorrect $\propto (1-\epsilon_t)$; multiply correct $\propto \epsilon_t$為原則進行。

(ϵ_t: weighted incorrect rate)，實作上以($\blacklozenge t$： optimal re-weighting factor) 實現:

$$[\![y_n \neq g_t(x_n)]\!] \text{(incorrect examples)}: \ u_n^{(t+1)} \leftarrow u_n^{(t)} \cdot \blacklozenge t$$
$$[\![y_n = g_t(x_n)]\!] \text{(correct examples)}: \ u_n^{(t+1)} \leftarrow u_n^{(t)} / \blacklozenge t$$

$$\blacklozenge t = \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \propto (1-\epsilon_t) \text{ , and } \epsilon_t = \frac{\sum_{n=1}^{N} u_n^{(t)} [\![y_n \neq g_t(x_n)]\!]}{\sum_{n=1}^{N} u_n^{(t)}}, \text{ P.12~15}$$

2) 小弱分類器g_t:

在 Adaptive Boosting 演算法中，是經由演算法 A，利用資料 D 與權重$u^{(t)}$最小化分類錯誤學習小弱分類器。實際上在程式碼裡面，演算法 A 就是直接利用 decision stump 當小弱分類器，挑選 feature, threshold, direction 一刀切下去看看何種組合的錯

誤率最小，這樣簡單的演算法 A。

3) 每個小弱分類器結合為最終分類器的權重:

每個小弱分類器結合為最終分類器的權重α_t會希望令表現較好的模型g_t能有較大的權重，即模型的 blending weight 會與正確率呈 monotonic 變化，故以 $\alpha_t = \ln(\blacklozenge t)$實現。

$$\epsilon_t = \frac{1}{2} \Rightarrow \blacklozenge t = 1 \Rightarrow \alpha_t = 0 \text{ (bad } g_t \text{ zero weight)}$$

錯誤率高，$\blacklozenge t = 1$，$\alpha_t = 0$，該模型權重小。

$$\epsilon_t = 0 \Rightarrow \blacklozenge t = \infty \Rightarrow \alpha_t = \infty \text{ (super } g_t \text{ superior weight)}$$

錯誤率低，$\blacklozenge t = \infty$，$\alpha_t = \infty$，該模型權重大。 P.16

(b) 資料權重在下一輪調整後，能讓前一輪的模型 g 的表現像是 random，因此每一輪都能學出不一樣的模型。

13. a) (2%) 請用一句話解釋 Support Vector Machine 是什麼? b) (2%) 請說明什麼是 Support Vector? c) (2%) SVM 與 PLA 相比的優點是什麼?

(a) Large-Margin Separating Hyperplane 最大邊界分隔超平面。

(b) 會被用來決定 Large-Margin Separating Hyperplane 的那些資料點，也就是落在 Large-Margin 上的那些資料點。

(c) 優點一:對資料雜訊的抵抗程度

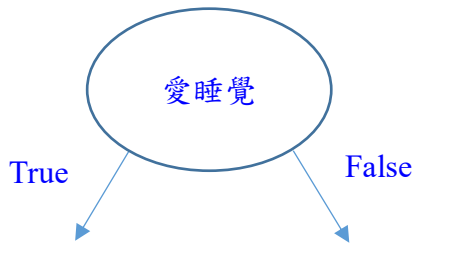
因 large margin 特性，SVM 的強韌度(robustness)會優於 PLA，且 SVM 可用在線性不可分割的資料上，但 PLA 不行。

優點二:Large-Margin Hyperplane 的個數較 Hyperplane 少

14. a) (8%)試以下列資料建構一棵決策樹。b) (2%)若有一位同學，其特徵為不愛睡覺、愛打 LOL、上課 13 次，試以 a)題模型預測，其 ML 課程是否 PASS?

特徵			預測結果
愛睡覺	愛打 LOL	上課次數	ML 課程是否有 PASS
Y	Y	2	N
Y	N	15	Y
Y	Y	18	Y
N	N	5	N
N	Y	10	N

(a)



PASS: PASS:

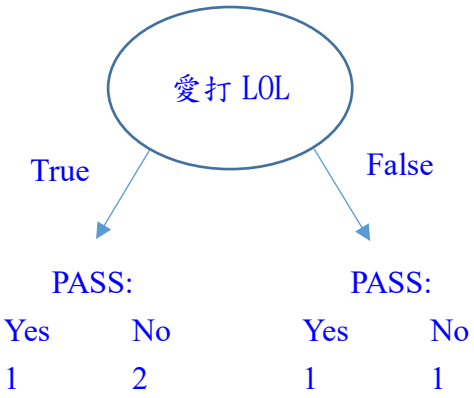
Yes	No	Yes	No
2	1	0	2

計算 Gini Impurity:

$$1 - \left[\left(\frac{2}{2+1}\right)^2 + \left(\frac{1}{2+1}\right)^2\right] = \frac{4}{9}$$

$$1 - \left[\left(\frac{0}{0+2}\right)^2 + \left(\frac{2}{0+2}\right)^2\right] = 0$$

$$\left(\frac{3}{3+2}\right) \times \left(\frac{4}{9}\right) + \left(\frac{2}{3+2}\right) \times 0 = \frac{4}{15}$$

