

機器學習 01

簡介

楊智淵 2025/9/5

關於老師



23/8- 長庚大學人工智慧學系助理教授



16/8-19/12 台大智慧聯網創新研究中心博後



09/8-15/6 加州大學美熹德分校電機資訊博士



08/8-09/7 中研院資科所研究助理



07/4-08/7 華晶科技演算法工程師



06/9-07/3 中研院資科所研究助理



03/2-05/6 台大資工所



99/10-03/1 原相科技軟體工程師



93/9-97/6 台大數學系

股票研究2

股票研究1

旅行與寫作

兵役

本門課學習範圍

問題、任務

圖像分類

客戶分類

盜刷偵測

人臉偵測

詐保偵測

商品推薦

語音辨識

圖像辨識

傳染病模型

設備故障預測

資料集

安德森鳶尾花

CIFAR-10/100

MNIST

SNLI

演算法

線性回歸

線性分類

類神經網路

支持向量機

K-Means

Ada-boosting

Markov
Random Field

隨機森林

Monte Carlo

Principle
Component Analysis

Hidden
Markov Model

K-NN

t-SNE

工具與
framework

Scikit-learn

NumPy

Pandas

Matlab

PyTorch

TensorFlow

OpenCV

libSVM

Keras

Matplotlib

Seaborn

程式語言

Python

C/C++

Matlab

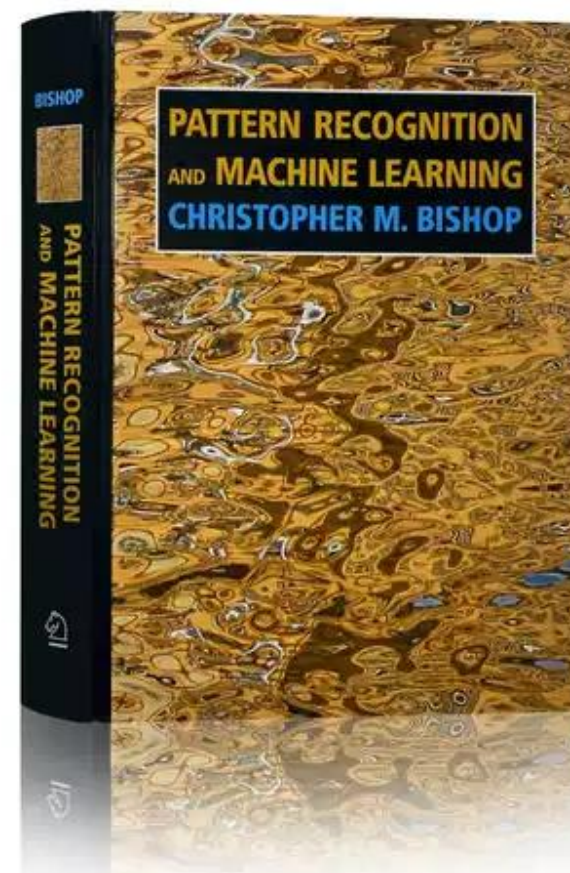
Java

去年各組期末專題題目

| 題目 | 資料來源 | 目標 |
|----------------------|------------------|-------------------|
| 極限配速預測與訓練計劃生成 | 自己戴手表，自己去跑、去收集 | 算出最佳配速的組合 |
| 厄瓜多爾 Favorita 商店銷售預測 | Kaggle競賽 | 預測出未來的銷售額、銷售量 |
| MLB大聯盟冠軍預測 | MLB官網 | 預測出下一屆的冠軍 |
| 鐵達尼號生存預測 | Kaggle競賽 | 由乘客的屬性估計生存的機率 |
| 影評情緒智析平台 | Kaggle的IMDB資料集 | 分類影評屬於正評還是負評 |
| DeepRacer | 系上有四台DeepRacer小車 | 由賽道路況影像和車子的參數操控車上 |
| MNIST手寫字辨識 | MNIST 1998年釋出 | 分類0到9共10個手寫數字 |
| Kaggle房價預測競賽 | Kaggle競賽 | 由房子的屬性為房子估價 |

教科書

- Pattern Recognition and Machine Learning by Christopher M. Bishop (2006)，著重在數學原理，可從作者官網([連結](#))下載正版pdf檔([連結](#))，習題選題解答([連結](#))，唯一的問題是這份pdf是初版的(2006)，堪誤還沒完整。
- 甚至有MIT學生寫出自己版本的整套習題解答([連結](#))
- 校圖有紙本，天瓏書局仍有貨，一本\$3500 NTD
- 這本書有簡體中文版，2014譯成，很多錯誤已經修掉了，網路非常容易找到別人分享的pdf檔，但請注意版權問題。



PRML的優點與缺點

優點

- 大名鼎鼎，知道這本書的人非常多，而且有柏克萊大學的教授，幫它釋成簡中版，以餉簡中的讀者。
- 用非常嚴謹數學講，清清楚楚，令讀得懂的人拍手叫好。
- 作者自己釋出PDF檔，免費。

缺點

- 內容又深又廣，多翻幾頁，就跳到研究所的等級了。
- 幾乎整本書都是數學式，又是線代、又是微積分、又是機率，而你們現在才正在學線代，機率要等到二下才學。
- 作者釋出的PDF檔是2006初版，錯誤還沒校訂掉，需要配合堪誤表看。

課程網頁

- <https://yangchihyuan.github.io/courses/MachineLearning2025>
- 行事曆、office hour、助教資訊、參考書籍、期末專期
- 之後會放上的有課程錄影、作業、期中考、期末專題

The screenshot shows a web browser displaying the course page for 'Machine Learning 2025' at CGU AICV Lab. The page includes a navigation bar with links to 'people', 'publications', 'courses', 'talks', 'cv', and 'contact'. The main content area is titled '機器學習 2025' and contains several sections: '課程資訊' (Course Information), '修課需求' (Prerequisites), '時間與地點' (Time and Location), '助教與答問時間' (TA and Q&A Time), and '教師與答問時間' (Professor and Q&A Time). At the bottom, there is a '課程大綱' (Course Outline) table.

| 週次 | 日期 | 主題 | 投影片 | 錄影 | 預定事項 |
|----|------|------|-----|----|------|
| 1 | 9/5 | 簡介 | | | |
| 2 | 9/12 | 機率分佈 | | | |

生成式AI工具

- 學校核給我們\$94000元訂閱生成式AI工具。
- 班上目前有48人選課，加上我和兩位助教。
- 每人分配 $94000/51=1843$ 元，加退選之後，這個金額再確定。
- 大家自由去訂ChatGPT、Gemini、Copilot等等工具，自己選。
- 報帳方式和限制
 - 你只能以月為單位訂閱，不能以年為單位訂，也不能購買點數儲存下來
 - 使用期間必需在12/31日之前，超過這個日期沒有輔助
 - 大家自己先付，將刷卡明細和收據列出來交給助教(或是截圖，目前規則不詳)，我和助教會整理好之後交給教務處教學資源中心，學校會計室核准後，會將總額一筆給我，我和助教再將現金一筆一筆發給同學。

訂閱範例

- ChatGPT，每月\$20美金，台灣銷售稅5%=\$1美金。21美金約\$644台幣，再加國際刷卡手續費1.5%，約\$654台幣。訂三個月，約\$1962。
- AI易付卡、Monica AI等々はLLM服務批發商，他們向大LLM公司購買大量點數，再轉售給消費者，消費者可以付得少，而且用到多種的模型，甚至非語言的模型。缺點：他們有一套自己的使用者界面，你要看看符不符合你的需求。

Q&A (1)

- 可以不訂嗎？可以。其實現在免費的模型像Gemini 2.5 Flash已經很好用了，而且幾乎可以無限發問。ChatGPT 5會限制你的用量，但就算暫時降到ChatGPT 4，仍然很利害。
- 那為什麼學校要核這筆經費給我們？
 - 1. 校長的政策。如果長庚是以AI為中心的學校，鼓勵學生熟用AI工具很合理。
 - 2. 減少不公平。ChatGPT非常強大，付費和免費有差別，有人每月繳600元，就繳出了得98分的作業，而有人因為沒有錢而沒買，寫了很久，只得70分，你會覺得不公平嗎？

Q&A (2)

- 什麼不可以訂？不符合「生成式AI工具」的不可以，像是影劇服務或是音樂串流服務。
- 可以半路換另一款嗎？可以，一樣繳收據和刷卡明細就可以。
- 刷ChatGPT時，會有國外刷卡手續費，會補償嗎？會。這是為麼要請各位繳刷卡明細的關係。
- 如果我12月中才訂一個月，會有超過12/31的部份，那我會收到多少補償？不知道，教學資源中心沒有明說，可能要看會計室怎麼判斷了。請儘量避免這種模糊的狀況。

Q&A (3)

- 這堂課的作業和專題可以用AI生成嗎？可以，對我來說，你用的是AI工具，我很鼓勵你熟用這些工具，這也是你們競爭力的一部份，但生成的內容你得負責，會反映在你的得分上。
- 我可以用這堂課訂閱的AI工具去寫其他課的嗎？我這邊沒有問題，在我看來這是外溢效用，反正月租費已經付了，儘量用沒關係。但你得問其他課的老師允不允許你用AI去寫作業。

我自己用這些生成式AI工具 (1)

- ChatGPT 5：上傳一份英文的PDF論文稿件，問有沒有錯別字和文法錯誤？我發現ChatGPT 4不能分析PDF，而Gemini 2.5 Flash認不太出來錯別字和文法錯誤。
- Gemini 2.5 Flash：將寫程式遇到的看不懂錯誤訊息丟進去，它生成解釋和建議，讀著讀著，我常常就懂了。什麼大大小小寫程式的問題我都問，它給的提示和解釋常常都很好，也可以直接生成程式碼給我的範例。
- Grammarly Education：寫論文、寫評論、寫Teams訊息時，自動挑出拼寫文法錯誤，並能建議適當的用字和介係詞。

我自己用這些生成式AI工具 (2)

- Copilot GitHub：與MS Code整合，自動生成我將要寫的程式碼或註解，常常猜對，拉高我的寫程式速度，並且可以出現我沒想到，但更好的寫法。
- Microsoft Copilot：生成照片，品質不錯，但構圖不見得符合我想要的，多試幾次，再加上用繪圖軟體編輯，修到可用的結果，作為我研討會海報和研究案投稿的素材。

我期望你們怎麼使用LLM

- 把它當作一個機器人，看過全世界所有的書、網頁、和對話記錄，你有問題，隨時可以問他。
- 他說的話，可能有錯，你若懷疑時，就去查證。
- 你要為最後繳出來的作業和專題負責。

