Machine Learning

Sophie Chen

Business Data Analytics 2021

Today's Agenda

- 課程特色|亮點
- 機器學習概述
  - 人工智能概述
  - 什麼是機器學習
  - 機器學習算法分類
  - 機器學習開發流程
  - 學習框架和資料介紹
- 安裝、實作

陳俞君

Introduction

Data Analysis

Data Visualization

陳九中

Classification

Clustering

Jason

Association Analysis

Text Mining

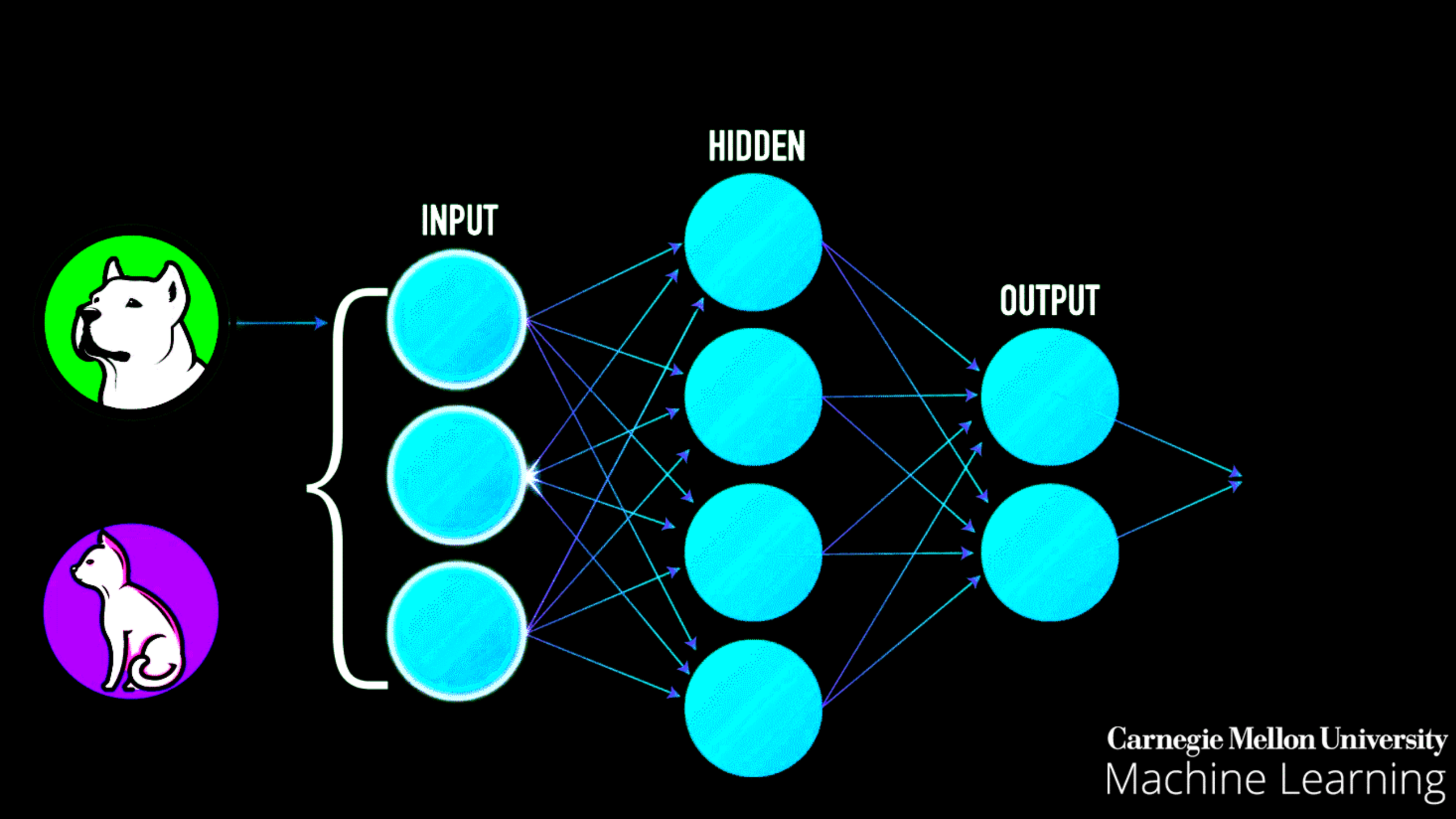
Deep Learning

Expectation & outcomes

10堂課快速入門
機器學習+深度學習

3步輕鬆
實作Scikit-Learn機器學習庫

課程結束產出1+完整專案

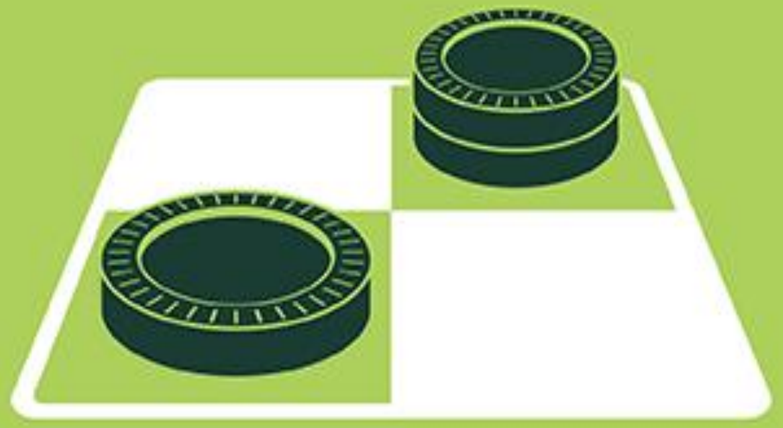

HIDDEN
INPUT
OUTPUT

Brief Introduction

