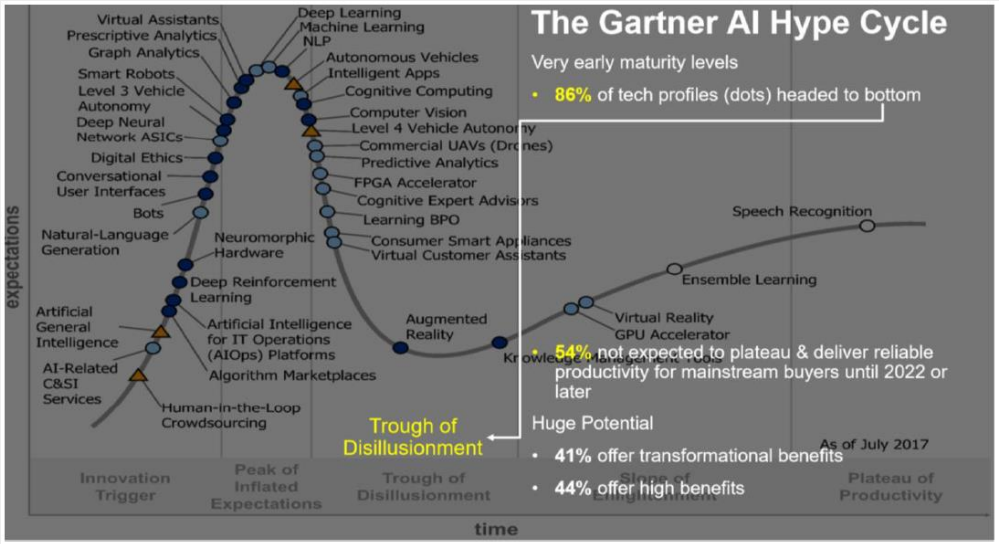
初探Google Tensorflow
人工大智慧

AI Hype Cycle：全球已邁入快速發展期

- 以技術面而言諸如「語音辨識」、「整體學習」等技術已算成熟，兩年內應可普及應用；另如「自然語言處理」、「虛擬個人助手」、「深度強化學習」等，預計再5-10年可成為主流技術
- 超過86%的科技還沒有越過「低谷期」(Trough of Disillusionment)，代表人們目前可能對於這些科技的期望過高
- 總體來說，最終將各有4成以上的科技，可為所屬產業帶來「變革性的利益」或「較高的利益」



資料來源：Gartner (2018/03)

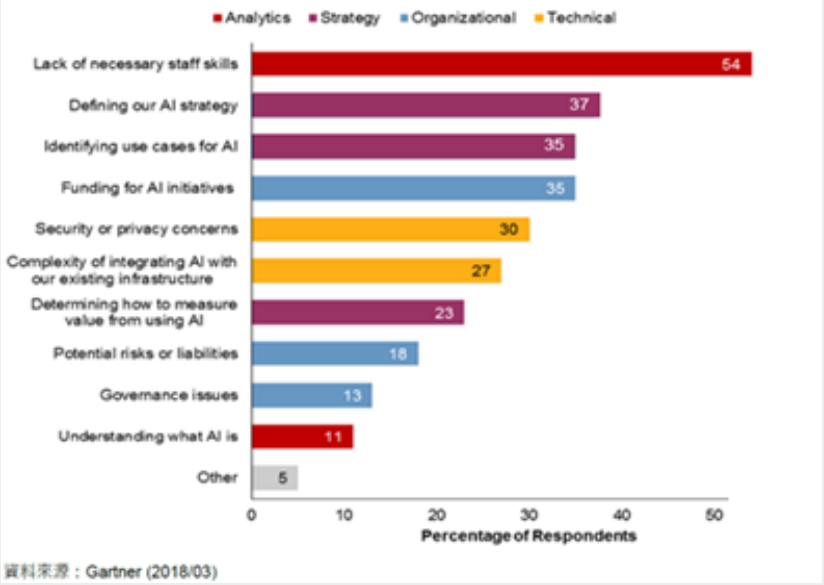
人工智慧在產業的導入現況

- 2017/07，Gartner針對83家產業客戶進行調查，將AI導入分為「知識蒐集/調查/策略發展中」、「試驗性」、「佈署中」、「已佈署/使用中」四個前後階段
  - 高達60%：初期的「知識蒐集/調查/策略發展中」階段
  - 25%：在「試驗性」階段
  - 6%：達「佈署中」或「已佈署/使用中」狀態
    - Gartner預測即使在一年後，也僅會由6%提升至10%

- AI技術是熱門話題，許多產業也都在積極導入；能夠真正在商業經營上佈署運用此技術者，現今仍不到10%；此調查對象主要是大型產業，更不用談以中小企業為主的某些產業，是否能快速導入AI

產業導入AI時面臨的挑戰

- Gartner將發生的挑戰類型整理為分析面、策略面、組織面、及技術面四個構面
  - 54 : 引入AI技術時「缺乏必須的員工技能」(分析面)
  - 35~37 : 「如何定義AI策略」、「辨別AI使用案例」(皆為策略面)
  - 35 : 「經費/投資議題」(組織面)
  - 30 : 「安全與隱私考量」(技術面)
  - 「了解AI是什麼」仍是一項挑戰的產業僅剩11



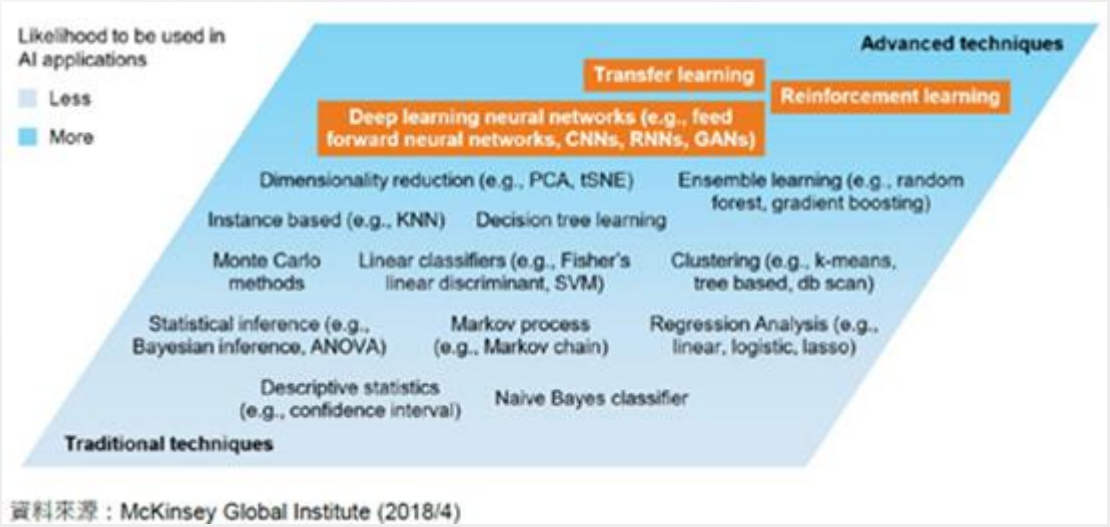
資料來源：Gartner (2018/03)

