

機器學習

1. 機器學習是甚麼：

1-1 聚類

1-2 回歸

1-3 分類

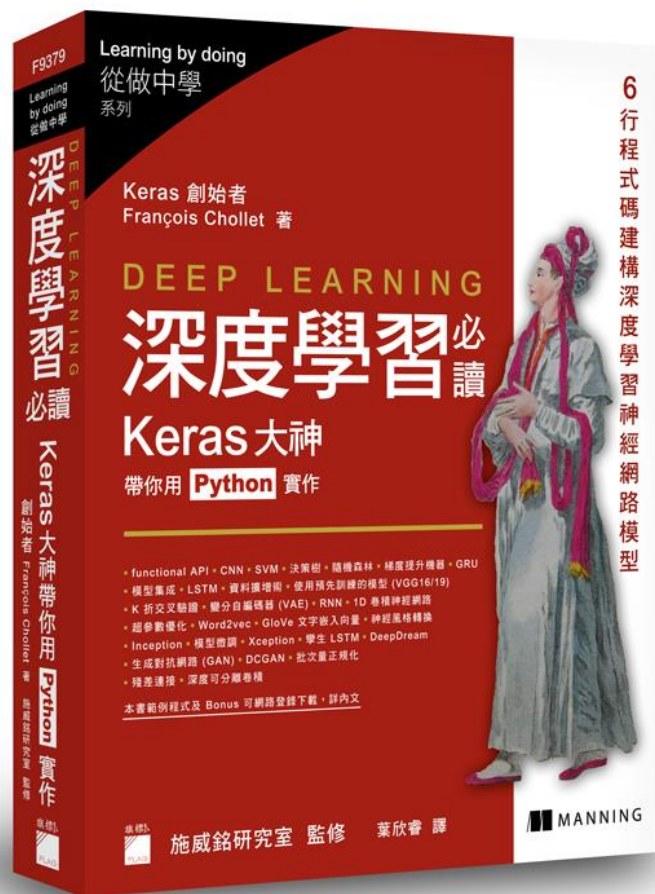
1-4 歸一化

1-5 L1 regularization

1-6 L2 regularization (權重衰減)

1-7 Dropout

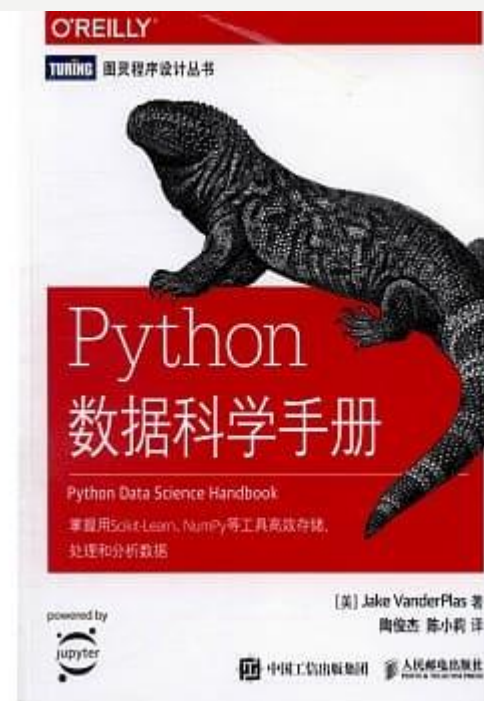
參考書目



<https://www.flag.com.tw/books/product/F9379>



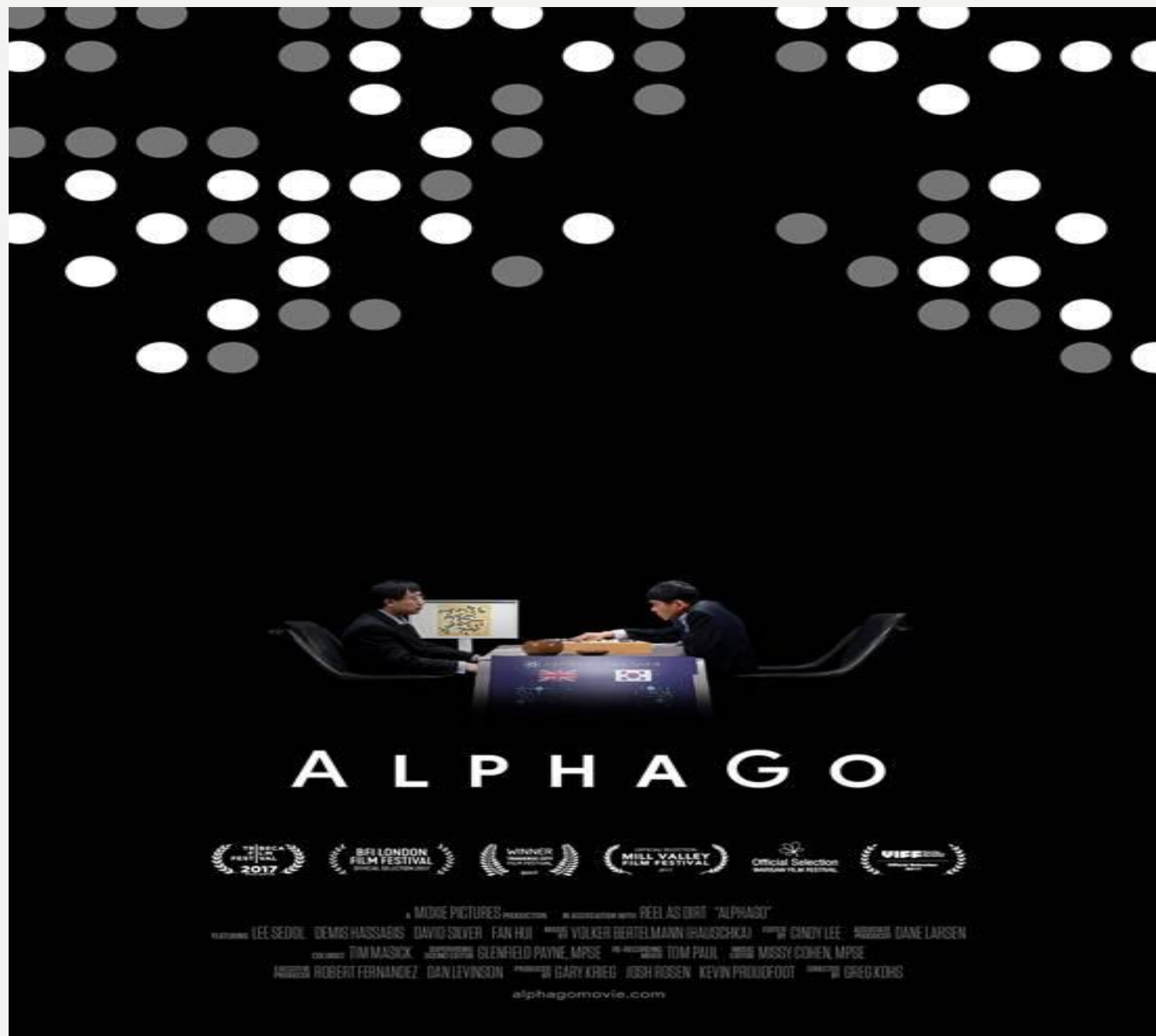
<https://www.books.com.tw/products/0010826415>



