Microsoft

Reactor

Machine Learning 入門

Map

Meetup link: meetup.com/Microsoft-Reactor-London/

我係

Kinfey Lo – (盧建暉)

Microsoft Cloud Advocate

前微軟MVP、Xamarin MVP和微軟RD，擁有超過10年的雲原生、人工智能和流動應用程式開發經驗，為教育、金融和醫療提供應用。 在微軟，為技術人員和不同行業宣講技術和相關應用場景。

Love Coding(Python， C# , TypeScript , Swift , Rust , Go ）

專注於人工智能，雲原生，流動平台移動開發

Github : https://github.com/kinfey

Email : kinfeylo@microsoft.com **Blog :** https://dev.to/kinfey

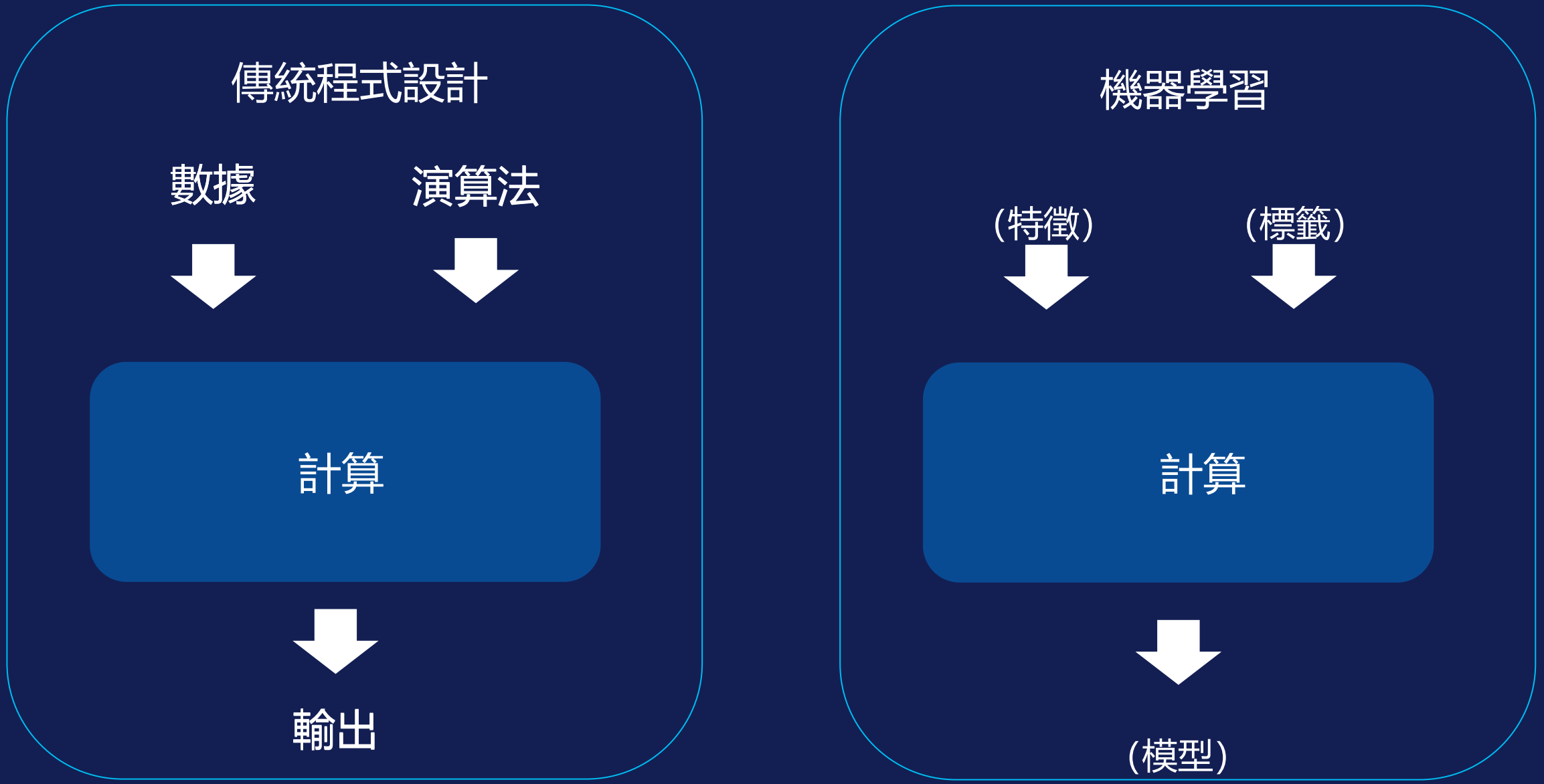
Twitter : @Ljh8304

Machine Learning 係 ...

