資料探勘 HW2

R07725010 徐嘉琪, R07725043 陳奕瑄, R07725021 洪靖雯, R07725013 張睿君

1. 研究動機與目的

不只在商業界蓬勃發展，資料分析也能運用於醫療、氣象、體育等多元領域，其中癌症診治便是醫療領域中一個應用的重點。癌症若於早期發現，治癒率能達到 65% 以上，在各種癌症當中，乳癌診治的進步可算相當迅速，而乳癌診斷技術的進步又比治療的進步還快。但截至目前為止，即使有高超的乳房理學檢查技巧，配以精密的乳房影像儀器，仍無法在沒有病理組織證實是乳癌的情況下，就對病患採取乳癌的治療。病患辛苦地做了切片手術後，除了醫生本身的經驗判斷，若可以在大量切片影像資料的基礎下使用分析模型，便能幫助醫生提供更快速精準的癌症腫瘤良惡性判斷。

醫療領域的研究關乎著人類的生命健康，平時較無機會接觸此領域，藉由選擇相關資料，不僅能找出這樣的資料能嘗試的分析方法，也能了解該領域資料的意義、型態，對應用資料分析於該領域略窺一二。

2. 資料描述

與 HW1相同，對於初次接觸相關領域，選擇較不複雜且容易取得的 Scikit Learn內建 Wisconsin 乳癌資料集，由細針抽吸細胞診斷 (fine needle aspiration cytology, FNAC) 的切片影像轉換而成，共有半徑、紋理、圓滑度、緊密度等 10 種乳房腫瘤的特徵。總共有 569 筆腫瘤資料，212 筆屬於惡性腫瘤，357 筆屬於良性腫瘤。每筆資料皆有 30 個特徵欄位，以診斷結果作為分析應變數，30 個特徵作為自變數，關於各變數的定義如下：

	變數名稱	變數說明
Y	無（資料儲存形式為 numpy）	惡性(M)： 1 良性(B)：0
平均值相關自變數	mean radius	平均半徑
	mean texture	平均紋理（灰階的標準偏差）
	mean perimeter	平均周長
	mean area	平均區域大小

	mean smoothness	平均平滑度
	mean compactness	平均緊密度
	mean concavity	平均凹度（輪廓凹部的嚴重性）
	mean concave points	平均凹點數量（輪廓的凹入部分的數量）
	mean symmetry	平均對稱度
	mean fractal dimension	平均碎形維數
標準差相關自變數	radius error	半徑標準差
	texture error	紋理（灰階的標準偏差）標準差
	perimeter error	周長標準差
	area error	區域大小標準差
	smoothness error	平滑度標準差
	compactness error	緊密度標準差
	concavity error	凹度（輪廓凹部的嚴重性）標準差
	concave points error	凹點數量（輪廓的凹入部分的數量）標準差
	symmetry error	對稱度標準差
	fractal dimension error	碎形維數標準差
最大值相關自變數	worst radius	最大半徑
	worst texture	最大紋理（灰階的標準偏差）
	worst perimeter	最大周長
	worst area	最大區域大小
	worst smoothness	最大平滑度
	worst compactness	最大緊密度
	worst concavity	最大凹度（輪廓凹部的嚴重性）

	worst concave points	最大凹點數量（輪廓的凹入部分的數量）
	worst symmetry	最大對稱度
	worst fractal dimension	最大碎形維數

3. 實驗方法

a. 使用語言及套件

i. 語言：python

ii. 套件：
sklearn.model_selection-train_test_split;
sklearn.tree-DecisionTreeClassifier;
sklearn.tree-export_graphviz,pydotplus; IPython.display-Image;
sklearn.externals.six-StringIO;
sklearn.neural_network-MLPClassifier; sklearn.svm-SVC;
numpy; matplotlib.pyplot; matplotlib.mpl;
sklearn.metric-accuracy_score, recall_score, precision_score, f1_score, confusion_matrix, roc_curve, auc