<https://www.books.com.tw/products/CN11517291>

課程參考連結

<https://github.com/jumbokh/intro-computers>



人工智慧

- 2016年3月15日下午,舉世矚目的圍棋人機大戰在圍棋世界冠軍李世石與人工智慧博弈軟體AlphaGo之間展開第五局對戰。
- 人工智慧的概念是由以麥卡賽、明斯基、羅切斯特和香農等為首的一批科學家在1956年提出的。
- GPU(影像處理器)的高速發展,使得大規模深度學習成為可能。
- 人工智慧應用場景中,有無人駕駛汽車穿行於車水馬龍,有智慧型機器人探索宇宙太空,有面部識別系統精確定位、尋蹤於茫茫人海,還有AlphaGo智慧博弈軟體與人類圍棋世界冠軍激戰

Machine Learning ≈ Looking for a Function

- Speech Recognition

$$f(\text{audio waveform}) = \text{"How are you"}$$

- Image Recognition

$$f(\text{cat image}) = \text{"Cat"}$$

- Playing Go

$$f(\text{Go board state}) = \text{"5-5" (next move)}$$

- Dialogue System

$$f(\text{"Hi" (what the user said)}) = \text{"Hello" (system response)}$$

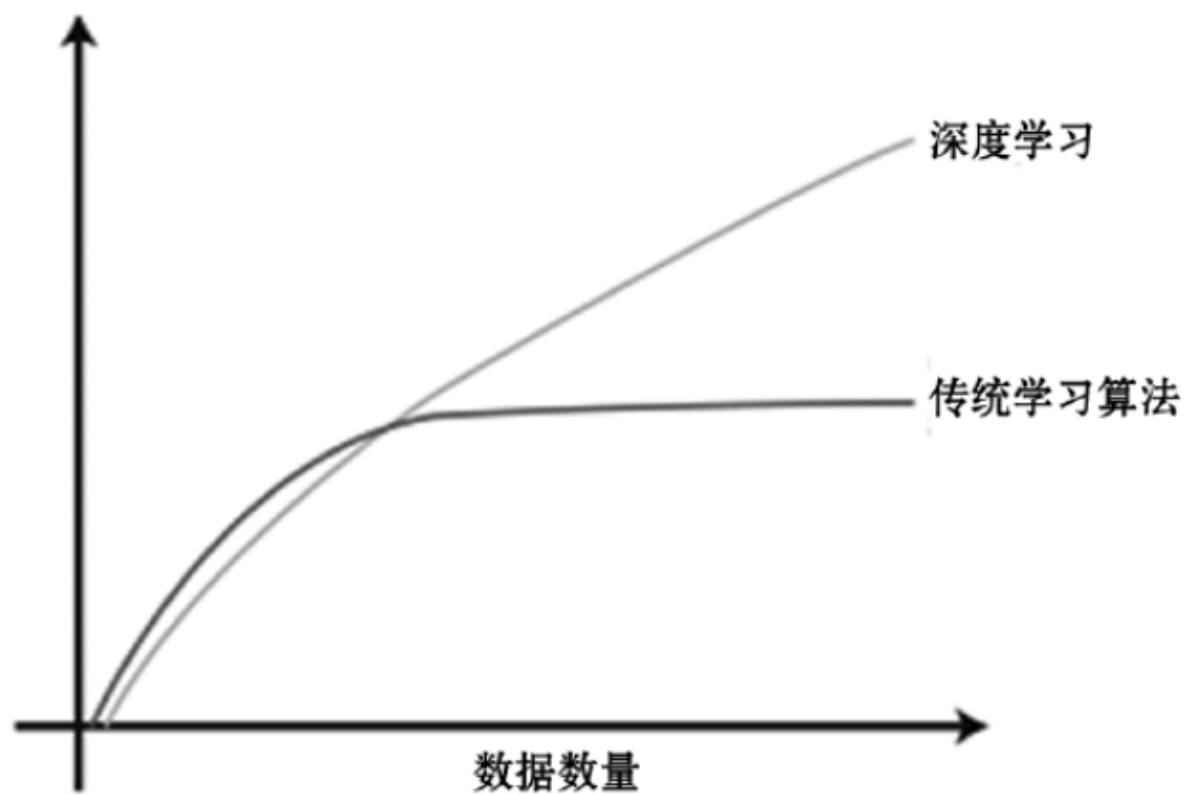


图 2-1 基于神经网络的深度学习算法和传统学习算法效果图

機器學習四大類

- A. 監督學習 (Supervised Learning)
- B. 非監督學習 (Unsupervised Learning)
- C. 半監督式學習 (Semi-supervised learning)
- D. 強化式學習 (Reinforcement learning)

聚類(CLUSTERING)

聚類分析(**clustering analysis**)是依據資料相似度或相異度而將資料分群歸屬到數個聚類(**clusters**)的方法;使得同一群內的資料或個體相似程度大,而各群之間的相似程度小。同一組樣本有時會因為不同的目的、資料登錄方式、所選的分群特徵或資料屬性,形成不同的分群結果。

分類**(CLASSIFICATION)**

分類(classification)則是根據已知或所給定目標資料的類別,找出其分類屬性,建立分類規則或模式,將資料分類至所對應的目標類別。

聚類分析

聚類分析是分群以找出各子聚類資料背後可能隱藏的特徵、樣型或關聯現象。聚類分析事先並不知道聚類數目,而分群結果的特徵及其所代表的意義僅能事後加以解釋。因此,聚類分析可視為無監督式學習

聚類分析應用的領域

- 根據顧客基本資料和交易資料將顧客分群,定義並分析不同類型顧客的消費行為模式,以設計定制化的行銷方案;或是通過聚類分析將信用卡使用行為分為不同群組樣型,以分析信用卡異常消費的情形,避免盜刷所造成的損失。
- 在製造業,可依據機台的特徵、功能等的相似程度,將機台分為可以相互替代和備援(backup)的聚類,以提升作業效率並維持良率(Chien&Hsu,2006)。
- 在網路行銷中,可將性質或特性相仿的網頁予以分類,增快網頁搜索速度,並根據流覽行為和客戶聚類分析作客戶消費行為預測和搭配行銷。

聚類分析的階段

聚類分析主要包括以下四個階段:

- (1)**資料準備與分群特徵選取**:根據問題特性、資料類型及所選擇的分群演算法等,自搜集的變數中選取具代表性的變數作為分群特徵屬性。
- (2)**相似度計算**:選擇衡量相似度的方式,如距離、相關係數等。在選擇衡量相似度的方式時,需考慮資料的類型以及後續使用的分群演算法,例如,在類別尺度中,選用歐氏距離可能會造成資料尺度的誤用。
- (3)**分群演算法**:為整個聚類分析中最重要的階段,主要為利用分群演算法將資料分組,有些分群演算法可能需要自行決定群數,例如,劃分聚類分析演算法可由使用者自行決定或利用其他方式決定適當的分群個數。
- (4)**分群結果評估與解釋**:當分群結束後需檢查分群結果是否合理。例如,聚類間的距離是否過大、該資料是否適用所選用的分群演算法,若發現有不合理的地方,則需重新審視前三個階段是否有問題。另外,由於分群後的結果可能作為另一個方法的輸入資料,因此可能需要對聚類結果進行定義或命名。