機器學習的應用場景

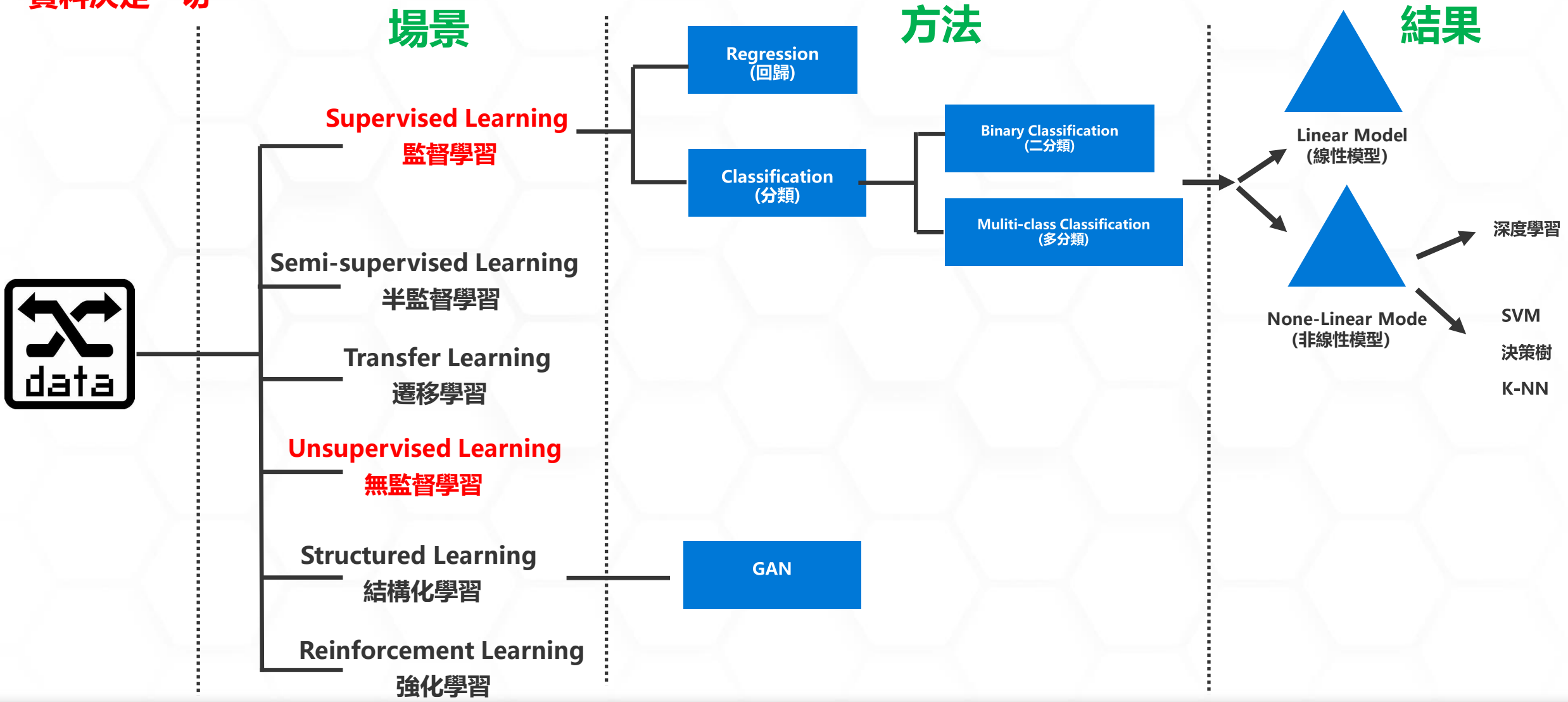
人工智慧，機器學習，深度學習的關係

- **人工智慧 (AI)** 是一種讓電腦能夠模仿人類智慧的技術。它包括機器學習。
  - **機器學習 (ML)** 是人工智慧的一部分，它包括讓電腦能夠依靠經驗更好地處理任務的多項技術（例如深度學習）。
    - **深度學習 (DL)** 又是機器學習的一部分，它以人工神經網路為基礎，讓電腦能夠自我訓練。


傳統程式設計
數據
演算法
計算
輸出
機器學習
(特徵)
(標籤)
計算
(模型)

機器學習大全

資料決定一切



Machine Learning 環境配置

常規依賴的package

庫

pip3 install numpy scikit-learn matplotlib pandas

Jupyter Notebook

pip3 install jupyter

Numpy 介紹

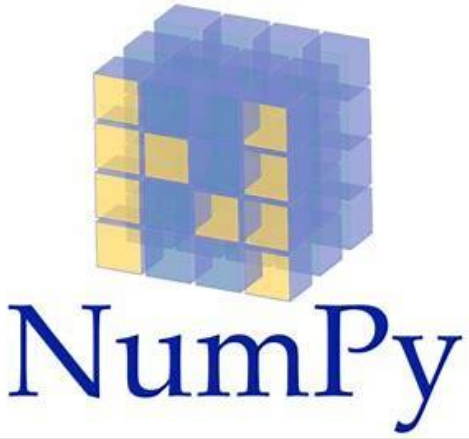
NumPy(Numerical Python) 是 Python 語言的一個擴展程式庫，支援大量的維度數組與矩陣運算，此外也針對陣列運算提供大量的數學函式程式庫。

NumPy（Numeric Python）提供了許多高級的數值程式設計工具，如：矩陣資料類型、向量處理，以及精密的運算庫。專為進行嚴格的數文書處理而產生。多為很多大型金融公司使用

```
import numpy as np

arr = np.array([1, 2, 3, 4, 5])

print(arr)
```



Pandas 介紹

pandas 是基於NumPy 的一種工具，該工具是為解決資料分析任務而創建的。Pandas 納入了大量庫和一些標準的資料模型，提供了高效地操作大型資料集所需的工具。pandas提供了大量能使我們快速便捷地處理資料的函數和方法。你很快就會發現，它是使Python成為強大而高效的資料分析環境的重要因素之一。

Pandas 適合處理一個規正的二維資料（一維也可以，應用較少），即有 N 行 N 列，類似於 SQL 執行後產出的，或者 無合併儲存格Excel 表格 這樣的資料。它可以把多個檔的資料合併在一起，如果結構不一樣，也可以經過處理進行合併。

```
import pandas as pd
```

```
data={'state':['Ohi','Ohi','Ohi','Nev','Nev','Nev'],
      'year':[2000,2001,2002,2003,2004,2005],
      'pop':[1.5,3.4,3.0,1.2,2.9,3.2]}
```

```
frame=pd.DataFrame(data)
```

```
frame
```

