機器學習導論–期中報告

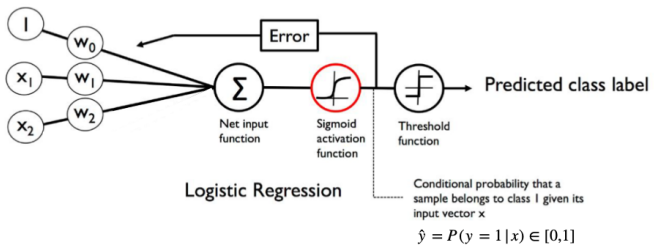
Part A. 計算題

Part A. 計算題（將數字四捨五入至小數點後兩位）

1. 假設有一筆訓練資料如下：

i	x_1	x_2	y
1	10	2	1
2	3	5	0
3	25	3	1
4	15	20	0
5	5	12	0
6	8	5	1

考慮使用下列的邏輯迴歸 (logistic regression)演算法來建立模型，且若 $\hat{y} \geq 0.5$ 則輸出 1，而 $\hat{y} < 0.5$ 則輸出 0。若目前建立了兩個模型：模型A的參數為 $w_0 = -15, w_1 = 1, w_2 = 1$；模型B的參數為 $w_0 = -5, w_1 = 1, w_2 = -1$。



(1) [20 分] 請計算兩個模型的**自然對數**–概似函數值 (log–likelihood function)

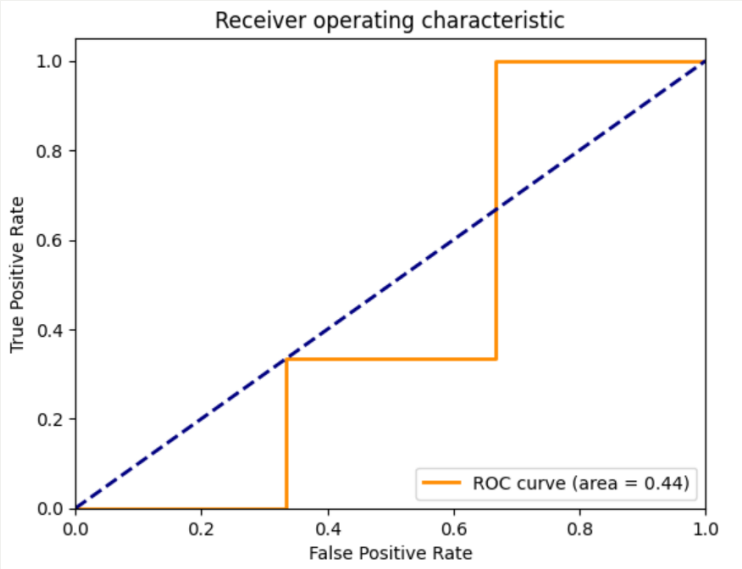
$$\sum_{i=1}^{6} y^{(i)} \ln\left(\hat{y}^{(i)}\right) + \left(1 - y^{(i)}\right) \ln\left(1 - \hat{y}^{(i)}\right),$$

先算Z(標準分數，standard score)再算y^得到機率。最後算log-liklihood(不要忘了加總)。根據預測結果以及log-liklihood，可以知道B模型較佳。

