邏輯回歸 Logistic Regression

國立東華大學電機工程學系 楊哲旻

Outline

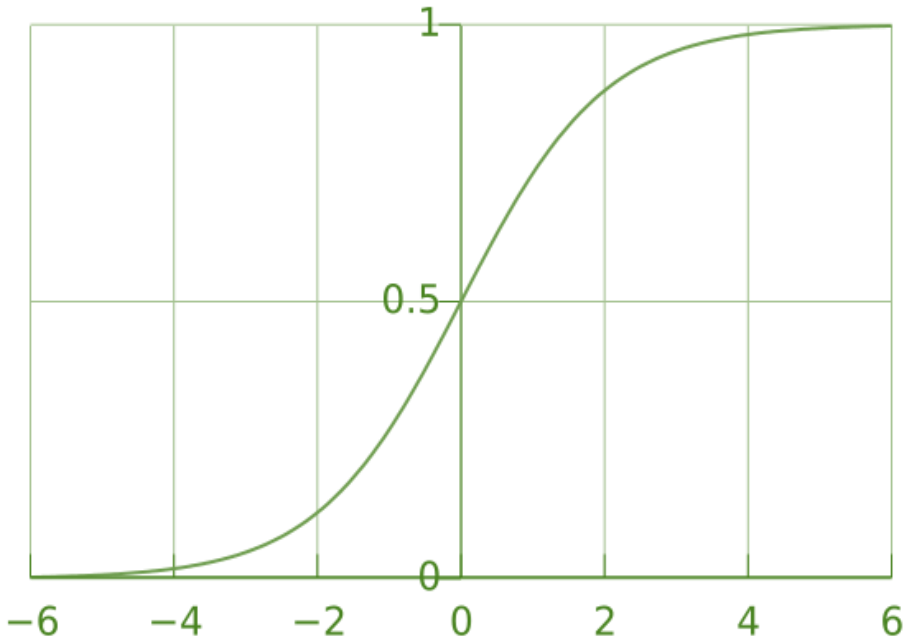
1. 邏輯回歸
2. 鳶尾花資料集
3. 線性回歸處理分類問題
4. 邏輯回歸處理分類問題
5. 邏輯函數
6. 損失函數
7. 勝算比
8. 正則化
9. 邏輯回歸實作

機器學習 – 邏輯回歸 01. 邏輯回歸

邏輯回歸為分類模型，是由線性回歸模型經邏輯函數輸出的模型。訓練過程是從訓練集中以梯度下降法來確定權重與偏差。

本教學以四小節來講解：

1. **單變數的二元分類**
2. **單變數的多元分類**
3. 多變數的二元分類
4. 多變數的多元分類



機器學習 – 邏輯回歸　02. 鳶尾花資料集

鳶尾花資料集

Iris dataset 是非常著名的生物資訊資料集之一，取自美國加州大學歐文分校的機器學習資料庫 http://archive.ics.uci.edu/ml/datasets/Iris，資料的筆數為150筆：

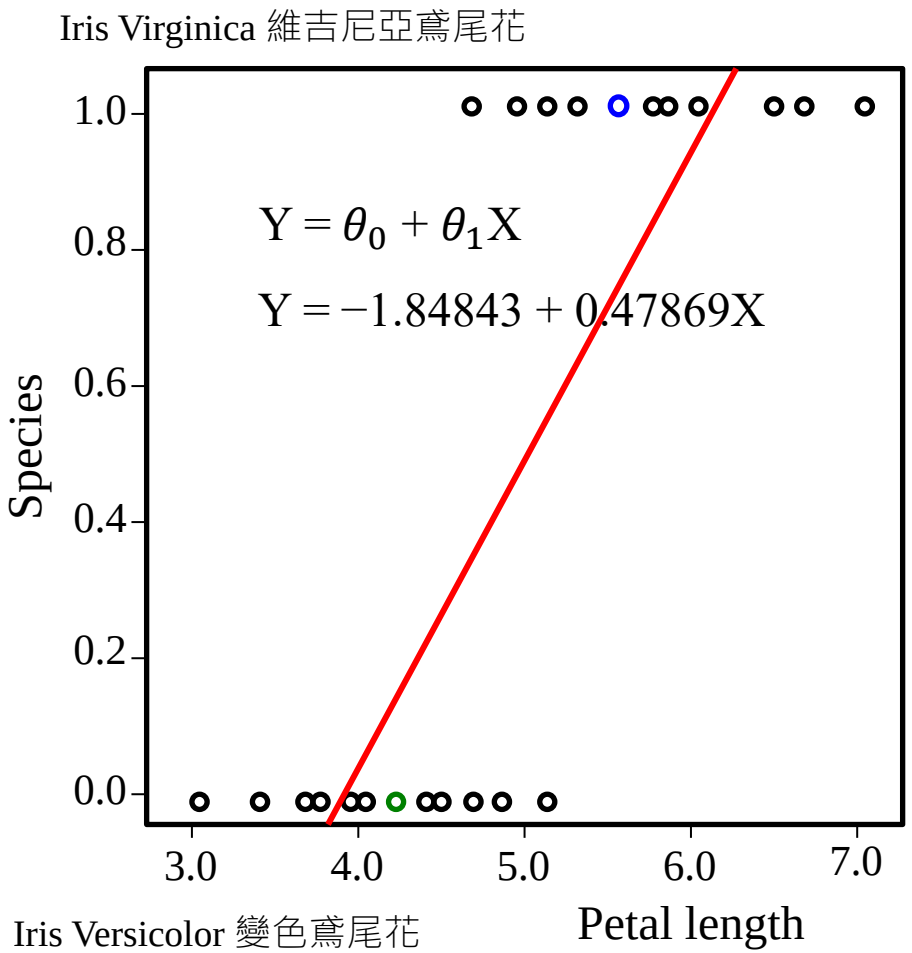
特徵 (Feature)：

1. 花萼長度 (Sepal Length) (cm)
2. 花萼寬度 (Sepal Width) (cm)
3. 花瓣長度 (Petal Length) (cm)
4. 花瓣寬度 (Petal Width) (cm)

