<http://moocs.nccu.edu.tw/course/123/intro> python入門

常用套件

* + Numpy
  + Pandas
  + Matplotlib
  + Scikit-Learn

Python 共有33個保留的關鍵字 請見「lecture1」

寫程式根據需求定義關鍵字

變數(var)、函數(func)、類別(class)

資料預處理

遺漏值得處理🡪數值形:平均數 次序:中位數 類別:眾數

but要看資料離散太強不適合

回歸

如何去衡量哪條線性模型是最佳的?(最佳化)

回歸就是要去算點到線的距離(誤差)

想要估計的就是距離平方和(最小平方法)

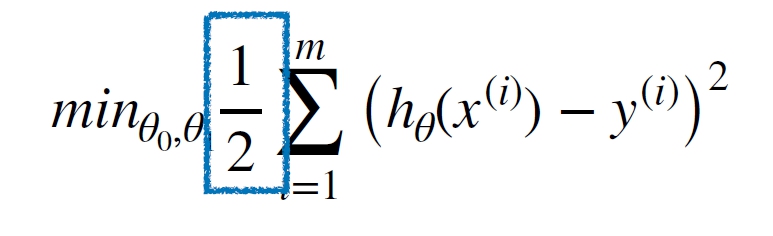
Pearson相關

不存在線性相關= 0 並非無相關

最好的相關線🡪與全部的點距離最近

使用ML的時機=>預測在經過訓練後效果能提升的情況下

最小平方法

cost-function

最小平方法是回歸常用的方法，目的是為了讓觀察值與估計值之誤差平

方和最小。

梯度下降法，是為了找到最佳的目標函數組合。

讓目標函數(θ0、θ1)越小越好(極值，用微積分找位置)

線性模型🡪y=ax+b、hθ(x)= θ0 + θ1x (找最好的θ0、θ1)

乘上1/2是為了要讓微分(求最佳)後的數字可以是整數(比較好看)

缺點:如果只用統計學的方式，加入新資料的話就要重算一遍，效率差

所以就需要機器學習。

同樣是讓cost-function越小越好，找出最小值(利用微積分找到極值)

θ<1和θ>1都不是最小值，因此讓程式在方程式中往返函數去走，就能收

歛。(參考regression PPT 17頁)