台灣市場調查結果

- 法人單位針對國內六個產業二十多家企業進行面訪
  - 不清楚可利用AI解什麼議題，也不確定自身是否具足夠資料量；此項類似Gartner調查中排名第三的「辨別AI使用案例」
  - 找不到所需的專業人才，同時也缺乏可靠的招募管道；此等同於Gartner調查中排名第一的「缺乏必須技能的員工」項目
  - 缺乏相關實驗場域；企業表示即使招募到所需人才，完成的AI演算法因為其成效難被驗證，也不敢直接上線進行商業化，害怕造成反效果；此類似於Gartner調查中排名第七的「如何衡量使用AI的價值」項目
  - 進行跨業合作時，對於個資處理常常缺乏處理經驗，特別是所擁有的資料來自於終端消費者的相關產業(如電信與金融)；此點等同調查中排名第五的挑戰「安全與隱私考量」項目
  - 「投資成本考量」也是一項挑戰；一套AI解決方案所需之初置成本與維護成本皆不低，是否能有效回收須審慎評估；此點等同Gartner調查中排名第五的挑戰「經費/投資議題」項目

AI可行性的實證研究

- 針對AI的可行性，McKinsey分析400個現有AI應用個案，進行歸納性整理



資料來源：McKinsey Global Institute (2018/4)

- 資料分析：
  - 傳統分析技術
    - 敘述性統計、迴歸分析、蒙地卡羅模擬等
  - 「人工智慧」技術
    - 使用到深度學習神經網路的相關技術(如FNN、CNN、RNN、GAN等)
    - 強化學習(Reinforcement Learning)遷移學習(Transfer Learning)兩種

- 從「全球市場規模」來看，估計整體資料分析市場的營收一年約9.5兆到15.4兆美元，佔此市場份額的40%，且以FNN、CNN、RNN三種為主流
- McKinsey分析的400個現有AI應用個案中，AI能提供「顯著效益」提升者占了69%；另在16%個案中，AI甚至為提升效益的「唯一有效方式」；而在剩下的最終15%個案裡，AI所能造成的效益提升則非常有限。
- AI技術對各產業的價值提升貢獻，McKinsey估算平均可達62%；其中旅遊業、運輸物流業、零售業、汽車組裝業、高科技為前五名，可利用AI技術達到高度價值提升者(皆可達到八成以上)
- 藥物與醫療產品、保險、先進電子/半導體、國防太空等四項產業，藉AI進行價值提升的貢獻度將較低；但AI技術並非全然無用，對產業效益提升仍可達三成到四成之間，只是距離前述的平均水準有段距離

環境介紹與安裝

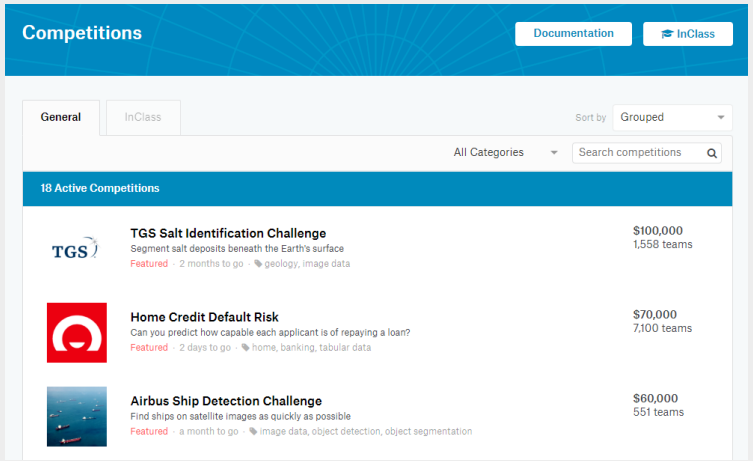
- Anaconda : Python的懶人包
  - Python, 資料分析, 機器學習, 視覺化套件
    - Numpy: Python做多維陣列（矩陣）運算時的必備套件，比起Python內建的list，Numpy的array有極快的運算速度優勢
    - Pandas：Pandas可以讓Python很容易做到幾乎所有Excel的功能，像是樞紐分析表、小記、欄位加總、篩選
    - Matplotlib：基本視覺化工具，可以畫長條圖、折線圖等等...
    - Seaborn：另一知名視覺化工具，畫起來比matplotlib好看
    - SciKit-Learn: 關於機器學習的model基本上都在這個套件，像是SVM, Random Forest...
    - Notebook(Jupyter notebook): 一個輕量級web-base 寫Python的工具，在資料分析這個領域很熱門，雖然功能沒有比Pycharm, Spyder這些專業的IDE強大，但只要code小於500行，用Jupyter寫非常方便，Jupyter也開始慢慢支援一些Multi cursor的功能了，可以讓你一次改許多的變數名稱
  - 一鍵安裝完， 涵蓋90%一般人這輩子會用到的Python套件；剩下的再用pip install個別去安裝即可

- Google推出免費使用GPU的深度學習雲計算平台 Google Colaboratory
  - Google Colab提供的是免費Tesla K80 GPU，可以玩Keras、Tensorflow、PyTorch或者Mxnet等
  - Google Colab 是完全免費！不過一次只能使用12個小時
- Keras
  - !git clone https://github.com/wxs/keras-mnist-tutorial.git
    - !wget https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/datasets/Titanic.csv -P /
    - import pandas as pd
      titanic = pd.read_csv('drive/app/Titanic.csv')
      titanic.head(5)
  - !pip install -q keras
    import keras