三個品種 - 類別/標籤 (Class/Label)：

1. 山鳶尾(Setosa)　2. 變色鳶尾(Versicolor)　3. 維吉尼亞鳶尾(Virginica)

機器學習 – 邏輯回歸　03. 線性回歸處理分類問題

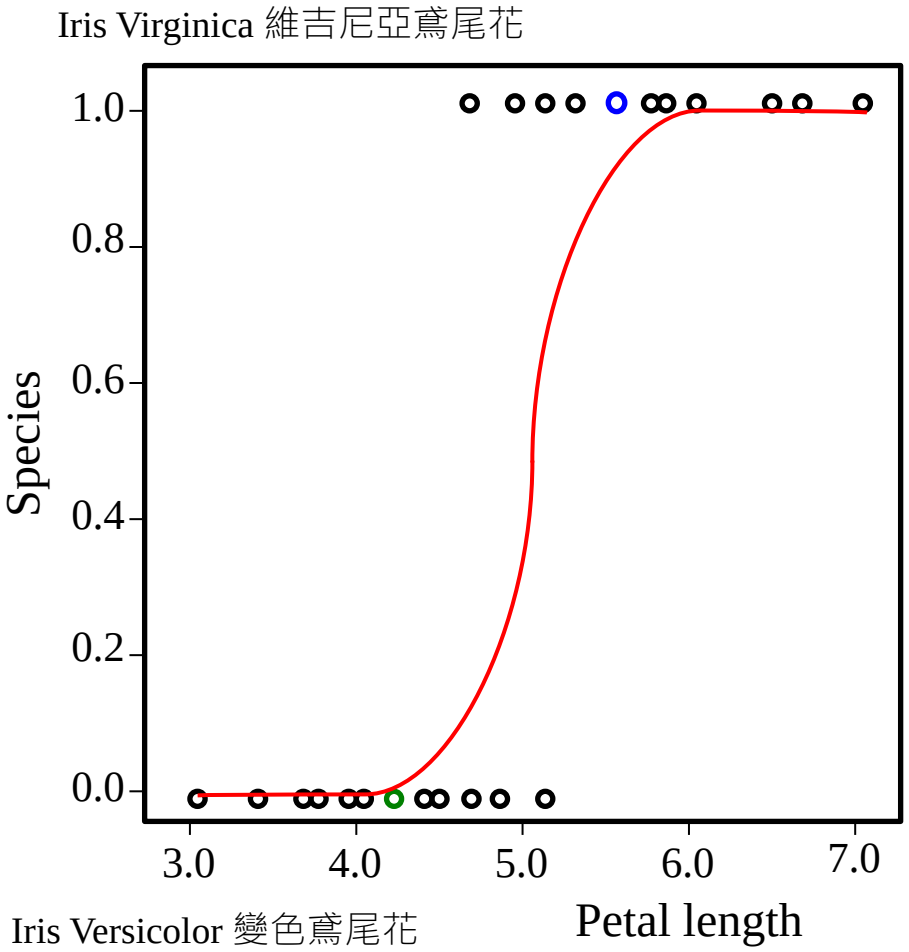


線性回歸作分類任務時

	X	Y
	Petal length	Species
113	5.5	維吉尼亞 Virginica (1)
98	4.3	變色Versicolor (0)
121	5.7	維吉尼亞 Virginica (1)
52	4.5	變色 Versicolor (0)
58	3.3	變色 Versicolor (0)
139	4.8	維吉尼亞 Virginica (1)
94	3.3	變色 Versicolor (0)
⋮	⋮	⋮

