類神經網路期末報告

- 王斯暘
- 徐文彥
- 林泓志

壹、緒論

貳、方法簡介

a. 單純貝式分類器 Naive Bayes' Classifier
b. 支持向量機 Support Vector Machines
c. 反向傳播網路Backpropagation Neuron

參、研究方法

a. 資料來源與處理
b. 評估指標
c. 研究方法

肆、討論與結論

參考論文

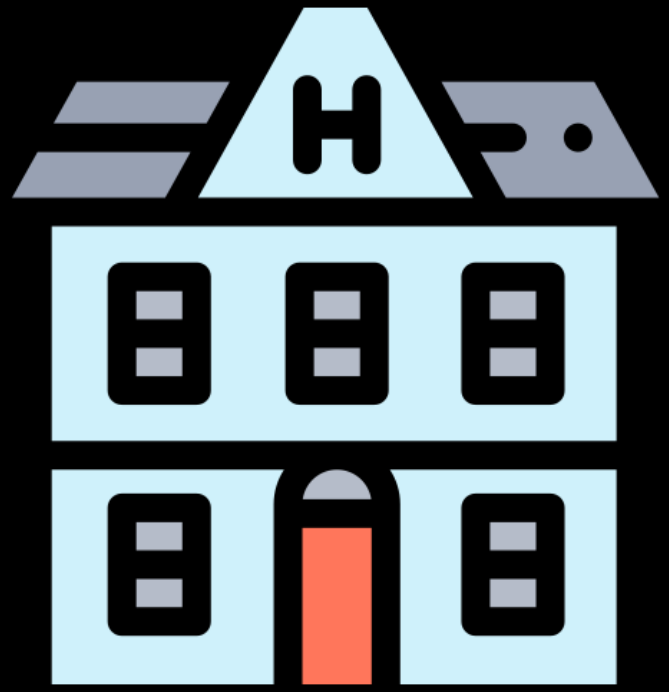
A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services