給分與出席

- 作業40%
- 期中考30%，考大的概念，不考很細節的東西，close book，要看各位同學腦中吸收多少
- 期末專題報告30% (2到5人一組，若題目大可以加入。要上台簡報、繳程式碼和書面報告。助教、我的評分各佔42.5%，其他組互評平均佔15%，我的評分原則：專題的精采程度一半，能夠幫助其他同學學習這門課知識一半。助教和其他同學的評分原則不必和我相同。互評時，只要評其他組即可，不用評自己組。)
- 不點名、沒有缺席懲罰，上課會錄影，放在課程網頁上，可以事後看。
- 課堂中的參與，不會影響你們的分數，但會影響我的印象，這會影響我在推薦函裏的描述。

算力

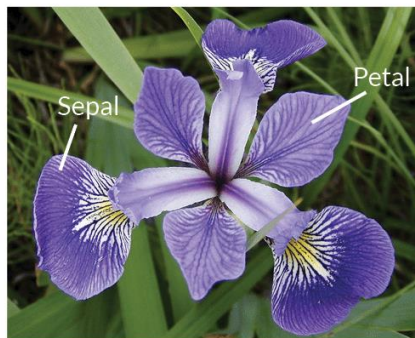
- 在這堂課，你們還不會使用的GPU，到三上的深度學習時，你會大量的使用GPU。
- 平常的作業，算力需求不大，用自己的筆電或是Colab，都夠了。
- 在期末的時候，如果你們挑的資料集大、你們想要套用的演算法比較重，你們想測的超參數又多的話，可能會發生筆電算一整天都算不完的狀況，去年8組裏有2組有發生這種狀況。
- 學校的AI中心會給開帳號給你們，因為你們不需要和別人搶GPU，應該都分配得到CPU。

什麼是機器學習？

- 給定一堆資料，從中建立一支演算法，讓它從這堆資料中「學」到某些東西，可以用來達到預測或是判斷的任務。
- 這裏的「某些東西」，往往指的是參數的值。
- 我們來看一個例子：安德森鳶尾花資料集

安德森鳶尾花資料集

- 三個不同的品種，150筆資料，1936年由埃德加·安德森從加拿大加斯帕半島實際收集來的。
- 有四個變數：花萼長、花萼寬、花瓣長、花瓣寬



Iris Versicolor
變色鳶尾



Iris Setosa
山鳶尾



Iris Virginica
維吉尼亞鳶尾

Sepal:花萼 Petal:花瓣

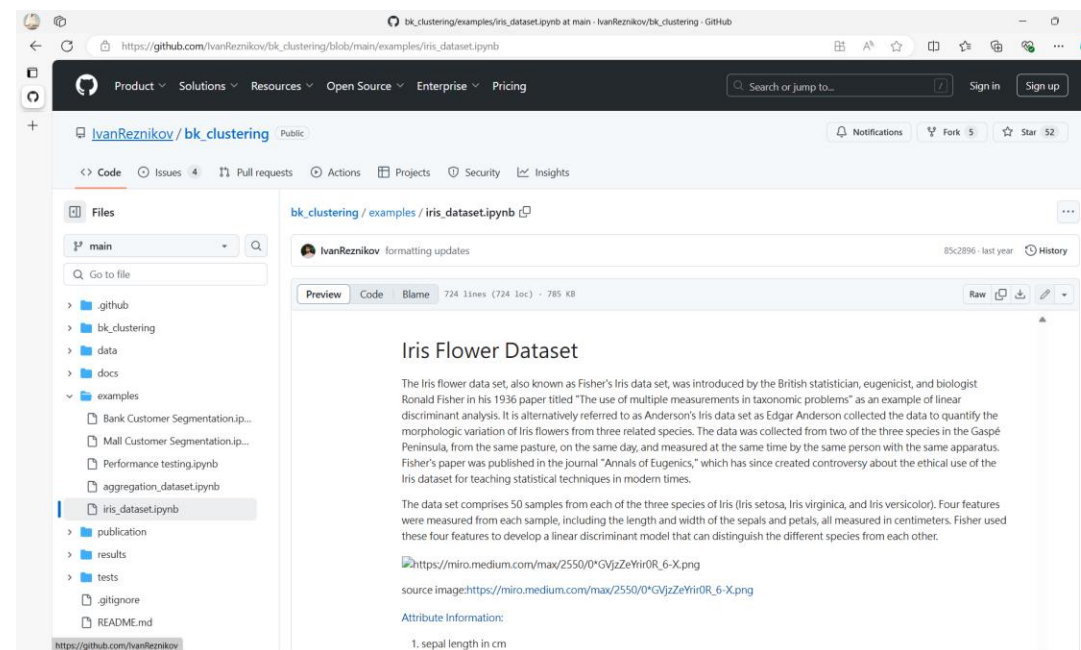
| | A | B | C | D | E |
|----|--------------|-------------|--------------|-------------|---------|
| 1 | sepal_length | sepal_width | petal_length | petal_width | species |
| 2 | 5.1 | 3.5 | 1.4 | 0.2 | Setosa |
| 3 | 4.9 | 3 | 1.4 | 0.2 | Setosa |
| 4 | 4.7 | 3.2 | 1.3 | 0.2 | Setosa |
| 5 | 4.6 | 3.1 | 1.5 | 0.2 | Setosa |
| 6 | 5 | 3.6 | 1.4 | 0.2 | Setosa |
| 7 | 5.4 | 3.9 | 1.7 | 0.4 | Setosa |
| 8 | 4.6 | 3.4 | 1.4 | 0.3 | Setosa |
| 9 | 5 | 3.4 | 1.5 | 0.2 | Setosa |
| 10 | 4.4 | 2.9 | 1.4 | 0.2 | Setosa |
| 11 | 4.9 | 3.1 | 1.5 | 0.1 | Setosa |
| 12 | 5.4 | 3.7 | 1.5 | 0.2 | Setosa |
| 13 | 4.8 | 3.4 | 1.6 | 0.2 | Setosa |
| 14 | 4.8 | 3 | 1.4 | 0.1 | Setosa |
| 15 | 4.3 | 3 | 1.1 | 0.1 | Setosa |
| 16 | 5.8 | 4 | 1.2 | 0.2 | Setosa |
| 17 | 5.7 | 4.4 | 1.5 | 0.4 | Setosa |
| 18 | 5.4 | 3.9 | 1.3 | 0.4 | Setosa |
| 19 | 5.1 | 3.5 | 1.4 | 0.3 | Setosa |
| 20 | 5.7 | 3.8 | 1.7 | 0.3 | Setosa |
| 21 | 5.1 | 3.8 | 1.5 | 0.3 | Setosa |
| 22 | 5.4 | 3.4 | 1.7 | 0.2 | Setosa |

這個資料集的特性

- 非常小，才150筆。
- 維度也少，才4維，可以說是剛剛好，因為2維和3維就簡單到畫得出來了，4維剛好畫不出來，需要想像一下。
- 以當前的計算力說來，是個玩具等級的資料集。
- 但很直覺，大家都看過花，知道什麼是花瓣、花萼。
- 所以非常適合拿來當作入門的第一個資料集。

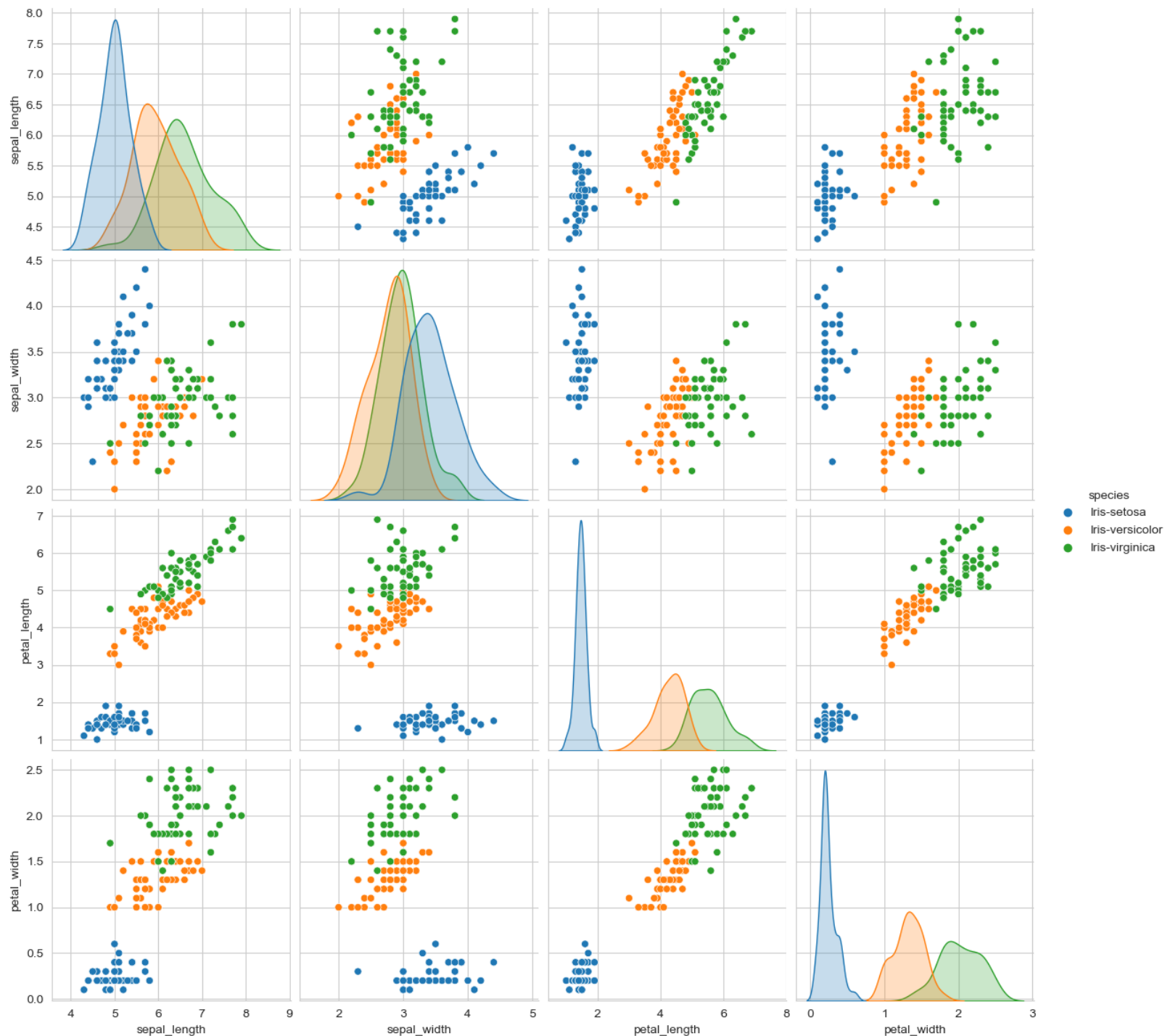
我們來視覺化它

- 你可以叫LLM生視覺化的程式碼給你，也可以直接看別人寫好的程式碼，如右。[連結](https://github.com/IvanReznikov/bk_clustering/blob/main/examples/iris_dataset.ipynb)