b. 資料分割

與 HW1 相同，將原先資料隨機以 2:1 的比例切成 training 和 test dataset，並存成 X_train、y_train、X_test、y_test，X 表示用來預測的 feature，y 表示預測值也就是是否為惡性腫瘤。

c. 分類演算法

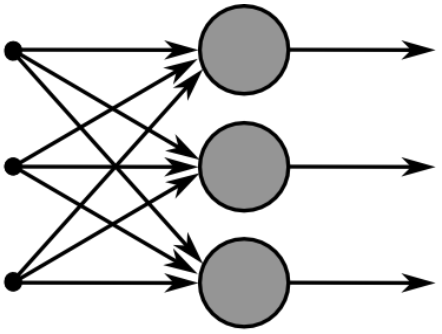
i. Decision Tree Induction

決策樹是現今數一數二執行起來非常有效率，模型又相當簡單的監督式學習模型，以 binary tree 為基底，將資料依據資料特徵做分類，讓相似的資料被歸為同一類。

我們運用 SKlearn 的 DecisionTreeClassifier 套件來做決策術分類，將 X_train 與 y_train 丟入模型中學習。使用上課提到的 entropy 來做分類依據，並限制決策樹高度最高只能長到 3 層，以避免樹過度歪斜造成 overfitting。

ii. Multilayer Neural Network

Neural Network 是最基本的神經元網路形式，由有限個神經元構成，所有神經元的輸入向量都是同一個向量。由於每一個神經元都會產生一個純量結果，所以單層神經元的輸出是一個向量，向量的維數等於神經元的數目。



output layer

運用 SKlearn 的 MLP 套件來做分類，可以做多層的神經元，此時就會分成

- 輸入層（Input layer），眾多神經元接受大量非線形輸入訊息。輸入的訊息稱為輸入向量。
- 輸出層（Output layer），訊息在神經元連結中傳輸、分析、權衡，形成輸出結果。輸出的訊息稱為輸出向量。
- 隱藏層（Hidden layer）是輸入層和輸出層之間眾多神經元和連結組成的各個層面，隱層可以有一層或多層。

隱層可以有一層或多層。隱層的節點數目不定，但數目越多神經網路的非線性越顯著，從而使神經網路的 robustness 更顯著。

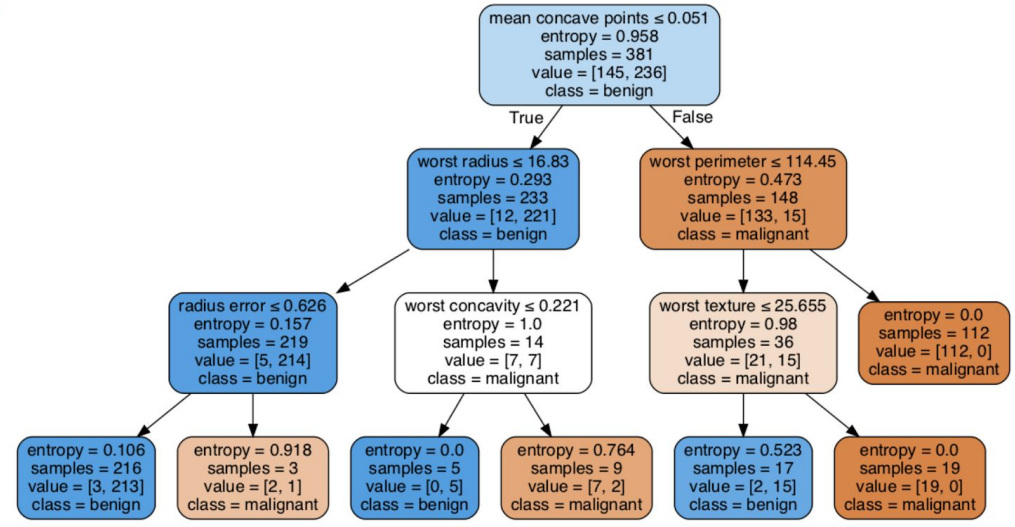
iii. SVM

SVM 是一種監督式的學習方法，將資料投影到高維度，使用不同 kernel function 找出一個分類的超平面，並找出一個決策邊界讓兩類之間的 margins 最大化，使其可以完美區隔開來。本研究運用 SKlearn 的 SVM 套件來做分類，考量資料特性，並在嘗試過不同 kernel function 後，我們使用表現最好的 linear kernel function 做線性切割。首先利用 training data 訓練出一個 SVM 模型，接著將 test data 帶入此模型預測資料的分類。