採用Naïve Bayes' Classifier：

- 針對旅客入住旅館的線上評論進行分析

分為正向評論或負面評論

- 找出評論中對旅客最具影響力之特徵以利旅館管理者找出問題並改進



貳、方法簡介

a. 單純貝式分類器 Naive Bayes' Classifier

b. 支持向量機 Support Vector Machines

c. 反向傳播網路 Backpropagation Neuron

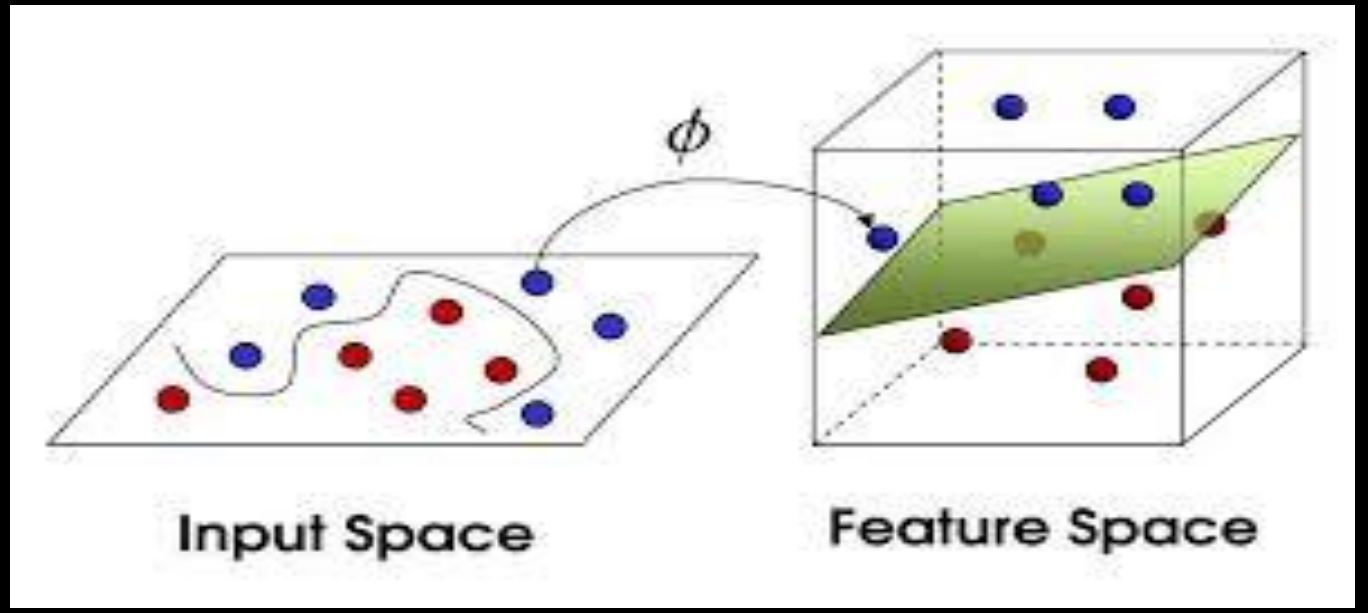
貳、方法簡介

採單純貝式分類器 NAIVE BAYES' CLASSIFIER

- 基於貝氏定理 (Bayes’ Theorem)
- 假設特徵彼此獨立
- 優點：計算量低、實作簡單
- 缺點：特徵不獨立時易受影響

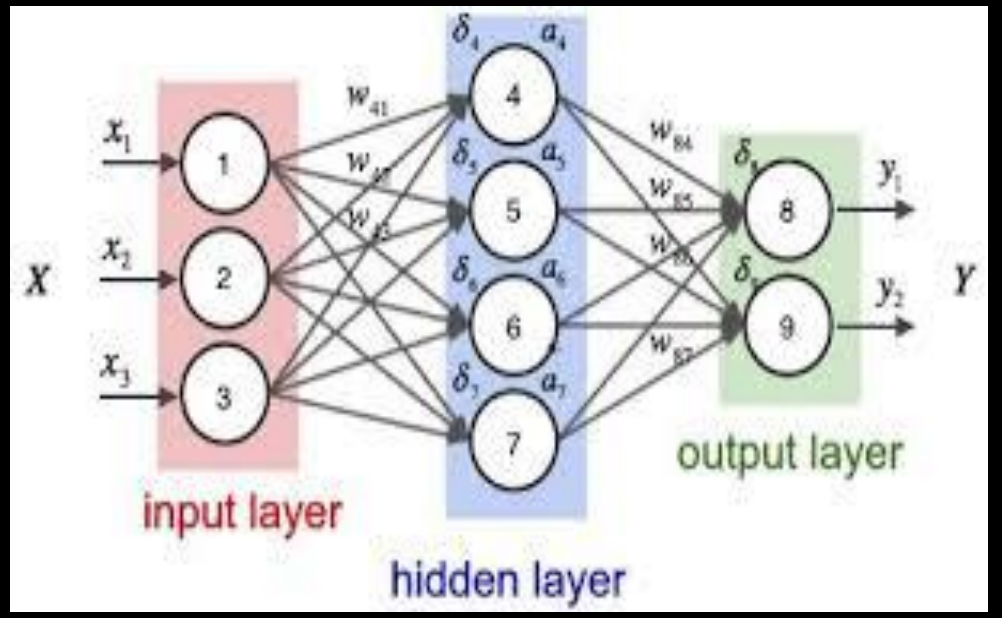
支持向量機 SUPPORT VECTOR MACHINES

- 在高維空間找最佳分割超平面
- 常見函式：Linear, Poly, RBF, Sigmoid
- 優點：對高維/小樣本表現良好
- 缺點：參數(theta, kernel)需調整，花時間