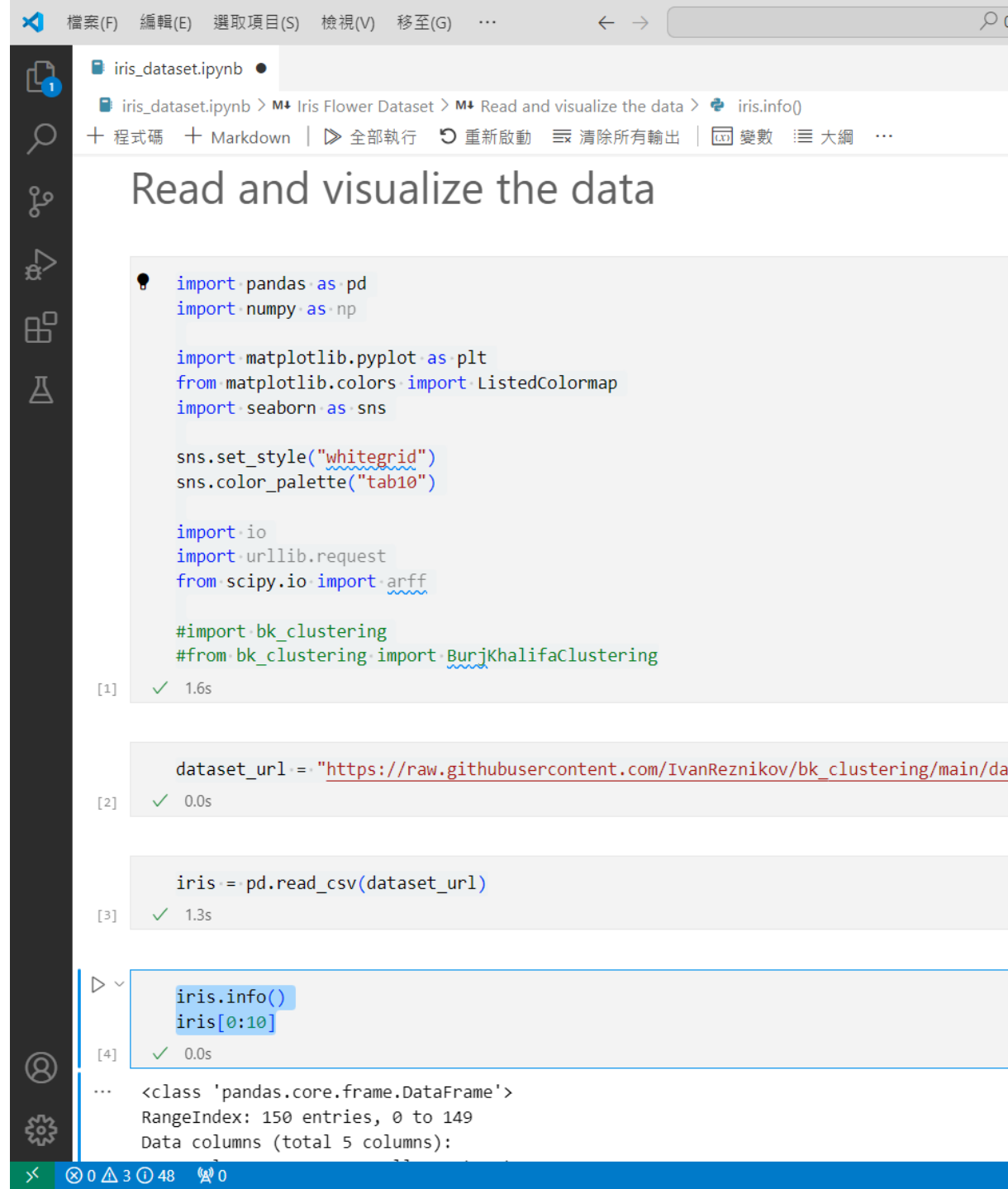
視覺化結果

- 我們只看四維中的兩維的話，會有 $C_2^4 = 6$ 種組合。
- 所以說(1,2)和(2,1)是一樣的，只是x,y軸對調而已。
- 也可以只看四維裏的單獨一維，就是對角線上的圖。



程式碼

- Python 語言
- Pandas 讀csv檔
- Seaborn 視覺化
- 我們在課裏不教程式，假設你們在大一已經學過了



The screenshot shows a Jupyter Notebook titled "iris_dataset.ipynb" with the following content:

```
iris_dataset.ipynb > Iris Flower Dataset > Read and visualize the data > iris.info()
+ 程式碼 + Markdown | ▶ 全部執行 ↺ 重新啟動 🗑 清除所有輸出 | 📄 變數 📖 大綱 ...
```

Read and visualize the data

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
import seaborn as sns

sns.set_style("whitegrid")
sns.color_palette("tab10")

import io
import urllib.request
from scipy.io import arff

#import bk_clustering
#from bk_clustering import BurjKhalifaClustering
```

[1] ✓ 1.6s

```
dataset_url = "https://raw.githubusercontent.com/IvanReznikov/bk_clustering/main/dataset.csv"
```

[2] ✓ 0.0s

```
iris = pd.read_csv(dataset_url)
```

[3] ✓ 1.3s

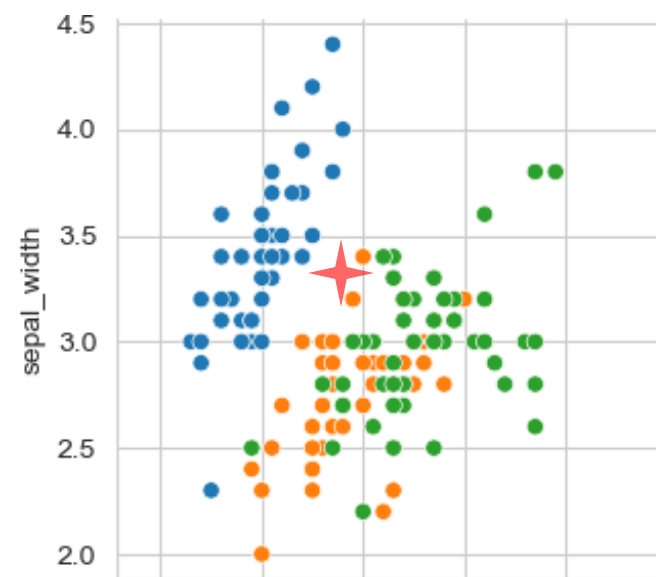
```
iris.info()
iris[0:10]
```

[4] ✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
```

任務

- 如果給一筆新的資料(4維：花萼長、花萼寬、花瓣長、花瓣寬)，但是沒有給你是那一種花，要你猜，你要怎麼猜？
- 有一個很直覺的解法，就是跟訓練資料集比，看比較接近哪一種。
- 但這只是一個概念，你必需定義什麼叫「接近」。
- 以(2,1)這個圖為例。如果新資料的位置在紅十字處，你覺得是哪一種？



回到機器

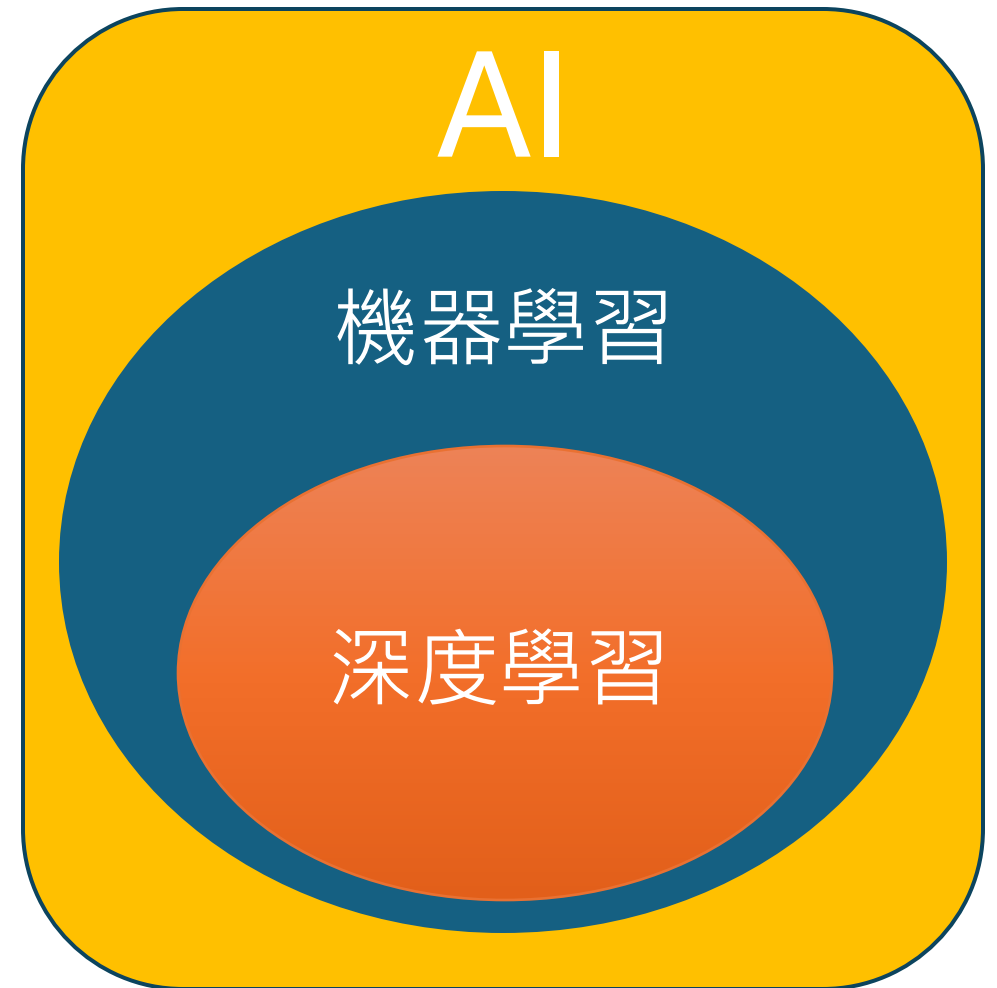
- 現在，你不要用自己的頭腦去判斷，而是要用機器去判斷，你要寫程式讓機器能達到這個任務，這就是你們在這堂要學習的知識和經驗了。

機器學習/深度學習/AI的關係

各位在這堂課裏學的，是2006年之前的機器學習，比較直觀，參數較少、計算量較少、資料較少。

深度學習其實就是多層的類神經網路，在1998的LeNet就出現了，參數較多、計算量較大、也需要更多的資料來訓練，當時的電腦沒那麼快，資料也沒那麼多，所以用途不廣，2012的AlexNet撼動了電腦視覺界，從此大爆發，非常多人投入研究，一路成長到現在2025年。