機器學習 – 邏輯回歸

04. 邏輯回歸處理分類問題

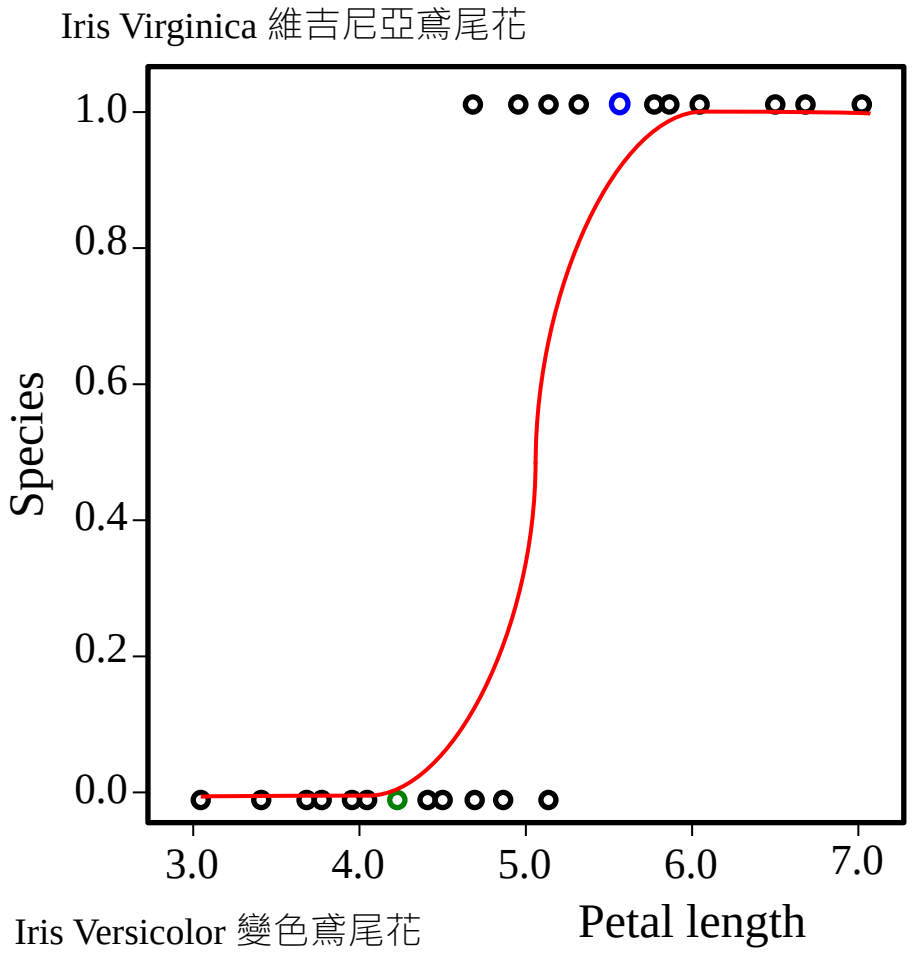


邏輯回歸作分類任務時

	X	Y
	Petal length	Species
113	5.5	維吉尼亞 Virginica (1)
98	4.3	變色Versicolor (0)
121	5.7	維吉尼亞 Virginica (1)
52	4.5	變色 Versicolor (0)
58	3.3	變色 Versicolor (0)
139	4.8	維吉尼亞 Virginica (1)
94	3.3	變色 Versicolor (0)
⋮	⋮	⋮

1. 單變數的二元分類

機器學習 – 邏輯回歸 05. 邏輯函數

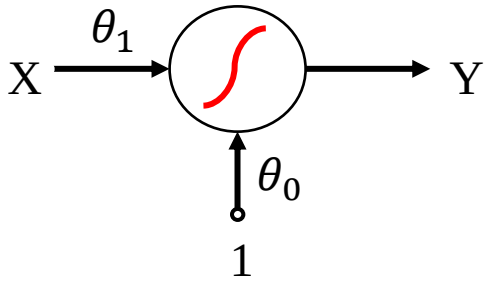


邏輯/乙狀函數 (Logistic / Sigmoid Function)

線性回歸 $t = \theta_0 + \theta_1 X$

邏輯回歸 $Y = f(t) = f(\theta_0 + \theta_1 X)$

$$f(t) = \frac{e^t}{1 + e^t} = \frac{1}{1 + e^{-t}}$$



$$\Pr(Y=1 \mid X) = \frac{e^{\theta_0 + \theta_1 X}}{1 + e^{\theta_0 + \theta_1 X}} = f(\theta_0 + \theta_1 X)$$

1. 單變數的二元分類

機器學習 – 邏輯回歸 05. 邏輯函數

$$\Pr(Y=1 \mid X) = \frac{e^{\theta_0 + \theta_1 X}}{1 + e^{\theta_0 + \theta_1 X}} = f(\theta_0 + \theta_1 X)$$

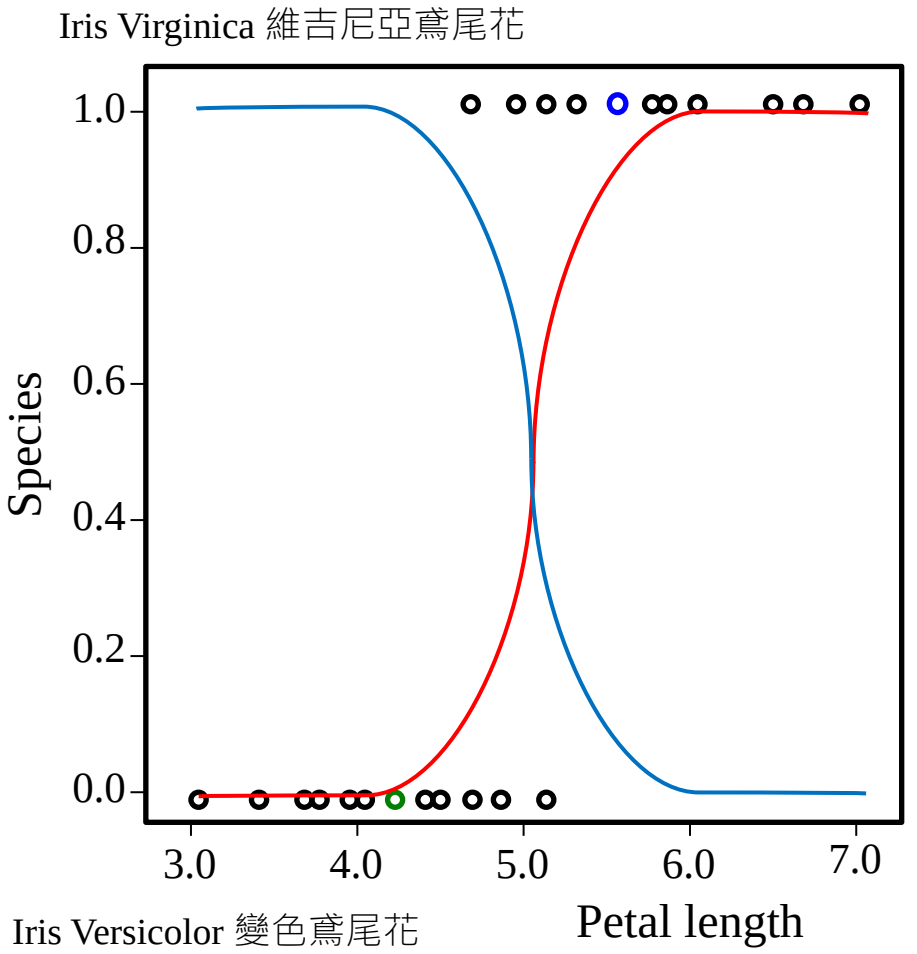
$$\Pr(Y=0 \mid X) = 1 - \Pr(Y=1 \mid X)$$