這個就是梯度下降法

但為了讓計算(程式)能往反方向走(一直來回)達到收斂。要乘上η(learning

rate)控制

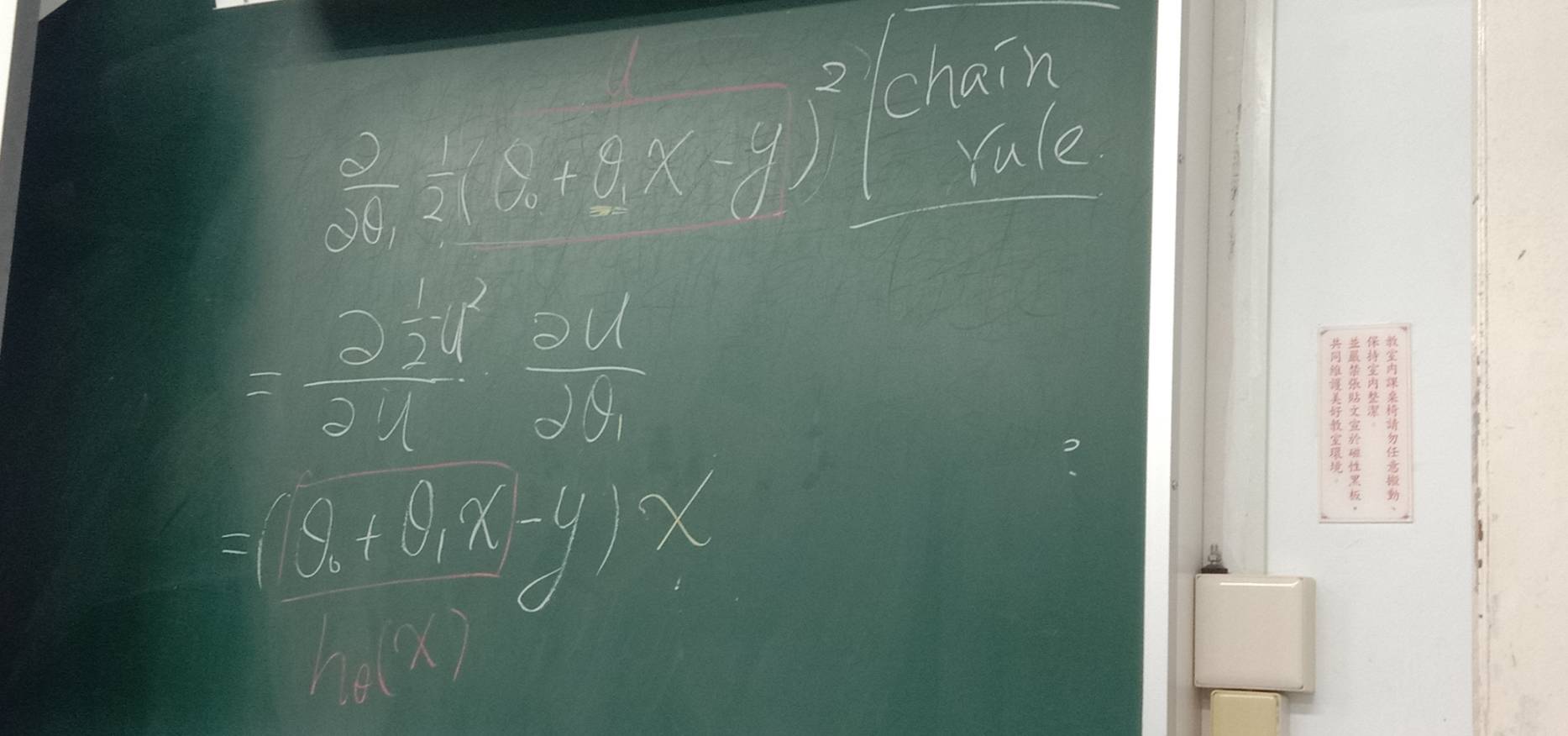
η的大小要靠常識與經驗，如果太大在靠近極值的時候就會超過(直接跳到

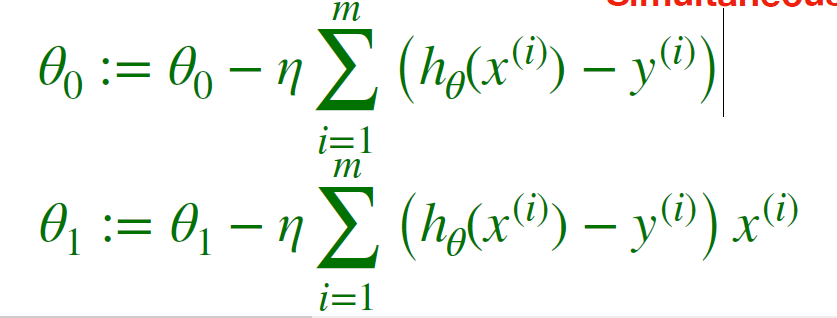
對面)導致無法收斂

邏輯思回歸也是一種梯度下降法

從原來的點(θ0,θ1)各往返函數方向更新一步(點的移動)