反向傳播網路BACKPROPAGATION NEURON

- 多層感知器 (MLP) + 誤差倒傳遞
- 可處理非線性映射
- 優點：表現靈活
- 缺點：訓練時間較長、易陷局部極小



參、研究方法

a. 資料來源與處理

b. 評估指標介紹

c. 研究方法

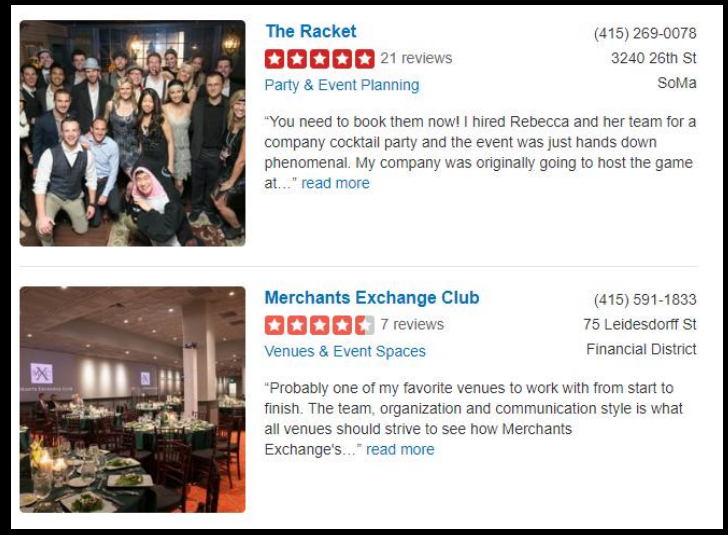
資料來源與處理

公司：Yelp

特色：成熟的評論社群

豐富的數據庫

成就：2018世界品牌500強

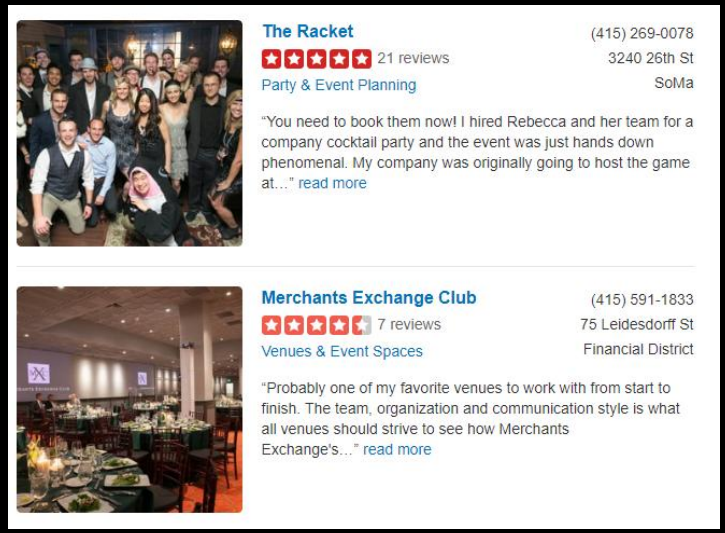


資料來源與處理

資料屬性 ：［評論］，［類別］

資料筆數 ： 26251筆

分析目標 ： 滿意度為正或負



資料來源與處理

```
In [91]: #NBC
         Naive = naive_bayes.MultinomialNB()
         Naive.fit(Train_X_Tfidf,Train_Y)
         predictions_NB = Naive.predict(Test_X_Tfidf)
         print("Naive Bayes Accuracy Score -> ",accuracy_score(predictions_NB, Test_Y)*100)

         Naive Bayes Accuracy Score ->  1.9447287615148412

In [92]: #SVM
         SVM = svm.SVC(C=1.0, kernel='linear', degree=3, gamma='auto')
         SVM.fit(Train_X_Tfidf,Train_Y)
         predictions_SVM = SVM.predict(Test_X_Tfidf)
         print("SVM Accuracy Score -> ",accuracy_score(predictions_SVM, Test_Y)*100)

         SVM Accuracy Score ->  1.9191402251791199
```