人工智能演變，機器學習
與深度學習可以做什麼?
什麼是機器學習?

ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.



MACHINE LEARNING

Machine learning begins to flourish.

DEEP LEARNING

Deep learning breakthroughs drive AI boom.

1950's 1960's 1970's 1980's 1990's 2000's 2010's

應用場景

傳統預測:
店鋪銷量預測、量化投資、廣告推薦..

圖像辨識:
車牌辨識、人臉辨識..

自然語言處理:
情感分析、Siri、聊天機器人..

Data

Training

New Data → Input → 模型 → Predict → ?

經驗

歸納

問題 → Input → 規律 → Predict → 未來

機器學習是從數據中自動分析獲得模型，並利用模型對未知數據進行預測。

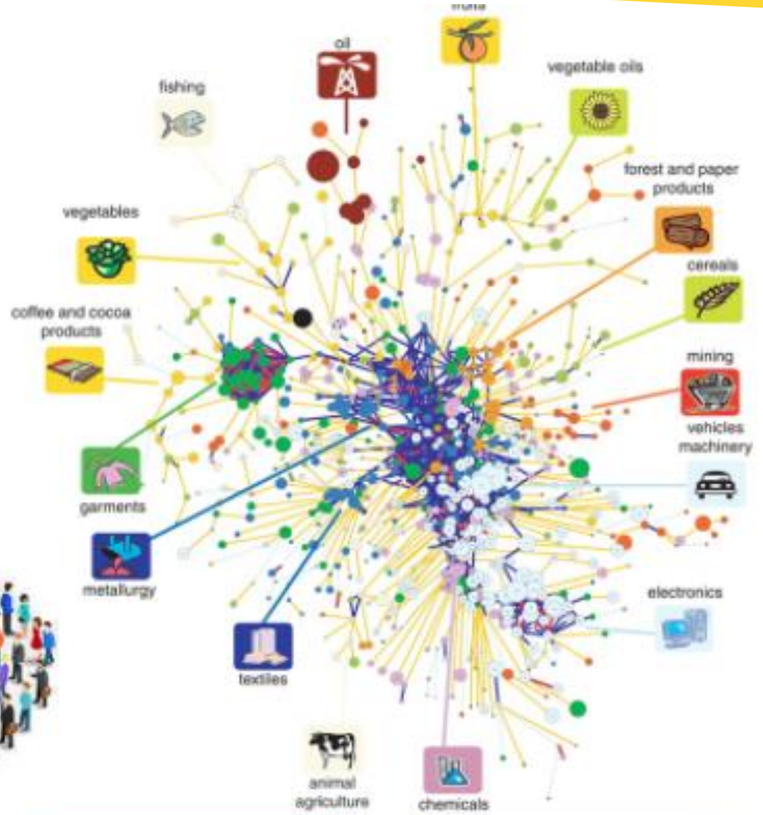
貓? 狗?

結構: 特徵值+目標值

房屋價格?