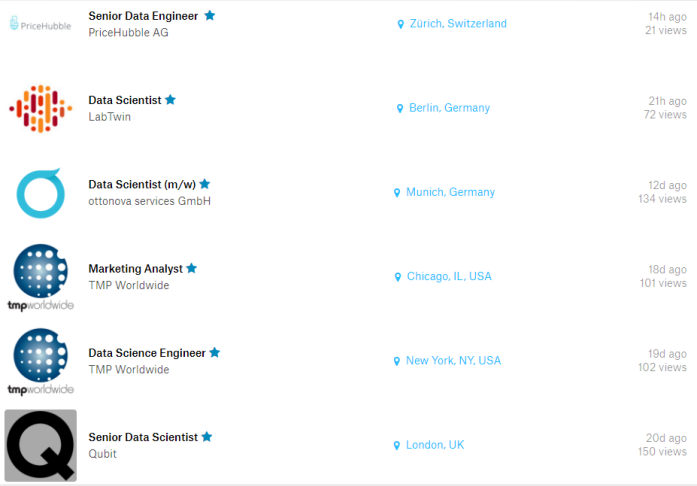
Kaggle介紹

kaggle

- Kaggle是全世界公認最大的資料科學社群
  - ✓進行各種資料分析的比賽，比賽提供高額的獎金
  - ✓吸引全世界優秀的資料科學家前來比賽
  - ✓許多熱愛分享的科學家於在比賽過後，討論區留下他們的當初思考問題的邏輯以及解題的脈絡
- Kaggle上還有一個Dataset專區
  - ✓ 分析時最缺的就是整理好的資料，如果要在自己用API或是爬蟲去抓資料，會耗費許多的時間跟精力，Dataset可以看到許多已經被整理好的資料提供大家下載



- kaggle上面還會有徵才訊息，可以在這邊大概推敲出各科技公司需要什麼樣的資料科學人才，了解目前資料科學的趨勢，對於往後如果要申請工作/人才需求，這邊是一個很好的風向指標

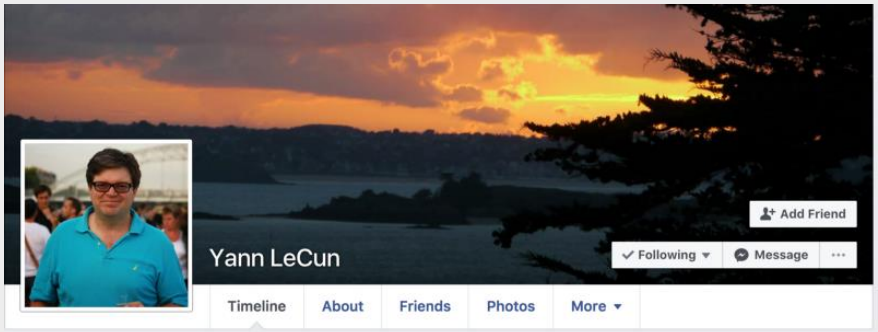


資料參考:
http://www.jianshu.com/p/eb0b37500229
http://blog.csdn.net/u012162613/article/details/41929171
http://blog.csdn.net/willtongji/article/details/52874773

資料科學領域的專家、網站

- 關注這些專家，長期來說可以讓你對這整個產業的近況更了解
  - **吳恩達(AndrewNG)**
    - Facebook: https://www.facebook.com/andrew.ng.96
    - 這個領域的大牛、教父。史丹佛大學副教授、Coursera創辦人、Google Brain創辦人、前百度首席資料科學家...。
    - 全世界公認最好的Machine Learning課程就是他在Coursera上面的machine Learning課程

- **揚·勒丘恩 (Yann LeCun)**
  - Facebook:https://www.facebook.com/yann.lecun
  - 為Facebook AI研究院院長，同時也是美國紐約大學的終身教授。
  - 研究包括機器學習、計算機視覺、移動機器人以及計算神經學等。
  - 他因著名且影響深遠的卷積神經網絡（CNN）相關的工作而被人稱為CNN之父



- **Awesome Machine learning**
  - https://github.com/josephmisiti/awesome-machine-learning
  - 熱心的網友將Machine learning相關的資源都整理在這個github
  - 相關的書籍、課程、部落格、聚會、各語言的Machine learning套件都在這邊可以找到
- **Hacker news for Data science**
  - http://www.datatau.com/news
  - 可以了解目前最新的資料科學相關知識

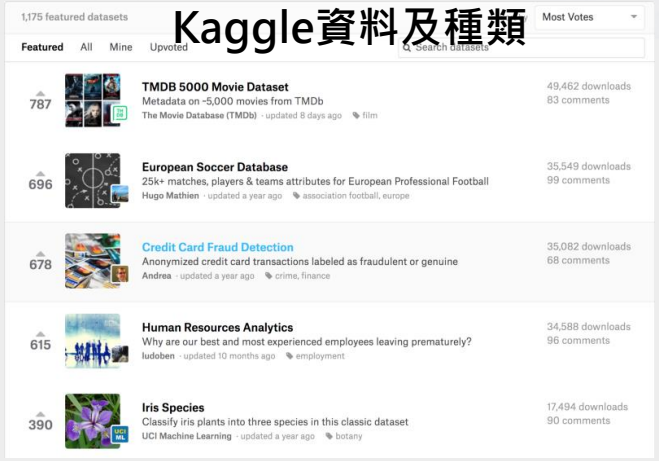
Sklearn內建資料集

資料分析示範說明

- 如何獲取資料？
  - Sklearn內建資料集
- 如何獲取資料？
  - Google Map API
- Pandas 基本function介紹
  - Series, DataFrame, Selection, Grouping
- 資料前處理
  - Missing data, One-hot encoding, Feature Scaling
- 資料視覺化
  - Matplotlib, Seaborn, Plotly

Sklearn內建資料集

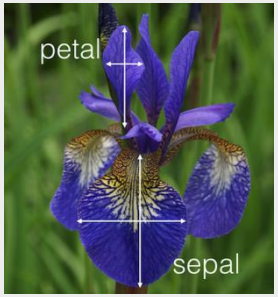
- 新手一開始不知道去哪裡取得資料進行分析，有許多的第三方網站會提供許多資料, ex : kaggle
  - 裡面提供許多已經整理好的資料集，讓資料科學家可以不需要花太多時間去做資料的前處理才開始分析
    - 電影資訊的資料集
    - 歐洲足球比賽的資料
    - 信用卡盜刷偵測
    - 人力資源分析資料等等
    - Iris資料集
      (後面會以Iris dataset作為示範)

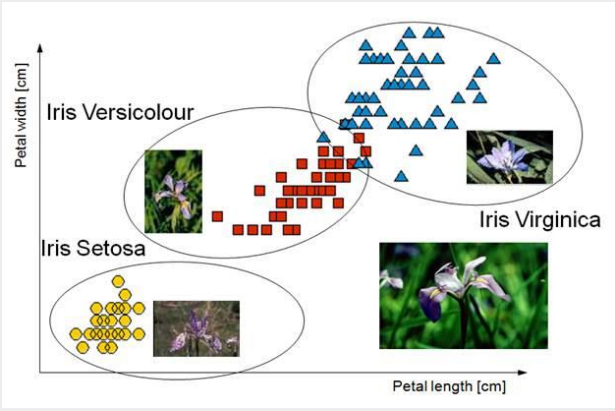


  - 這些資料都是非常具有高度分析價值的，而且也可以在上面跟其他的資料科學家做交流以及分享

- 使用scikit-learn內建的資料
  - 內建的資料集用起來非常簡單，只要一行指令就可以載入資料
  - scikit-learn 提供的dataset : http://scikit-learn.org/stable/datasets/index.html
- 以Iris dataset為例，鳶尾花資料集是非常著名的生物資訊資料集之一，取自美國加州大學歐文分校的機器學習資料庫http://archive.ics.uci.edu/ml/datasets/Iris，資料的筆數為150筆，共有五個欄位：
  1. 花萼長度(Sepal Length)：計算單位是公分。
  2. 花萼寬度(Sepal Width)：計算單位是公分。
  3. 花瓣長度(Petal Length) ：計算單位是公分。
  4. 花瓣寬度(Petal Width)：計算單位是公分。
  5. 類別(Class)：可分為Setosa，Versicolor和Virginica三個品種。