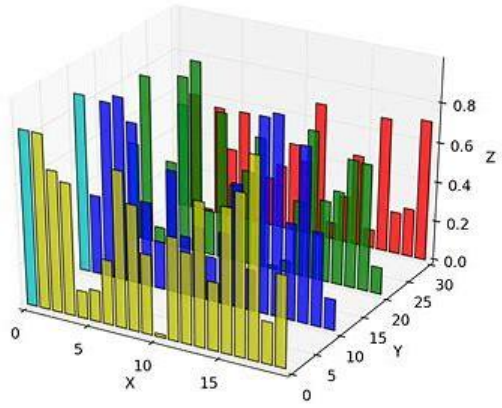
	state	year	pop
0	Ohi	2000	1.5
1	Ohi	2001	3.4
2	Ohi	2002	3.0
3	Nev	2003	1.2
4	Nev	2004	2.9
5	Nev	2005	3.2

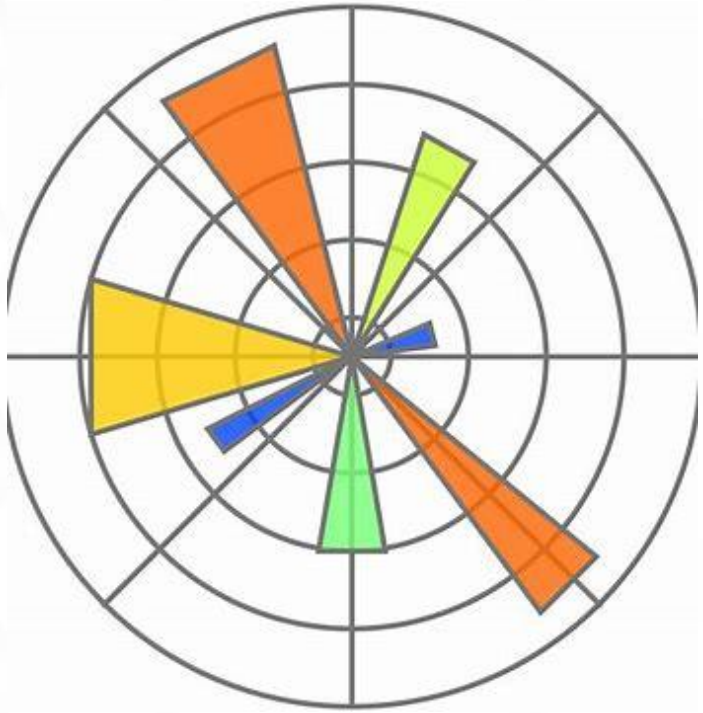
Matplotlib 介紹

Matplotlib是一個Python的2D繪圖庫，它以各種硬拷貝格式和跨平臺的互動式環境生成出版品質級別的圖形。通過Matplotlib，開發者可以僅需要幾行代碼，便可以生成繪圖。一般可繪製折線圖、散點圖、柱狀圖、圓形圖、長條圖、子圖等等。Matplotlib使用Numpy進行陣列運算，並調用一系列其他的Python庫來實現硬體交互。

```
import matplotlib.pyplot as plt

fig = plt.figure()
```





scikit-learn介紹

scikit-learn(簡記sklearn)，是用python實現的機器學習演算法庫。sklearn可以實現資料預處理、分類、回歸、降維、模型選擇等常用的機器學習演算法。sklearn是基於NumPy, SciPy, matplotlib的。

1.簡單高效的資料採擷和資料分析工具

2.可供大家在各種環境中重複使用

3.建立在 NumPy ， SciPy 和 matplotlib 上

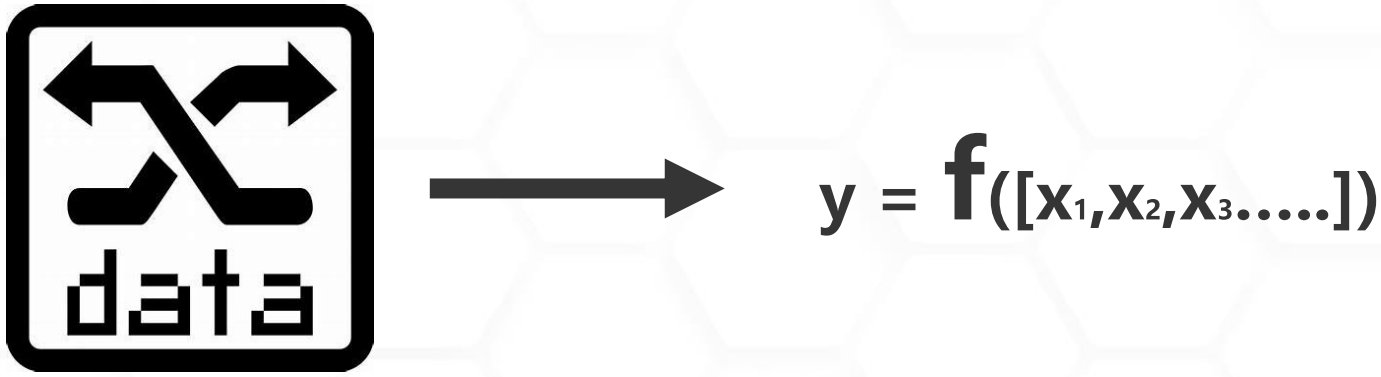
4.開源，可商業使用 - BSD許可證

Sample

三. 監督學習介紹

一切從監督學習談起

Supervised Learning – 預測未來



經驗決定一切，整合大量的標注資料，應用在預測價格，判斷分類等場景上

需要找到一個基於特徵資料，生成結果的方法

一般都有多個特徵資料

x值擬合到計算中，從而為訓練資料集中的所有情況合理準確地生成y。

監督學習的常見解決方法

	Regression (回歸)	**Classification (分類)**
	預測明天的氣溫是多少度	**預測明天是陰、晴還是雨**
輸出	連續資料	離散數據
目的	定量- 找到最佳擬合	定性- 決策邊界
評價	擬合度	精度
場景	預測房價，天氣	垃圾郵件，物品分類