歐氏距離與曼哈頓距離

$$D_{(y_1, y_2)} = \sqrt{\sum_{j=1}^P (x_{1j} - x_{2j})^2}$$

$$D_{(y_1, y_2)} = \sum_{j=1}^P |x_{1j} - x_{2j}|$$

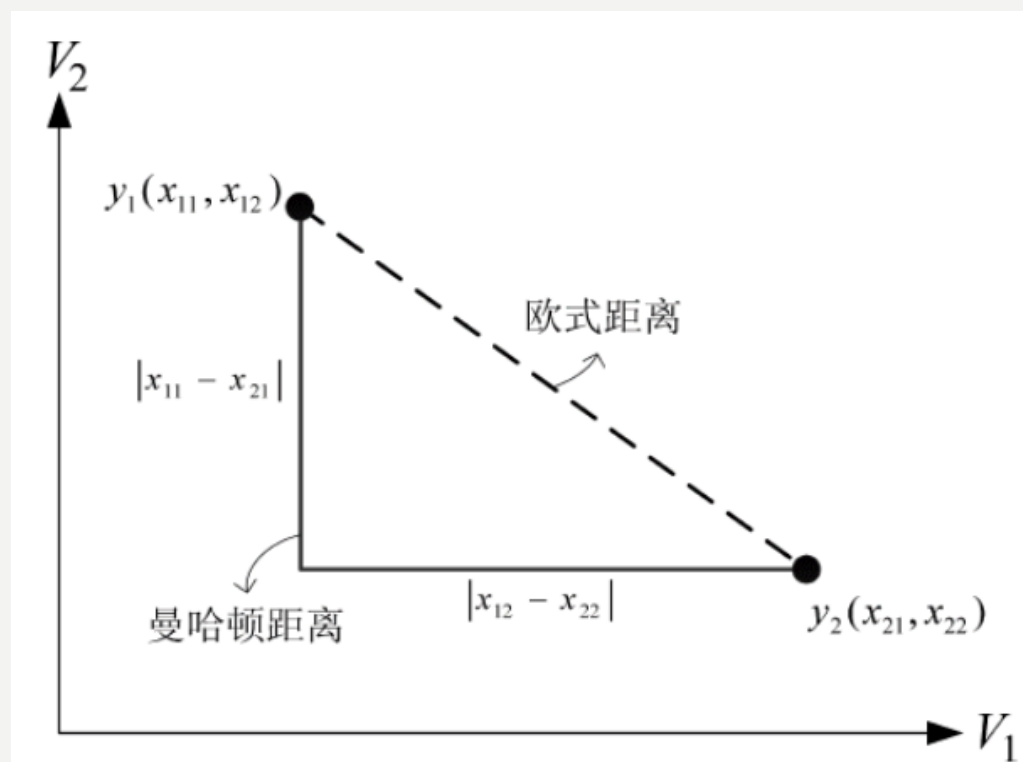


图 6.2 欧式距离与曼哈顿距离示意图

皮爾遜相關係數

PEARSON CORRELATION COEFFICIENT

- 定義：兩個變數之間的皮爾遜相關係數定義為兩個變數之間的共變異數和標準差的商

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

相關性

相關係數與單位無關;且相關係數介於-1到+1之間。

當 $r(V1, V2) > 0$

表示 $V1$ 增加時, $V2$ 也增加;