$$= 1 - \frac{e^{\theta_0 + \theta_1 X}}{1 + e^{\theta_0 + \theta_1 X}}$$

$$= \frac{1}{1 + e^{\theta_0 + \theta_1 X}}$$

$$\Pr(Y=1 \mid X) = \frac{e^{\theta_0 + \theta_1 X}}{1 + e^{\theta_0 + \theta_1 X}}$$

$$\Pr(Y=0 \mid X) = \frac{1}{1 + e^{\theta_0 + \theta_1 X}}$$



1. 單變數的二元分類

機器學習 – 邏輯回歸　05. 邏輯函數

$$\Pr(Y=1 \mid X) = \frac{e^{\theta_0 + \theta_1 X}}{1 + e^{\theta_0 + \theta_1 X}} \qquad \Pr(Y=0 \mid X) = \frac{1}{1 + e^{\theta_0 + \theta_1 X}}$$

$$\Pr(Y=y_{113}=1 \mid X=x_{113}) = \frac{e^{\theta_0 + \theta_1 x_{113}}}{1 + e^{\theta_0 + \theta_1 x_{113}}}$$

$$\Pr(Y=y_{98}=0 \mid X=x_{98}) = \frac{1}{1 + e^{\theta_0 + \theta_1 x_{98}}}$$

$$\Pr(Y=y_{121}=1 \mid X=x_{121}) = \frac{e^{\theta_0 + \theta_1 x_{121}}}{1 + e^{\theta_0 + \theta_1 x_{121}}}$$

$$\Pr(Y=y_{52}=0 \mid X=x_{52}) = \frac{1}{1 + e^{\theta_0 + \theta_1 x_{52}}}$$

$$\Pr(Y=y_{58}=1 \mid X=x_{58}) = \frac{e^{\theta_0 + \theta_1 x_{58}}}{1 + e^{\theta_0 + \theta_1 x_{58}}}$$

$$\Pr(Y=y_{139}=0 \mid X=x_{139}) = \frac{1}{1+e^{\theta_0 + \theta_1 x_{139}}}$$

$$\Pr(Y=y_{94}=1 \mid X=x_{94}) = \frac{e^{\theta_0 + \theta_1 x_{94}}}{1 + e^{\theta_0 + \theta_1 x_{94}}}$$

	X	Y
	Petal length	Species
113	5.5	維吉尼亞 Virginica (1)
98	4.3	變色Versicolor (0)
121	5.7	維吉尼亞 Virginica (1)
52	4.5	變色 Versicolor (0)
58	3.3	變色 Versicolor (0)
139	4.8	維吉尼亞 Virginica (1)
94	3.3	變色 Versicolor (0)
⋮	⋮	⋮

1. 單變數的二元分類

機器學習 – 邏輯回歸　06. 損失函數

尋找損失函數

$$P(\theta_0,\theta_1) = \prod_{y_i=1} \Pr(Y=y_i=1 \mid X=x_i) \prod_{y_i=0} \Pr(Y=y_i=0 \mid X=x_i)$$

$$= \prod_{i=1}^{n} \Pr(Y=y_i=1 \mid X=x_i)^{y_i} \Pr(Y=y_i=0 \mid X=x_i)^{1-y_1}$$

$$= \prod_{i=1}^{n} \Pr(Y=y_i=1 \mid X=x_i)^{y_i} (1-\Pr(Y=y_i=1 \mid X=x_i))^{1-y_1}$$

$$P(w) = \prod_{i=1}^{n} h_w(x_i)^{y_i} (1-h_w(x_i))^{1-y_1}$$

機器學習 – 邏輯回歸　06. 損失函數

尋找損失函數

$$P(w) = \prod_{i=1}^{n} h_w(x_i)^{y_i}(1 - h_w(x_i))^{1-y_1}$$

求$-P(w)$當作損失函數計算最低點是可以的？其實是不行，因為這$-P(w)$函數為非凸函數，無法使用梯度下降法(因為不可微)。因始要對最大似然函數取對數變換，轉為對數函數：

$$-log(P(w)) = -\sum_{i=1}^{m}[y_i log h_w(x_i) + (1 - y_i)log(1 - h_w(x_i))]$$

$$L(w) = -\frac{1}{m}\sum_{i=1}^{m}[y_i log h_w(x_i) + (1 - y_i)log(1 - h_w(x_i))]$$

機器學習 – 邏輯回歸　07. 勝算比

勝算比(Odds Ratio, OR)

單變數的二元分類

$$\Pr(Y=1 \mid X) = \frac{e^{\theta_0 + \theta_1 X}}{1 + e^{\theta_0 + \theta_1 X}}$$

$$\Pr(Y=0 \mid X) = \frac{1}{1 + e^{\theta_0 + \theta_1 X}}$$

$$\text{Odds ratio} = \frac{\Pr(Y=1 \mid X)}{\Pr(Y=0 \mid X)} = e^{\theta_0 + \theta_1 X}$$

單變數的三元分類

$$\text{Odds ratio} = \frac{\Pr(Y=1 \mid X)}{\Pr(Y=0 \mid X)} = e^{\theta_{10} + \theta_{11} X}$$

$$\text{Odds ratio} = \frac{\Pr(Y=2 \mid X)}{\Pr(Y=0 \mid X)} = e^{\theta_{20} + \theta_{21} X}$$