AI研究的是機器如何像人類一樣思考，機器學習擔任的是思考裏的預測與判斷的部份。



MNIST 資料集

- Modified National Institute of Standards and Technology
- 直譯：修改的美國國家標準暨技術研究院
- 意譯：美國中央標準局手寫數字圖像資料集修改版本

MNIST 資料集

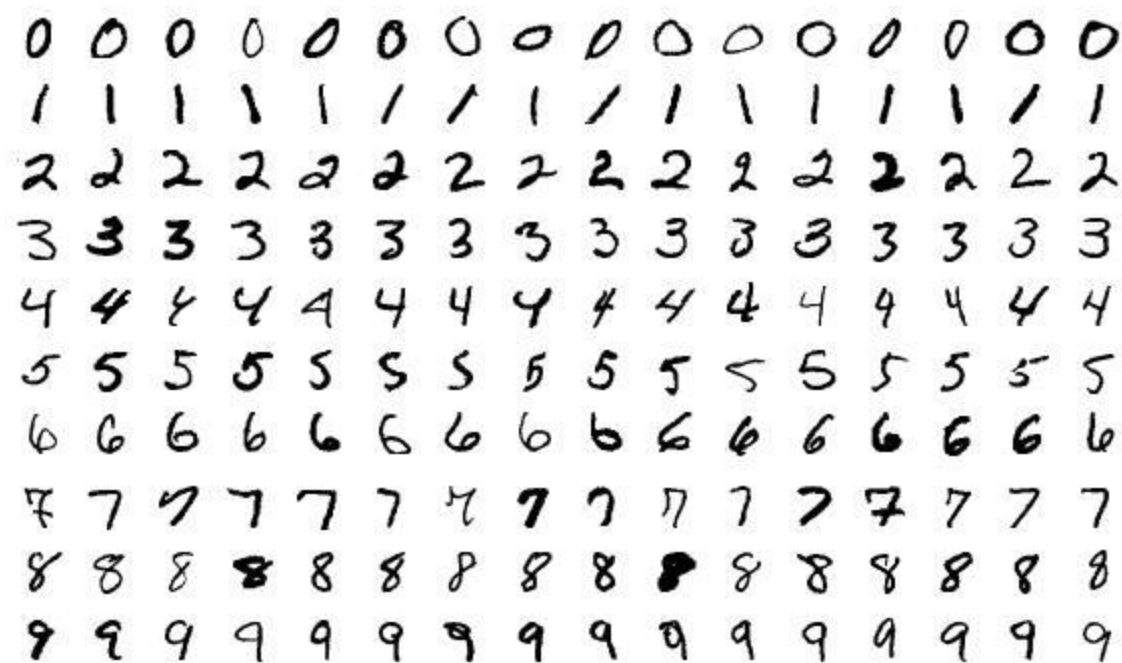
圖片解析度 28*28 圖素

維度 $28*28=784$

每一數字7000張

總資料集70000張

其中60000張用來訓練你的演算法，10000張用來測試你的演算法，看它的正確率有多高。



緣由

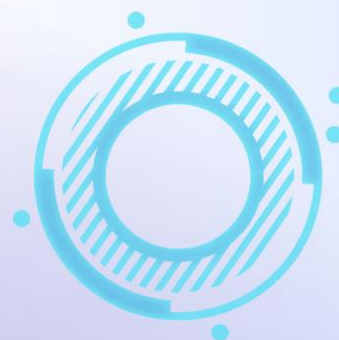
- 1980年起，美國人口普查局就一直很有興趣怎麼將人口普查表上的手字自動辨識出來(字包括字母、數字、和標點符號)。
- 他們聘請了國家標準局的影像辨識小組來評估發展一套光學字元辨識系統的可能性。
- 經過多年的累積，這個小組在1990年發表了SD1 (Special Database)，1992年SD3和7，1995年SD19。
- 1998年時，Yann LeCun取了SD1和SD3中的0~9數字，並將圖片縮到28x28圖素，各數字的量都控制在7000個，稱之為MNIST資料集。

動機

- 1998年時 Yan LeCun 38歲，是AT&T Labs-Research影像處理研究處的處長。
- 他們當時正在發展卷積神經網路，是類神經網路裏一種特別的型式，這種型式的網路，很適合拿來處理二維的影像，但計算量很大，而1998的電腦只有Pentium II等級，大圖片跑不動，彩色圖片的圖素量是灰階的三倍，不適合。
- 手寫數字的類別只有10種，又不需要彩色，用途又廣，郵遞區號、支票金額會用到，是個好目標。

去年期末專題其中一組，就是重測MNIST

- Yann LeCun在1998年的論文已經報告過了，卷積神經網路比起其他演算法，準確度更高。
- 但他們想知道，如果不用卷積神經網路，而是用其他演算法，效果會差多少？所以他們就自己重跑一遍。
- 因為2024的電腦很快，類神經網路的函式庫也都建得很完整易用，他們的實驗很快就跑完了。



模型建立

我們共用了八種模型

1.LogisticRegression

2.SVM

3.RandomForest

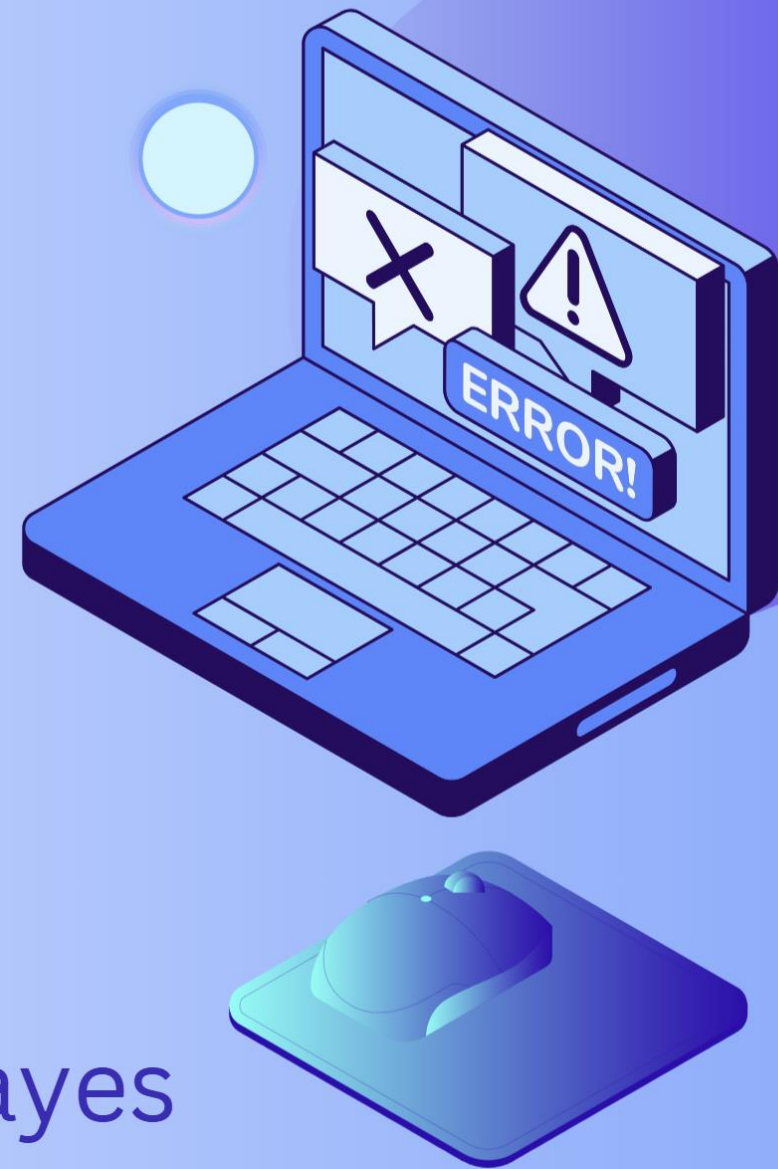
4.XGBoost

5.KNN

6.MLP

7.CNN

8.Naive Bayes



範例

- 因為現在的函式庫已經很完備了，對於我們這堂課會學的演算法，通常一兩行就算完了。
- 你不需要從零開始寫。
- 知道它的原理、性質、和用法，是你能獲得的基本能力。
- 如果你能自己寫出來，那是進階能力。



梯度提升樹 XGBClassifier

```
from xgboost import XGBClassifier

xgb_model = XGBClassifier(eval_metric='mlogloss', random_state=42)
xgb_model.fit(X_train, y_train)
xgb_accuracy = xgb_model.score(X_test, y_test)
print(f"XGBoost 準確率: {xgb_accuracy}")
```

XGBoost 準確率: 0.9739285714285715

最後結果比較

- 他們的實驗，也是CNN最好，但SVM，XGB、MLP也都到97%的準確度以上，也不差。
- 這裏沒有列出來的，是計算的成本。

| 模型名稱 | 準確率 |
|---------------------|---------------|
| Logistic Regression | 0.91952 |
| SVM | 0.97345 |
| Random Forest | 0.96548 |
| XGBoost Classifier | 0.97392 |
| KNN | 0.9649 |
| MLP (多層感知機) | 0.9752 |
| CNN (卷積神經網路) | 0.9911 |
| Naive Bayes | 0.8260 |

