機器學習課堂競賽

TEAM_1: 陳奕翔、陳昇達、林士登、謝安旭

前言 — 心臟病是人類重大疾病之一，如何防範和治療一直是人類關注的重要議題。近年來，隨著大數據和人工智慧等技術的興起，這些技術使得醫學界有了重大的發展和突破。本報告旨在運用課堂所學的知識和技術，從心電圖資料的分析開始，建構和訓練機器學習模型，以預測病患可能的疾病，探索機器學習在心臟疾病預測中的潛力。

I. 方法介紹

A. 資料前處理

A-1 降噪

利用 order = 4, lowcut = 20, fs = 500 的低通濾波器 (butterworth filter) 將高頻訊號去除。

A-2 心電圖標點

1. 利用 scipy 的內建 find_peaks 函式將心電圖中所有至高點 R 峰值找出來，函式內部參數 distance 防止抓到距離太近的峰值。

2. 抓到所有 R 峰值後利用最大 R 峰值的一半作為基準，若有 R 峰數值低於此基準則判定為偽 R 峰，將它從列表中刪除。

   - ✓ **T 峰**：從目前 R 峰後 30 個點開始，到當前和下一個 R 峰的中點的範圍尋找最大值。
   - ✓ **P 峰**：從目前和下一個 R 峰的中點後 30 個點開始，到下個 R 峰前 30 個點尋找最大值。
   - ✓ **S 峰**：從當前 R 峰到 T 峰之最小值。
   - ✓ **Q 峰**：從 P 峰到下一個 R 峰的最小值。
   - ✓ **S'、T'、L'、P'特殊點**：在給定範圍內根據轉折點或斜率最小值的點來決定。

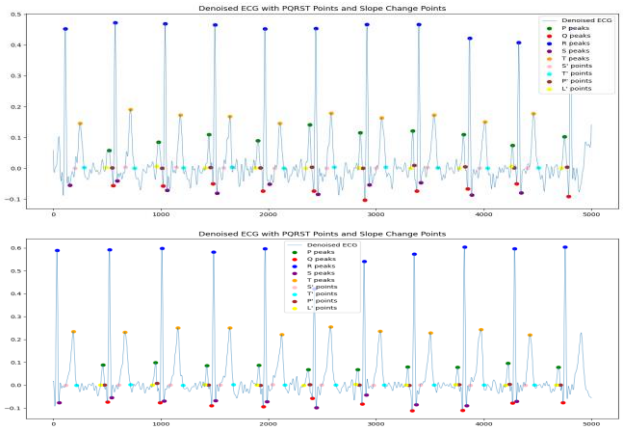

圖 1. 特徵標點示意圖

A-3 錯誤點標示處理

雖然標示點 Norm 有九成都是正確的，但剩下的一成及一些患者的特徵點標示會因為心電圖波型異常而檢測錯誤，我們設計檢測異常的標準如下：

1. R 點的值超過 30 個或低於 5 個。

2. [P,Q,R,S,T] 中其中一組峰值標準差大於 0.15。

而當達到上述條件時將此人的特徵列表先全部填 0，待後續所有人的特徵都算出來後取平均填入列表值都為 0 的人。

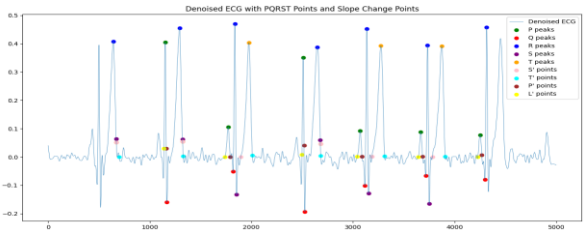

圖 2. 異常抓點示意圖

B. 特徵過濾

B-1 特徵的種類

根據課堂競賽講義，我們最初選取了 21 個關鍵心電圖特徵。為了創建更多特徵，我們將這些原有特徵的 X 軸和 Y 軸數值對調，從而生成新的特徵。例如，原本取 R 和 Q 點的 Y 軸距離的特徵，我們改為取它們的 X 軸距離。這樣操作後，理論上我們應有 42 個特徵。然而，由於部分原有和新生成的特徵重複（比如 RS 的 X 軸和 Y 軸距離原本就包含在 21 個特徵中），我們移除了重複的特徵，最終保留了 34 個獨特的心電圖特徵。