機器學習 – 邏輯回歸

$$\text{Odds ratio} = \frac{\Pr(Y=1 \mid X)}{\Pr(Y=0 \mid X)} = e^{\theta_{10} + \theta_{11}X}$$

$$\Pr(Y=1 \mid X) = \Pr(Y=0 \mid X)\,(e^{\theta_{10} + \theta_{11}X})$$

$$\text{Odds ratio} = \frac{\Pr(Y=2 \mid X)}{\Pr(Y=0 \mid X)} = e^{\theta_{20} + \theta_{21}X}$$

$$\Pr(Y=2 \mid X) = \Pr(Y=0 \mid X)\,(e^{\theta_{20} + \theta_{21}X})$$

代入

$$\Pr(Y=0 \mid X)+\Pr(Y=1 \mid X)+\Pr(Y=2 \mid X)=1$$

$$\Pr(Y=0 \mid X) + \Pr(Y=0 \mid X)\,(e^{\theta_{10} + \theta_{11}X}) + \Pr(Y=0 \mid X)\,(e^{\theta_{20} + \theta_{21}X}) = 1$$

$$\Pr(Y=0 \mid X)\left(1 + e^{\theta_{10} + \theta_{11}X} + e^{\theta_{20} + \theta_{21}X}\right) = 1$$

代回去

$$\Pr(Y=0 \mid X) = \frac{1}{1 + e^{\theta_{10} + \theta_{11}X} + e^{\theta_{20} + \theta_{21}X}}$$

$$\Pr(Y=1 \mid X) = \frac{e^{\theta_{10} + \theta_{11}X}}{1 + e^{\theta_{10} + \theta_{11}X} + e^{\theta_{20} + \theta_{21}X}}$$

$$\Pr(Y=2 \mid X) = \frac{e^{\theta_{20} + \theta_{21}X}}{1 + e^{\theta_{10} + \theta_{11}X} + e^{\theta_{20} + \theta_{21}X}}$$

機器學習 – 邏輯回歸

$$\Pr(\text{Species} = \text{Setosa(山鳶尾花)} \mid X)$$

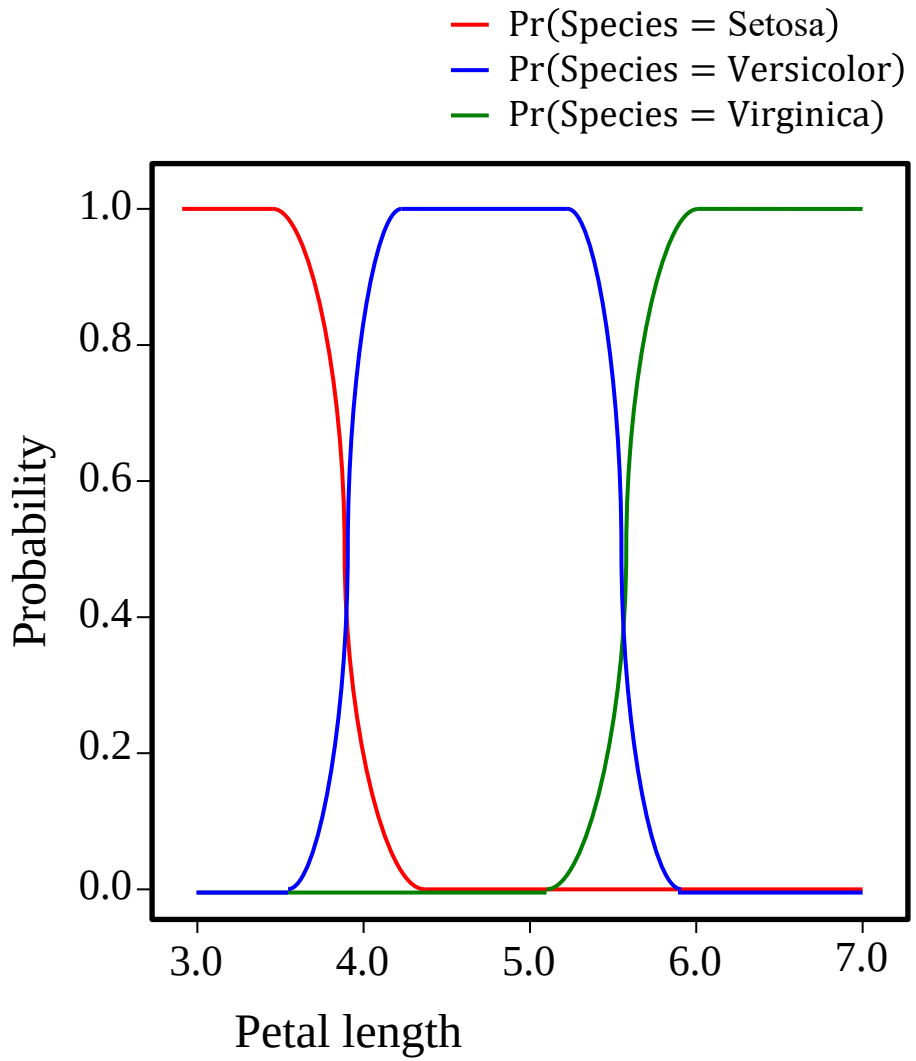
$$= \frac{e^{116.614 - 38.5797X}}{1 + e^{116.614 - 38.5797X} + e^{43.7809 - 9.0020X}}$$

$$\Pr(\text{Species} = \text{Versicolor(變色鳶尾花)} \mid X)$$

$$= \frac{e^{43.7809 - 9.0020X}}{1 + e^{116.614 - 38.5797X} + e^{43.7809 - 9.0020X}}$$