4. 實驗結果與分析

a. Decision Tree Induction

試著將決策樹結果視覺化，運用 SKlearn 的 export_graphviz，將決策樹模型丟入套件，並指定分類 class 的名稱。依據 SKlearn 的 cancer dataset 描述，資料本身會分成「malignant（惡性）」、「benign（良性）」兩類，以此做為 export_graphviz 套件的 class name。將 export_graphviz 的結果輸出成向量圖，並用 pydotplus、Image 套件來讀取向量圖並輸出成圖片。結果如下：



每個 node 會顯示依據何條件來做分類，由於所選資料僅分兩類，因此第一層即顯示結果，root node 依據 cancer data 的 mean concave

points ≤ 0.051 將資料分成兩類，true 為 benign、false 為 malignant；entropy 值為 0.958，總共有 381 筆資料，其中有 145 筆為 benign、236 筆為 malignant。

b. Multilayer Neural Network

使用架構為兩層 hidden layer 的神經網路結構，神經元數第一、二層各為 5 和 2，過程中測試了：

- 一層，神經元數 2~10
- 兩層，神經元數各 2~10 之組合

其中以兩層神經元數 (5, 2) 的結果最好因此採用此架構。

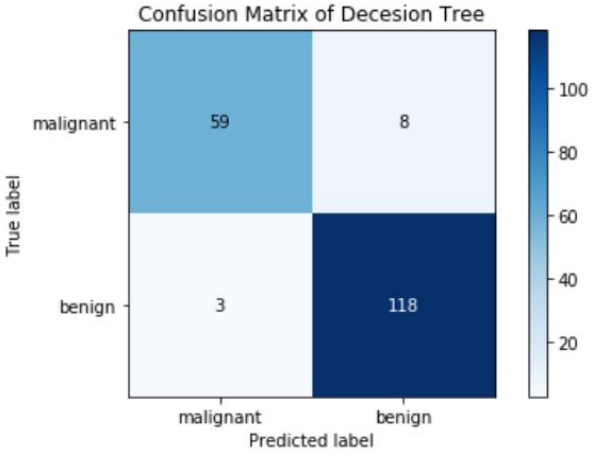
c. SVM

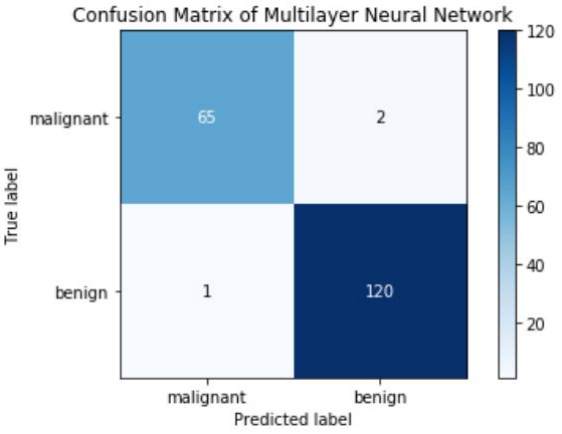
使用 Linear、RBF 及 Sigmoid 三種 kernel function 之比較如下表，因 Linear kernel function 表現最佳，最適合該資料，故採用 Linear kernel function 與其他分類方法比較。

