

Day 33

機器學習

機器如何學習？



出題教練

楊証琨

知識地圖 機器學習- 模型選擇 - 機器如何學習？

機器學習概論 Introduction of Machine Learning

監督式學習 Supervised Learning



非監督式學習 Unsupervised Learning



模型選擇 Model selection

概論

驗證基礎

預測類型

評估指標

基礎模型 Basic Model

線性回歸 Linear Regression

邏輯斯回歸 Logistic Regression

套索算法 LASSO

嶺回歸 Ridge Regression

樹狀模型 Tree based Model

決策樹 Decision Tree

隨機森林 Random Forest

梯度提升機 Gradient Boosting Machine

本日知識點目標

- 了解機器學習的原理
- 機器學習的模型是如何訓練出來的
- 過擬合 (Overfitting) 是甚麼，該如何解決

機器如何學習？

有三個步驟

1

定義好模型 (可以是線性回歸、決策樹、神經網路等等)

2

評估模型的好壞

3

找出讓訓練目標最佳的模型參數

機器如何學習 - 定義模型 (1/3)

- 一個機器學習模型中會有許多參數 (parameters)，例如線性回歸中的 w (weights) 跟 b (bias) 就是線性回歸模型的參數
- 當我們輸入一個 x 進到模型中，不同參數的模型就會產生不同的 \hat{y}
 - 希望模型產生的 \hat{y} 跟真實答案的 y 越接近越好
 - 找出一組參數，讓模型產生的 \hat{y} 與真正的 y 很接近，這個步驟就有點像學習的概念

Step 1: Model

$$y = b + w \cdot x_{cp}$$



w and b are parameters
(can be any value)

• $f_1: y = 10.0 + 9.0 \cdot x_{cp}$

$f_2: y = 9.8 + 9.2 \cdot x_{cp}$

$f_3: y = -0.8 - 1.2 \cdot x_{cp}$

..... infinite

圖片來源：[李宏毅ML Lecture 1: Regression - Case Study](#)

機器如何學習 - 評估模型的好壞 (2/3)

- 定義一個目標函數 (Objective function) 也可稱作損失函數 (Loss function)，來衡量模型的好壞
- 線性回歸模型我們可以使用均方差 (mean square error) 來衡量

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

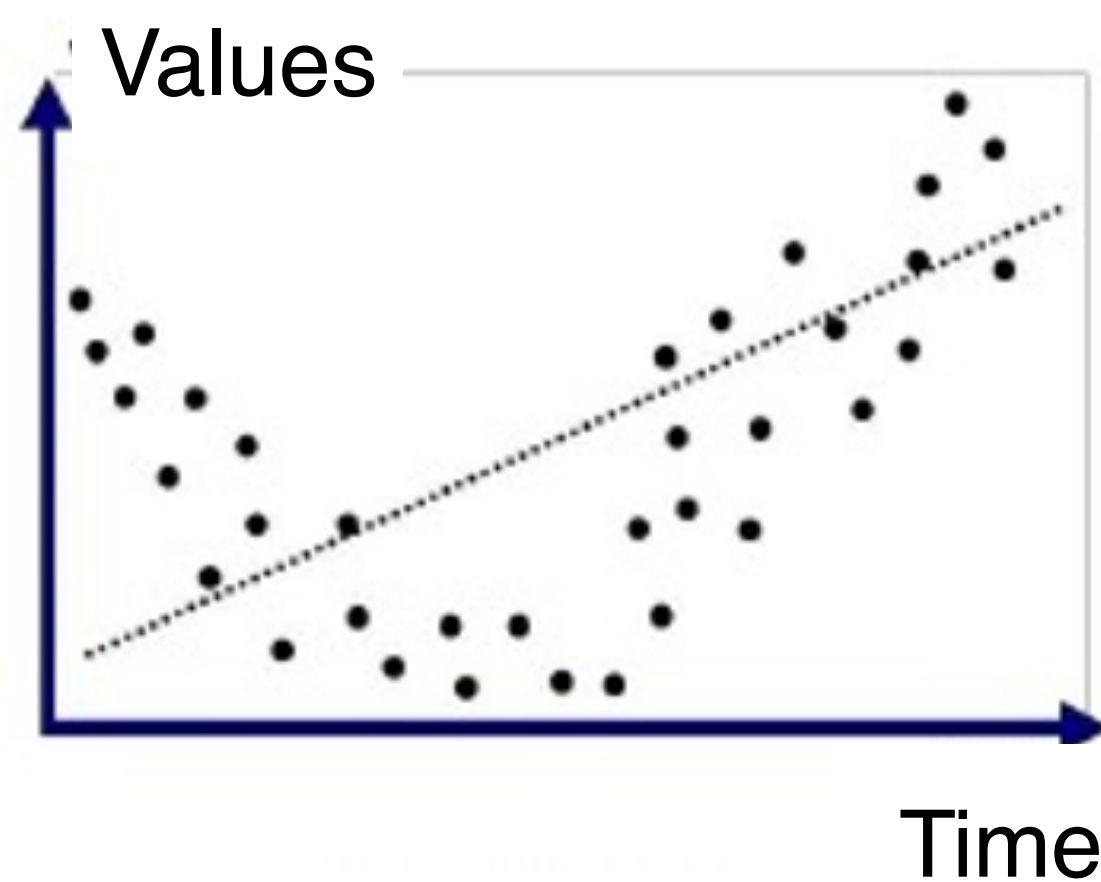
- Loss 越大，代表這組參數的模型預測出的 \hat{y} 越不準，也代表不應該選這組參數的模型