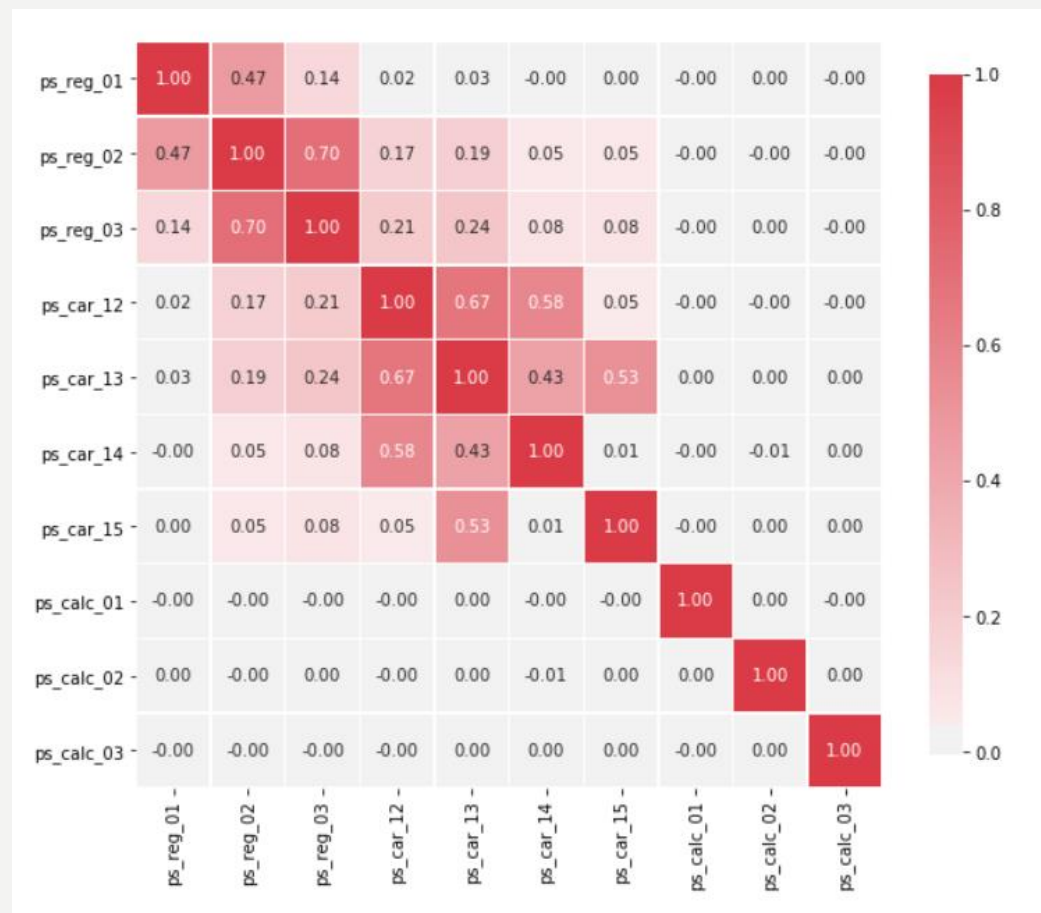
$r(V1, V2) < 0$ 表示 $V1$ 增加時, $V2$ 則減少。

$0 \leq |r(V1, V2)| \leq 0.3$ 表示兩變數為低相關性,

$0.3 \leq |r(V1, V2)| \leq 0.7$ 表示兩變數為中相關性,

$0.7 \leq |r(V1, V2)| \leq 1$ 表示兩變數為高相關性。

熱力圖



AI CUP - 愛文芒果影像辨識競賽



AI CUP 2020

愛文芒果影像辨識雙項競賽

總獎金高達 60 萬！

2 月 3 號，開始報名！

監督式學習

- 監督式學習的主要任務是根據物件的**特徵**來預測其**標籤**。相應的學習演算法需要對經驗進行學習。因此,訓練資料是監督式學習演算法的一個重要組成部分。訓練資料是隨機抽取觀測物件採集到的資料。這些資料簡稱為採樣,也稱為樣本。**每一條訓練資料都含有特徵與標籤**。例如,房價預測問題的訓練資料是過去一年的房屋成交記錄,它含有各種房屋的特徵以及售價。通過對訓練資料的學習,演算法能夠訓練出一個模型來預測標籤,並且根據給定的度量方法來檢驗模型預測的效果。
- 為了正式定義監督式學習,需要介紹以下幾個基本要素:**特徵、標籤、分佈、模型與損失函數**。下面逐一給出它們的正式定義。


回歸演算法

回歸方法是對數值型連續隨機變數進行預測和建模的監督學習演算法。其特點是標注的資料集具有數值型的目標變數。回歸的目的是預測數值型的目標值。最直接的辦法是依據輸入寫出一個目標值的計算公式，該公式就是所謂的回歸方程（ regression equation ）。求回歸方程中的回歸係數的過程就是回歸。

- 迴歸 (regression) 方法是一個分析變數和變數之間關係的工具
- 主要在探討自變數(x)與依變數(y)之間的線性關係
- 透過迴歸模型的建立，可以推論和預測研究者感興趣的變數(y)

Regression: Output a scalar

- Stock Market Forecast

$$f(\text{ ) = \text{Dow Jones Industrial Average at tomorrow}$$

- Self-driving Car

$$f(\text{ ) = \text{方向盤角度}$$

- Recommendation

$$f(\text{ 使用者 A } \quad \text{商品 B }) = \text{購買可能性}$$

前提假設

假設模型

- 估計式為: $Y=B_0+B_1X_1$

誤差項需滿足三大假設:

(1)**常態性(Normality)**: 若母體資料呈現常態分配(Normal Distribution)，則誤差項也會呈現同樣的分配。可採用常態機率圖(normal probability plot) 或 Shapiro-Wilk常態性檢定做檢查。