	坪數	公設比	目標值
房子一	20	40	100
房子二	60	10	120
房子三	50	30	200

機器學習算法分類



回歸 (Regression)

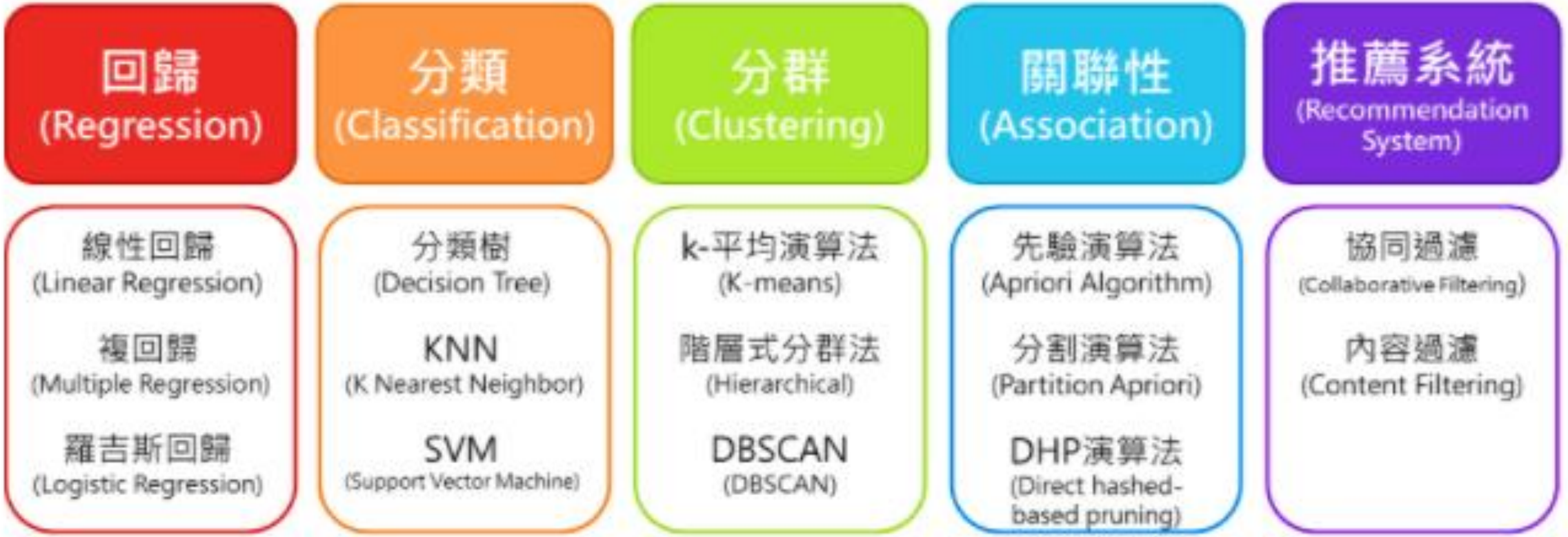
分類 (Classification)

分群 (Clustering)

關聯性 (Association)

推薦系統 (Recommendation System)

機器學習算法分類


回歸
(Regression)
分類
(Classification)
分群
(Clustering)
關聯性
(Association)
推薦系統
(Recommendation System)
線性回歸
(Linear Regression)
複回歸
(Multiple Regression)
羅吉斯回歸
(Logistic Regression)
分類樹
(Decision Tree)
KNN
(K Nearest Neighbor)
SVM
(Support Vector Machine)
k-平均演算法
(K-means)
階層式分群法
(Hierarchical)
DBSCAN
(DBSCAN)
先驗演算法
(Apriori Algorithm)
分割演算法
(Partition Apriori)
DHP演算法
(Direct hashed-based pruning)
協同過濾
(Collaborative Filtering)
內容過濾
(Content Filtering)

機器學習開發流程

Kaggle

- 大數據競賽平台
- 80萬科學家
- 真實數據
- 數據量巨大

UCI

- 收錄360個數據集
- 覆蓋科學、生活、經濟等領域
- 數據量幾十萬

scikit-learn

- 數據量較小
- 學習方便

Try and Learn

Activity Time

- Anaconda Installation
- Jupyter notebook
- Scikit-Learn(load_Iris)

Feature Engineering特徵工程

為什麼需要特徵工程?
機器大神吳恩達老師說: 數據和特徵決定了機器學習的上限，而模型和算法只是逼近這個上限而已。

什麼是特徵工程?
特徵工程是使用專業背景知識和技巧處理數據，使得特徵能在機器學習算法上發揮更好的作用的過程。

特徵工程包含:
- 特徵抽取
- 特徵預處理
- 特徵降維

目標:
- 了解數值型數據、類別型數據
- 應用MinMaxScaler實現對特徵數據進行歸一化
- 應用StandardScaler實現對特徵數據進行標準化

特徵預處理

特徵一	特徵二	特徵三	特徵四
90	2	10	40
60	4	15	45
75	3	13	46

特徵一	特徵二	特徵三	特徵四
1.	0.	0.	0.
0.	1.	1.	0.83
0.5	0.5	0.6	1.