- Iris 資料集算是最入門的機器學習演算法資料
  - 透過花瓣以及花萼的長與寬來預測是屬於哪一種類的Iris花(Setosa, Virginica, Versicolour)
  - 載入Iris 資料集
    - import sklearn的datasets
    - 使用load_iris()獲得資料，回傳的資料格式為dictionary
    - 資料處理轉為表格的形式，在python中有關表格的處理主要都使用pandas為主。





- from sklearn import datasets
  iris = datasets.load_iris()
  iris

dictionary格式


```
{'DESCR': 'Iris Plants Database\n====================\n\nNotes\n-----\nData Set Characteristics:\n    :Number of Inst
ances: 150 (50 in each of three classes)\n    :Number of Attributes: 4 numeric, predictive attributes and the class\n
    :Attribute Information:\n        - sepal length in cm\n        - sepal width in cm\n        - petal length in cm
\n        - petal width in cm\n        - class:\n                - Iris-Setosa\n                - Iris-Versicolour\n
                - Iris-Virginica\n    :Summary Statistics:\n\n    ============== ==== ==== ======= ===== ============
=======\n                    Min  Max   Mean    SD   Class Correlation\n    ============== ==== ==== ======= ===== ==
==================\n    sepal length:   4.3  7.9   5.84   0.83    0.7826\n    sepal width:    2.0  4.4   3.05   0.43
  -0.4194\n    petal length:   1.0  6.9   3.76   1.76    0.9490  (high!)\n    petal width:    0.1  2.5   1.20  0.76
   0.9565  (high!)\n    ============== ==== ==== ======= ===== ====================\n\n    :Missing Attribute Values:
None\n    :Class Distribution: 33.3% for each of 3 classes.\n    :Creator: R.A. Fisher\n    :Donor: Michael Marshall
 (MARSHALL%PLU@io.arc.nasa.gov)\n    :Date: July, 1988\n\nThis is a copy of UCI ML iris datasets.\nhttp://archive.ic
s.uci.edu/ml/datasets/Iris\n\nThe famous Iris database, first used by Sir R.A Fisher\n\nThis is perhaps the best know
n database to be found in the\npattern recognition literature.  Fisher\'s paper is a classic in the field and\nis ref
erenced frequently to this day.  (See Duda & Hart, for example.)  The\ndata set contains 3 classes of 50 instances ea
ch, where each class refers to a\ntype of iris plant.  One class is linearly separable from the other 2; the\nlatter
 are NOT linearly separable from each other.\n\nReferences\n----------\n   - Fisher,R.A. "The use of multiple measure
ments in taxonomic problems"\n     Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to\n     Mathe
matical Statistics" (John Wiley, NY, 1950).\n   - Duda,R.O., & Hart,P.E. (1973) Pattern Classification and Scene Anal
ysis.\n     (Q327.D83) John Wiley & Sons.  ISBN 0-471-22361-1.  See page 218.\n   - Dasarathy, B.V. (1980) "Nosing Ar
ound the Neighborhood: A New System\n     Structure and Classification Rule for Recognition in Partially Exposed\n
```


- iris.keys()
  print(iris[‘DESCR’])
  print(iris[‘data’])
- import pandas as pd
  x = pd.DataFrame(iris[‘data’],
  columns=iris[‘feature_names’])

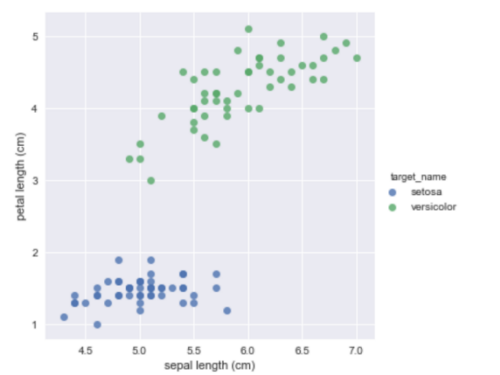
線性分類

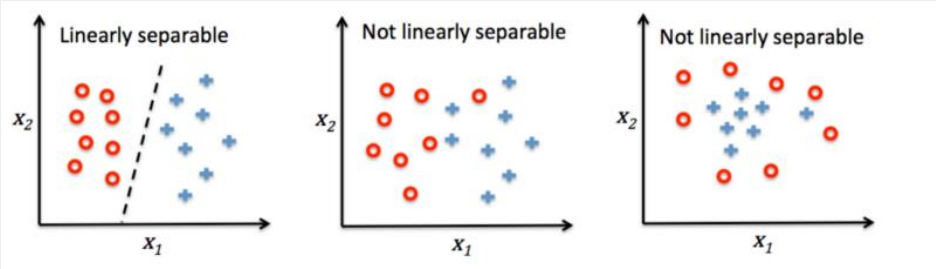
Scikit-learn介紹

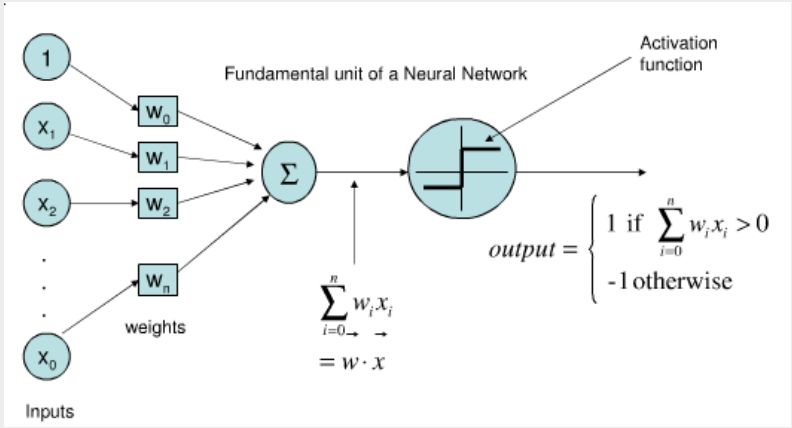
- Scikit-learn簡稱SKlearn(使用時為import sklearn)，SKlearn已內建在一開始安裝的Anaconda套件中，所以可以直接使用
- 除了包含許多知名的Machine learning的算法之外，我們在之前也提到SKlearn有內建許多的知名的dataset(比如說像Iris以及手寫辨識數字的資料)，讓你可以用兩三行的程式碼就將資料轉成pandas的格式
- 提供的功能分為六大塊，分別是監督式學習的分類(Classification)以及回歸(Regression)演算法、非監督式的分群演算法(Clustering)、還有降低維度(Dimensionality reduction)、模型選擇(Model selection)以及資料的前處理(Preprocess)
- 接下來的內容會以Classification演算法為主