E(θ0, θ1) = 1/2Σ (hθ(x(i)) − y(i))^2 做偏微分如下就會得到





最後會的到這兩個公式就可以進行計算

初始的θ0,θ1若皆為0，在圖的表示就是x軸。然後持續遞迴直到收斂。

注意!! θ0,θ1要同時更改，因為程式的值更新後下面的會採用更新的值

Local optimal VS Global optimal問題🡪但要看這個公式畫出來的圖有沒有這

個問題

解決:只要讓起始點不同多跑幾次就會找到整體最小了

可以設定停止點，讓更新差值很小的時候(自訂小於多少)就讓程式停止

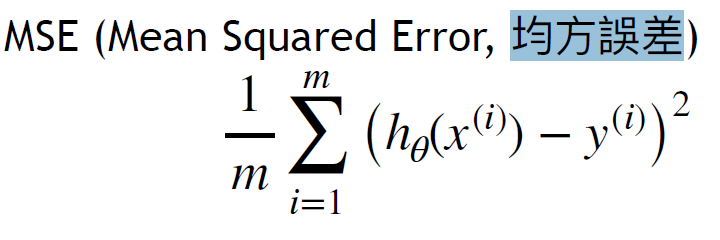
或是一開始就設定要讓程式跑幾次

10/15

MSE與R平方

均方誤差，點與線(模型)的距離

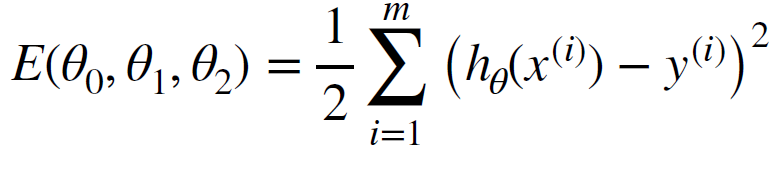
R平方越靠近1越好，越靠近0越差。代表公式有多少解釋力。



多項式回歸

如果一次方的線性模型解釋力不夠，可能需要更高次方，會擬合的越好。

但是可能有overfiting的問題(過度擬合)



觀測點到預期點的距離平方合。和一次方的一樣，只是多一個變項。

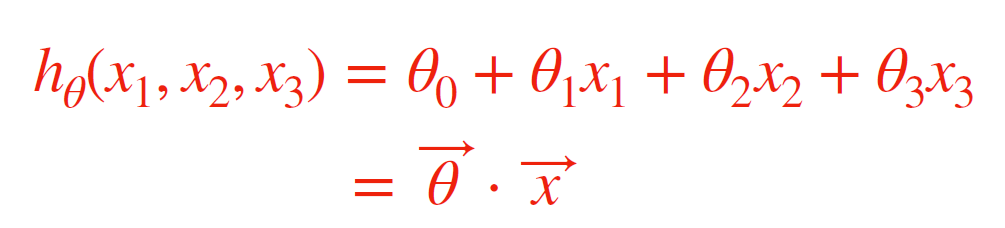
但是偏微分會變，多一個x平方的theta\_2

如果training資料的r^2大於測試的r^2就可能有overfiting的問題

做回歸前，要先看相關程度，可使用熱度圖。

複回歸

Y=θ0+θ1x+θ2x^2+θ3x^3 一般的回歸方程式



Normal equation 一樣是用微積分方法找最小誤差組合。但因為維度大所以

看起來非常複雜。更新時一樣利用梯度下降設定目前函數，然後用新數據

去跑就好。一樣讓函數不斷往反方向更新。

這個方法叫stachastic gradient descent常用在線上

補充:為何數據導入一定要做標準化(Normalization)?-->為了收斂變快

迴歸係數和相關係數會同號

倒函數不會一開始就正確地往最低點走(偏差相較標準化後較大)

數據分布範圍不同，所以可能會畫出很狹長的圖(曲面)

