機器學習期末報告

題目

慢性腎臟疾病預測

姓名：莊璧如

學號：10722453

**研究目的**:

* 大致描述你建立的模型能做甚麼

**使用資料集來源**:

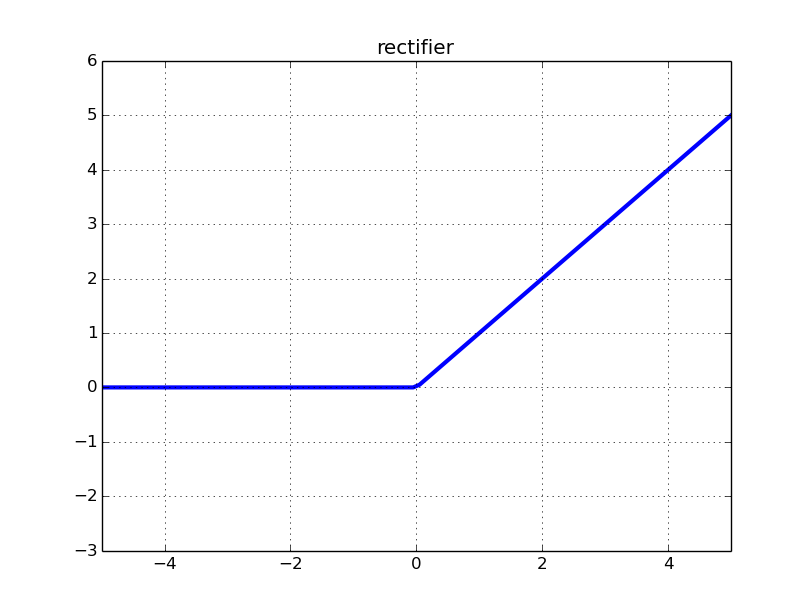
* Kaggle的網址<https://www.kaggle.com/datasets/mahmoudlimam/preprocessed-chronic-kidney-disease-dataset>
* 資料總筆數:401
* X有哪些欄位: Age (yrs)年齡/ Blood Pressure (mm/Hg)血壓/ Specific Gravity比重/Albumin白蛋白/Sugar糖/Blood Glucose Random (mgs/dL)隨機血糖(豪當量/公升)/ Blood Urea (mgs/dL) 血尿素/ Serum Creatinine (mgs/dL) 血清肌酐/ Sodium (mEq/L) 鈉/ Potassium (mEq/L) 鉀/ Hemoglobin (gms) 血紅蛋白 /Packed Cell Volume紅血球容積/White Blood Cells (cells/cmm)白血球細胞/ Red Blood Cells (millions/cmm)紅血球細胞/ Red Blood Cells: normal正常紅血球細胞/Pus Cells: normal膿細胞：正常/Pus Cell Clumps: present膿細胞團塊：現在/Bacteria: present細胞:現在/Hypertension: yes高血壓：是/Diabetes Mellitus: yes糖尿病：是/Coronary Artery Disease: yes冠狀動脈疾病：是/Appetite: poor食慾:不好/Pedal Edema: yes踏板水腫：是/Anemia: yes貧血：是
* Y 是標籤即類別，有幾類用甚麼表示?文字還是0 1 2? Chronic Kidney Disease: yes 慢性腎臟病:是 此資料以0:表示無/1:表示有
* 有無NAN?無
* 有沒有任何文字內容的欄位，需要處理的? 轉換成UINT8還是刪除?無

**使用方法**:

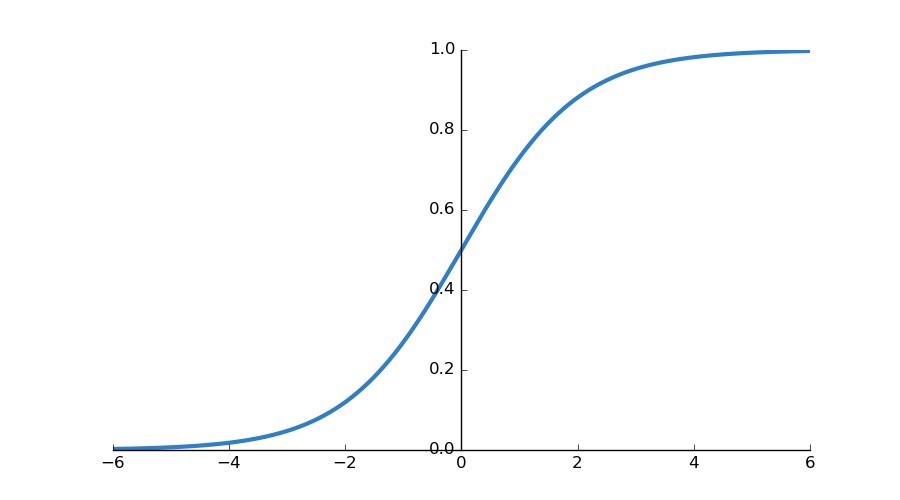
使用什麼演算法? ANN/KNN

大致介紹一下這個演算法，查一下列舉幾個和這個演算法類似的演算法名稱。

ANN 是Artificial Neural Network 會友輸入層，中間隱藏層，最後輸出層，隱藏層可以是多層，越多隱藏層，效果越好，值是輸入成以權重，中間會用激勵函數，像是relu或 sigmoid，如下圖，relu是y大於0我就保留，y小於0我就丟掉拿sigmoid來說，對sigmoid微分後，只有在[0-1]的範圍內有梯度更新，其餘都是0，所以在y > 1 or y < 0的區段是「沒有梯度的」，這就是 **梯度消失**