線性分類-感知器(Perceptron) 介紹

- 機器學習領域最早被開發出來的演算法：感知器Perceptron（也稱為Perceptron Learning Algorithm簡稱PLA）
  - Perceptron這個演算法只有在資料是線性可分的形況下才能正確分類（演算法才會停止）

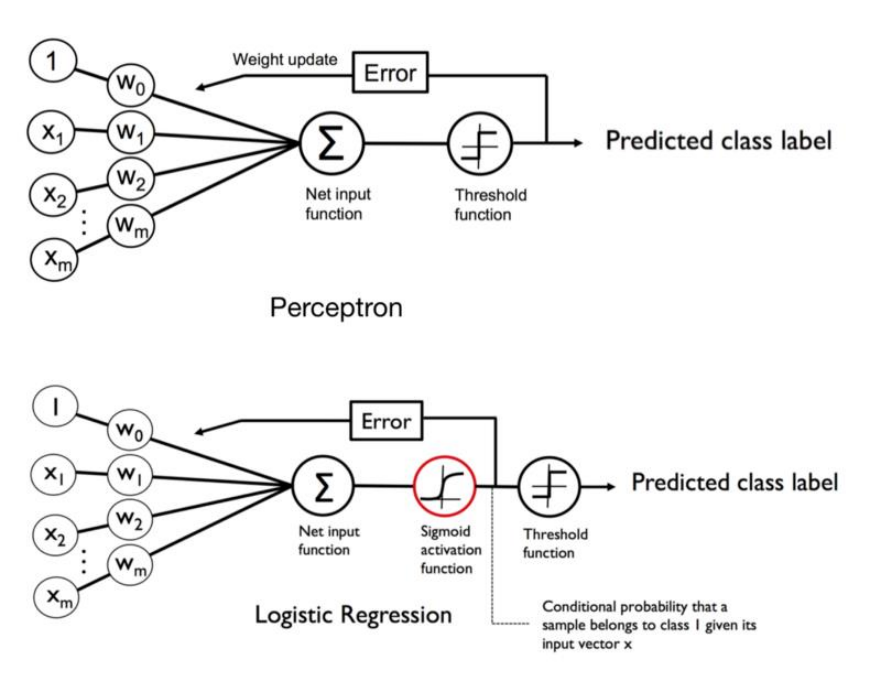


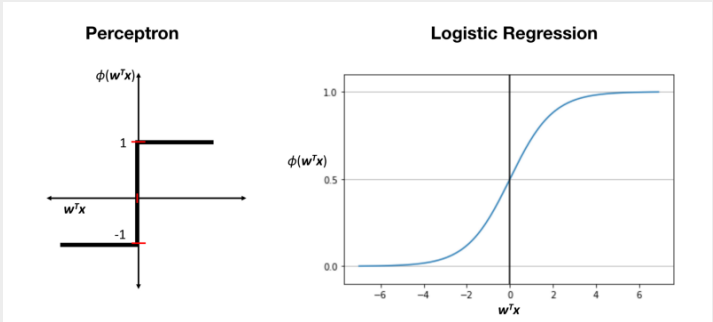




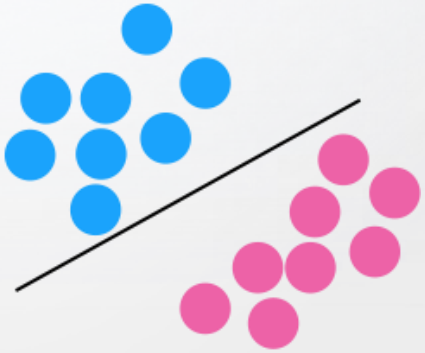
線性分類-邏輯斯回歸(Logistic Regression)

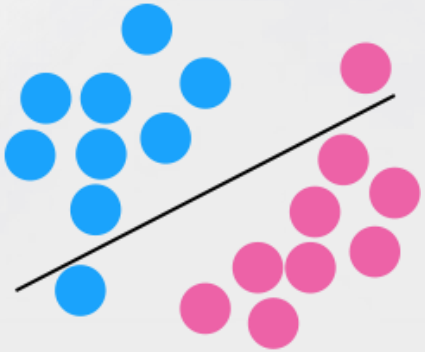
- Perceptron 能夠讓我們成功達成二元分類，但我們只能知道預測結果是A還是B，沒辦法知道是A、是B的機率是多少

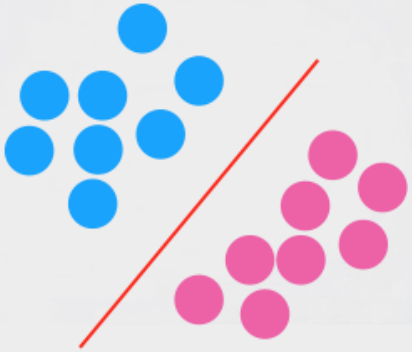


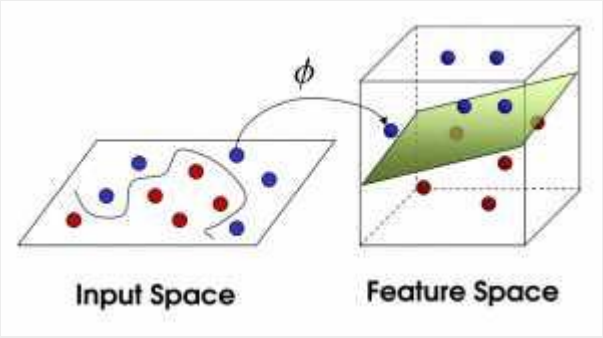


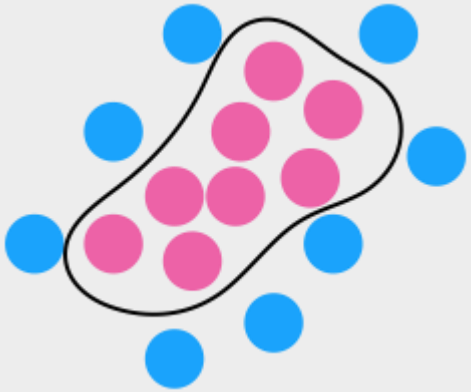
支援向量機(Support Vector Machine)