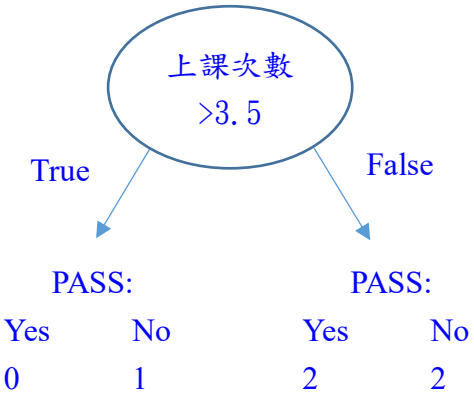


計算 Gini Impurity:

$$1-\left[\left(\frac{1}{1+2}\right)^2+\left(\frac{2}{1+2}\right)^2\right]=\frac{4}{9}$$

$$1-\left[\left(\frac{1}{1+1}\right)^2+\left(\frac{1}{1+1}\right)^2\right]=\frac{1}{2}$$

$$\left(\frac{3}{3+2}\right)\times\left(\frac{4}{9}\right)+\left(\frac{2}{3+2}\right)\times\frac{1}{2}=\frac{7}{15}$$

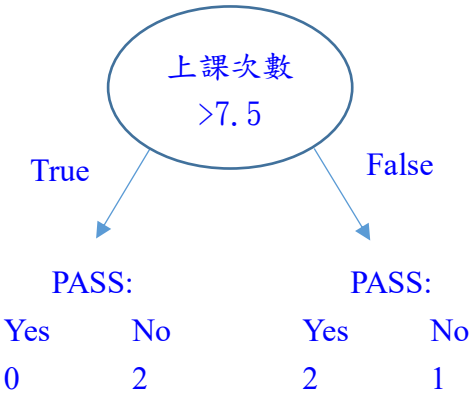


計算 Gini Impurity:

$$1-\left[\left(\frac{0}{0+1}\right)^2+\left(\frac{1}{0+1}\right)^2\right]=0$$

$$1-\left[\left(\frac{2}{2+2}\right)^2+\left(\frac{2}{2+2}\right)^2\right]=\frac{1}{2}$$

$$\left(\frac{1}{1+4}\right)\times 0+\left(\frac{4}{1+4}\right)\times\frac{1}{2}=\frac{2}{5}$$

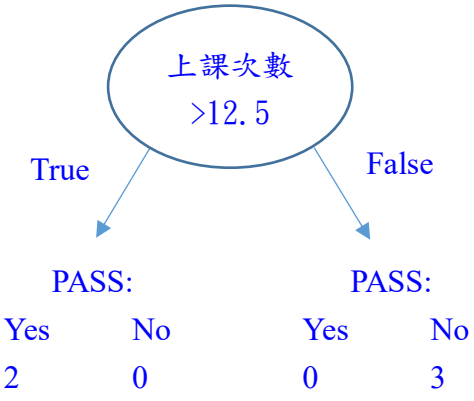


計算 Gini Impurity:

$$1-\left[\left(\frac{0}{0+2}\right)^2+\left(\frac{2}{0+2}\right)^2\right]=0$$

$$1-\left[\left(\frac{2}{2+1}\right)^2+\left(\frac{1}{2+1}\right)^2\right]=\frac{4}{9}$$

$$\left(\frac{2}{2+3}\right)\times 0+\left(\frac{3}{2+3}\right)\times\frac{4}{9}=\frac{4}{15}$$

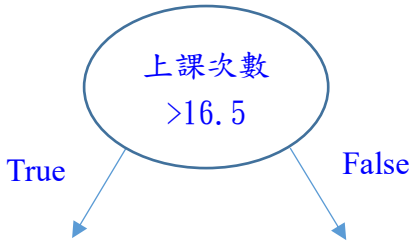


計算 Gini Impurity:

$$1-\left[\left(\frac{2}{2+0}\right)^2+\left(\frac{0}{2+0}\right)^2\right]=0$$

$$1-\left[\left(\frac{0}{0+3}\right)^2+\left(\frac{3}{0+3}\right)^2\right]=0$$

$$\left(\frac{2}{2+3}\right)\times 0+\left(\frac{3}{2+3}\right)\times 0=0$$



PASS:		PASS:	
Yes	No	Yes	No
1	0	1	3

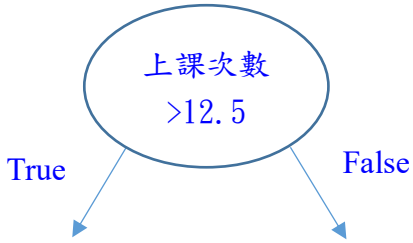
計算 Gini Impurity:

$$1-\left[\left(\frac{1}{1+0}\right)^2+\left(\frac{0}{1+0}\right)^2\right]=0$$

$$1-\left[\left(\frac{1}{1+3}\right)^2+\left(\frac{3}{1+3}\right)^2\right]=\frac{6}{16}$$

$$\left(\frac{1}{1+4}\right)\times 0+\left(\frac{4}{1+4}\right)\times\frac{6}{16}=\frac{3}{10}$$

選擇 Gini Impurity 最小的作為 branching criteria，即上課次數>12.5。



PASS:		PASS:	
Yes	No	Yes	No
2	0	0	3

依據 C&RT termination 規則，當出現下列兩種情形會使 C&RT algorithm 停止:

1) 當輸入的所有資料其 Label y_n均相同
   (all y_n the same: impurity = 0 $\Rightarrow g_t(x) = y_n$)
2) 當輸入的資料其 feature x_n均相同
   (all x_n the same: no decision stumps) P.12

因觸發條件一，故建樹過程到此停止。

(b) 依照(a)之 decision tree 來看，該同學上課次數超過 12.5 次，模型會將其預測為 PASS。