在測試的過程中，為了提升準確度，其中一個方向是去增加諸如 heartbeat 和 spectrum 相關的特徵，總共涵蓋表 1 所列出的以下 17 種。

表 1. heartbeat 和 spectrum 相關的特徵介紹

Heartbeat		Spectrum	
bpm	每分鐘心跳次數	PSD	功率頻譜之密度
ibi	心跳間隔的平均時間	RF	頻譜中最大頻率
sdnn	心跳間隔的標準差	LFP	頻率介於5~15Hz的功率密度合
sdsd	R-R之間連續差異標準差	HFP	頻率介於15~40Hz的功率密度合
rmssd	R-R之間連續差異均方根	PR	低頻功率和高頻功率的比
pnn20	每對相鄰心跳時間間隔>20 ms的數目	THP	基頻及諧波功率之合
pnn50	每對相鄰心跳時間間隔>50 ms的數目	AHP	基頻及諧波功率之平均值
		MHP	基頻及諧波中最大的單個諧波功率
		HPP	基頻功率及諧波功率之比值
		SE	功率密度的熵值

最終我們依抓點的準確度（表 2.），取 Lead0、1、4、5、9、10、11 共 7 個導程的資料，因此，綜合上述的特徵數量以及乘上導程數量共$(34+17)\times 7=357$ 個特徵。

表 2. 每個導程在 Train 和 Test 資料未能抓到點的資料(筆)

Lead	Train	Test	Lead	Train	Test
lead0	477	341	lead6	4126	2599
lead1	723	684	lead7	5182	2362
lead2	2602	2472	lead8	3504	1811
lead3	5088	2965	lead9	1499	1096
lead4	691	843	lead10	777	696
lead5	1226	1436	lead11	645	509

B-2 篩選特徵

我們最初嘗試透過散布圖的方式找尋有用的特徵，然而 357 個特徵的資料量過於龐大，故不適用此方法。後來也嘗試過做 PCA，但用 PCA 篩選出的特徵也無法達到較高的準確率，我們推測可能原因為異常值的影響、數據分布的特性、相關性的過度消除等，而造成有用的特徵被過濾掉。因此，我們引進多種選取特徵的方法做結合（圖 3.），選出最終的有用特徵。

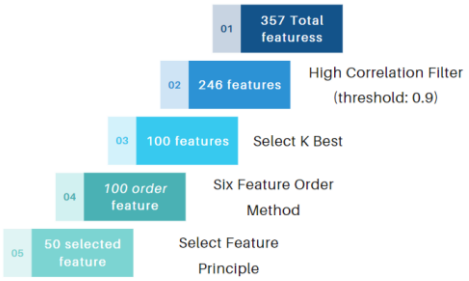

圖 3. 選取特徵的完整流程圖

第一步使用 high correlation filter 篩除高度相關性的特徵，避免造成數據的多重共線性；第二步為使用 select K best 的方法，選出前 100 重要的特徵。Select K best 的原理為透過計算的 F 值排序特徵的重要度：

$$F=\frac{SSB/(類別數-1)}{SSW/(樣本總數-類別數)}$$

(SSB: 組間平方和；SSW: 組內平方和)

第三步為使用六種排序特徵重要度的方法將此 100 個特徵重新做排序（圖 4.），前三種為數學的計算，找出能區分四種 Label 的特徵；後三種為透過隨機森林找出每個特徵的重要度。(隨機森林的原理會在分類器部分介紹)

圖 4. 六種排序特徵重要度的方法

排序完後，我們可以得到全新的特徵重要度（圖 5.），並透過平均編號的排序、特徵出現次數、特徵的散佈圖、lead 的編號（從表 2.可發現取用的導程中 Lead 5、9 較多抓不到的資料），篩出較重要的約 50 個特徵。最終，實際的將不同數量的特徵放入各式分類器中嘗試，得出貝氏分類器約 35 個特徵、SVM 約 40 個特徵、隨機森林約 45 個特徵達到最高的準確度。

圖 5. 透過上述六種排序方法後所得的全新排序

B-3 數據增強 (data augmentation)

我們透過增加疾病數據的數量來平衡四種類別的比例，將疾病數據複製後加上雜訊（noise），雜訊是隨機生成平均值為 0，標準差為 0.05 的常態分佈數值。會選擇用常態分佈生成雜訊是根據中央極限定理，當樣本數夠多時，隨機變量（random variable）的平均值會近似常態分佈，而雜訊可以當作多個獨立的變因組合成的，所以我們用常態分佈來生成雜訊，並設定雜訊的平均值為 0，避免造成整體的偏差（bias）。

C. 模型介紹

C-1 貝氏分類器 (Naïve Bayesian Classifier)