但標準化後圖比較好辨認

迴歸係數越大，代表該變項每增加一個單位，就影響模型解釋力越大

所以迴歸係數越大的比較值得投入。正負代表相關方向。

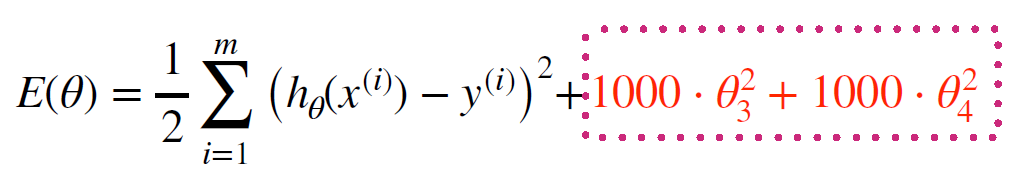
一定要記得檢查模型的合理程度。利用相關係數與迴歸係數。

如何避免共線性問題?-->資料預處理

**資料轉換、只留下獨立的變數、脊迴歸、主成分分析**

主成分分析缺點:降維後可能導致模型沒有解釋性

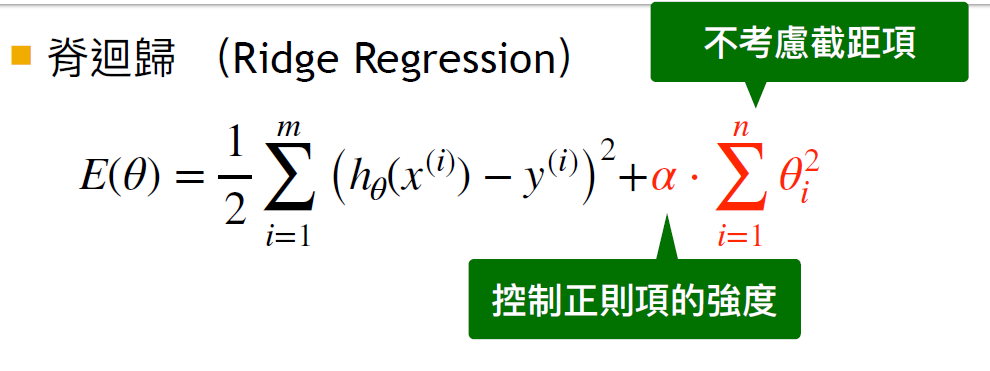
目標:希望模型的cost要小同時係數越小越好(正確率與overfiting的權衡)