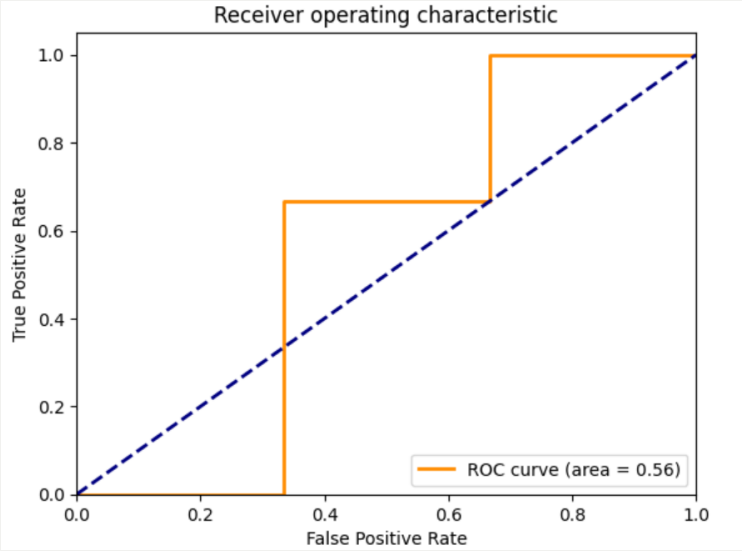
X1	X2	y	z(A)	z(B)	y^ (A)	Y_pred (A)	y^ (B)	Y_pred (B)	Z公式(z = w0 + w1*X1 + w2*X2)
10	2	1	-3	3	0.047425873	0	0.952574	1	A模型(z = -15 + 1*X1 + 1*X2)
3	5	0	-7	-7	0.000911051	0	0.000911	0	B模型(z = -5 + 1*X1 + (-1)*X2)
25	3	1	13	17	0.99999774	1	1	1	y^ = 1 / (1 + e^(-z))
15	20	0	20	-10	0.999999998	1	4.54E-05	0	log-liklihood = y*ln(y^)+(1-y)ln(1-y^)
5	12	0	2	-12	0.880797078	1	6.14E-06	0	
8	5	1	-2	-2	0.119202922	0	0.119203	0	
A log-liklihood	B log-liklihood				y^ (A) RE				
-3.048587352	-0.048587352				0.952574127				
-0.000911466	-0.000911466				0.999088949				
-2.26033E-06	-4.13994E-08				2.26032E-06				
-19.99999997	-4.53989E-05				2.06115E-09				
-2.126928011	-6.14419E-06				0.119202922				
-2.126928011	-2.126928011				0.880797078				
-27.30335707	-2.176478414								

(2) [20 分] 請計算模型 A 的 AUC 值。若模型 A 的 AUC < 1/2，請修正此模型，並計算新模型的 AUC 值。

這是原版的，確實只有0.44。



有一句話是這麼說的，當你考0分相當於你考了100分。在二元分類上可以直接把答案(機率)顛倒，就可以達成相反的效果。



Part B. 觀念題

2. [10 分] 請說明線性迴歸 (linear regression) 和邏輯迴歸 (logistic regression)有哪些異同點。

💡 兩者都是在計算模型估計和真實值的相似性，且都使用 gradient descent 尋求最佳參數解。線性迴歸主要用於預測數值任務，預測輸出為實數值，Loss 的計算方式為 MSE、MAE 等；邏輯迴歸主要用於預測分類任務，利用 sigmoid 將預測值轉為機率值並輸出，Loss 的計算方式為 cross entropy。
BTW:邏輯迴歸是可以做多元分類的，可以下參數(multiclass = ovr)。多元問題會用Softmax函數。
（Lecture 2. Regression）（Lecture 3 人工神經元）

3. [15 分] 請說明什麼是欠擬合 (underfitting)現象和過擬合(overfitting)現象，並舉出幾種降低欠擬合 (underfitting)和過擬合(overfitting)的方法。

💡 欠擬合是指模型無法充分擬合訓練數據，通常是因為模型過於簡單，無法擬合數據的複雜性。欠擬合的結果是模型在訓練集和測試集上都表現不好，無法達到很高的準確度。

過擬合是指模型在訓練集上表現很好，但是在測試集上表現很差，即模型過於複雜，導致對新數據的泛化能力差。通常是因為模型過於複雜，或者訓練數據過少，或者訓練數據和測試數據的分佈不一致等原因。

解決**欠擬合**的方法包括：增加模型複雜度、增加特徵數量、調整超參數（例如：降低正則化係數，當正則化係數太高時，模型的權重會降至接近零，進而限制了模型的複雜度）等。
解決**過擬合**的方法包括：減少特徵數量、增加訓練數據量、正則化等降低模型複雜度。