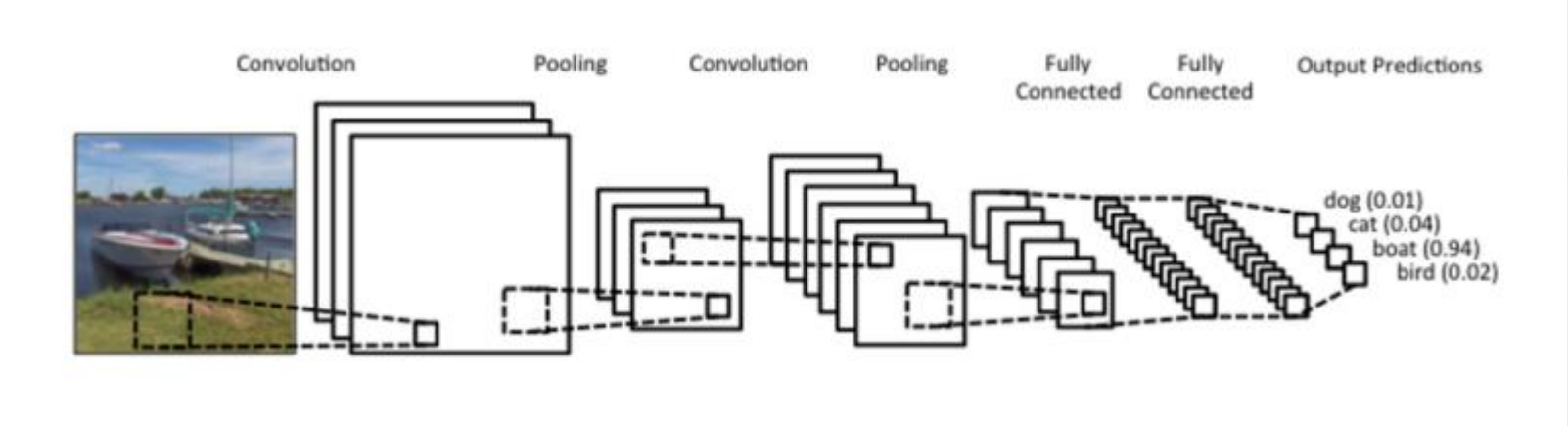
卷積神經網絡介紹
(Convolutional Neural Network)

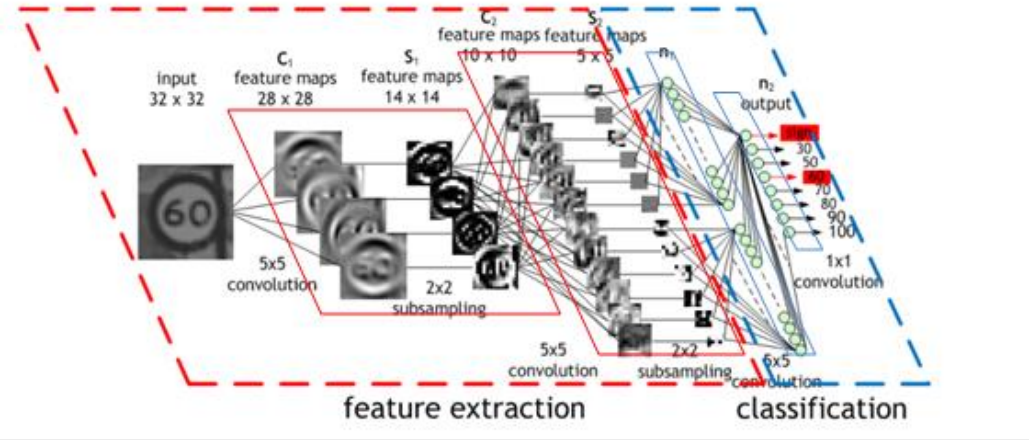
CNN卷積神經網絡

- CNN是所有深度學習課程、書籍必教的模型(Model)
- CNN在影像識別上，實用性及其應用處理非常強大，許多影樣辨識的模型也都是以CNN的架構為基礎做延伸
- CNN模型是參考人的大腦視覺組織來建立的深度學習模型，學會CNN之後，對於學習其他深度學習的模型會很大的幫助

CNN的概念圖

- 圖片經過各兩次的Convolution, Pooling, Fully Connected

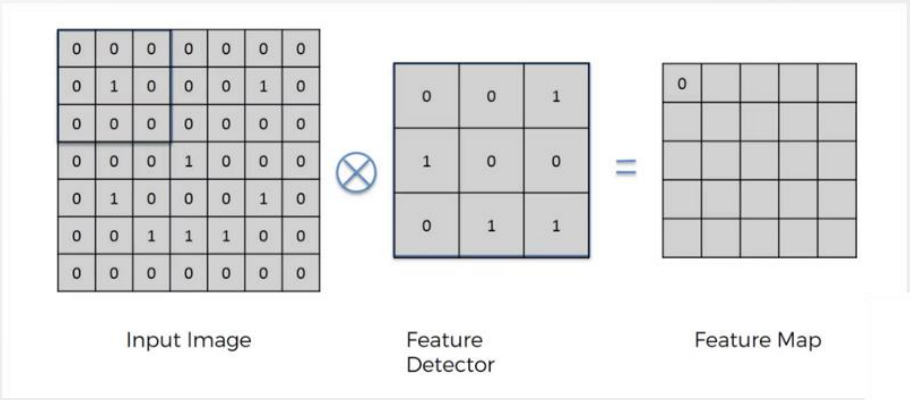




Convolution Layer卷積層

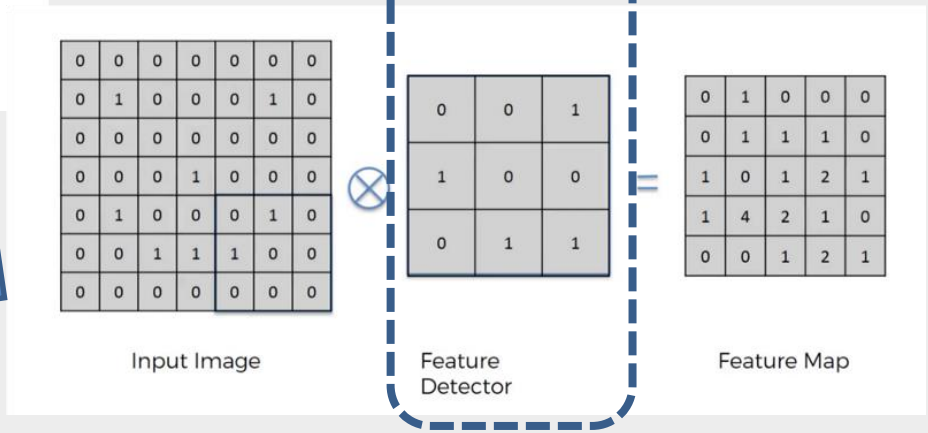
- 卷積運算就是將原始圖片的與特定的Feature Detector(filter)做卷積運算(⊗)