加上懲罰項

模型越複雜正確率越好，但可能會overfiting

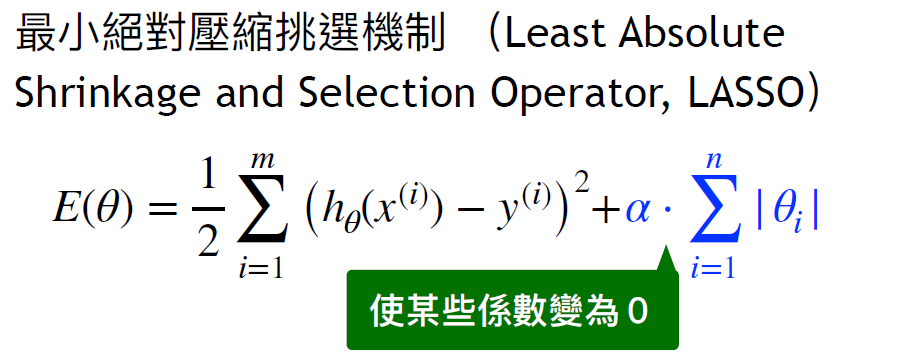
模型越簡單複雜度越低，但可能正確率不好



複雜度越低的模型，泛化能力會比較好

如果在意的是泛化能力的話，就應該要用脊回歸

LASSO🡪因為脊回歸還是會留下所有的解釋因子，因此用這個方法來排除



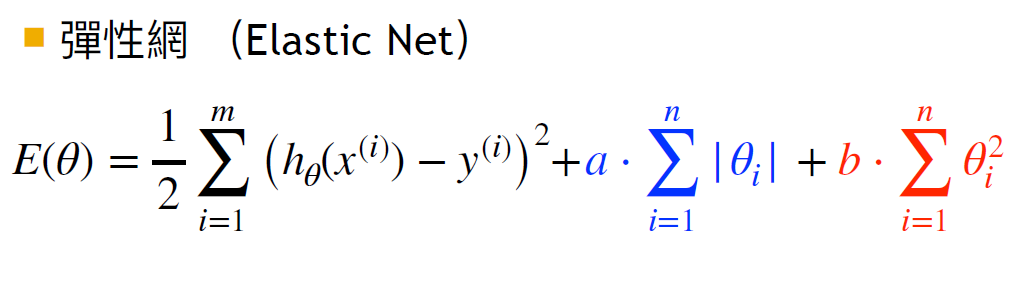
會導致解釋力變差，但是泛化能力會變好，然而要面對underfiting的問題

Ridge v.s. LASSO

* 實作時，Ridge通常是首選，因為LASSO在移除變數的同時，會犧牲模型的正確性
* 但如果特徵太多，且只有一小部分是真正重要的，那應該選擇LASSO
* 如果須解釋模型，LASSO也更更好理理解，因為使用較少特徵

那有沒有辦法同時解決希望有LASSO和Ridge的優點??--> Elastic Net

Elastic Net



L1 norm:藍色 L2 norm:紅色

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html> 參數調整參閱

迴歸總結

優點

簡單、直覺、易於運算

迴歸係數能得到有用的訊息

缺點

易受異常值影響

相關預測因子的權重會被扭曲

容易受曲線趨勢影響(一定要先檢查是否為線性關係，不是的話要轉換)

預測因子並非是因果關係，只有相關關係

分類

下載鳶尾花資料集

<https://archive.ics.uci.edu/ml/datasets/Iris>

術語用法Attribute=feature=variable=dimension

這四個詞不同領域使用習慣不同，但意思一樣

X常用大寫是因為X通常是矩陣(變數)，y只是向量(目標變數)

下載完資料要看欄位和目標變數的位置

目標:從花的長度和寬度來分辨屬於哪個亞屬

**ML process**

**準備資料 （包含資料預處理理）**

**選擇演算法**

**調整參參數**

**評估結果**

KNN

用測試集找出距離最近的來預測是哪一個種類

距離用歐式距離K設定範圍

將K個樣本多數決

K越大會容易有underfitting 越小容易overfitting

**交叉驗證**:為了避免產出的結果是極端結果

例如:這次隨機組成的test資料使正確率變成100%但重新跑之後可能就不

是了

所以從整個資料裡面切成需要的等份，每次都用其中一份作為test資料來

用。

最後把產生的所有結果取平均🡪CV\_score

Logistic regression

用來處理類別資料，可以二元到多元

預測值要介於0~1之間

若要讓復回歸的公式轉換成的結果介在0~1之間(theata介在0~1)

 就要用Sigmod function

假設跑出結果，我們該如何去決定什麼是Yes什麼是No?

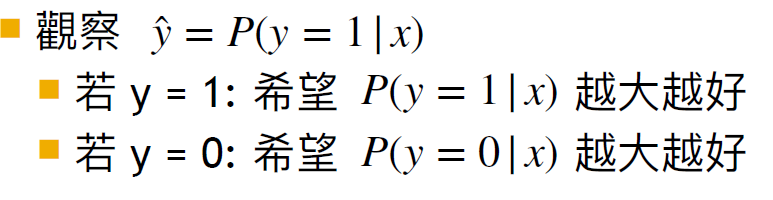
多少的值要是Yes多少要是No?-->**決策邊界(分區塊)** 「**θ⋅ x=0」**

例如:降雨機率80%VS降雨機率20%哪個算是會下雨?