4. [15 分] 考慮要預測某地區是否會發生地震的二元分類問題：0 代表不會發生，1 代表發生地震。請問若要評估此分類演算法，使用正確率（accuracy）是否是一個合理的評估指標？如果正確率不是一個合理的指標，那應該要選擇哪種評估指標呢？試說明原因。

💡 使用**正確率（accuracy）**做為評估指標並不一定是一個合理的選擇，尤其是當正負樣本數量不平衡時，即使分類器完全沒有學習到任何有用的特徵，也有可能會得到高正確率的結果。例如，如果有 99% 的樣本是負樣本，那麼一個總是預測為負樣本的分類器仍然可以達到 99% 的正確率。**(以實際例子來說，沒發生地震的天數比發生地震的天數多太多了)**
可以使用**混淆矩陣 (confusion matrix)** 和相關指標如**精確度 (precision)**、**召回率 (recall)**、**F1 分數 (F1 score)** 等來評估模型性能。在這個地震預測問題中，可以根據實際情況選擇適當的指標，例如，如果更關心減少漏報 (false negative)，可以選擇較高的召回率；如果更關心減少誤報 (false positive)，可以選擇較高的精確度。
更注重召回率在這個案例的意義是，代表我們更關心的是模型能夠找到所有真正發生地震的事件，不要地震真的來了卻沒反應。
更注重精確率在這個案例的意義是，代表我們更關心的是模型預測為發生地震的事件中，有多少是真正發生地震的，才不會動不動就給你發國家警報擾民。

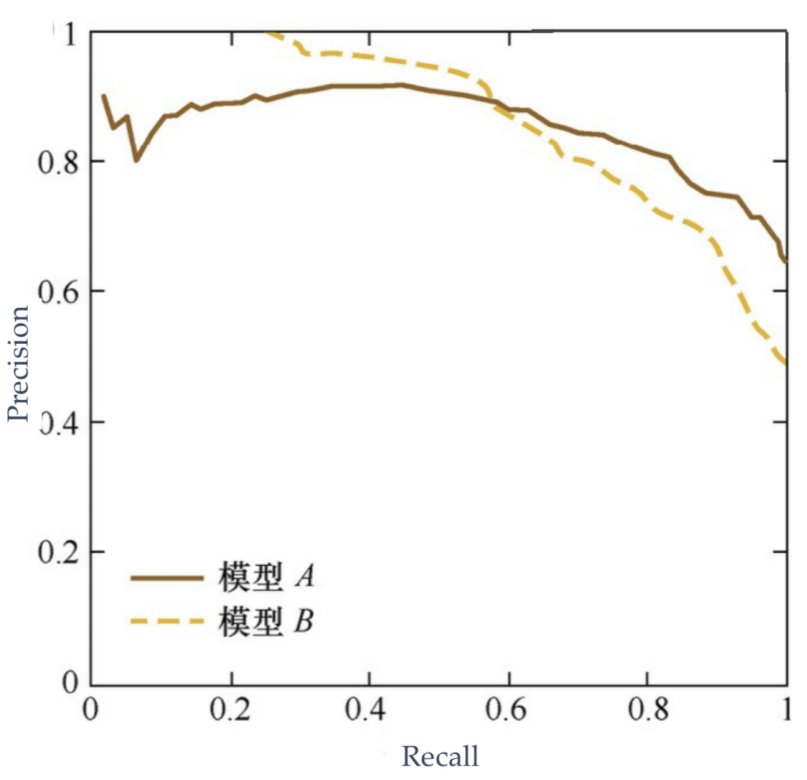
5. [20 分] 網站的模糊搜尋（Fuzzy search）功能指的是當使用者輸入關鍵字查詢時，系統不僅會查找符合關鍵字的內容，也會自動查詢與輸入關鍵字意義相似但卻不一定完全相同的內容，因此即使當使用者拼錯關鍵字，模糊搜尋也可以找到相關的內容。假設某購物網站提供的模糊搜尋功能，系統會傳回與輸入關鍵字相關程度最高的 5 個商品，雖然這種作法的的精準度（precision）很高，但實際上，使用者卻還是常常反應找不到他們想要物品，特別是冷門商品。請問可能是哪個環節出了問題呢？