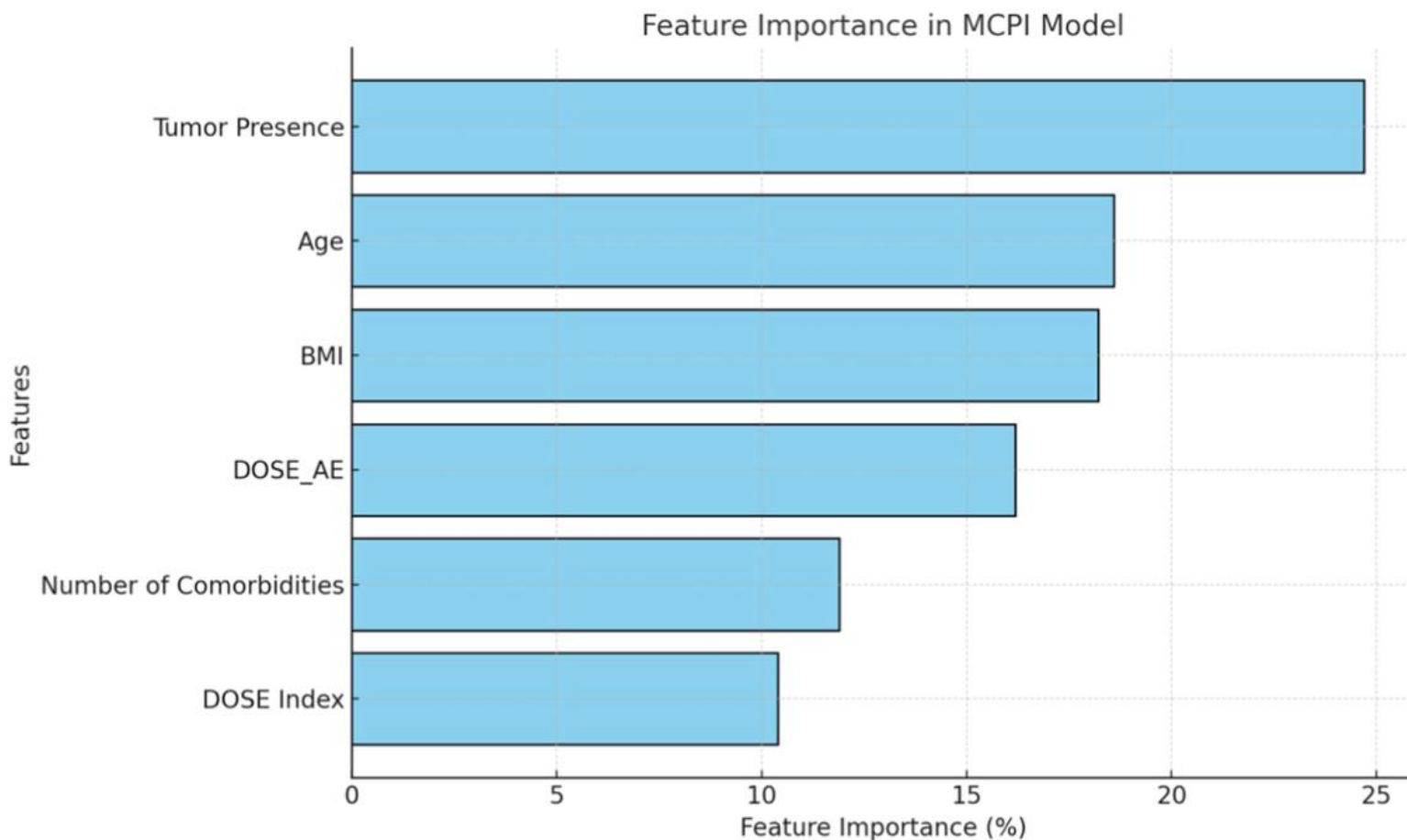
我剛畢業的碩士生的論文([連結](#))

- 396筆高雄長庚醫院胸腔科慢性阻塞性肺病病人資料。
- 資料內容：年紀、性別、BMI、是否吸煙、吸煙量、最大吸氣量、一秒最大呼氣量、GOLD指標、6分鐘步行距離、BODE指數、DOSE指數、ADO指數、共病數、急性惡化程度、惡性腫瘤數、總花費金額。
- 目的：建立一個演算法，可以不要用到所有的變數，而只要找出少數幾個重要的變數，希望能預測出總花費金額，愈準愈好。

資料準備

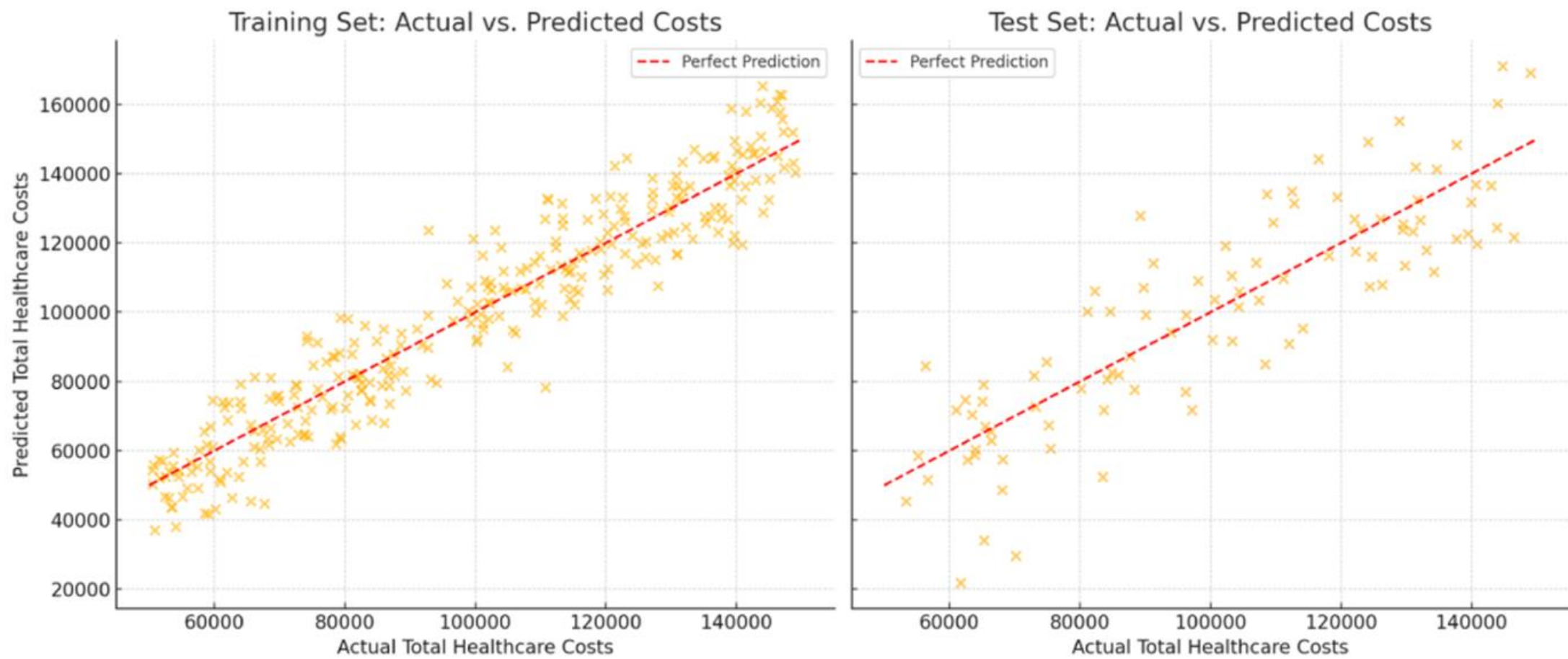
- 他們原本的總筆數是1063筆，是2015/1/31到2017/8/31在高雄長庚胸腔科就診的病人。
- 他們濾掉不滿40歲的病人、不符合慢性阻塞性肺病的病人、臨床資料不完整的病人。
- 總花費金額要花很多年的累積才能知道，所以他們2025年才寫出這篇論文。
- 他們用的方法都是別人已經發展好的，研究裏最珍貴的部份是資料，那很難獲得。
- 他們拿方法測資料，將結果報告出來，成為一篇期刊論文，而且很實用。
- 醫生和健保局可以用他們的模型，去預準每一個慢性阻塞性肺病，會花掉多少的治療費用。

實驗結果



- 最重要的變數：有無腫瘤、年紀、BMI，急性惡化次數、共病數，DOSE指數

視覺化預測結果

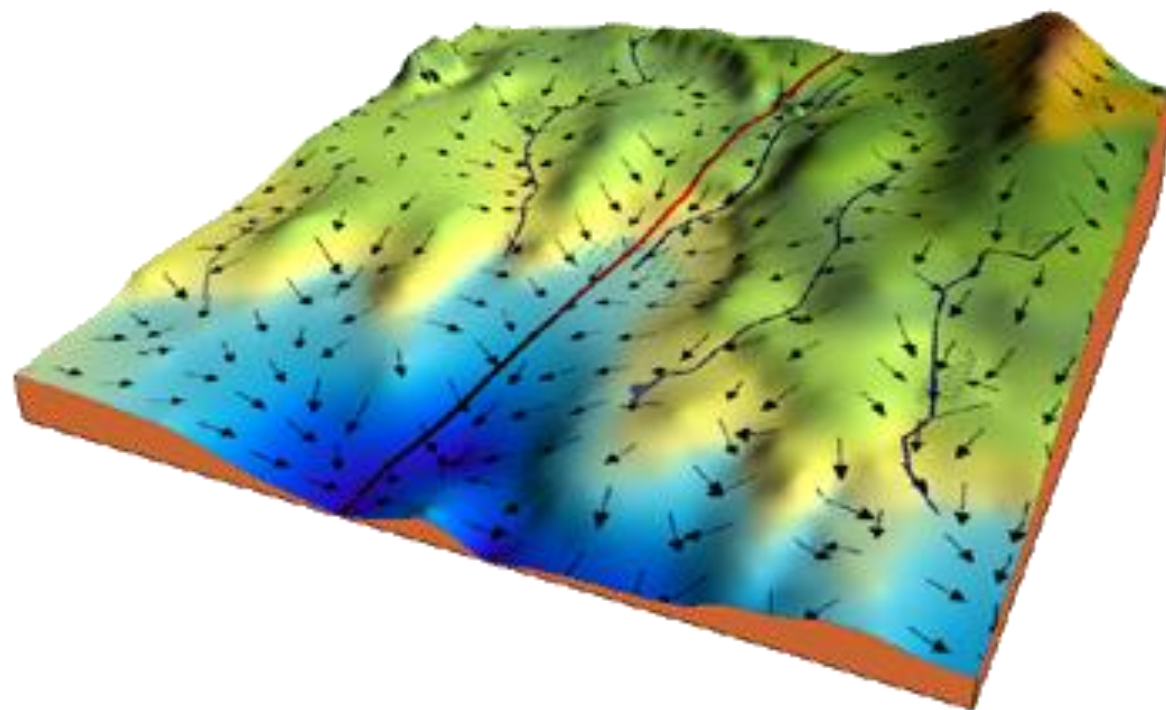


演算法

- 他們選的是Gradient Boosting，中文叫梯度提升。
- 這個演算法的想法是，先建第一棵決策樹，將318筆訓練資料(396的80%)跑一遍，一定會錯得很離譜
- 然後我們建第二棵決策樹，去預目標值和第一棵樹之間的殘差，這樣第一棵樹加第二棵樹，就比單獨只有第一棵樹準。
- 然後再訓練第三棵樹，去預測 $GT - (T1 + T2)$ 的值，這樣 $T1 + T2 + T3$ 又會比 $T1 + T2$ 更準一點。
- 每一次都比前一次好一點，所以叫作提升。

梯度提升

- 為什麼它叫作梯度呢？這其實是一個數學術語，指的是在多維函數時，函數值最下降最快的方向。
- 在這裏，我們希望資料集的總誤差值(將我們的演算法看到一個函數)，下降得愈快愈好。



梯度提升演算法的特性

- 很適合處理非線性資料，像剛才的病人資料集裏，有癌症和沒有癌症，就不是線性的。
- 其他非線性的資料，但不在剛才的病人資料集裏，像是上船的港口、房屋所在縣市、道路名。
- 它必需一棵訓練完才能訓練下一棵，沒辦法平行計算。
- 樹本身有也有些超參數，像是可以長到幾層、結點裏至少要有幾筆資料等等，總共要長幾棵樹、要去回補殘差的幾成，這些都要跑實測才能知道好壞。