我們設計的貝氏分類器基於高斯混和模型(Guassian Mixture Model, GMM)來進行分類，其基本的概念是利用訓練資料來估計每個類別的機率分布，並在測試資料上計算每個數據點屬於每個類別的機率，最終選擇機率最大的類別作為預測結果，以下為實作步驟。

1. 對特徵進行標準化
2. GMM 假設數據由多個高斯分布混和而成，每個高斯分布對應一個簇(cluster)，而選擇初始化參數透過 K-means 快速找到適合的值，參數如下。

$$\begin{cases}\pi_k：第\ k\ 個高斯分布的權重\\ \mu_k：第\ k\ 個高斯分布的平均向量\\ \Sigma_k：第\ k\ 個高斯分布的協方差矩陣\end{cases}$$

又高斯分布 PDF 如下

$$N(x|\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

每個數據點 x_i 的生成機率為：

$$p(x_i)=\sum_{k=1}^{K}\pi_k N(x_i|\mu_k,\Sigma_k)$$

3. 利用期望最大化演算法(EM algorithm)，在 GMM 中透過 Expectation step 及 Maximization step 來最大化資料的 likelihood。

Expectation step:

$$\gamma_{ik}=\frac{\pi_k N(x_i|\mu_k,\Sigma_k)}{\sum_{k=1}^{K}\pi_k N(x_i|\mu_k,\Sigma_k)}$$

計算每個資料點屬於每個高斯分布的 posterier。

Maximization step:

根據上個步驟算出的 posterior 去更新π_k、μ_k、Σ_k，更新後的公式如下。

$$\begin{cases}\pi_k=\dfrac{1}{n}\sum_{i=1}^{n}\gamma_{ik}\\ \mu_k=\dfrac{\sum_{i=1}^{n}\gamma_{ik}x_i}{\sum_{i=1}^{n}\gamma_{ik}}\\ \Sigma_k=\dfrac{\sum_{i=1}^{n}\gamma_{ik}(x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^{n}\gamma_{ik}}\end{cases}$$

重複迭代 E. & M. step 直到參數收斂或達到最大自訂 iteration 次數。

4. 基於估計出來的 GMM 參數使用基本貝氏定理方法計算並選擇每個類別最大的 posterior 作為預測結果

5. 使用 K-Fold Cross Validation 檢驗訓練模型的穩健度，取平均準確率作為模型性能指標。

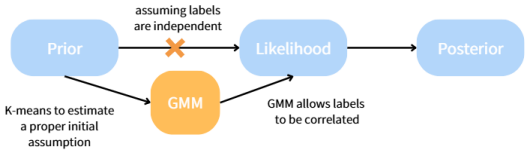

圖 6. 加入 GMM 的貝氏分類器流程

C-2 SVM 分類器 (Support Vector Machine)

我們利用線性核和一對多方法來分類。其主要目的是找到一個能最大化類別間邊際的決策邊界 $w^Tx+b=0$，其中 w 是權重向量，b 是偏差項。邊際的最大化涉及最小化$\|w\|^2$，並確保每個數據點x_i都正確分類，即 $y_i(w^Tx+b)\geq 1$。以下為實作步驟：

1. 特徵標準化。

2. 將權重向量初始值設置為零。

3. 對於給定的訓練數據集，進行多次迭代以逐步調整權重向量。每個迭代中，遍歷所有訓練樣本，對每個樣本檢查其與決策邊界的距離，並根據錯誤分類來更新權重。權重更新方法如下：

   使用一個簡單的梯度下降法：如果 $y_i(w^Tx+b)<1$（代表 x_i處於邊際內或被錯誤分類），則更新權重為 $w\leftarrow w+\eta(y_ix_i-\lambda w)$，其中 η 是學習率，λ是正則化參數，為控制權重增長的速率和防止過擬合。

4. 透過實際訓練模型，嘗試η 從1~10^{-12}，每 0.1 倍做試驗，得出$\eta=10^{-7}$有最好的準確率；嘗試 λ 從 1~0.0001，每 0.1 倍做試驗，得出$\lambda=0.01$附近有較高的準確率，接著，試驗0.01~0.09，得出$\lambda=0.02$有最高準確率。

5. 利用多個 SimpleSVM class 進行一對多的訓練，每個分類器負責將一個類別與其他類別區分。在預測階段，每個分類器對一個樣本給出一個得分，最終預測為得分最高的類別。

6. 使用訓練好的模型對驗證數據集進行預測，並計算模型的準確率，以映模型在驗證數據上的整體表現。

C-3 隨機森林 (Random Forest)