主要是沒考慮 **Recall**，也有可能是 **Top N** 考慮太少。
這題如果只寫關鍵字對應或搜索演算法問題之類的只會給一些分數。但最主要的問題是這個檢索系統只列出相關程度最高的5個商品，這代表這個系統考慮的可能性過少。這也代表了Recall非常低。同時由於Precision的評判標準是【回傳結果是否符合使用者需求】，因此在冷門商品的搜索上如果都很難找到，其實在意義上也是無法符合使用者需求的，所以precision反而會變低。然而對系統而言他只是把使用者輸入的關鍵字最相關的結果回傳而已。這就會有系統運作與人類認知上的落差。所以回過頭來要考慮使用者體驗的話就必須做到P和R都要兼顧才能讓使用者找到自己需要的商品。
精準度是指，檢索出的所有結果中，與查詢相關的結果的比例。
召回率是指，在所有與查詢相關的結果中，系統所檢索到的比例。

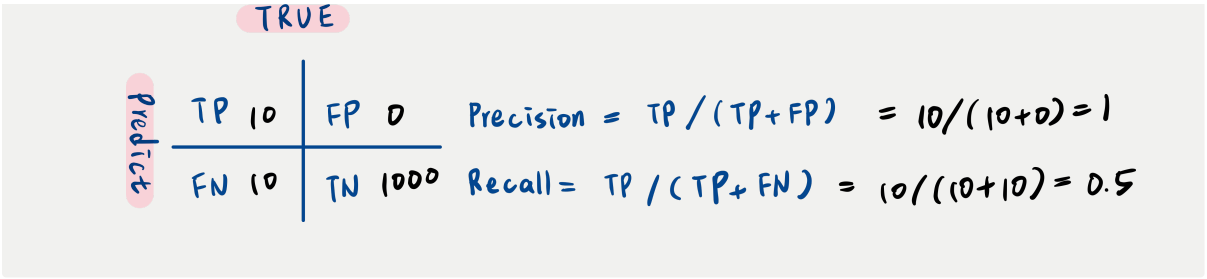
除此之外也可以參考 **P-R 曲線**，P-R 曲線是由不同閥值下的 Precision 和 Recall 所繪製而成，閥值是模型在預測時預測為正值的最小機率值。透過繪製 P-R 曲線可以觀察模型在不同閥值下的表現，並可以用來決定最佳的閥值，以最大化 Precision 和 Recall，從而達到最佳的分類效果。

以下是一些閾值和 P-R 曲線中對應的數值和其意義的舉例說明（參考圖中模型B）：

- 閾值=0.1：P=0.5、R=0.9。在這種情況下，模型會錯過一些正樣本，但是檢測到的正樣本很可能是真正的正樣本，因此 Recall 很高。然而，由於檢測到的正樣本太多，因此 Precision 很低。
- 閾值=0.5：P=0.8、R=0.7。在這種情況下，模型經過平衡，正確檢測到了大部分的正樣本並減少了錯誤的檢測，因此 Precision 和 Recall 都相對較高。
- 閾值=0.9：P=0.9、R=0.3。在這種情況下，模型很保守，只檢測出很少的正樣本，但是它檢測出的正樣本通常是真正的正樣本，因此 Precision 很高。然而，由於模型漏檢了許多正樣本，因此 Recall 很低。



經由上述 P-R 曲線的介紹，可以知道 P-R 曲線因為主要考慮 Precision 與 Recall，所以容易受到正負樣本的分佈影響。也就是說，當真實資料的正樣本很少（數據不平衡），預測時只要預測出少數正確的正樣本，Precision 就會大幅提升。因此 P-R 曲線比較常用於同一批資料在不同模型之間的表現評估。


TRUE
predict
TP 10
FP 0
FN 10
TN 1000
Precision = TP / (TP + FP) = 10/(10+0) = 1
Recall = TP / (TP + FN) = 10/(10+10) = 0.5