網站原始資料做分類 ->準確度極低

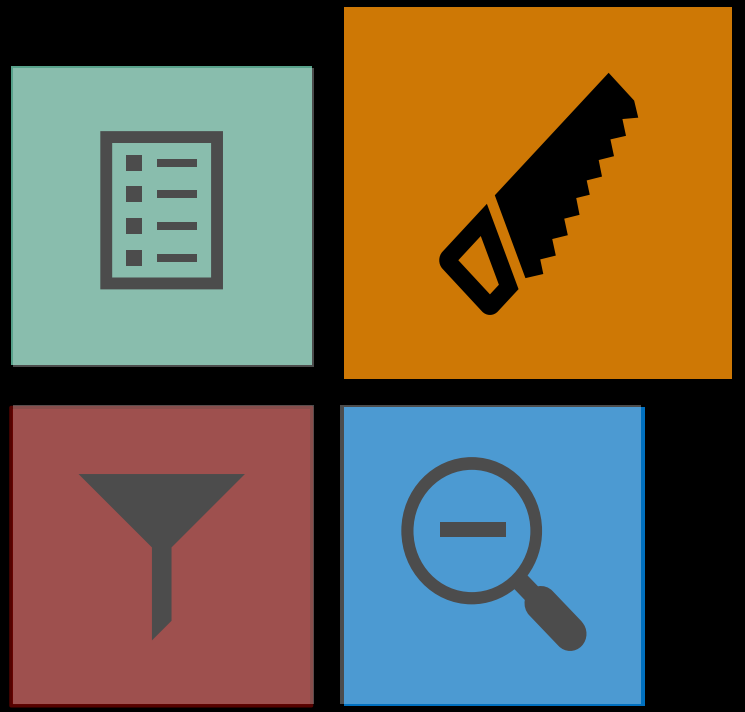
無意義的斷詞

冗餘的文字特徵量

資料處理步驟

1. 將評論中英文大寫之部分，全部轉換為小寫
2. 將無用之單字去除，除了系統內建之過濾單字外，我們還新增原始資料中常見之無效單字之過濾網
3. 針對過濾完之資料進行正規表示式之處理，此方法可將資料中之數字及標點符號除去
4. 將資料中單字間之空格去除

資料處理步驟



評論斷詞

使用NLTK套件

以單詞為單位

1. 將評論中英文大寫之部分，全部轉換為小寫
2. 將無用之單字去除，除了系統內建之過濾單字外，我們還新增原始資料中常見之無效單字之過濾網
3. 針對過濾完之資料進行正規表示式之處理，此方法可將資料中之數字及標點符號除去
4. 將資料中單字間之空格去除

資料處理步驟

轉換大小寫

意義不同

格式統一

1. 將評論中英文大寫之部分，全部轉換為小寫
2. 將無用之單字去除，除了系統內建之過濾單字外，我們還新增原始資料中常見之無效單字之過濾網
3. 針對過濾完之資料進行正規表示式之處理，此方法可將資料中之數字及標點符號除去
4. 將資料中單字間之空格去除

資料處理步驟

文字過濾

自訂義字袋 X NLTK

過濾冗餘字詞

1. 將評論中英文大寫之部分，全部轉換為小寫
2. 將無用之單字去除，除了系統內建之過濾單字外，我們還新增原始資料中常見之無效單字之過濾網
3. 針對過濾完之資料進行正規表示式之處理，此方法可將資料中之數字及標點符號除去
4. 將資料中單字間之空格去除

資料處理步驟

正規表示式

過濾隱藏符號與數字

1. 將評論中英文大寫之部分，全部轉換為小寫
2. 將無用之單字去除，除了系統內建之過濾單字外，我們還新增原始資料中常見之無效單字之過濾網
3. 針對過濾完之資料進行正規表示式之處理，此方法可將資料中之數字及標點符號除去
4. 將資料中單字間之空格去除

評估指標介紹

- **混淆矩陣**，混淆矩陣是用來評價算法或者說分類器的結果分析表。其每一列代表預測值，每一行代表的實際值。

	P	N
P'	TP	FP
N'	FN	TN

- True Positive(真正，TP)：將正類預測為正類數
- True Negative(真負，TN)：將負類預測為負類數
- False Positive(假正，FP)：將負類預測為正類數誤報 (Type I error)
- False Negative(假負，FN)：將正類預測為負類

評估指標介紹

- Accuracy:

為分類之準確度，計算方法為分類正確之資料筆數佔全部測試資料筆數之百分比，是評估分類的一項重要指標。

- Precision:

指在所有被預測為正的樣本中實際為正的樣本的概率，

Precision=TP / (TP+FP)，即在所有被預測為陽性的測試資料中，真正是陽性的比率。

- Recall:

指在實際為正的樣本中被預測為正樣本的概率，計算公式為:

Recall=TP / (TP+FN)，

即在所有實際為陽性的測試資料中，真正是陽性的比率。

評估指標介紹

F1 score ：

綜合Precision與Recall指標

找一平衡點，達到最大值

$$F_\beta = (\beta^2 + 1) * \frac{PR}{\beta^2 * P + R}$$

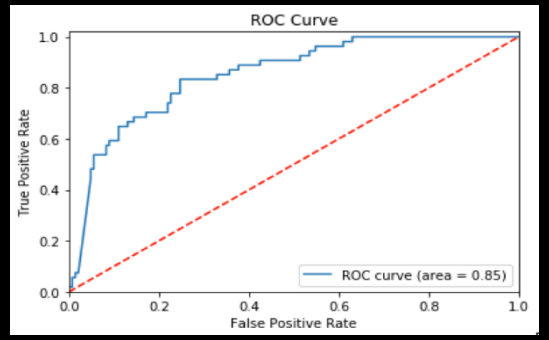
挑整參數β提升

精確率的權重

AUC-ROC ：

TPR與FPR之間的關係

越往左上敏感度越高



評估指標介紹

Matthews correlation coefficient :

測量二分類的分類性能的指標，其本質上是描述實際分類與預測分類之間的相關係數。

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Kappa係數:

是一種衡量分類精度的指標判斷是否類別不平衡。

$$\kappa = \frac{(p_0 - p_c)}{(1 - p_c)}$$

		Actual			
		1	2	3	sum
Predicted	1	99.29%	0.20%	0.50%	99.99%
	2	0.00%	0.00%	0.00%	0.00%
	3	0.01%	0.00%	0.00%	0.01%
	sum	99.30%	0.20%	0.50%	

研究方法

- 單純貝式分類器
  - 採用Multinominal模型，算每一個關鍵字在單一樣本點出現的次數。
  - 文件分類中另有Binominal，以0,1進行資料分析
  - 由於論文中未詳細提及使用方法，因此我們先採用分類效果較好的Mulinominal模型。

	MaxFeature = 150	MaxFeature = 5000
Accuracy	88.7%	95.5%
Precision	87.1%	95.6%
Recall	91.8%	95.8%
AUC	88.6%	95.5%
MCC	0.7727	0.911
Kappa	0.7739	0.911
F-Measure	89.3%	95.7%

研究方法

- 支持向量機(kernel=poly)
  - SVM方面我們設定theta(Error penalty)為預設值。
  - 預期在準確率若低於論文的標準時再進行額外調整，Kernel從Poly測起。

Kernel = poly	MaxFeature = 150	MaxFeature = 5000
Accuracy	51.7%	95.5%
Precision	87.1%	95.6%
Recall	91.8%	95.8%
AUC	88.6%	95.5%
MCC	0.0	0
Kappa	0.0	0.0
F-Measure	68.2%	95.7%

研究方法

- 支持向量機(kernel=RBF)

Kernel = RBF	MaxFeature = 150	MaxFeature = 5000
Accuracy	90.1%	95.5%
Precision	90.1%	95.6%
Recall	90.2%	95.8%
AUC	90.0%	95.5%
MCC	0.8009	0.9101
Kappa	0.8010	0.9101
F-Measure	90.4%	95.7%

研究方法

- 支持向量機(kernel=sigmoid)

Kernel = Sigmoid	MaxFeature = 150	MaxFeature = 5000
Accuracy	89.6%	51.7%
Precision	90.0%	51.7%
Recall	90.0%	100%
AUC	89.7%	0.5%
MCC	0.7925	0.0
Kappa	0.7925	0.0
F-Measure	90%	68.2%

研究方法

- 支持向量機(kernel=linear)

Kernel = Linear	MaxFeature = 150	MaxFeature = 5000
Accuracy	90.6%	96.7%
Precision	91.1%	96.9%
Recall	90.7%	96.7%
AUC	90.6%	96.7%
MCC	0.81213	0.934
Kappa	0.81213	0.91009
F-Measure	90.9%	96.8%

研究方法

- 倒傳遞神經網路
  - Input layer我們採用Tf-idf取得之Max_Feature做為Input node個數。
  - 找出每個樣本出現的關鍵字個數。
  - Hidden layer則是先參考實做經驗與課本的建議，依照問題複雜程度與經驗逐一測試50、75、100、125個Hidden nodes之測試結果。
  - Hidden layer數量目前測試到兩層。

研究方法

BPN Model ; Feature = 150	Hid.lay = 1 ; nodes = (50)	Hid.lay = 1 ; nodes = (75)
Accuracy	47.9%	56.4%
Precision	47.3%	54.8%
Recall	37.9%	72.6%
MCC	-0.04288	0.135298
Kappa	0.168053	-0.47929
F-Measure	42.1%	62.5%