機器如何學習 - 找出最好的模型參數 (3/3)

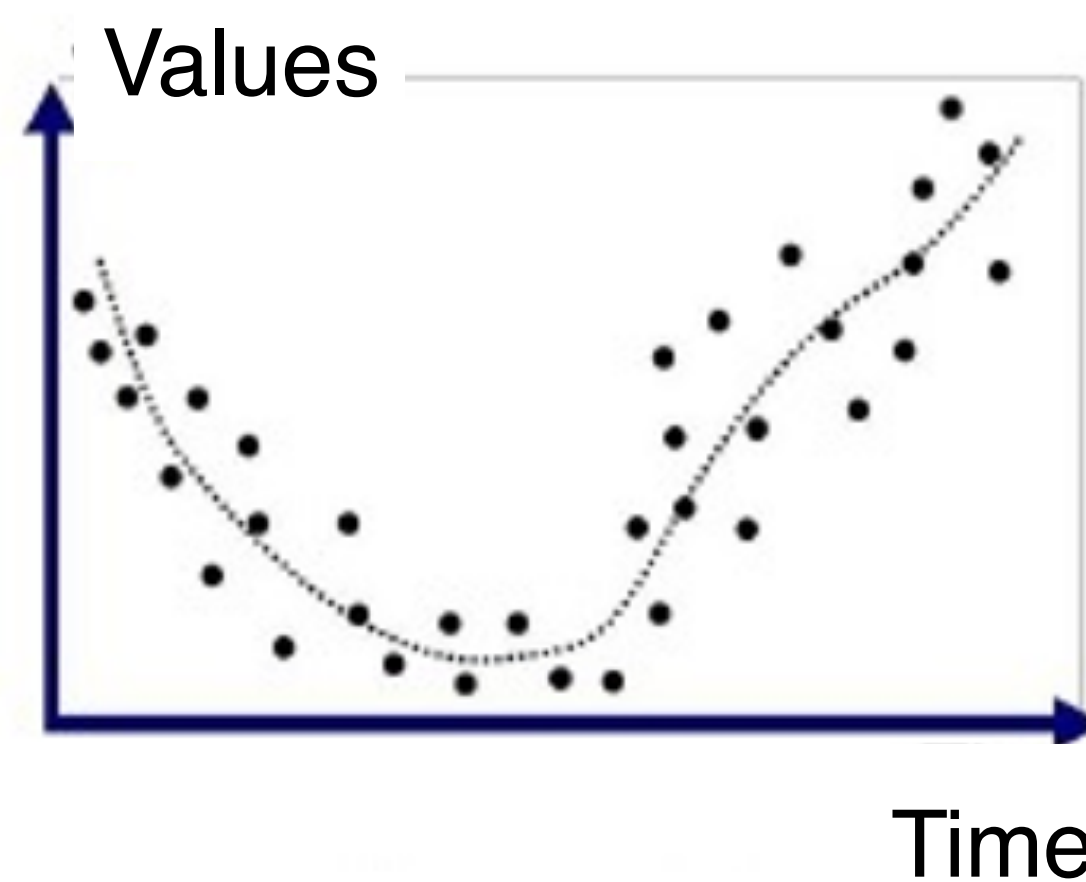
- 模型的參數組合可能有無限多組，我們可以用暴力法每個參數都試看看，從中找到讓損失函數最小的參數
- 但是這樣非常沒有效率，有許多像是梯度下降 (Gradient Descent)、增量訓練 (Additive Training) 等方式，這些演算法可以幫我們找到可能的最佳模型參數