隨機森林(random forest)是一種監督式學習(supervised learning) 廣泛的用在分類問題上。隨機森林是建立在決策樹(decision tree)上，透過隨機樣本與隨機特徵來建構多棵的決策樹，每一棵決策樹會對輸入的樣本產出一個分類，最後使用多數決的方式決定最後結果，其想法像是結合多個分類器組合建構成一個強的分類器，也就是集成學習(Ensemble Learning)，透過結合多種模型的表現，提升最後預測/分類的結果。

在隨機森林中最主要使用基尼不純度(Gini Impurity) Gini Impurity 主要是在計算分類器分錯的機率，愈小代表錯誤機率愈小。其數學公式：$Gini(y)=1-\sum_{i=1}^{C}(\frac{n_i}{N})^2$，其中 n_i 是類別 i 的樣本數、N 是節點中的總樣本數、C 是類別的總數。我們將排序完成的 List 透過迴圈去計算每個特徵值的 Gini Impurity，再去比較大小，若結果比原先的小，就會更新為新的 threshold 值，直到迴圈結束。

我們的隨機森林主要有以下參數可進行調整，分別為樹的數量、最大深度以及特徵數量。樹的數量（n_estimators）最後我們選擇 200，我們有嘗試到 500 甚至 1000 以上，但能得到最好結果的區間是 100-300 之間。最大深度（max_depth）：最後選擇 35，這個參數是為了控

制每棵決策樹的最大層數，若是太淺可能有很差的預測結果，太深則會過度擬和。特徵數量：45，也是透過過不斷測試所得出的最好結果。

■ 優點

1. 高準確率：通過集成多棵決策樹，可以減少過擬合的風險，從而提高模型的穩健性和準確率。
2. 抗過擬合：通過對數據的隨機抽樣與選擇特徵訓練多棵決策樹，因此面對高維數據和噪音數據時能有效地減少過擬合。
3. 不需要大量的參數調整：與其他機器學習算法相比，隨機森林的超參數調整較為簡單，主要需要調整樹的數量和最大深度。

■ 缺點

計算量大：訓練隨機森林需要訓練多棵決策樹，因此在訓練時間和內存消耗上會比單一決策樹高，尤其在處理大規模數據集時。

II. 結果與討論

我們不斷地嘗試優化我們的分類器模型，像是 Bayesian Classifier 中，我們透過 GMM optimization 的方法提升模型的準確度（表 3.）；SVM Classifier 中，我們不斷地嘗試調整學習率和正規化參數，並且由準確度明顯的提升可看見模型優化的效果（表 4.）；Ramdom Forest Classifier 中，我們嘗試不同樹的數目、深度等參數，優化出較佳的分類器模型，以提升整體的準確率（表 5.）。

我們嘗試了以上的幾種方法以後，並採取選定一整組特徵丟入 Bayesian Classifier 以及 SVM Classifier 中觀察準確率的結果，再去評估是否將其用於 Ramdom Forest Classifier 中，以節省不必要的時間浪費。並最後使用 Ramdom Forest Classifier 中 Public Score 最高的作為最終提交結果。我們最終的 Public Score 為 0.646，排名為第五；Private Score 為 0.644，排名為第六。此外，由 Public Score 和 Private Score 的結果相近，Private Score 並沒有掉很多可確認我們的模型沒有過度擬合。

表 3. 使用貝氏分類器結果

Bayesian Classifier	Private Score	Public Score
Basic Model	0.487	0.506
GMM optimization	0.575	0.588

表 4. 使用 SVM 分類器結果

SVM Classifier	Private Score	Public Score
未調過參數 $(\eta = 1, \lambda = 1)$	0.285	0.280
調過參數 $(\eta = 10^{-7}, \lambda = 0.02)$	0.610	0.609

表 5.使用隨機森林分類器結果

Random Forest (200/35)	Private Score	Public Score
特徵:50 個	0.638	0.643
特徵:45 個	0.644	0.646

(為最終提交結果)

III. 結論

在這次的課堂競賽中，我們學習到了完整的機器學習過程。從一開始的資料處理，由於許多不正常的圖造成找點失敗，我們嘗試了許久，才找出可以成功準確抓到點的方法。接著選取特徵，在過程中我們也搜索了許多相關資料，來嘗試除了常見的 21 fiducial 以外，還有無其他特徵可以作為幫助，最後我們也利用了頻譜分析等特徵成功的提升準確率。