研究方法

BPN Model ; Feature = 150	Hid.lay = 1 ; nodes = (100)	Hid.lay = 1 ; nodes = (125)
Accuracy	46.1%	48.7%
Precision	45.5%	48.8%
Recall	39.8%	53.2%
MCC	-0.07863	-0.02611
Kappa	0.111901	-0.0989
F-Measure	42.5%	50.9%

研究方法

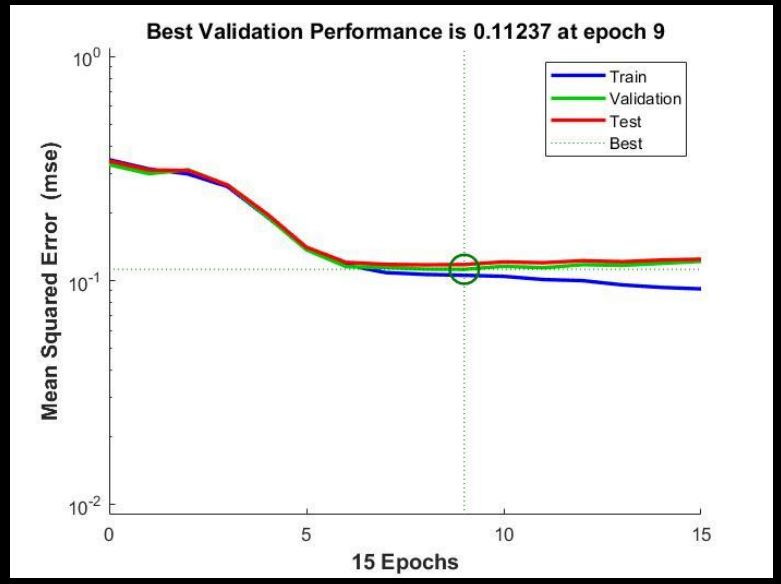
BPN Model ; Feature = 150	Hid.lay = 2 ; nodes = (75,25)	Hid.lay = 2 ; nodes = (100,25)
Accuracy	48.4%	52.7%
Precision	48.6%	52.3%
Recall	57.6%	61.2%
MCC	-0.03256	0.054798
Kappa	-0.022549	-0.20482
F-Measure	52.7%	56.4%

研究方法

BPN Model ; Feature = 46	Hid.lay = 1 ; nodes = (75)	Hid.lay = 2 ; nodes = (75,10)
Accuracy	77.1%	71.1%
Precision	78.7%	68.4%
Recall	74.0%	78.2%
MCC	0.540975	0.42632
Kappa	0.136604	-0.1655
F-Measure	76.8%	73%

討論與結論

- 測試結果下，SVM由Linear Kernel的運算結果最優。
- 效率上則是Naïve Bayes Classifier最快。
- BPN多次測試結果下，MSE皆無法低於一定水準。



討論與結論

- NBC
  - 在MAX Feature數量上升之後，準確度會顯著提升。
  - Feature之間違反條件獨立假設的可能上升，準確度會隨之下降。
- SVM
  - Poly 在Max Feature較少時，準確度會不高。
  - Sigmoid 可能因為Feature數量上升而有過度擬合的問題。
  - Linear 在不同Feature數量下，指標都在一定水準之上。
  - RBF 在不同Feature數量下，指標也都很穩健，不太會受到Feature數量的影響。

討論與結論

- BPN
  - 特徵值會影響訓練成果。
  - 當使用較少特徵值進行測試時，分類效果較好。
  - 總共測試八種模型架構，最佳結果為(46-75-1)。

討論與結論

- 在我們所使用的三個方法中，Naive Bayes 和 SVM 的分類準確率皆達到不錯的效果。
- BPN可能須要做額外前處理才能有更優秀的學習效果。

	NBC ; Feature = 5000	SVM.Linear ; Feature = 5000	BPN(46-75-1) ; Feature = 46
準確度	95.5%	96.7%	77.1%
Precision	95.6%	95.6%	78.7%
Recall	95.8%	96.7%	74.0%
AUC	95.5%	95.5%	None
MCC	91.1%	93.4%	0.540975
Kappa	91.1%	93.4%	0.136604
F1	95.7%	96.8%	76.8%

類神經網路期末報告

Thank you