監督學習- 回歸示例

問題導入

期望已有共用單車的的資訊，來預測某個季節，某個天氣條件的投放的單車數量

監督學習- 回歸資料

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	rentals
1	1/1/2011	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331
2	1/2/2011	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131
3	1/3/2011	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120
4	1/4/2011	1	0	1	0	2	1	1	0.2	0.212122	0.590435	0.160296	108
5	1/5/2011	1	0	1	0	3	1	1	0.226957	0.22927	0.436957	0.1869	82
6	1/6/2011	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.0895652	88
7	1/7/2011	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148
8	1/8/2011	1	0	1	0	6	0	2	0.165	0.162254	0.535833	0.266804	68
9	1/9/2011	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.36195	54
10	1/10/2011	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41
11	1/11/2011	1	0	1	0	2	1	2	0.169091	0.191464	0.686364	0.122132	43
12	1/12/2011	1	0	1	0	3	1	1	0.172727	0.160473	0.599545	0.304627	25
13	1/13/2011	1	0	1	0	4	1	1	0.165	0.150883	0.470417	0.301	38
14	1/14/2011	1	0	1	0	5	1	1	0.16087	0.188413	0.537826	0.126548	54
15	1/15/2011	1	0	1	0	6	0	2	0.233333	0.248112	0.49875	0.157963	222
16	1/16/2011	1	0	1	0	0	0	1	0.231667	0.234217	0.48375	0.188433	251
17	1/17/2011	1	0	1	1	1	0	2	0.175833	0.176771	0.5375	0.194017	117
18	1/18/2011	1	0	1	0	2	1	2	0.216667	0.232333	0.861667	0.146775	9
19	1/19/2011	1	0	1	0	3	1	2	0.292174	0.298422	0.741739	0.208317	78
20	1/20/2011	1	0	1	0	4	1	2	0.261667	0.25505	0.538333	0.195904	83
21	1/21/2011	1	0	1	0	5	1	1	0.1775	0.157833	0.457083	0.353242	75
22	1/22/2011	1	0	1	0	6	0	1	0.0591304	0.0790696	0.4	0.17197	93
23	1/23/2011	1	0	1	0	0	0	1	0.0965217	0.0988391	0.436522	0.2466	150
24	1/24/2011	1	0	1	0	1	1	1	0.0973913	0.11793	0.491739	0.15833	86
25	1/25/2011	1	0	1	0	2	1	2	0.223478	0.234526	0.616957	0.129796	186

Sample

監督學習- 分類

分類是有監督機器學習的一種形式，在這種學習中，您訓練模型使用特徵（我們函數中的x值）預測標籤（y），該標籤計算屬於多個可能類別的觀察案例的概率，並預測適當的標籤。最簡單的分類形式是二進位分類，其中標籤為0或1，表示兩個類中的一個；例如，“真”或“假”“內部”或“外部”“是”或“否”；等等

Binary Classification（二分類）

Multi - Classification（多分類）

多類分類可以看作是多個二進位分類器的組合。解決問題有兩種方法：

一對一（OVR），其中為每個可能的類值創建一個分類器，對於預測為此類的情況，結果為正，對於預測為任何其他類的情況，結果為負

一對一（OVO），其中為每個可能的類對創建一個分類器。

監督學習- 二分類示例

問題導入

根據身體指標，判斷是否患有糖尿病

監督學習- 二分類資料

PatientID	Pregnancies	PlasmaGlucc	DiastolicBloc	TricepsThickı	SerumInsulir	BMI	DiabetesPed	Age	Diabetic
1354778	0	171	80	34	23	43.5097259	1.21319135	21	0
1147438	8	92	93	47	36	21.2405757	0.15836498	23	0
1640031	7	115	47	52	35	41.5115235	0.07901857	23	0
1883350	9	103	78	25	304	29.5821919	1.28286985	43	1
1424119	1	85	59	27	35	42.6045359	0.54954187	22	0
1619297	0	82	92	9	253	19.7241602	0.1034245	26	0
1660149	0	133	47	19	227	21.9413567	0.17415978	21	0
1458769	0	67	87	43	36	18.2777226	0.23616494	26	0
1201647	8	80	95	33	24	26.6249289	0.44394739	53	1
1403912	1	72	31	40	42	36.8895757	0.10394364	26	0
1943830	1	88	86	11	58	43.2250409	0.23028462	22	0
1824483	3	94	96	31	36	21.2944794	0.25902048	23	0
1848869	5	114	101	43	70	36.4953197	0.07919016	38	1
1669231	7	110	82	16	44	36.0892934	0.28127616	25	0
1683688	0	148	58	11	179	39.1920755	0.16082901	45	0
1738587	3	109	77	46	61	19.847312	0.20434527	21	1
1884264	3	106	64	25	51	29.0445728	0.58918802	42	1
1485251	1	156	53	15	226	29.7861916	0.20382353	41	1
1536832	8	117	39	32	164	21.230996	0.08936275	25	0
1438701	3	102	100	25	289	42.1857203	0.17559283	43	1
1359971	0	92	84	8	324	21.8662604	0.25833233	33	0
1631185	0	118	95	7	276	42.5008866	0.08355755	24	0
1061812	1	82	55	18	165	36.6282491	0.17161962	23	0
1218879	1	124	82	42	266	34.9857724	0.08333502	25	0
1940297	2	44	81	46	146	34.5340823	0.69350217	55	1
1710438	9	104	68	42	40	51.8554011	0.18293782	21	1
1139740	6	135	91	31	14	45.2741129	0.70716268	21	1
1398321	3	163	87	42	428	18.5711882	0.77701581	25	0
1975790	0	119	50	52	16	45.3911208	0.27056708	22	0
1721341	0	70	64	9	16	20.9852233	0.13739311	33	0
1121857	11	75	89	8	541	29.4227539	0.08373162	47	1
1117458	8	152	83	42	46	18.9095449	0.60258169	34	0
1083679	0	149	50	8	67	46.3205974	1.07158383	42	0