過擬合 (Over-fitting)

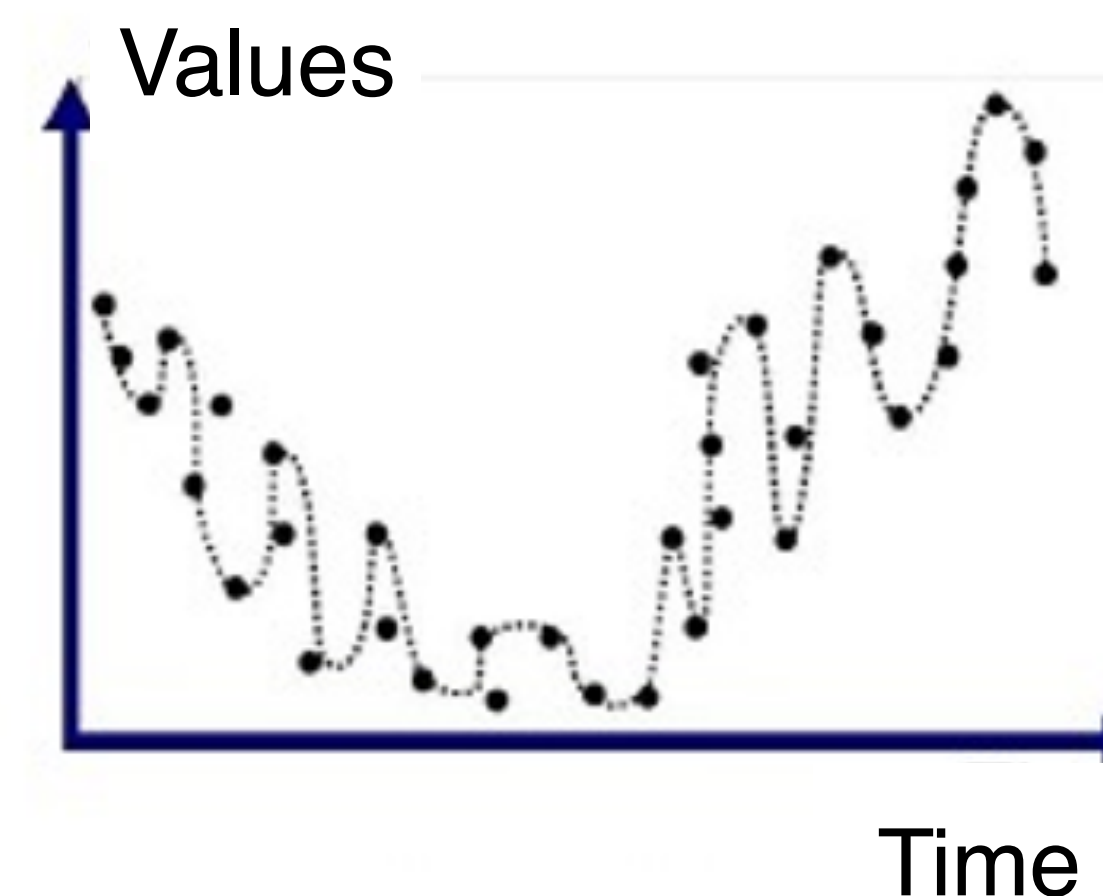
- 模型的訓練目標是將損失函數的損失降至最低
- 過擬合代表模型可能學習到資料中的噪音，導致在實際應用時預測失準