在篩選特徵時，是我們遇到的最大困難，由於我們原始的特徵較多，需進行篩選才能有效的分類，然而我們的特徵在 PCA 後，並無法達到理想的準確率，因此最後我們透過融合多種選取特徵的方法，來篩選出重要的特徵。

在分類器的部分我們也嘗試了多種方法，包含了貝氏、SVM、隨機森林。而在手刻分類器時，我們在過程中理解了其數學原理，以及各個參數的重要性，最後我們是透過隨機森林的分類器來達到最好的準確率。

在這次的 project 中，我們利用上課所學到的內容成功地發想實作；而在最後的課堂報告，我們得到教授與助教的反饋，了解到了哪些部分需要最改進，像是頻譜分析得到的特徵可以嘗試看看不要標準化，依此來保留差異性等。除此之外，在聆聽其他組的報告時，我們也得到一些收穫，比如有組別使用較為簡單的分類器，卻也可以達到相當不錯的準確率。我們觀察到他們在取特徵的部分做得很好，因此未來我們想優化結果，也可以嘗試重新檢視選特徵的過程。最後感謝教授與助教們提供了這次學習機會，讓我們對 Machine Learning 領域有了一些初步的理解。

IV. 參考資料

[1] 心電圖 heartbeat 相關特徵：

https://hackmd.io/@ncnucsie/HyYzfPc6q

[2] 高斯混合模型：

https://blog.csdn.net/qq_52466006/article/details/127186276

[3] 高斯混合模型相關理論：

https://ithelp.ithome.com.tw/articles/10317928

[4] SVM 分類器模型：

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%AC3-4%E8%AC%9B-%E6%94%AF%E6%8F%B4%E5%90%91%E9%87%8F%E6%A9%9F-support-vector-machine-%E4%BB%8B%E7%B4%B9-9c6c6925856b

https://medium.com/chung-yi/ml%E5%85%A5%E9%96%80-%E5%8D%81%E4%B8%80-%E6%94%AF%E6%8F%B4%E5%90%91%E9%87%8F%E6%A9%9F-support-vector-machine-svm-c8c1bb1c970f

[5] 隨機森林分類器模型：

https://medium.com/chung-yi/ml%E5%85%A5%E9%96%80-%E5%8D%81%E4%B8%83-%E9%9A%A8%E6%A9%9F%E6%A3%AE%E6%9E%97-random-forest-6afc24871857

[6] Gini Impurity 相關理論：

https://johnny-chuang.medium.com/cart-randomforest-88e7139e035c

[7] A. Sharma and K. P. S. Rana, "A Novel Approach for Gaussian Mixture Model Clustering Based on Soft Computing Method," *2020 International Conference on Soft Computing and Pattern Recognition (SoCPaR)*, Hanoi, Vietnam, 2020, pp. 45-50, doi: 10.1109/SoCPaR49749.2020.00013.

[8] K. Netti and Y. Radhika, "A Novel Method for Minimizing Loss of Accuracy in Naive Bayes Classifier," *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Madurai, India, 2015, pp. 1-6, doi: 10.1109/ICCIC.2015.7435801.

[9] R. P. Browne, P. D. McNicholas, and M. D. Sparling, "Model-Based Learning Using a Mixture of Mixtures of Gaussian and Uniform Distributions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 814-817, Apr. 2012, doi: 10.1109/TPAMI.2011.199

[10] Y. Wang, F. Agrafioti, D. Hatzinakos, and K. N. Plataniotis, “Analysis of human electrocardiogram for biometric recognition,” EURASIP J. Adv. Signal Process., vol. 2008, no. 1, Dec. 2007, Art. no. 148658.

[11] J. M. Irvine, S. A. Israel, W. T. Scruggs, and W. J. Worek, “eigenPulse:Robust human identification from cardiovascular function,” Pattern Recognit., vol. 41, no. 11, pp. 3427–3435, 2008.

[12] A. D. C. Chan, M. M. Hamdy, A. Badre, and V. Badee, “Wavelet distance measure for person identification using electrocardiograms,” IEEE Trans. Instrum. Meas., vol. 57, no. 2, pp. 248–253, Feb. 2008.

[13] S.-C. Wu, P.-T. Chen, and J.-H. Hsieh, “Spatiotemporal features of electrocardiogram for biometric recognition,” Multidimensional Syst. Signal Process., vol. 30, no. 2, pp. 989–1007, Apr. 2019.

[14] J. Surda, S. Lovas, J. Pucik, and M. Jus, "Spectral Properties of ECG Signal," presented at the 2007 17th International Conference Radioelektronika, 2007.