(2)**獨立性(Independency)**: 誤差項之間應該要相互獨立，否則在估計迴歸參數時會降低統計的檢定力。我們可以藉由Durbin-Watson test來檢查。

(3)**變異數同質性(Constant Variance)**: 變異數若不相等會導致自變數無法有效估計依變數。我們可以藉由殘差圖(Residual Plot)來檢查。

常用的回歸方法

- 線性回歸：使用超平面擬合資料集
- 最近鄰演算法：通過搜尋最相似的訓練樣本來預測新樣本的值
- 決策樹和回歸樹：將資料集分割為不同分支而實現分層學習
- 集成方法：組合多個弱學習演算法構造一種強學習演算法，
如隨機森林（**RF**）和梯度提升樹（**GBM**）等
- 深度學習：使用多層神經網路學習複雜模型

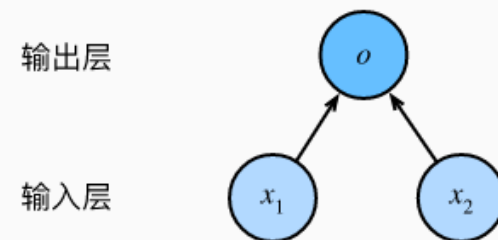


图 3.1 线性回归是一个单层神经网络

如何應用

- **收集資料**：可以使用任何方法。
- **準備數據**：回歸需要數值型資料，標稱型資料將被轉換成二值型資料。
- **分析資料**：繪出資料的視覺化二維圖將有助於對資料做出理解和分析，在採用縮減法求得新回歸係數之後，可以將新擬合線繪在圖上作為對比。
- **訓練演算法**：找到回歸係數。
- **測試演算法**：使用 R^2 或者預測值和資料的擬合度，來分析模型的效果。
- **使用演算法**：使用回歸，可以在給定輸入的時候預測出一個數值，這是對分類方法的提升，因為這樣可以預測連續型資料而不僅僅是離散的類別標籤。

優缺點

- 優點：結果容易理解，計算上不複雜。
- 缺點：對非線性的資料擬合不好。
- 適用資料範圍：數值型和標稱型。

在許多實際問題中,物件的特徵組與其標籤之間存在一定的關係。例如,在房價預測問題中,一個地區的房價與該地區的地理位置、人口數、居民收入等諸多特徵有著密切的關係。在監督式學習中,這種關係就稱為回歸關係。

如果特徵與標籤之間的關係是近似線性的,就可以用一個線性模型來擬合這種回歸關係。用線性模型來擬合特徵組與標籤之間回歸關係的方法就稱為線性回歸。


● 線性回歸

Example Application

- Estimating the Combat Power (CP) of a pokemon after evolution

$$f(x) = y$$

The input vector x is represented by the following Pokemon data:



Attribute	Value	Variable Label
CP	14	x_{cp}
Species	Bulbasaur	x_s
HP	10 / 10	x_{hp}
Type	Grass / Poison	
Weight	11.62 kg	x_w
Height	0.88 m	x_h

The output y is the CP after evolution.