Underfitted



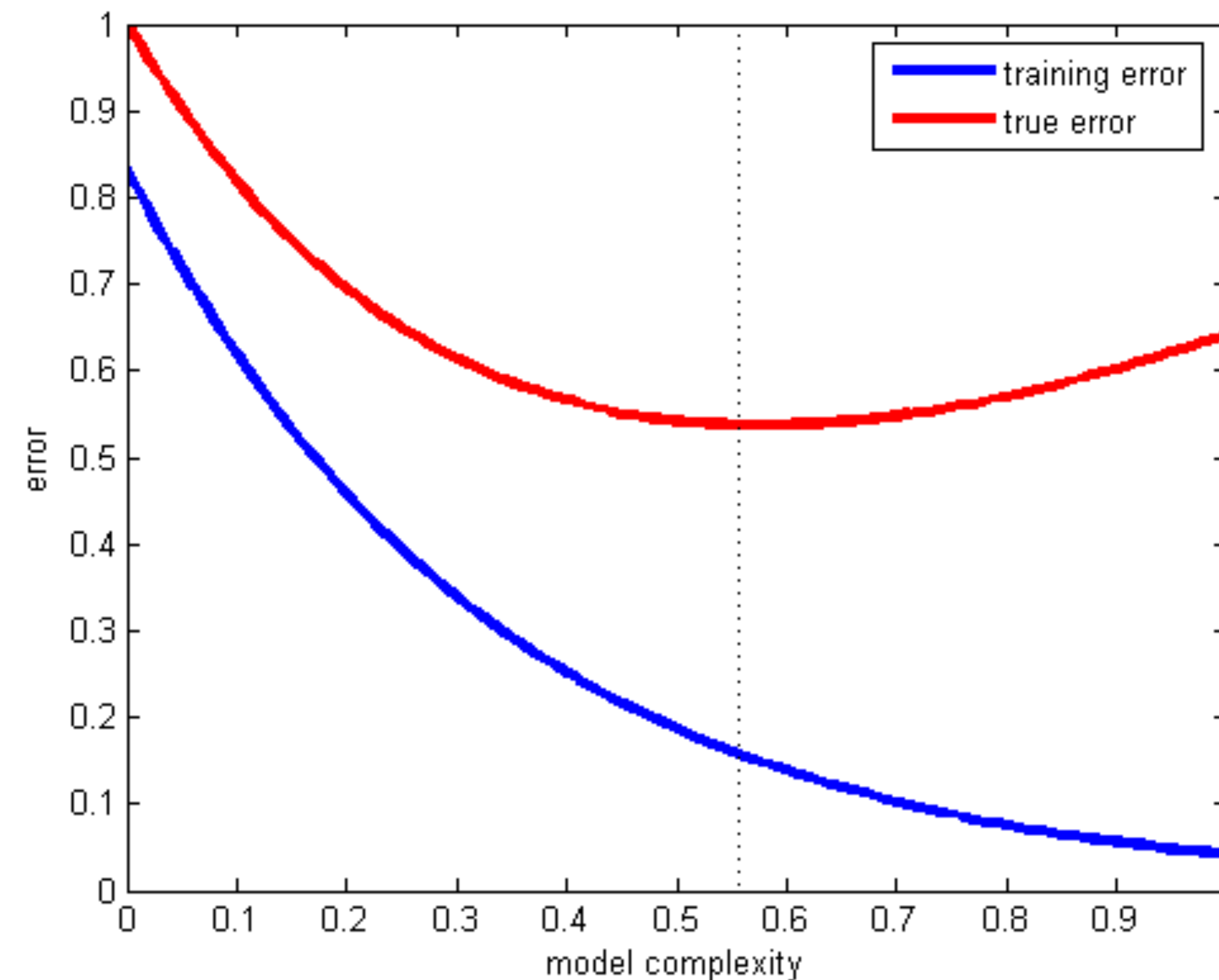
Good Fit/Robust



Overfitted

學習曲線 Learning curve

- 如何知道模型已經過擬合了？
- 保留一些測試資料，觀察模型對於訓練資料的誤差與測試資料的誤差，是否有改變的趨勢



圖片來源：[CIS 520 Machine learning](#)

如何解決過擬合或欠擬合

- 過擬合
 - 增加資料量
 - 降低模型複雜度
 - 使用正規化 (Regularization)
- 欠擬合
 - 增加模型複雜度
 - 減輕或不使用正規化



Q：前三天的課程作業好像都沒有程式碼？

A：機器學習的概念非常重要，我們希望學員先對機器學習有基本且正確的認識，後續再開始實作程式碼，避免因為撰寫程式碼卡關而對機器學習概念沒有正確的認知。



Q：過擬合在實務上經常發生嗎？跟所選的模型有關？

A：當資料沒有很大量時，過擬合實務上非常容易發生，正確了解是否有過擬合以及如何解決是非常重要的。所選模型也跟是否容易過擬合有關，像是後面課程會學到的決策樹模型就是個非常容易過擬合的模型，必須透過適當的正規化來緩解過擬合的情形。

解題時間 It's Your Turn

請跳出PDF至官網Sample Code & 作業
開始解題