0*0 + 0*0 + 0*1+ 0*1 + 1 *0 + 0*0 + 0*0 + 0*1 + 0*1 =0

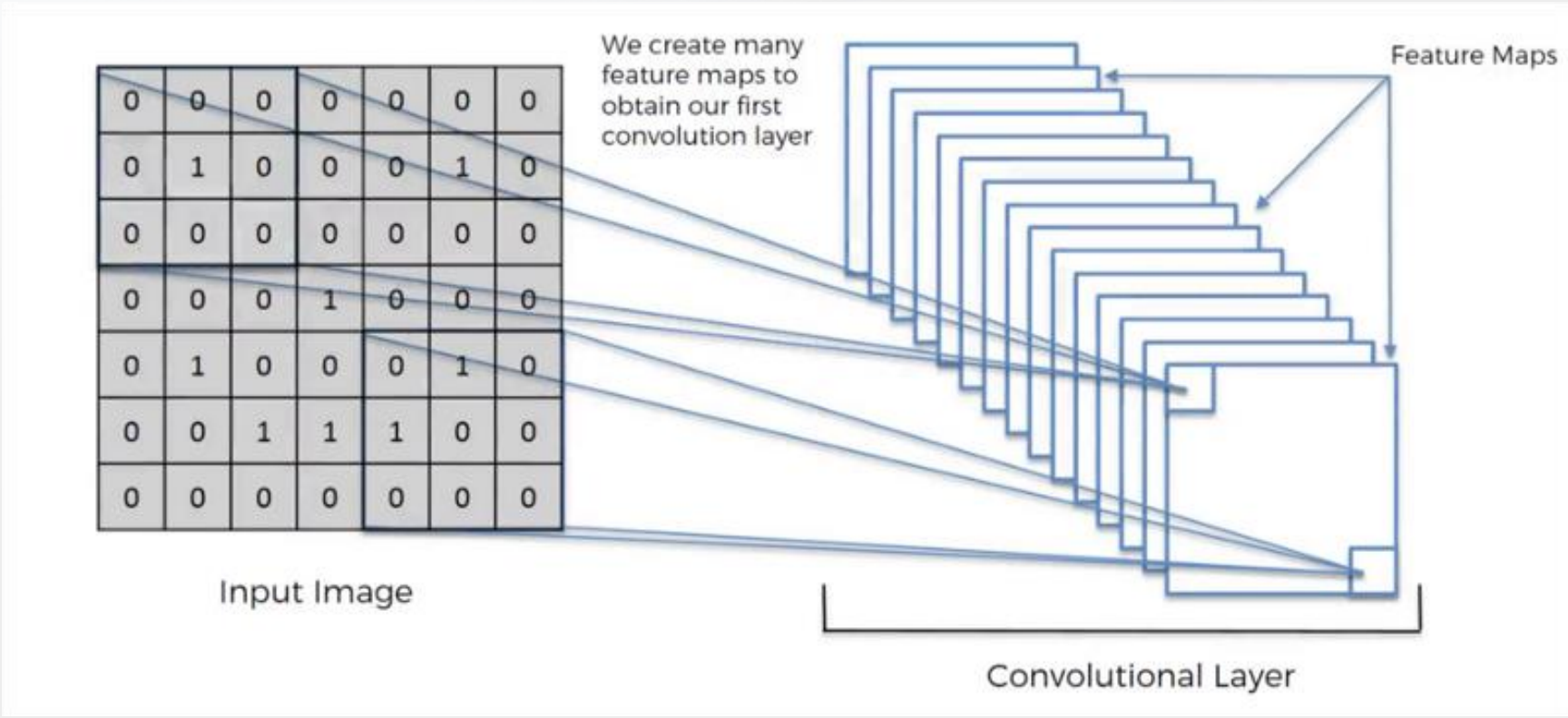


隨機產生
ex: 16種

目的:
萃取圖片的特徵



Input Image -> Feature Maps

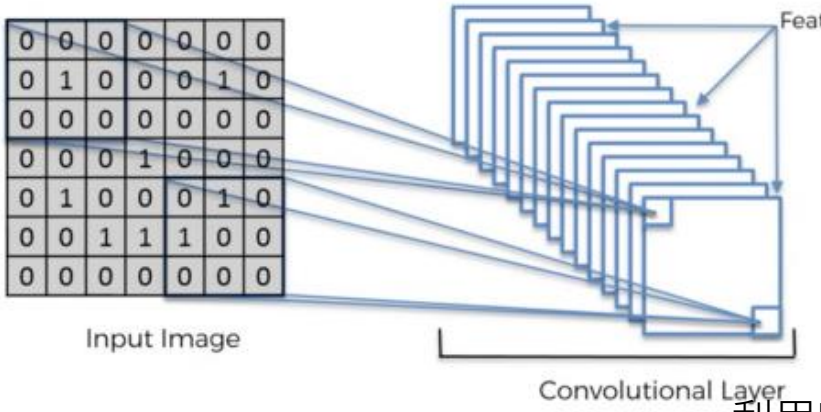


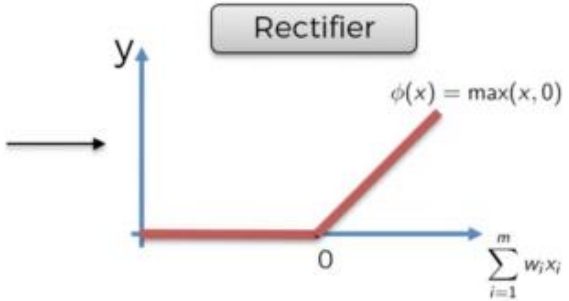
Feature Detector : 邊界(edge)