分類問題

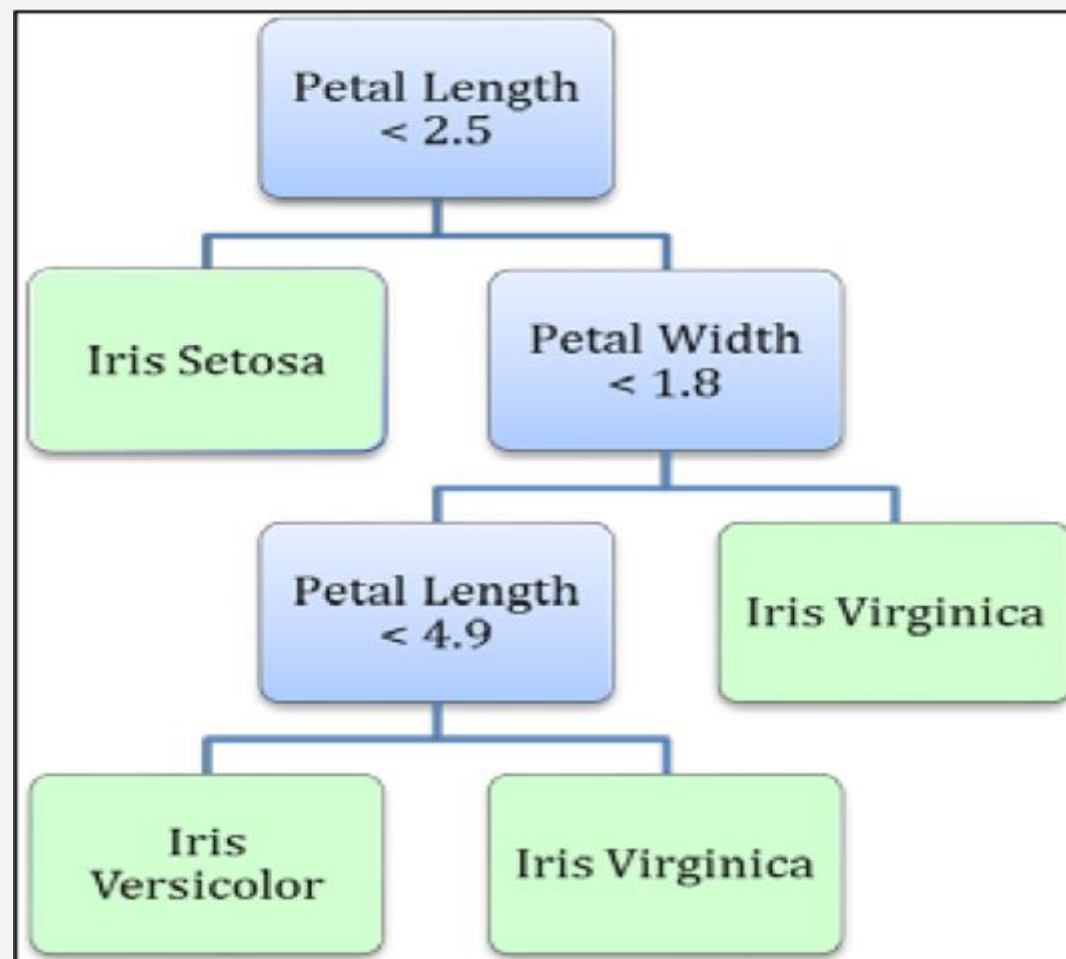
6. 假设一分类问题,输入变量为 I_1 与 I_2 ,输出变量为 O ,对应的数据如下表:

I_1	I_2	O
-1	1	1
0	0	0
1	-1	0
1	0	1
0	1	1
1	1	0

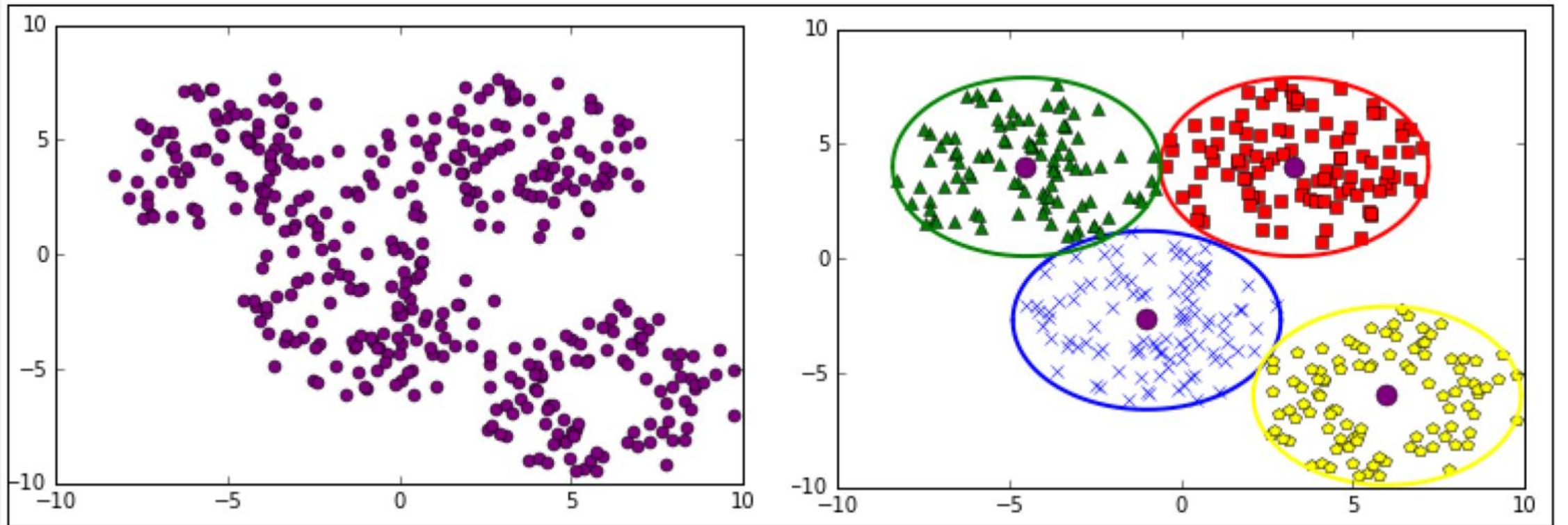
- (1) 请画出上表中的网络图,并给定相关的初始参数值。
- (2) 请利用反向传播人工神经网络说明一次学习的过程。