CIFAR-10



圖片解析度 32*32 圖素

維度 $32*32*3=3072$ ，因為每個圖素有紅、綠、藍三個通道

圖片數量：全部60000張

類別數：10

每組圖片數量：6000張

特性

- 圖片很小，跟MNIST差不多。
- 資料量也和MNIST差不多。
- 但維度愈比較大，784 vs 3072，將近4倍
- 但是是自然影像，同類別間的變化很大

機器學習的任務

- 你要怎麼設計一支演算法，給你50000張這種小圖，而且知道他們的類別，這演算法要從中學一點判斷力，再給你10000張這種小圖，但不告訴你它們的類別，你的演算法要分辨它們？
- CIFAR是加拿大高等研究院 (Canadian Institute for Advanced Research) 的縮寫，這個資料集是他們建的。
- 這個資料集很有名，它在2009年發佈的，是類神經網路爆發的前三年，Hinton當時用這個小資料集作研究，要看怎麼用類神經網路解這個問題。三年後的AlexNet (Alex是Hinton的學生)，處理的是一樣的問題，只是圖片較大、類別較多。

CIFAR-100

- 大小不變
- 類別從10個變成100個
- 難度提高，更能測出各演算法的極限

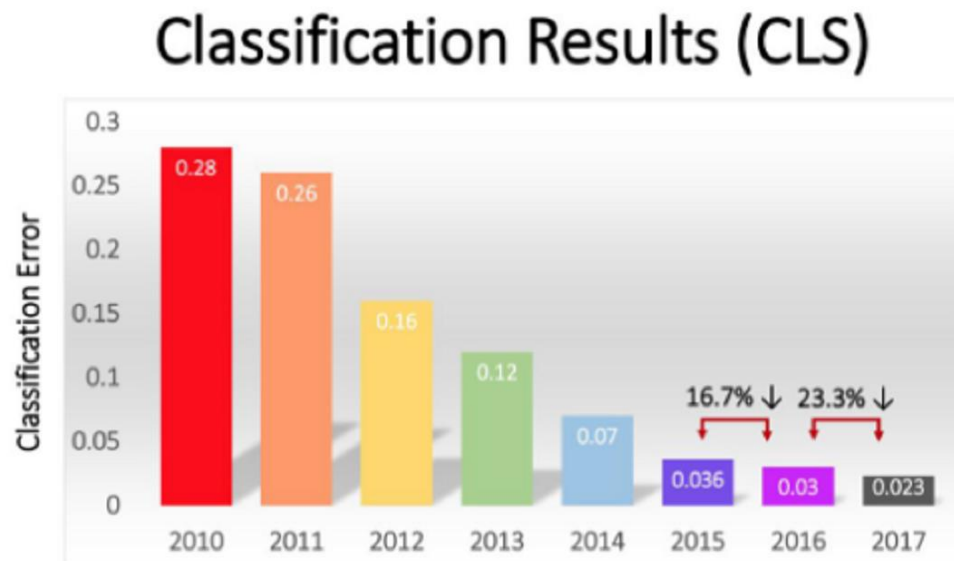


ILSVRC

- 全名是ImageNet Large Scale Visual Recognition Challenge (ImageNet大規模視覺辨識挑戰賽)
- 大小變成很大，而且不固定
- 種類變成1000種



歷史成績



- 2012年的大降，就是Hinton的AlexNet帶來的進步。
- 之後大家全用CNN了。
- 2017年是最後一年，這個挑戰已經被解完了，之後就不再辦下去了。

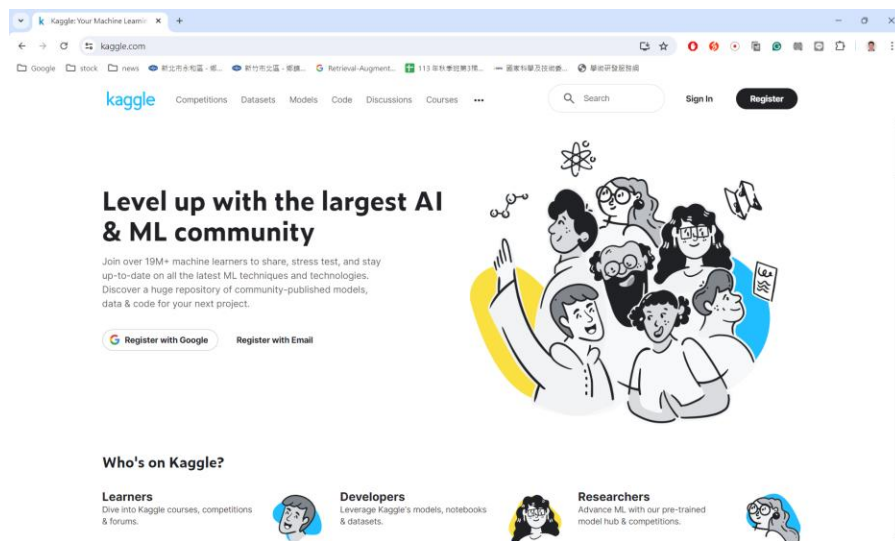
小回顧

- 機器學習是研究如何建立演算法，從資料中學習，作到預測、分類等等任務。
- 我們在這堂裏，學的是比較古典的演算法，當年發明這些演算法的人，他們的電腦比較慢，預設要處理的資料筆數較少小、維度也不太高，就像鳶尾花。
- 後來電腦愈來愈強大，資料也愈來愈多，深度學習效果很好，於是很多人投入研究，發明了各種各樣的方法，一年後你們會去學。

古典演算法的資料集哪裏找？

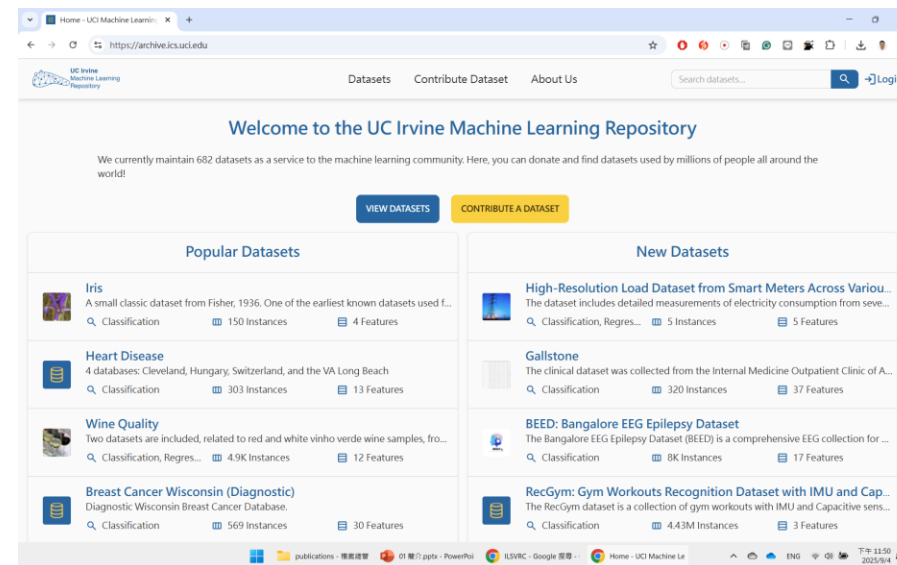
Kaggle

- 問題與任務的資料集分享平台
- 機器學習競賽



UC Irvine Machine Learning Repository

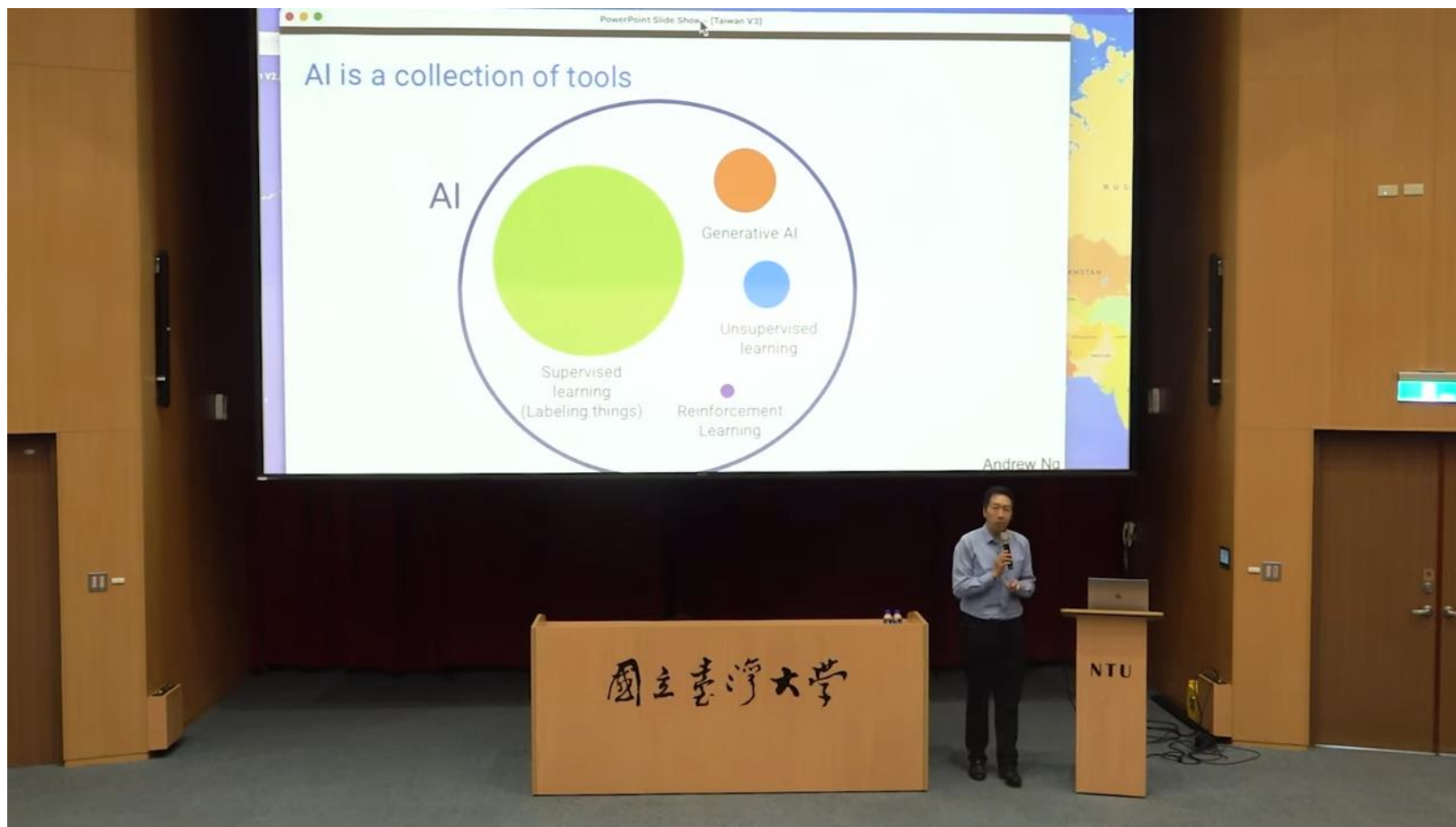
- 有682個已經整理好的資料集



機器學習的四大分類

- 監督式學習
 - 訓練資料都有標記，機器要學習從資料去預測標記，像剛剛的Iris、MNIST、CIFAR、ImageNet、肺病總金額，都屬這一類。
- 非監督式學習
 - 沒有標記，大多是非結構化的資料，機器要學習識別模式和資料的關聯性。例如說分群問題，如果你是個行銷經理，手上有幾萬筆的客戶資料，你要怎麼將這些客戶分成不同的類別，好制定各類的行銷策略。
- 半監督式學習
 - 部份資料有標記、部份沒有，通常是資料量太大，或標記極為昂貴時會發生。
- 強化式學習
 - 讓代理人在環境中學到作最佳決策的方法，用獎勵方式在訓練它。就像是在訓練寵物一樣，作對了，給一塊餅乾，作錯了，打一下。