示例

監督學習- 多分類示例

問題導入

基於企鵝的生理指標判斷企鵝類型，'Adelie', 'Gentoo', 'Chinstrap'

數據		場景		方法
企鵝的生理指標判斷企鵝類型	判斷類別 →	監督學習	預測類型 →	多分類

監督學習- 多分類資料

PatientID	Pregnancies	PlasmaGlucc	DiastolicBloc	TricepsThickı	SerumInsulir	BMI	DiabetesPed	Age	Diabetic
1354778	0	171	80	34	23	43.5097259	1.21319135	21	0
1147438	8	92	93	47	36	21.2405757	0.15836498	23	0
1640031	7	115	47	52	35	41.5115235	0.07901857	23	0
1883350	9	103	78	25	304	29.5821919	1.28286985	43	1
1424119	1	85	59	27	35	42.6045359	0.54954187	22	0
1619297	0	82	92	9	253	19.7241602	0.1034245	26	0
1660149	0	133	47	19	227	21.9413567	0.17415978	21	0
1458769	0	67	87	43	36	18.2777226	0.23616494	26	0
1201647	8	80	95	33	24	26.6249289	0.44394739	53	1
1403912	1	72	31	40	42	36.8895757	0.10394364	26	0
1943830	1	88	86	11	58	43.2250409	0.23028462	22	0
1824483	3	94	96	31	36	21.2944794	0.25902048	23	0
1848869	5	114	101	43	70	36.4953197	0.07919016	38	1
1669231	7	110	82	16	44	36.0892934	0.28127616	25	0
1683688	0	148	58	11	179	39.1920755	0.16082901	45	0
1738587	3	109	77	46	61	19.847312	0.20434527	21	1
1884264	3	106	64	25	51	29.0445728	0.58918802	42	1
1485251	1	156	53	15	226	29.7861916	0.20382353	41	1
1536832	8	117	39	32	164	21.230996	0.08936275	25	0
1438701	3	102	100	25	289	42.1857203	0.17559283	43	1
1359971	0	92	84	8	324	21.8662604	0.25833233	33	0
1631185	0	118	95	7	276	42.5008866	0.08355755	24	0
1061812	1	82	55	18	165	36.6282491	0.17161962	23	0
1218879	1	124	82	42	266	34.9857724	0.08333502	25	0
1940297	2	44	81	46	146	34.5340823	0.69350217	55	1
1710438	9	104	68	42	40	51.8554011	0.18293782	21	1
1139740	6	135	91	31	14	45.2741129	0.70716268	21	1
1398321	3	163	87	42	428	18.5711882	0.77701581	25	0
1975790	0	119	50	52	16	45.3911208	0.27056708	22	0
1721341	0	70	64	9	16	20.9852233	0.13739311	33	0
1121857	11	75	89	8	541	29.4227539	0.08373162	47	1
1117458	8	152	83	42	46	18.9095449	0.60258169	34	0
1083679	0	149	50	8	67	46.3205974	1.07158383	42	0