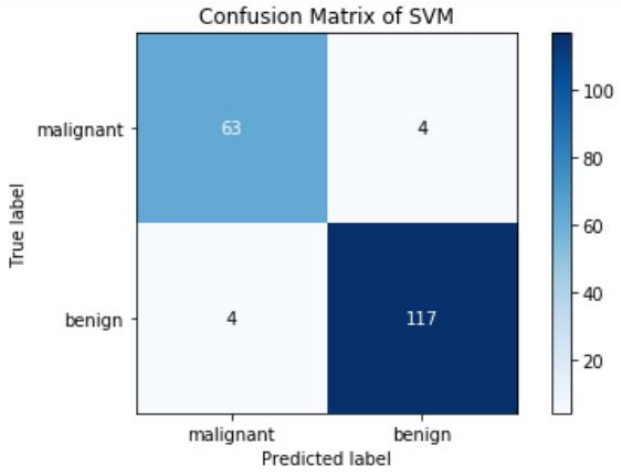
	Linear	RBF	Sigmoid
Accuracy	0.9574	0.6436	0.6436
Precision	0.9669	0.6436	0.6436
Recall	0.9669	1.0	1.0
F1-measure	0.9669	0.7832	0.7832

d. 分類結果比較

利用三種方法將資料分類，並分析分類後的 Confusion matrix、Accurracy、Precision、Recall、F1-measure 以及繪出 ROC Curve，分析結果統整如下：

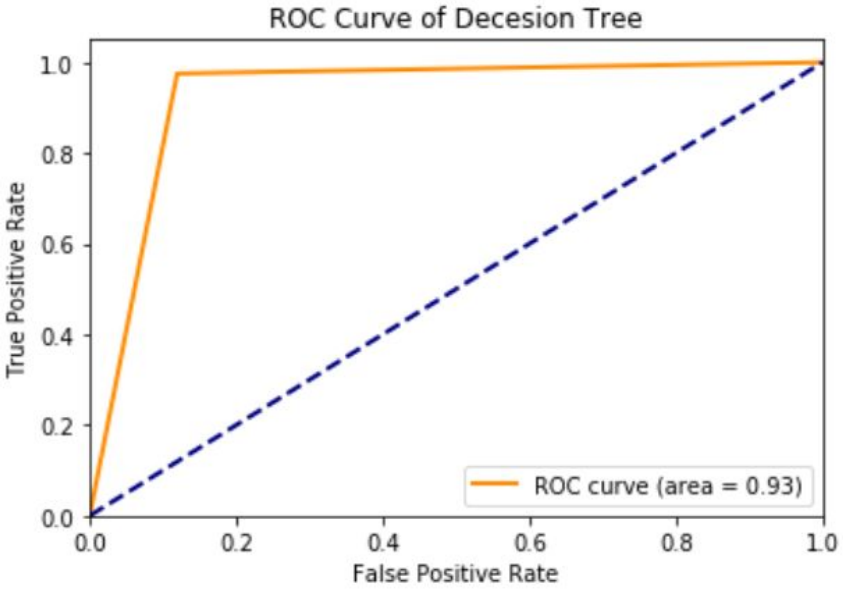




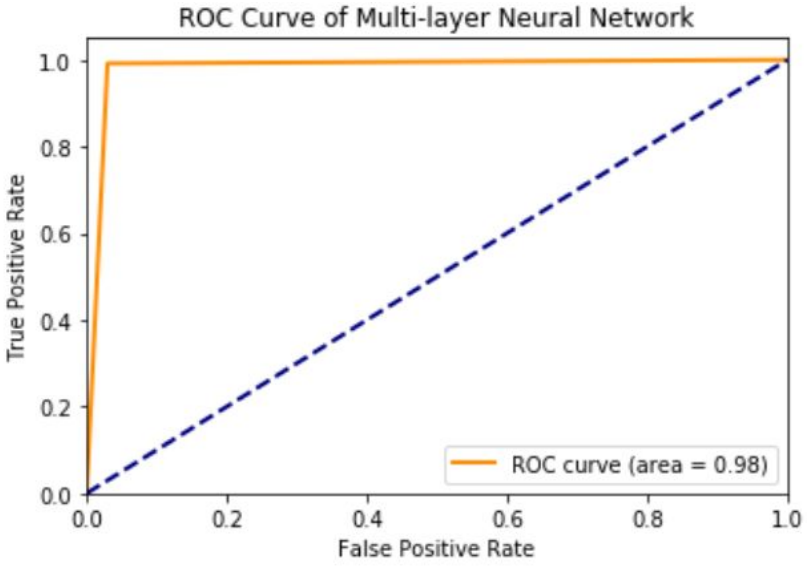


	Decision Tree Induction	NN	SVM
Accurracy	0.9414	0.9840	0.9574
Precision	0.9365	0.9836	0.9669
Recall	0.9752	0.9917	0.9669
F1-measure	0.9555	0.9877	0.9669