假設以50%作為基準，就可得出80%會下雨20%不會下雨

但不一定所有情況都能精準知道範圍

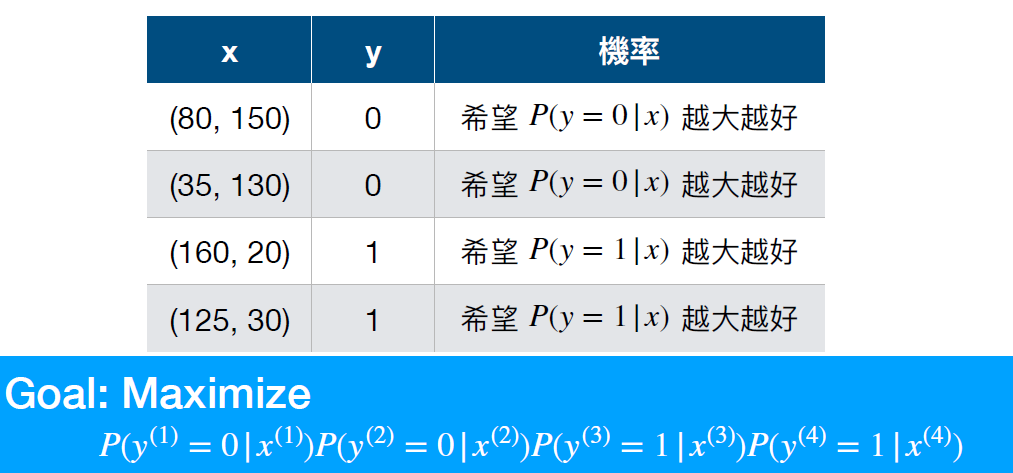
因此要找到theata的最佳組合來區分出範圍。



Y=1的時候，希望結果越接近1(是1的結果的機率要越大越好)

Y=0的時候，希望結果越接近0(是0的結果的機率要越大越好)

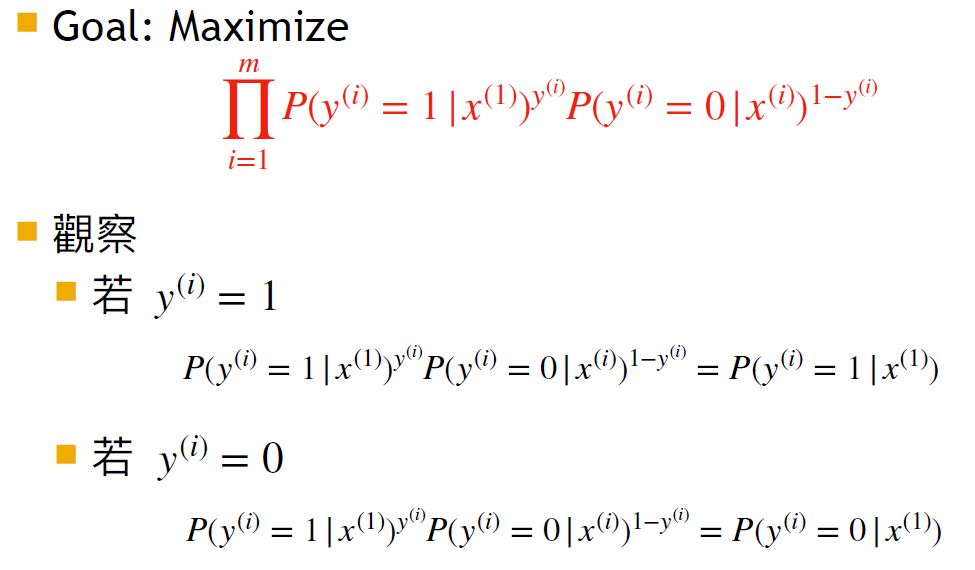
範例如下



最終希望的結果，是每個機率同時發生的結果，因此要相乘。

並且我要找出讓這個公式結果越大越好的組合。

如何將公式簡化呢?



