Final Project Proposal

Team member

109061517 邱俊嘉 / 109061520 陳俊宇 / 109061807 吳亞澤

Deadline: 11/24

Proposal format:

PDF

-Title

臺灣海洋廢棄物預測

-Methods

1. **資料處理與分析**
   1. **train dataset**

下圖為AIea提供的train dataset之數量分布



總共有319筆sample，共10種level(海廢等級)，各level的數量分布不均勻，因此會做upsample讓每個level的數量都複製增加到51筆，所以最後總train data數會是510筆，這樣可以避免data imbalance的問題，不然model會傾向猜數量最多的那類

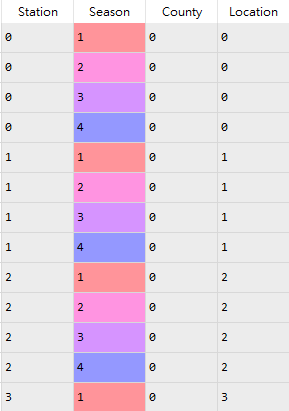
* 1. **缺失值以-1取代**

某些sample的部分特徵值可能會有缺失，像是下圖中的紅框處，此時我們會以-1來取代這種缺失的特徵值

 🡺 

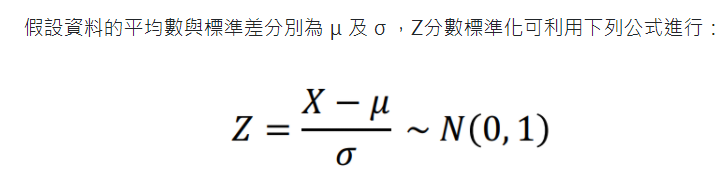
* 1. **特徵值若為中文或英文就以整數代號取代**

station、location和country這三個feature的特徵值為中文或英文，這會讓model無法學習，所以我們將其改成整數代號，如下圖所示

 🡺 

* 1. **將所有feature做個別的z-score standardization**

因為不同feature的數值分布範圍差異甚大，某些feature只有0或1，某些feature從0到87，如果不做standardization的話，某些feature值會dominate整個model的學習，導致model學不起來，所以我們會做z-score standardization，其公式如下圖所示，每個feature轉換後都會變成平均值為0、標準差為1的標準高斯分布