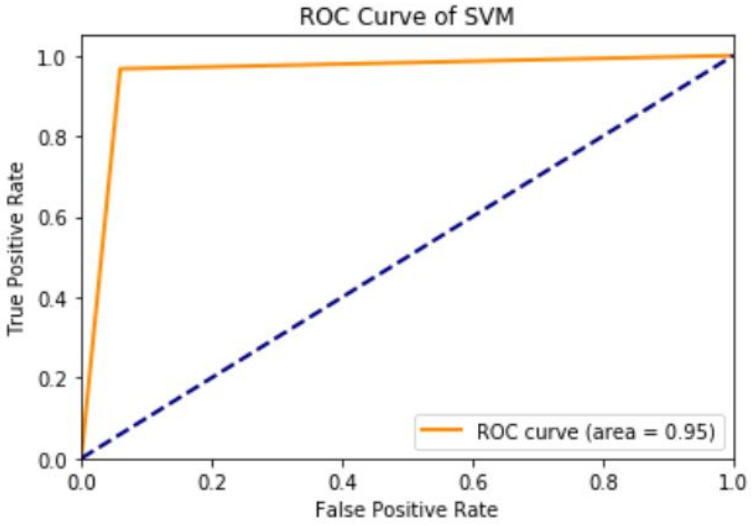
- Decision Tree Induction ROC curve



- Multilayer Neural Network ROC curve



- SVM ROC curve



5. 結論

根據結果，不同分類方式皆有超過 0.9 的準確率，其中使用 Decision Tree Induction 方法的表現較差，推測因為所選資料只分成兩類，故能作為分類判斷依據的 attribute 僅一個，無法充分發揮此方法的效用。而此次使用 multi-layer 的 Neural Network 分類器，由於多了 hidden layer 協助分類，且所選資料維度高，相較 SVM 分類器能提供更準確的分類結果。

無論是使用 Neural Network 或 SVM，皆需考量到分類相關參數，如 hidden layer 層數、每層節點數量、kernel function 種類，選擇參數不同，極可能造成分類結果相當大的差異，而誤判不同分類器的好壞。因此將資料分類時，除了嘗試不同分類方法，也應盡可能嘗試不同參數或 kernel function，才有辦法找出最適合的分類模型。

6. 參考資料

a. 【如何選擇取得細胞或組織的方法以診斷乳癌】

http://www.mmh.org.tw/taitam/gen_su/index4_2_1e.html

b. Breast Cancer Wisconsin (Diagnostic) Data Set

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

c. 決策樹(Decision Tree)以及隨機森林(Random Forest)介紹

https://medium.com/jameslearningnote/%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E7%AC%AC3-5%E8%AC%9B-%E6%B1%BA%E7%AD%96%E6%A8%B9-decision-tree-%E4%BB%A5%E5%8F%8A%E9%9A%A8%E6%A9%9F%E6%A3%AE%E6%9E%97-random-forest-%E4%BB%8B%E7%B4%B9-7079b0ddfbda

d. Creating and Visualizing Decision Trees with Python

https://medium.com/@rnbrown/creating-and-visualizing-decision-trees-with-python-f8e8fa394176

e. 人工神經網路

https://zh.wikipedia.org/wiki/%E4%BA%BA%E5%B7%A5%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C

f. SVM

https://medium.com/@chih.sheng.huang821/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E6%94%AF%E6%92%90%E5%90%91%E9%87%8F%E6%A9%9F-support-vector-machine-svm-%E8%A9%B3%E7%B4%B0%E6%8E%A8%E5%B0%8E-c320098a3d2e