relu 1



sigmoid 1

與ann類似的演算法有Convolutional Network(CNN)/ RNN/ LSTM/ GAN

cnn卷積神經網絡是一種特殊類型的人工神經網絡，它在至少一個層中使用稱為[卷積的數學運算代替一般矩陣乘法。](https://en.wikipedia.org/wiki/Convolution)它們專門設計用於處理像素數據，並用於圖像識別和處理。CNN 也是模仿人類大腦的認知方式，譬如我們辨識一個圖像，會先注意到顏色鮮明的點、線、面，之後將它們構成一個個不同的形狀(眼睛、鼻子、嘴巴...)，這種抽象化的過程就是CNN演算法建立模型的方式。

RNN 是一個擁有稱為 LSTM 這種活性資料記憶體的神經網路，可以用於一系列資料以猜測接下來會發生的內容。某些層的輸出內容透過 RNN 送回到前一層當成輸入項，如此一來便建立起回饋迴路。

生成對抗網路（英語：Generative Adversarial Network，簡稱GAN）是[非監督式學習](https://zh.wikipedia.org/wiki/%E9%9D%9E%E7%9B%91%E7%9D%A3%E5%BC%8F%E5%AD%A6%E4%B9%A0)的一種方法，通過讓兩個[神經網路](https://zh.wikipedia.org/wiki/%E4%BA%BA%E5%B7%A5%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C)相互[博弈](https://zh.wikipedia.org/wiki/%E5%8D%9A%E5%BC%88%E8%AE%BA)的方式進行學習。

實**驗設計:**

列舉要調整那些參數來觀察其產生結果的不同(各種可能的參數，越多越好．例如

* 訓練集和測試集不同比例

1. 訓練集比測試集(8:2) 正確率97.5%

2.訓練集比測試集(2:8)正確率98.4375%

* 有無使用cross validation以及其不同份量的切法

平均正確率0.99375

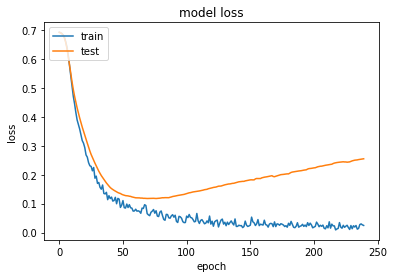
正確率標準差: 0.012499999999999999

* 其他你使用方法特定的參數

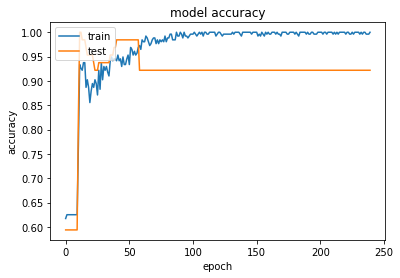
將epoch從240改為50，其他不變情況下，正確率65%

**實驗結果與討論：**

使用表格或圖形來視覺化你的實驗結果

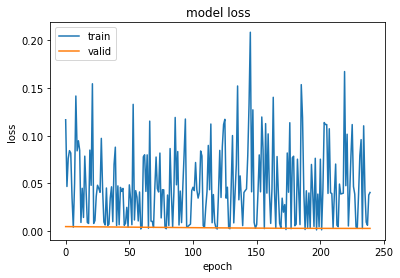
Loss曲線圖(訓練集比測試集8:2)

loss曲線圖 1

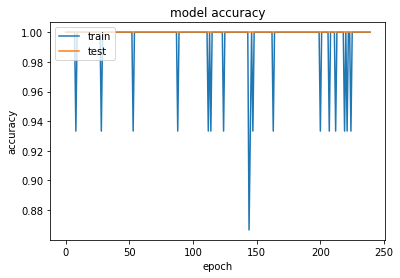


正確率 1

訓練集比測試集(2:8)



loss曲線圖 2

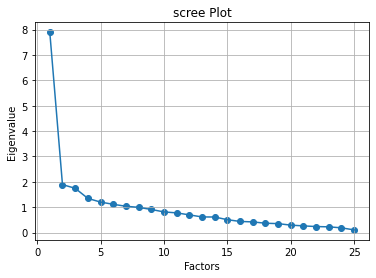


正確率 2

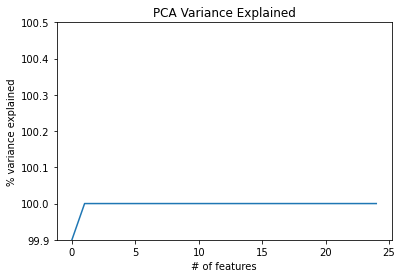
針對實驗結果提出一些想法

運算結果顯示，在訓練集:測試集=8:2時，當train 正確率的低點是在0.996，顯然是相當高的，有關於loss曲線圖，在epoch=50與誤差=0.1之後的train loss持續下降，然而test loss開始上升，這顯示了overfitting，而當訓練集比測試集為2:8時，epoch為0-250沒有overfitting的現象，代表這個模型適合用來做腎臟疾病的預測

scree plot



scree plot 1

PCA Variance Explained

PCA Variance Explained 1

接著看scree plot，在這個資料中，影響的特徵值大於等於1的有8個，總共有25個因子