Y=1或0的時候，只會留下對應的數值的機率

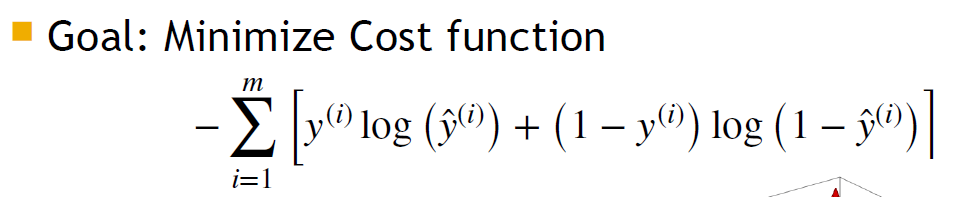
但問題是，這不能當作目標函數，因為所有結果都介於0和1之間。

所以會越乘越小，可以取Log解決

取Log之後公式就因其原理可轉換為相加

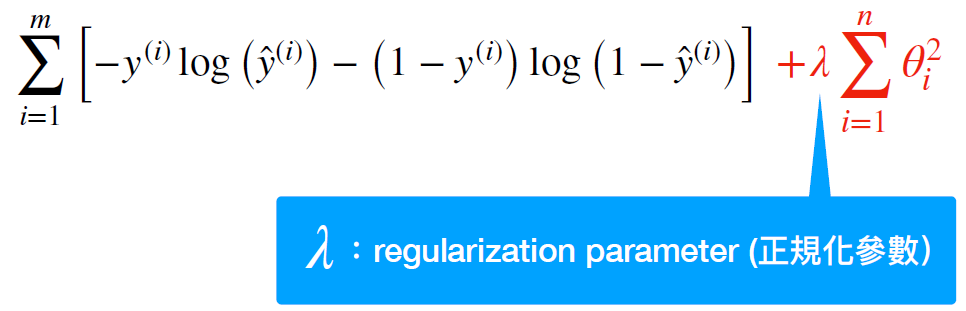
再乘上負號，目標要讓其越小越好。(cost function)

這樣就可以用梯度下降來實踐了!



結果就會呈現這樣。

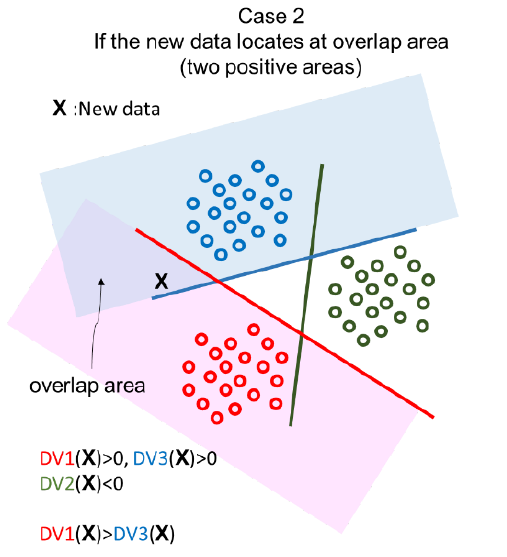
接下來進行偏微分，加上控制項(正規化參數控制平衡)



但邏輯思回歸一般只能處理二元分類的，若要變成多元分類要怎麼辦呢?

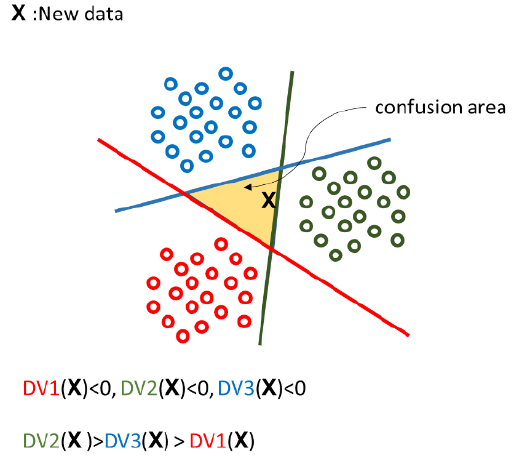
1 VS other方法🡪把每個類項個別當作基準(1)與其他，就可得出多種分類

但做完之後會有個問題，新資料進來的時候怎麼分類呢?以鳶尾花為例



距離分界越近的，越容易有誤差。所以要算這個點與兩條線距離。越遠越

代表他是某類的資料，以例題來看距離紅色比較遠，歸類為紅色。



這種情況則是新的點越靠近線越好，因為它不在三個區域任何一個裡

面。

Q:所以theta是什麼?X是什麼?

要拿來畫的公式為什麼求的是X2

一般回歸的結果是針對變數的解釋力。那羅吉斯的結果呢?