$*$

1	0	-1
2	0	-2
1	0	-1

利用Feature Detector萃取出物體的邊界



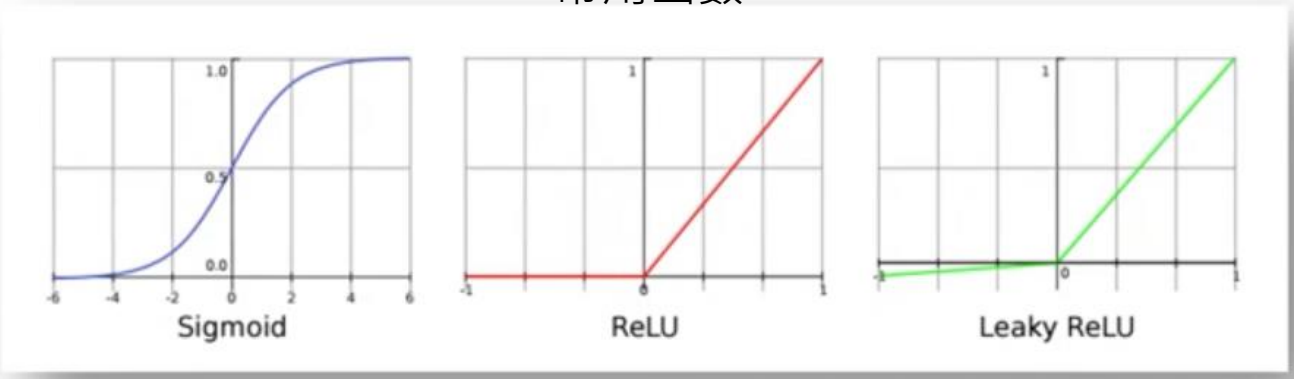


利用Relu函數去掉負值，更能淬煉出物體的形狀

Feature Detector : 邊界(edge)

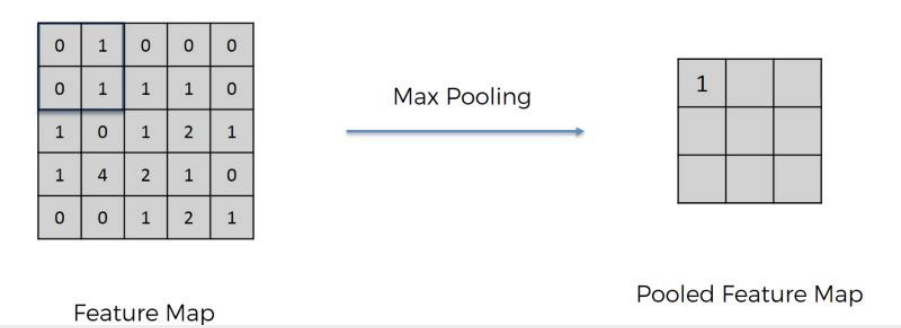
ReLU

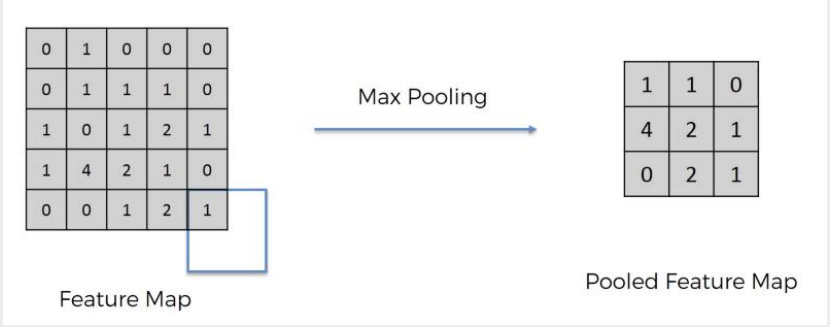
常用函數



Pooling Layer 池化層

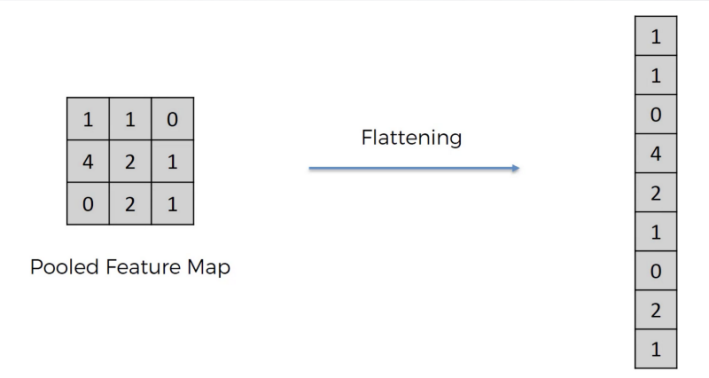
- 是一種形式的降採樣
- 採用Max Pooling，挑出矩陣當中的最大值，主要的好處是當圖片整個平移幾個Pixel的話對判斷上完全不會造成影響，以及有很好的抗雜訊功能
- 過去，平均池化的使用曾經較為廣泛，但是最近由於最大池化在實作的表現更好，平均池化已經不太常用。

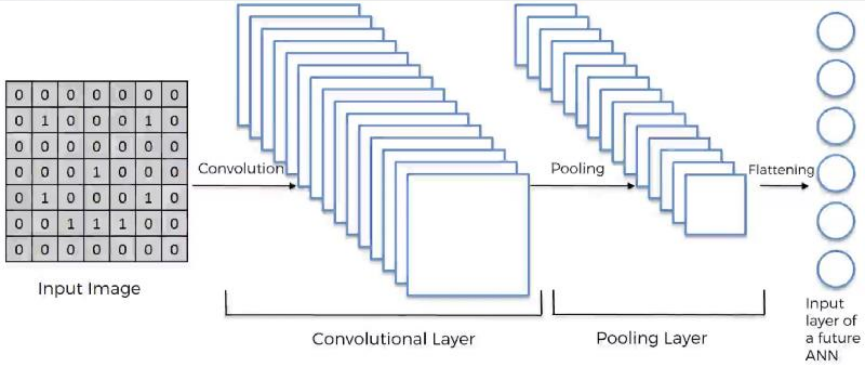


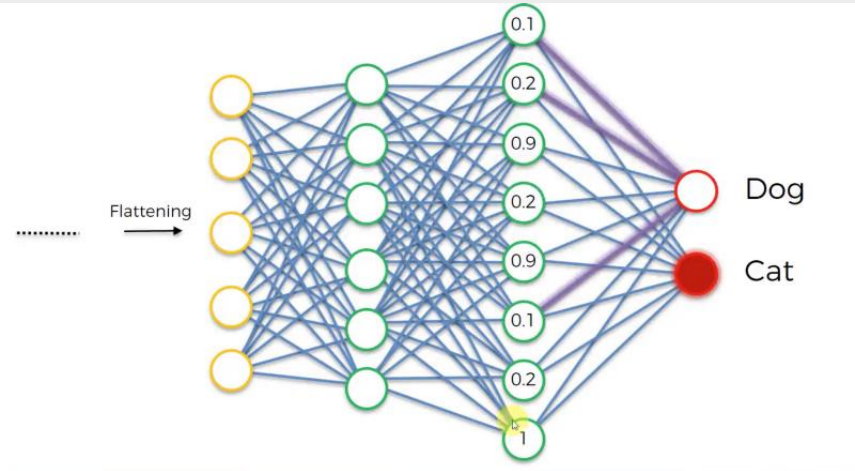


Fully Connected Layer 全連接層

- 全連接層就是將pooling的結果平坦化後，接到最基本的神經網絡







個案介紹