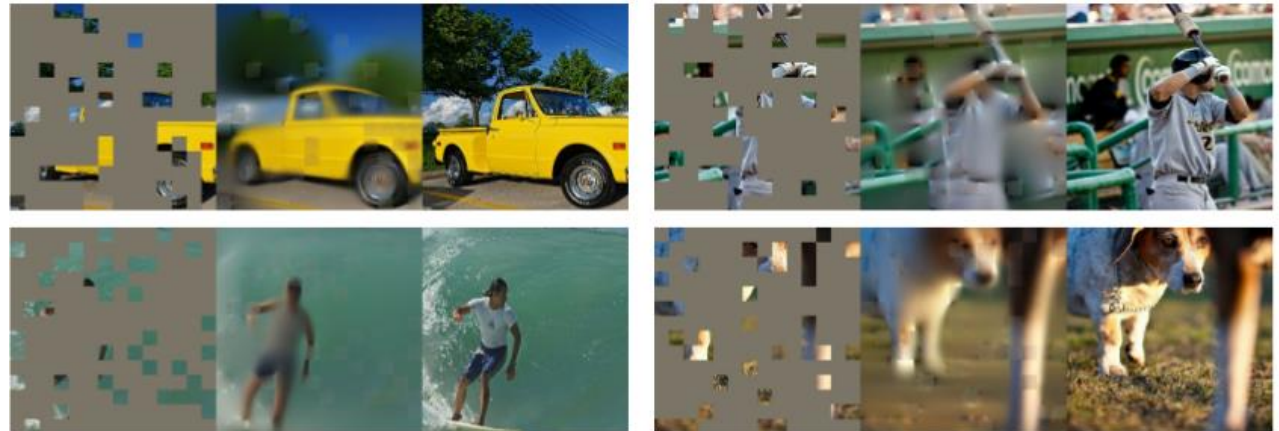
四大分類的應用廣度



擷取自YouTube
Opportunities in
AI (人工智慧的契
機) 4:10

其他機器學習的術語

- 自監督式學習：給定的資料是沒有標記的，但我們利用資料的結構來生出標記。
- 例如說我們把圖片的部份挖掉，用剩下的部份去預測失去的部份。

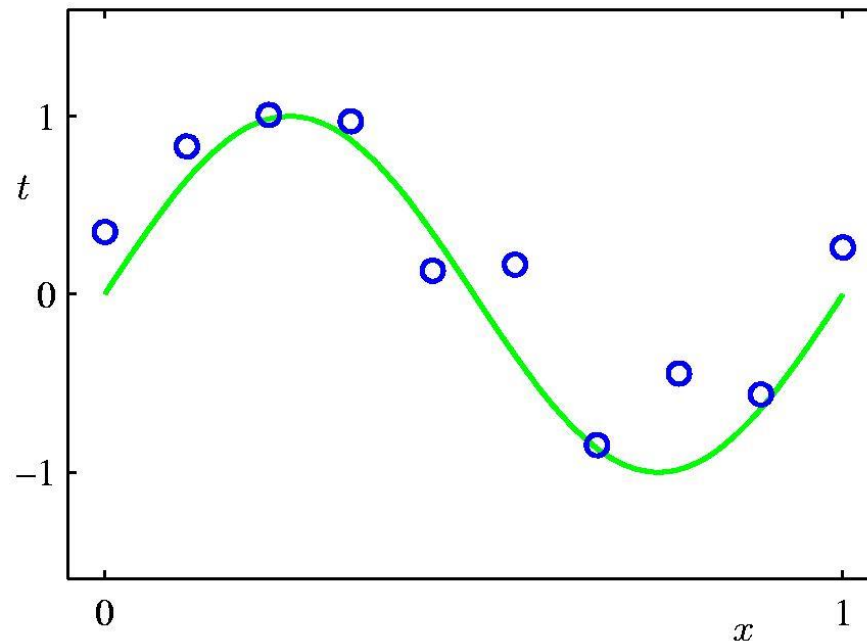


其他機器學習的術語

- 遷移學習：
- 機器學到的知識存在參數值(又稱權重)裏，我們在古典演算法裏的參數少，常常可以從零開始，用訓練資料去算出來。但到了深度學習的時候，參數多達千億個(例如說ChatGPT 3.5的參數量是1750億，LLama3是80億)，從零開始太耗力了，所以我們拿別人已經訓練好的參數來用，將知識遷移過來，再用我們自己準備的資料集去微調這些參數，可以省下很多的訓練成本。
- TAIDE 3.1就是微調LLama3-8B來的，讓這個LLM的術語、語氣、理解都更富有台灣知識。

多項式曲線擬合

- 課本上開場的演算法
- 資料維度：1維，符號 t
- 資料筆數：10筆
- 它們是 $\sin(2\pi x)$ 加上雜訊合成出來的數據
- 現在我們的演算法假設他們應該用一條一元多次多項式來模擬



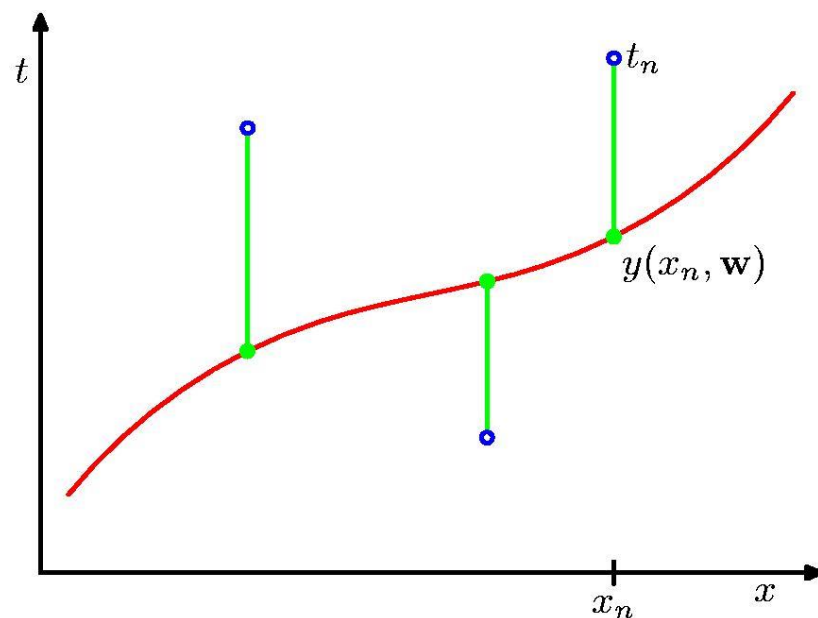
多項式的表示式

$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \dots + w_Mx^M = \sum_{j=0}^M w_jx^j \quad (1.1)$$

- 回想一下高中學過的一元二次和三次多項式
- $ax^2 + bx + c$
- $ax^3 + bx^2 + cx + d$
- (1.1)式是一般表示式，有M次，共M+1個參數，這M+1個參數就是我們要機器學習的知識。這M+1個參數的值，會由你的資料決定。而M該是多少，你必需事先決定，所以M稱為超參數，而w是參數，用w這個符號因為我們要決定他的權重(weight)。

誤差平方和函式

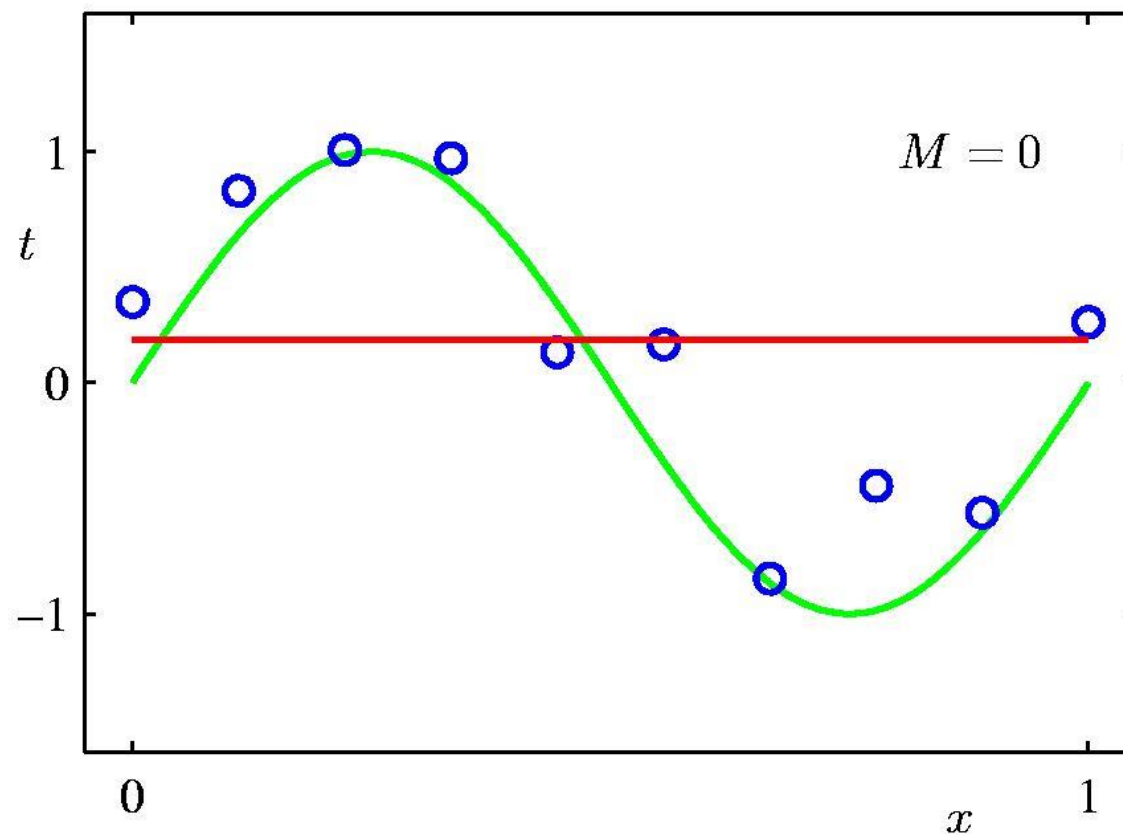
- 你有了一個假設的模型之後，如何決定權重？
- 一個很直覺的想法，就是讓錯誤最小。
- 所以我們要為錯誤下數學式的定義。
- 平方和是很常用的錯誤度量法。它平等對待每一筆資料，以偏離的平方計算錯誤，寧可筆筆小偏誤，也不要看到一筆大偏誤。