Sample

四. 無監督學習介紹

無監督學習

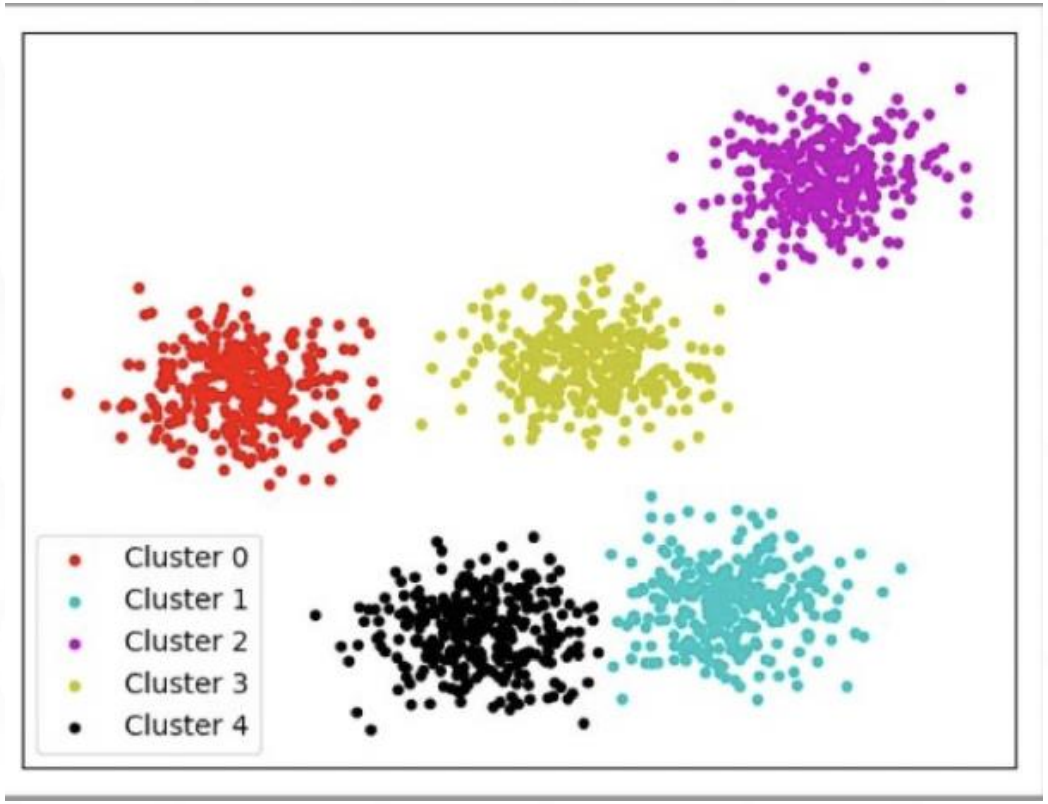
區別於監督學習，無監督學習希望機器做到无師自通，

在完全沒有任何標籤的情況下，機器到底能學到什麼樣的知識

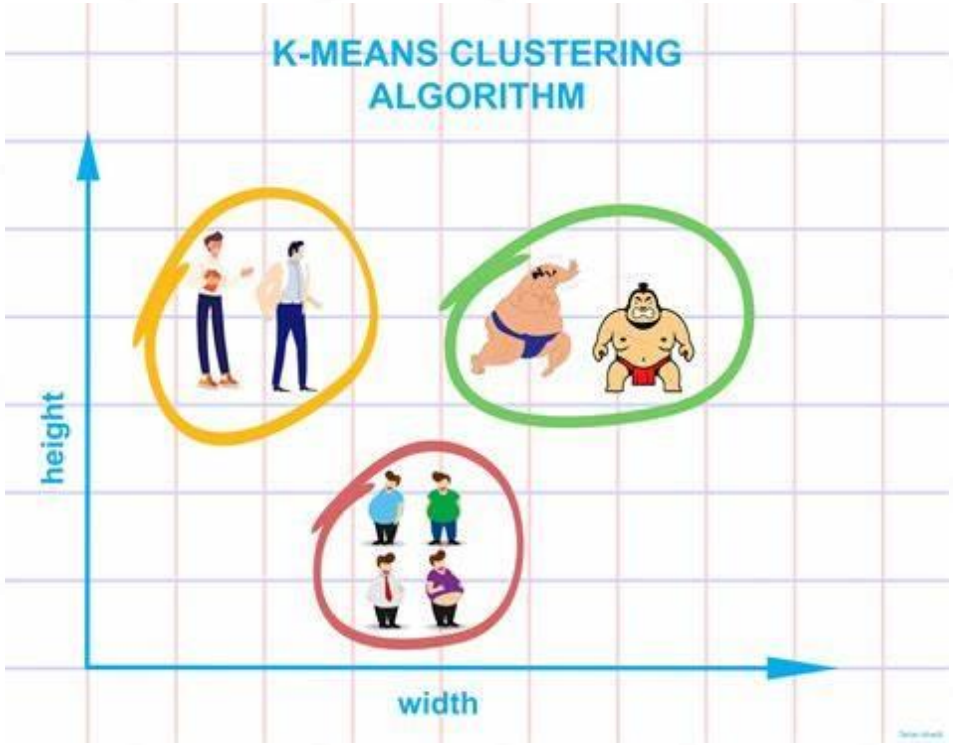
如：通過多維度的用戶行為分析，廣告投放

推薦系統

異常查找



無監督學習基本演算法 – 聚類分析 K-means

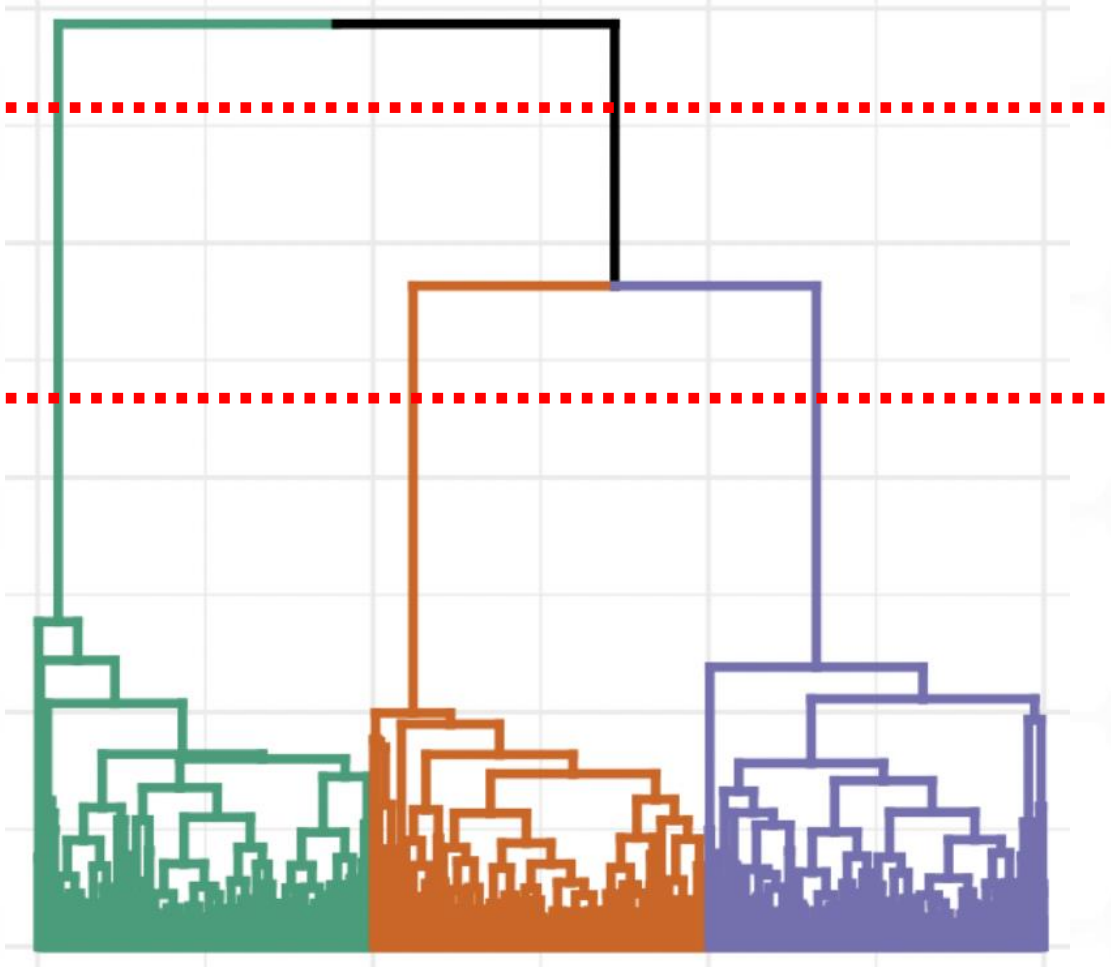


是在沒有給定劃分類別的情況下，根據資料的相似度進行分組的一種方法，分組的原則是組內距離最小化而組間距離最大化。

K-means演算法是典型的基於距離的非層次聚類演算法，在最小化誤差函數的基礎上將資料劃分為預定的K類別，採用距離作為相似性的評級指標，即認為兩個物件的距離越近，其相似度越大。