什麼是特徵預處理?
特徵預處理就是通過一些轉化函數將特徵數據轉換成更加適合算法模型的特徵數據過程(數學上稱為dimensionless)。

數值型數據的dimensionless: 歸一化/標準化

為何需要進行歸一化/標準化?
特徵的**單位或者大小相差較大，或是某特徵的方差相比其他的特徵要大出幾個數量級，容易影響(支配)目標結果**，使得一些算法無法學習到其他特徵。

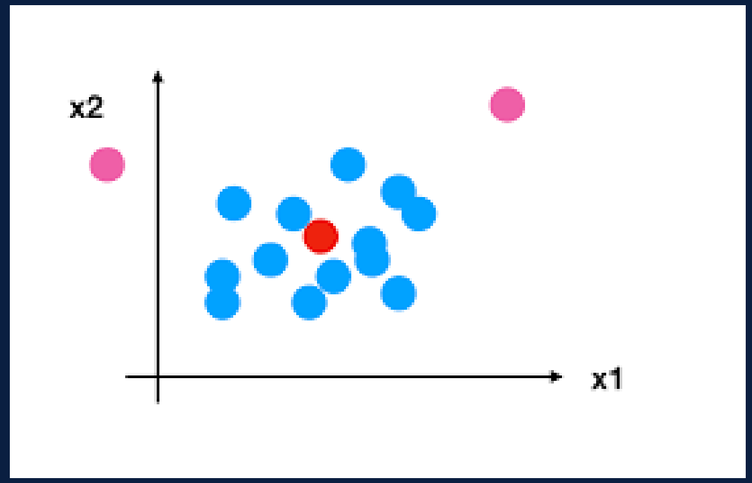
歸一化/標準化

定義:

- 歸一化: 通過對原始數據進行變換把數據映射到[0,1]之間。
- 標準化: 通過對原始數據進行變換到均值為0，標準差為1範圍內。

公式:

- 歸一化: $x' = \frac{x - \min}{\max - min}$ $X'' = X' * (Mx - m_i) + m_i$
- 標準化: $x' = \frac{x - mean}{\sigma}$

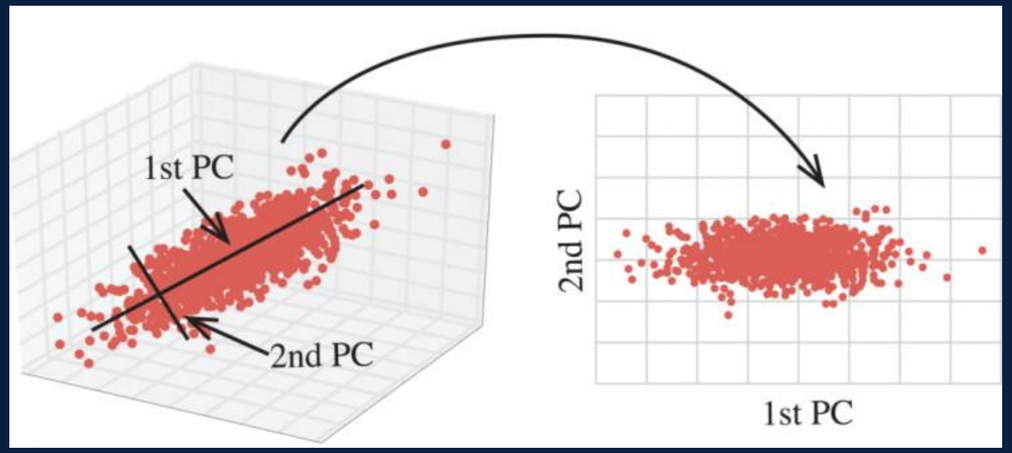


結論:

- 對歸一化來說，最大值與最小值是變化的，因此容易受異常點影響，所以這種方法的穩健性較差，只適合傳統精確小數據的場景。
- 對標準化來說，如果出現異常點，由於具有一定數據量，少量異常點對於平均值的影響並不大，因此方差改變較小。

降維

降維是指在某些限定條件下，**降低隨機變量(特徵)個數，得到一組”不相關”主變量**的過程。



相關特徵(correlation feature): 相對溼度與降雨量之間的相關…等等

正是因為在進行訓練的時候，我們都是使用特徵進行學習，如果**特徵本身存在問題或者特徵之間相關性較強，對於算法學習預測會影響較大**。

降維方式:

- 特徵選擇(主要講解**低方差特徵過濾式、相關係數過濾式**)->sklearn.feature_selection.VarianceThreshold
- 主成分分析(可以理解是一種特徵提取的方式)

Thank you for joining today's class.

TA: Sophie俞君
Email: 108421045@cc.ncu.edu.tw