$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 \quad (1.2)$$

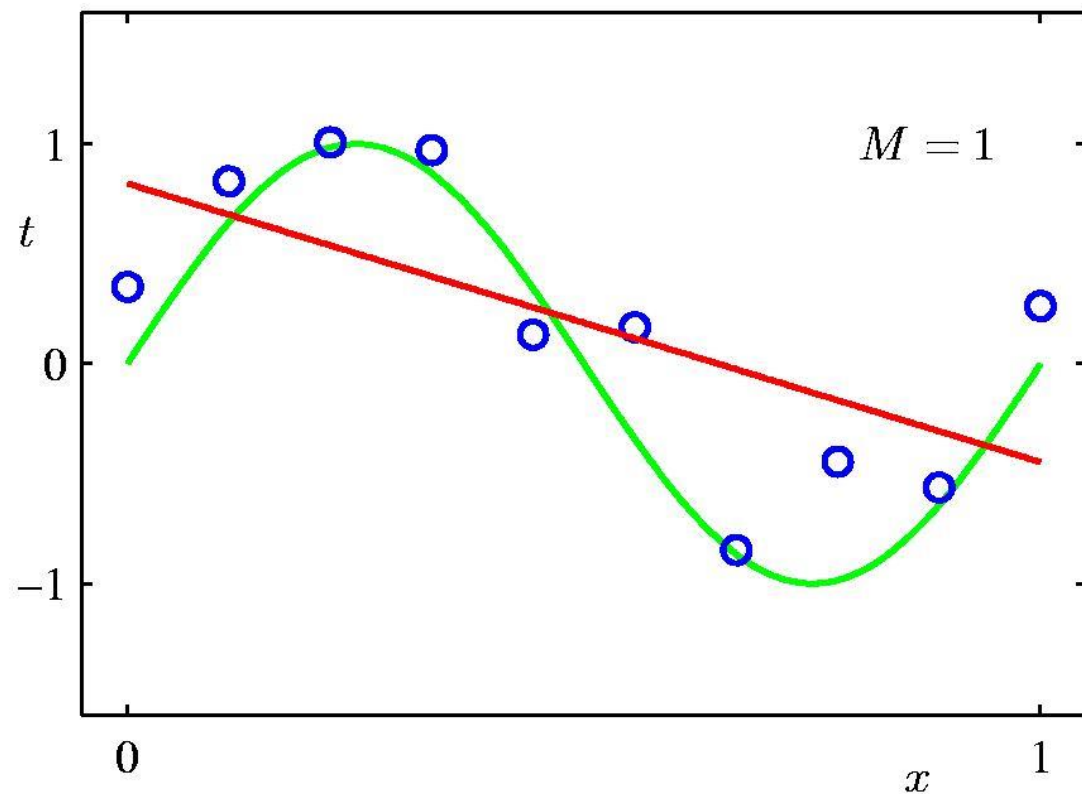
當 $M=0$ 時

- 我們只有一個參數，演算法產生的值是定值，在這裏例子，是0.2左右，會讓10個資料點的錯誤最小。



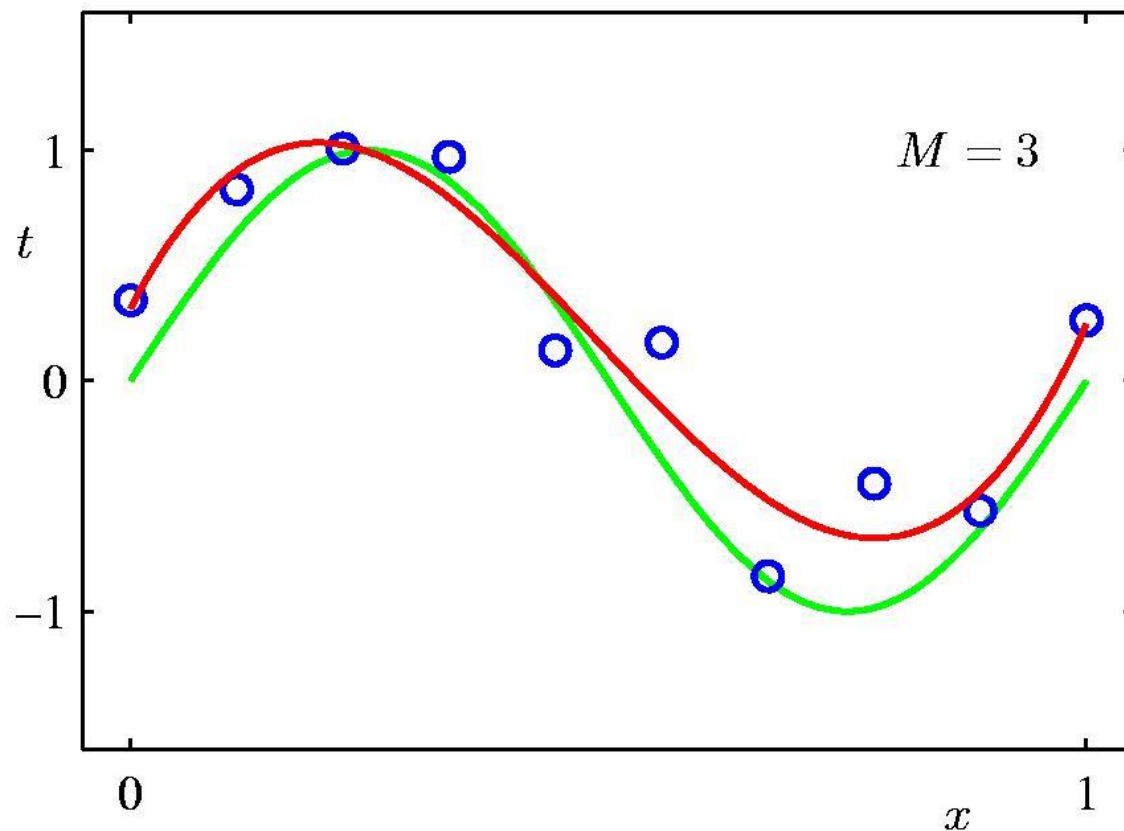
當 $M=1$ 時

- 一次多項式畫出來是一條可以有斜率的直線。
- w_0 大約是0.8， w_1 大約是-1.2，可以產出最小的 $E(w)$ 值。



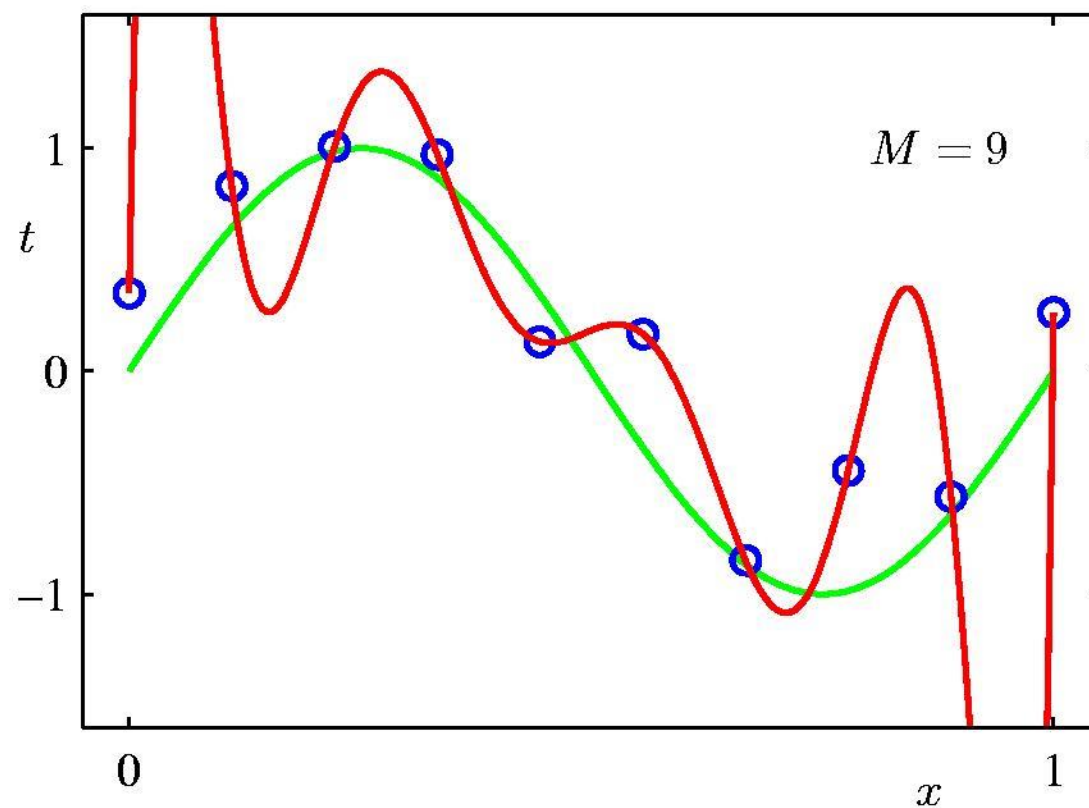
當 $M=3$ 時

- 看來和10個資料點貼合的不錯。
- 作者有把綠線畫出來，所以看起來三次是個好選擇，但在實際資料集時，我們並不知道資料點是由什麼模型產生的，我們只能假設我們演算法的模型，然後用資料集去測。



當 $M=9$ 時

- 因為你知道真正的資料是由綠線產生出來的，所以九次多項式離譜。
- 但九次多項式的錯誤值是0，因為每一個都穿過了，沒有任何一個點有偏離。



過度擬合

- 模型太過複雜，參數太多，造成為了最小化訓練資料的錯誤，權重太貼合訓練資料，而脫離了真實的來源模型。
- 而測試資料是從真實的來源模型來，但我們的演算法沒有看過，就會產生了很大的錯誤。

Root-Mean-Square 均方根

w^* 最佳參數組

$$E_{\text{RMS}} = \sqrt{2E(\mathbf{w}^*)/N} \quad (1.2)$$