$$\Pr(\text{Species} = \text{Virginica(維吉尼亞鳶尾花)} \mid X)$$

$$= \frac{1}{1 + e^{116.614 - 38.5797X} + e^{43.7809 - 9.0020X}}$$

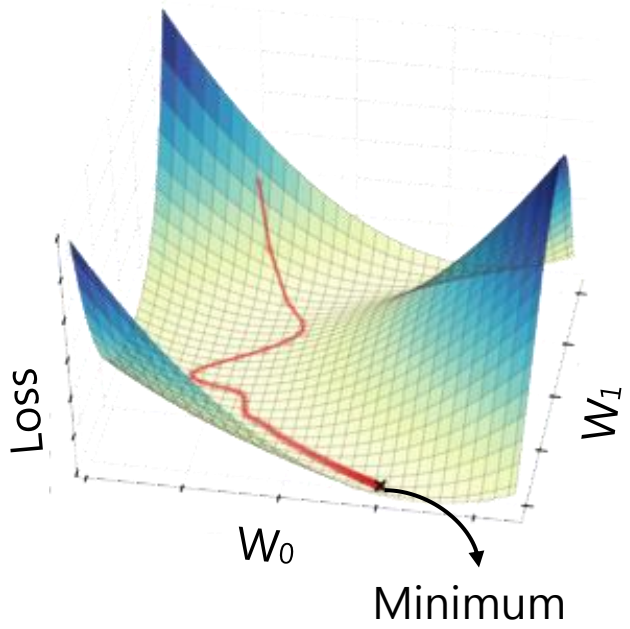


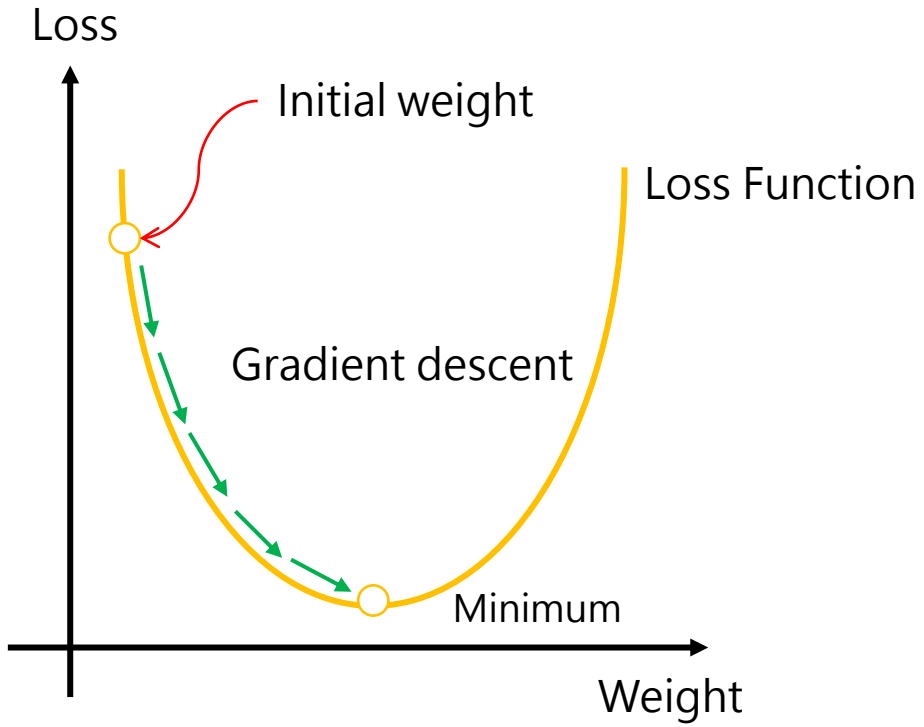
2. 單變數的多元分類

機器學習 – 邏輯回歸　08. 正則化

正則化 (Regularization)