1、隨機設置K個特徵空間內的點作為初始的聚類中心
2、對於其他每個點計算到K個中心的距離，未知的點選擇最近的一個聚類中心點作為標記類別
3、接著對著標記的聚類中心之後，重新計算出每個聚類的新中心點（平均值）
4、如果計算得出的新中心點與原中心點一樣，那麼結束，否則重新進行第二步過程

無監督學習基本演算法 – 聚類分析 Hierarchical Clustering



分層聚類通過分割方法或聚集方法創建聚類。除法是一種"自上而下"的方法，從整個資料集開始，然後逐步查找分區。凝聚聚類是一種"自下而上"的方法。

（1）初始化：把每個樣本各自歸為一類（每個樣本自成一類），計算每兩個類之間的距離，在這裡也就是樣本與樣本之間的相似度（本質還是計算類與類之間的距離）。

（2）尋找各個類之間最近的兩個類，把它們歸為一類（這樣，類的總數就減少了一個）

（3）重新計算新生成的這個類與各個舊類之間的距離（相似度）

（4）重複（2）（3）步，直到所有的樣本都歸為一類，結束。

無監督學習- 聚類示例

問題導入

劃分種子品質

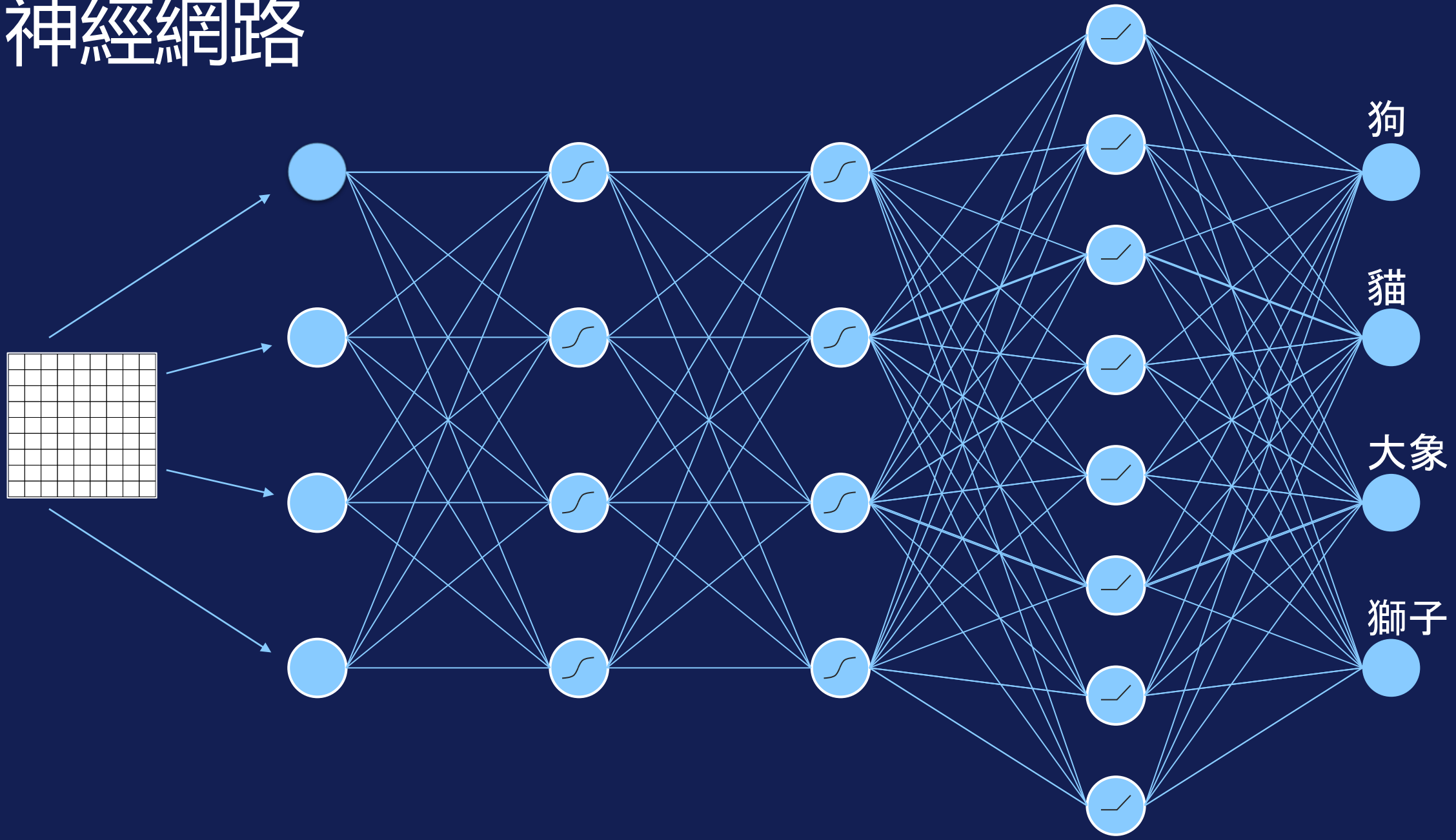
無監督學習- 聚類示例

area	perimeter	compactness	kernel_lengt	kernel_width	asymmetry_	groove_leng	species
15.26	14.84	0.871	5.763	3.312	2.221	5.22	0
14.88	14.57	0.8811	5.554	3.333	1.018	4.956	0
14.29	14.09	0.905	5.291	3.337	2.699	4.825	0
13.84	13.94	0.8955	5.324	3.379	2.259	4.805	0
16.14	14.99	0.9034	5.658	3.562	1.355	5.175	0
14.38	14.21	0.8951	5.386	3.312	2.462	4.956	0
14.69	14.49	0.8799	5.563	3.259	3.586	5.219	0
14.11	14.1	0.8911	5.42	3.302	2.7	5	0
16.63	15.46	0.8747	6.053	3.465	2.04	5.877	0
16.44	15.25	0.888	5.884	3.505	1.969	5.533	0
15.26	14.85	0.8696	5.714	3.242	4.543	5.314	0
14.03	14.16	0.8796	5.438	3.201	1.717	5.001	0
13.89	14.02	0.888	5.439	3.199	3.986	4.738	0
13.78	14.06	0.8759	5.479	3.156	3.136	4.872	0
13.74	14.05	0.8744	5.482	3.114	2.932	4.825	0
14.59	14.28	0.8993	5.351	3.333	4.185	4.781	0