決策樹

鳶尾花分類問題



K-MEANS



實現**K**近鄰算法

- 數據處理：從**CSV**文件導入數據集並分割成測試/訓練數據集。
- 相似度：計算兩個數據實例之間的距離。
- 近鄰：找到**k**個最相似的數據實例。
- 響應：從一組數據實例生成響應。
- 準確性：總結預測的準確性。
- 集成：把算法各部分集成在一起。

* 原文網址：<https://kknews.cc/code/3elmvxy.html>

維基百科上的鳶尾花數據集

http://en.wikipedia.org/wiki/Iris_flower_data_set

```
from sklearn.datasets import load_iris

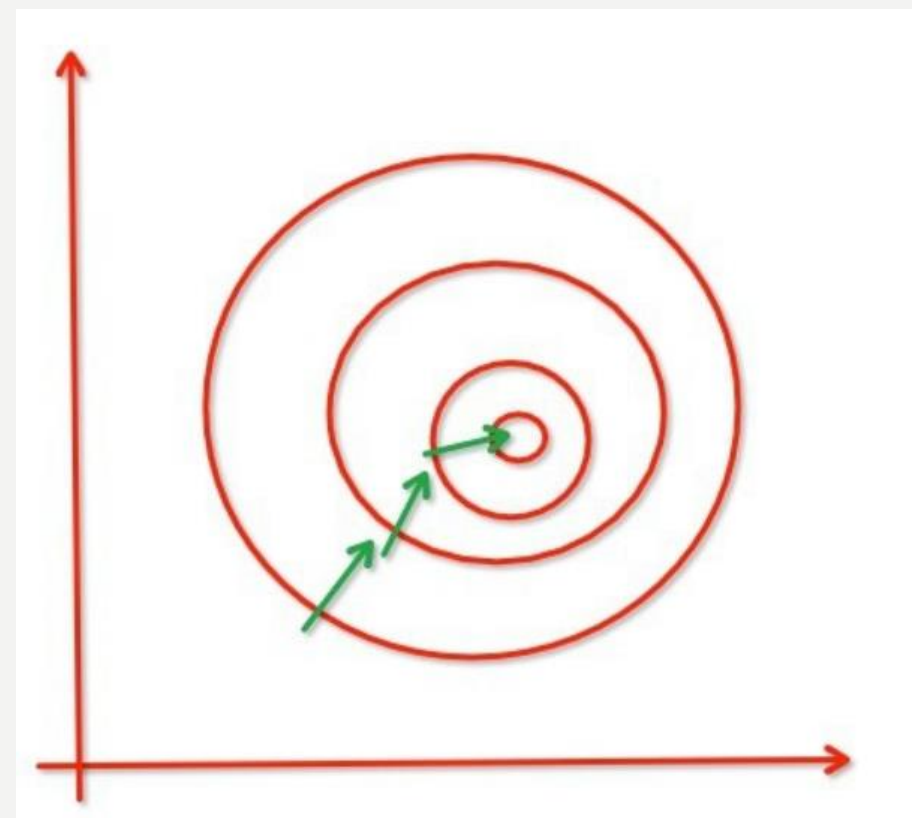
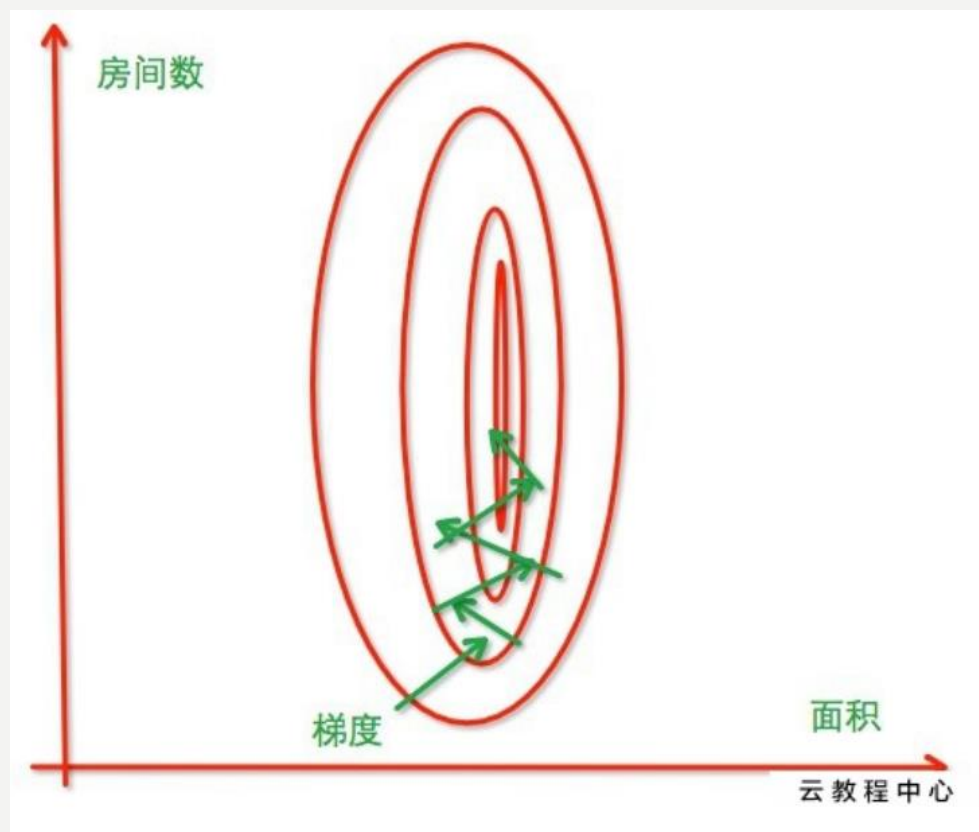
iris = load_iris()
iris
```