正則化是透過損失函數的改變，以避免下降至最低點，來預防過度擬和問題



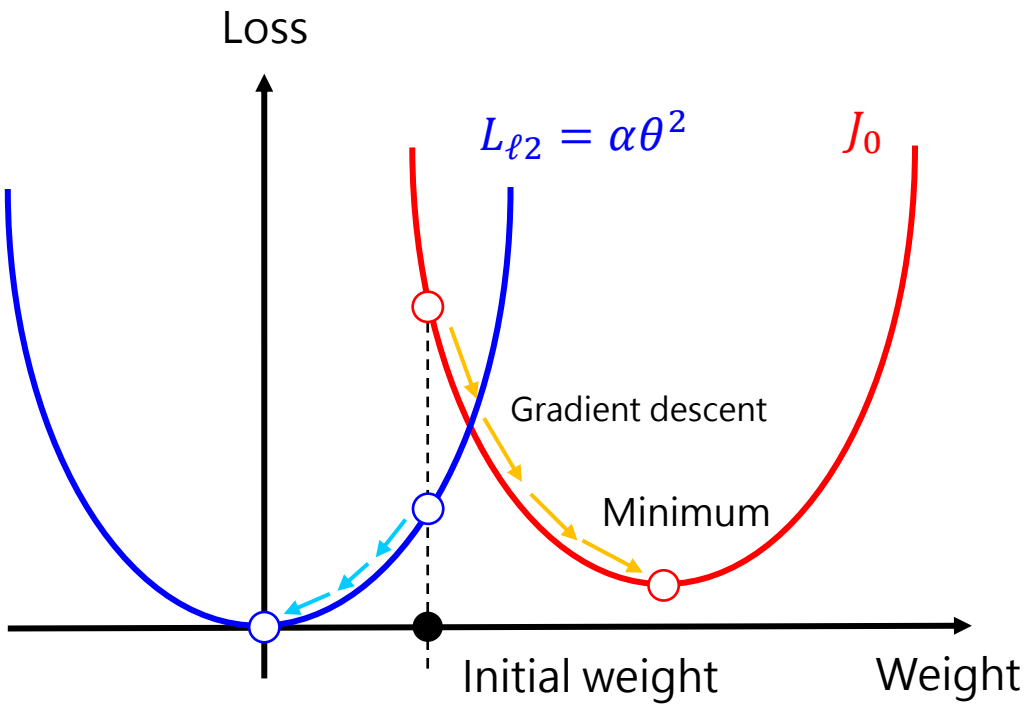


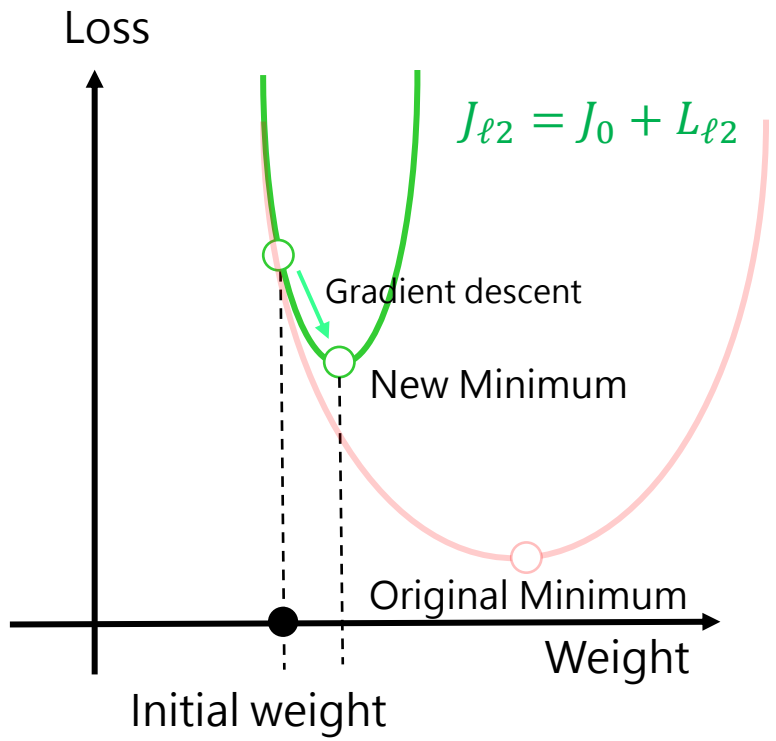
機器學習 – 邏輯回歸　08. 正則化

正則化 (Regularization)

$$J_{\ell 2} = J_0 + L_{\ell 2} = J_0 + \alpha \sum_{i=0}^{n} {\theta_i}^2$$

$$J_{\ell 2} = J_0 + L_{\ell 2} = J_0 + \alpha\theta^2$$

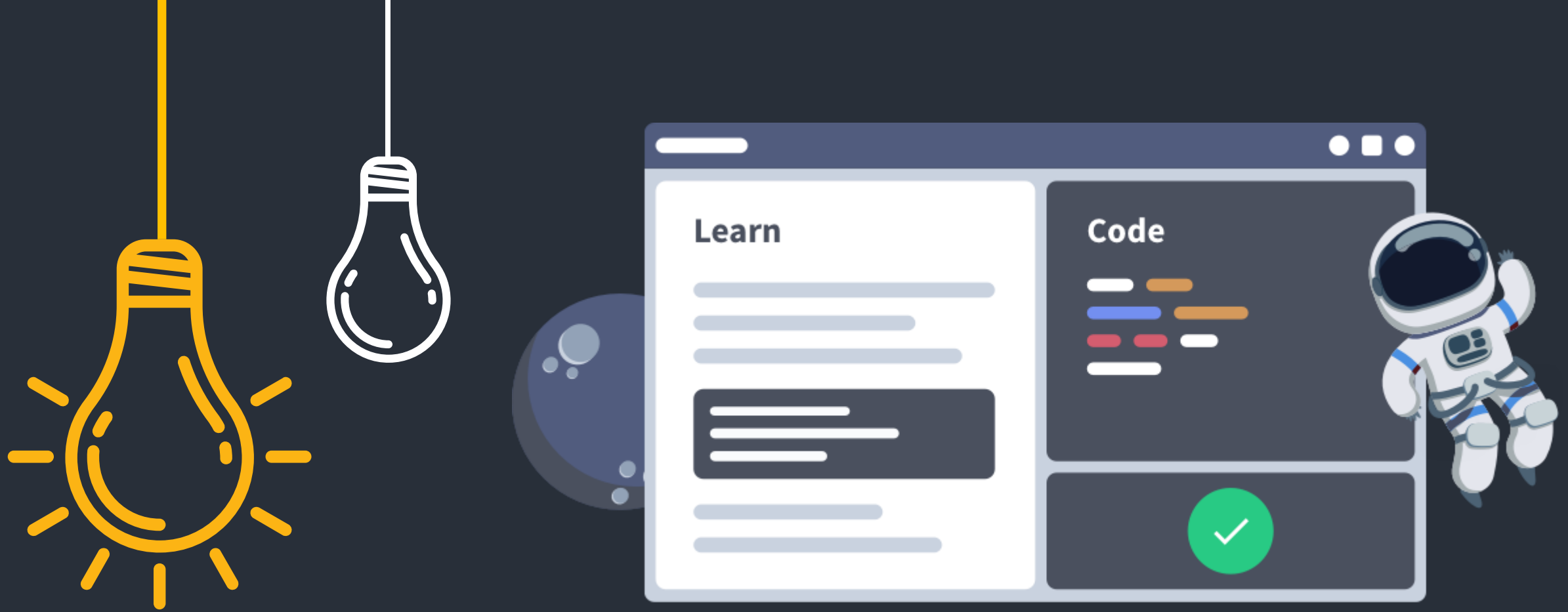




機器學習 – 邏輯回歸

邏輯回歸的超參數

- 學習速率 (Learning Rate)
- 批量 (Batch)
- 迭代次數 (Number of iterations)
- 正則化懲罰係數 (Regularization)