13.99	13.83	0.9183	5.119	3.383	5.234	4.781	0
15.69	14.75	0.9058	5.527	3.514	1.599	5.046	0
14.7	14.21	0.9153	5.205	3.466	1.767	4.649	0
12.72	13.57	0.8686	5.226	3.049	4.102	4.914	0
14.16	14.4	0.8584	5.658	3.129	3.072	5.176	0
14.11	14.26	0.8722	5.52	3.168	2.688	5.219	0
15.88	14.9	0.8988	5.618	3.507	0.7651	5.091	0
12.08	13.23	0.8664	5.099	2.936	1.415	4.961	0
15.01	14.76	0.8657	5.789	3.245	1.791	5.001	0
16.19	15.16	0.8849	5.833	3.421	0.903	5.307	0
13.02	13.76	0.8641	5.395	3.026	3.373	4.825	0
12.74	13.67	0.8564	5.395	2.956	2.504	4.869	0
14.11	14.18	0.882	5.541	3.221	2.754	5.038	0
13.45	14.02	0.8604	5.516	3.065	3.531	5.097	0

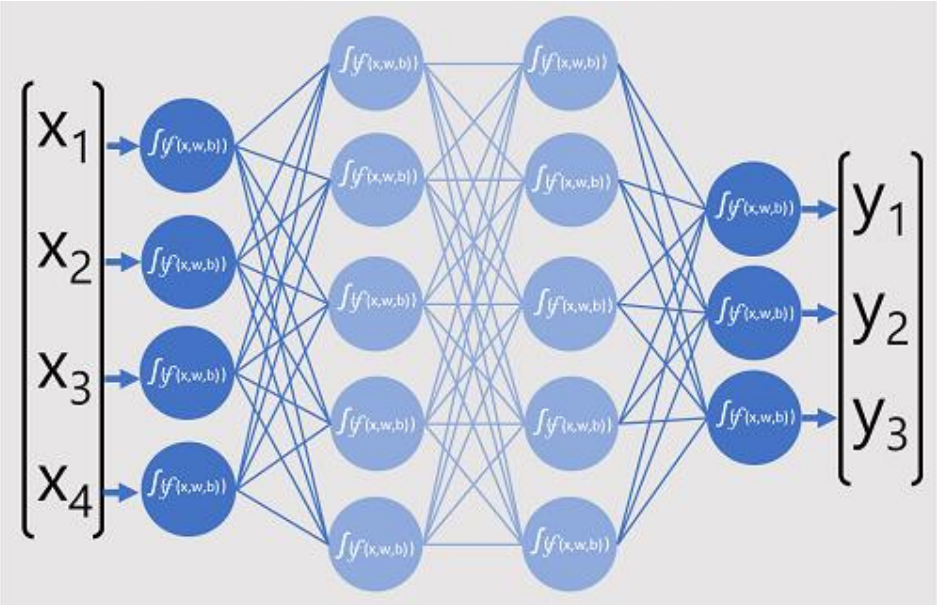
Sample

四. 深度學習介紹

神經網路



深度學習



深度神經網路的訓練過程包含多個稱為“時期”的反覆運算。 對於第一個時期，你首先要為權重分配隨機初始化值 (w) 和偏差 (b) 值。 然後，該過程如下所示：

1.具有已知標籤值的資料觀察特徵將提交到輸入層。 通常情況下，這些觀測值分組為多個批次（通常稱為小型批次處理）。

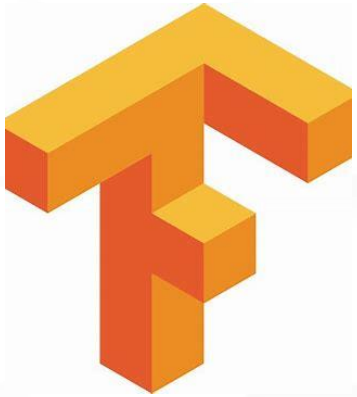
2.然後，神經元發揮其作用，並在啟動後將結果傳遞到下一層，直到輸出層生成預測。

3.將預測與實際的已知值進行比較，並對預測值和真實值之間的差異量（稱為“損失”）進行計算。

4.根據結果，計算權重和偏差值的修訂值以減少損失，並將這些調整反向傳播到網路層中的神經元。

5.下一個時期使用修改後的權重和偏差值重複批量訓練向前傳遞，有望通過減少損失來提高模型的準確性。

常用的深度學習庫



TensorFlow

PyTorch

PaddlePaddle

關於機器學習/深度學習 開發環境搭建

https://blog.csdn.net/kinfey/article/details/117635067

Sample

MS Learn 的學習模組推薦

https://aka.ms/HKPythonLearn003

五. 小結

Microsoft

Reactor

Thank You!