This code gives:

```
{'data': array([[5.1, 3.5, 1.4, 0.2],
                [4.9, 3. , 1.4, 0.2],
                [4.7, 3.2, 1.3, 0.2],
                [4.6, 3.1, 1.5, 0.2],
```

歸一化

經過歸一化,把各個特徵的尺度控制在相同的範圍內



歸一化

當度量單位在屬性之間不同時，某種屬性可能在對距離度量的貢獻中占主導地位。對於這些類型的問題，在計算相似性之前，您需要將所有數據屬性重新縮放到0-1範圍內（稱為歸一化）。

PYTHON程式碼實現

1.min-max標準化(Min-Max Normalization)

也稱為離差標準化,是對原始資料的線性變換,使結果值對映到[0 - 1]之間。

```
import numpy as np
```

```
def Normalization2(x):  
    return [(float(i)-np.mean(x))/(max(x)-min(x)) for i in x]
```

測試:

```
x=[1,2,1,4,3,2,5,6,2,7]
```

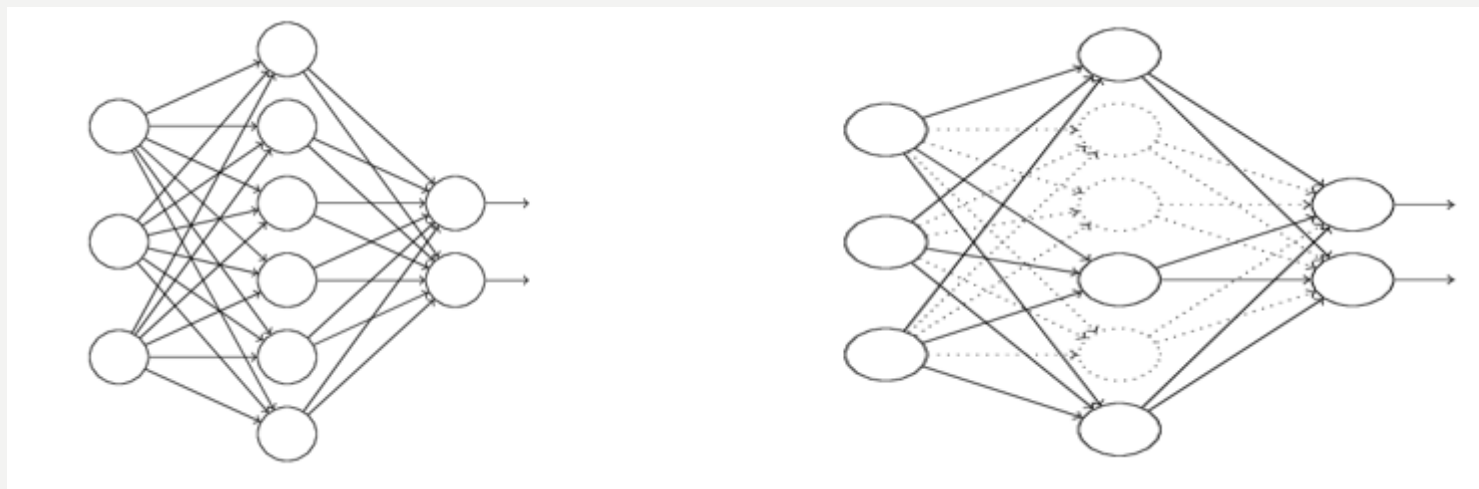
```
b=Normalization2(x)
```

Output:

```
[-0.3833333333333333, -0.21666666666666665, -0.3833333333333333, 0.11666666666666667, -  
0.049999999999999996, -0.21666666666666665, 0.28333333333333338, 0.45000000000000001, -  
0.21666666666666665, 0.6166666666666667]
```

DROPOUT

DropOut顧名思義丟棄，在第一次迭代的過程隨機地“刪除”一半的隱層單元，視它們為不存在直至訓練結束。



範例

https://drive.google.com/drive/u/1/folders/1X4kW1Xe8QQy1FSG9BTE4oYG2_WIN3gJn

[线性回归模型与诊断](#)

[5_6專題 線性回歸](#)

[Machine Learning with python](#)