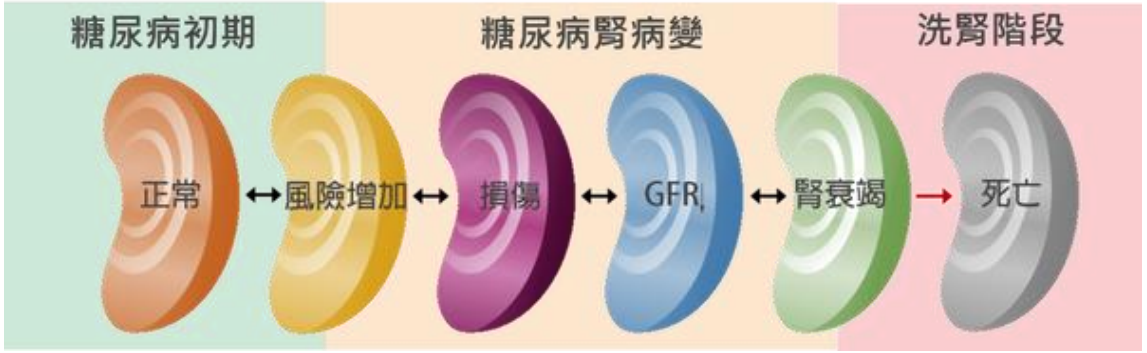


邏輯回歸 – 實作

邏輯回歸 – 實作　Pima Indians Diabetes Database

Kaggle Dataset

https://www.kaggle.com/uciml/pima-indians-diabetes-database



- 資料集來自美國國立糖尿病與消化與腎臟疾病研究所，預測患者是否患有糖尿病，患者均為皮馬印第安人血統20歲以上的女性。
- 輸入特徵為懷孕次數(Pregnancies)、葡萄糖(Glucose)、血壓(BloodPressure)、皮膚厚度(SkinThickness)、胰島素(Insulin)、身體質量指標(BMI)、糖尿病家族史(Diabetes Pedigree Function)、年齡(Age)，輸出特徵為有無罹患糖尿病(Diabetes)

邏輯回歸 – 實作 MNIST Database

Keras Dataset

- 資料集來自Keras中數字手寫資料集
- 輸入特徵維度大小是28×28，訓練集60000張，測試集10000張。數字手寫的標籤類別分別各為：

Label	0	1	2	3	4	5	6	7	8	9
Train	5923	6742	5958	6131	5842	5421	5918	7265	5851	5949
Test	980	1135	1032	1010	982	892	958	1028	974	1009

Keras

邏輯回歸 – 實作

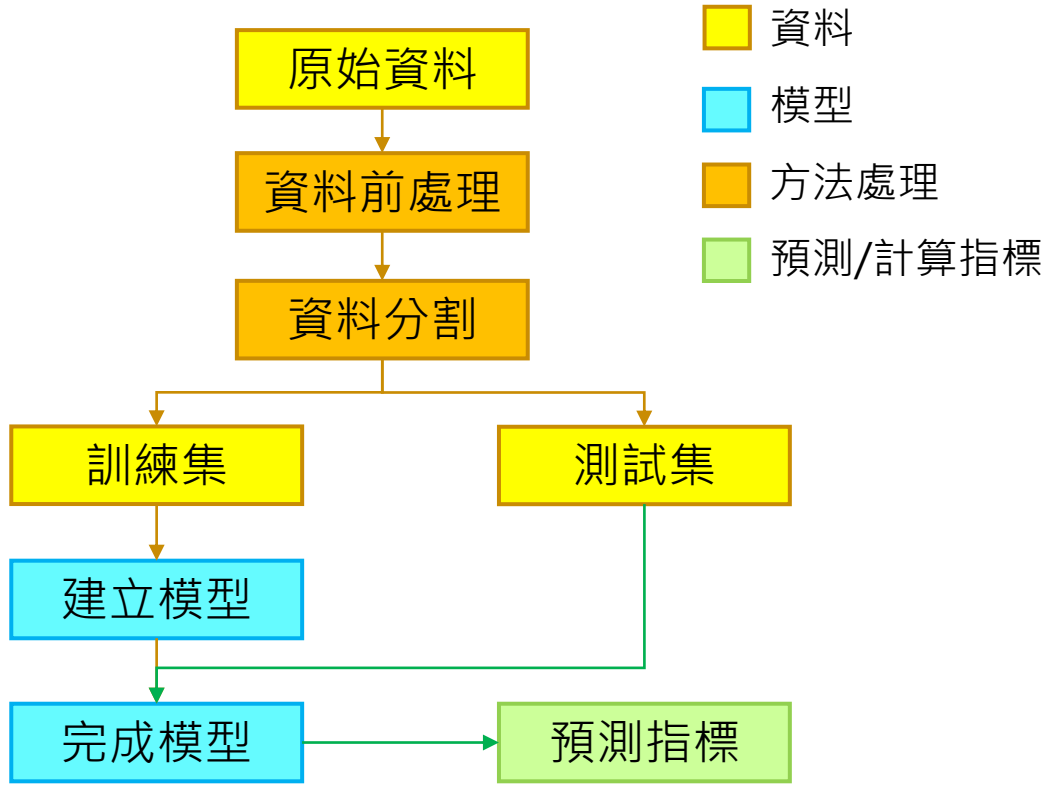
問題

1. 如何選擇模型的超參數？
2. 正則化超參數數值越大與越小，效果是如何？
3. 訓練時類別(標籤)比例若有懸殊是否會影響預測結果？

作業一 ★★★☆☆

1. 設計邏輯回歸預測灰階的數字手寫的模型，並儲存模型為.pkl檔案。設計GUI介面，可以插入圖片點擊確定後可以將該圖片帶入模型，並跳出視窗顯示該圖片數字為何。

流程圖



應用端