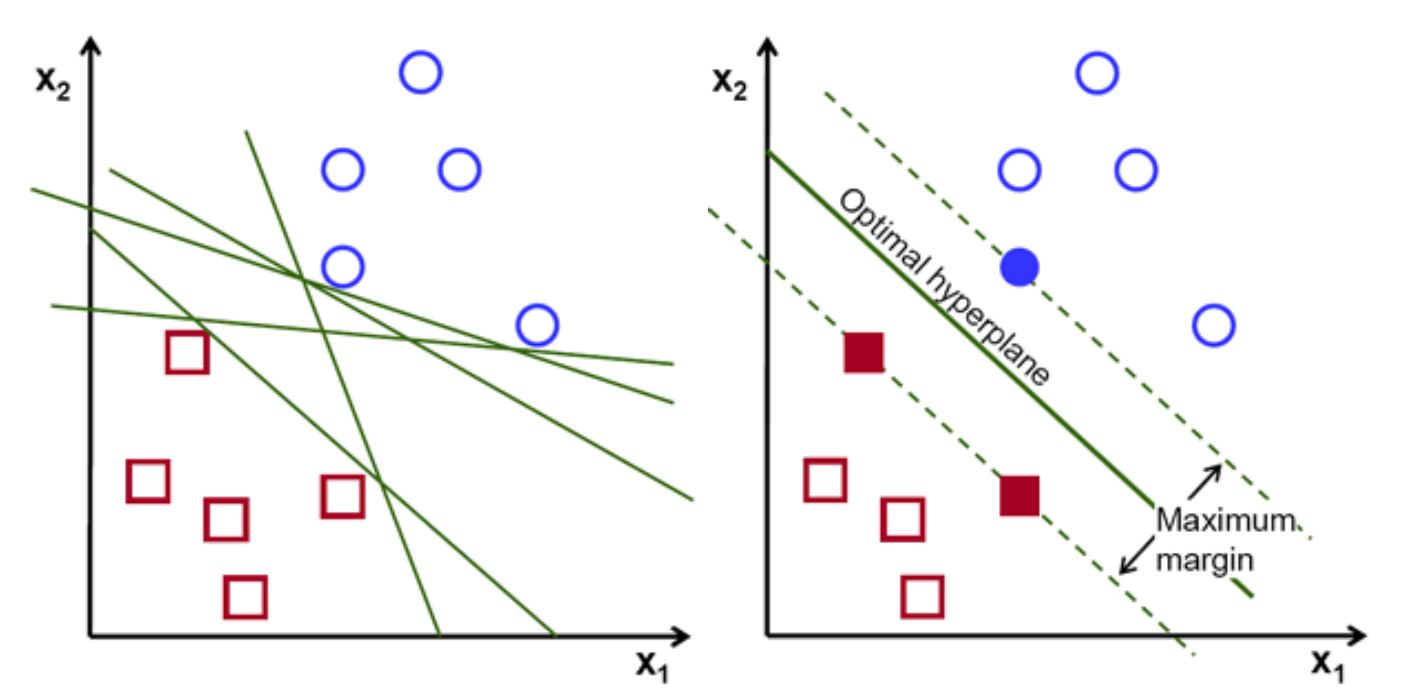
1. **模型架構**

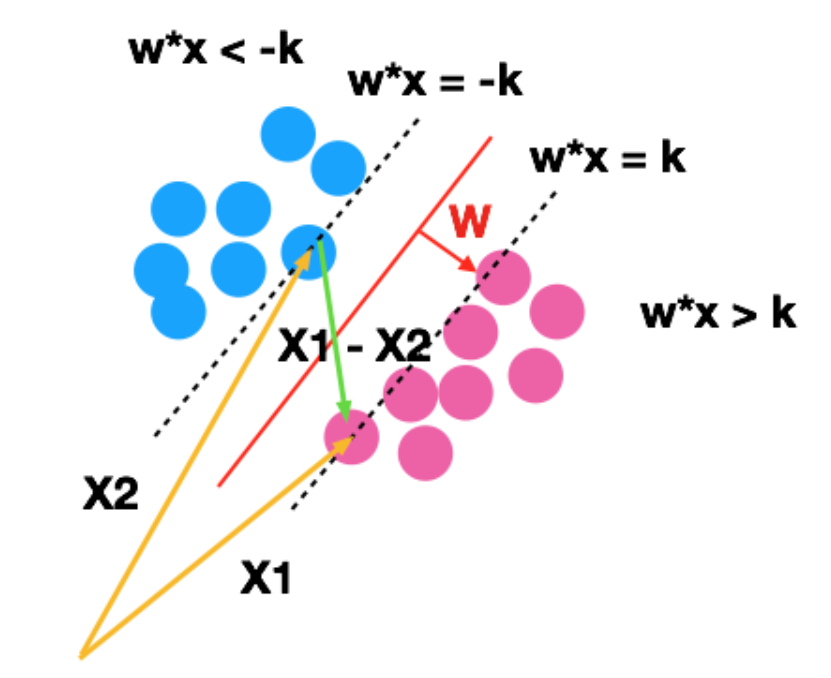
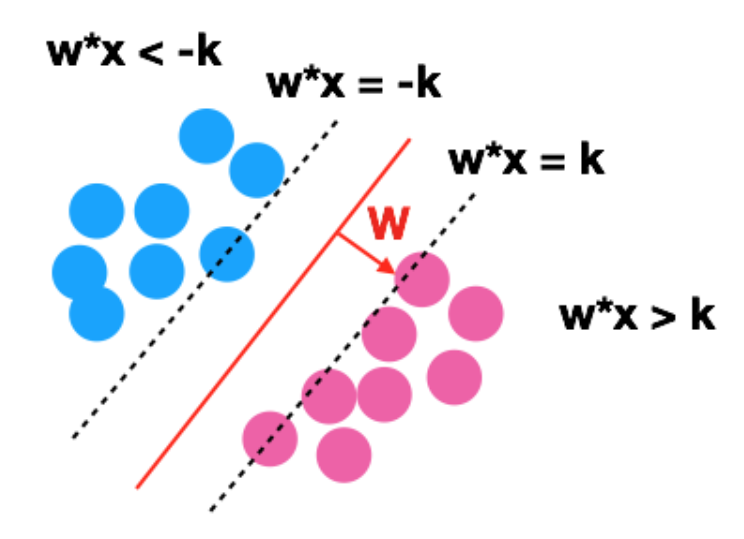
我們的model主要分成2大類，一類是機器學習方法，像是SVM、Random Forest，另一類是使用neural network的深度學習方法，像是DNN，下面舉幾個比較有名的架構來做細部說明

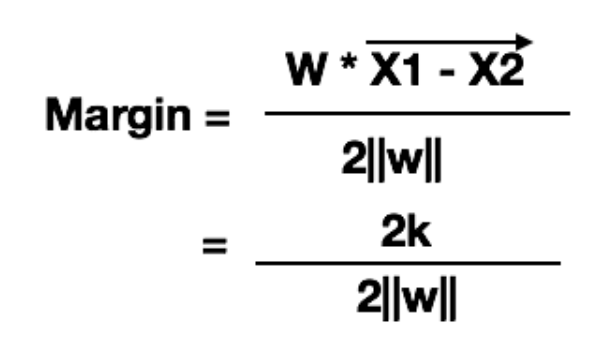
* 1. **支援向量機(Support vector machine)(SVM)**
     1. **目的**

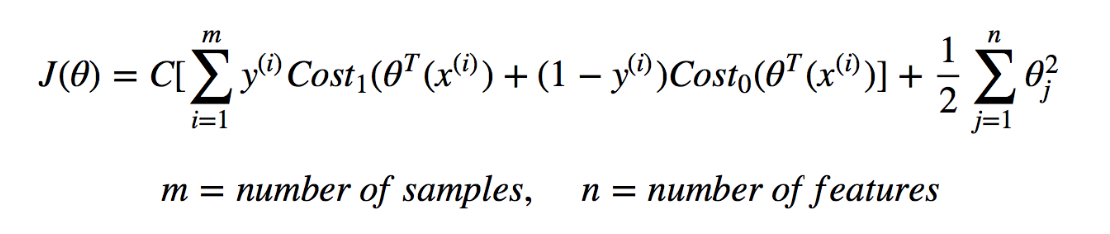
**SVM 是一種有堅實理論基礎的新穎的小樣本學習方法，它基本上不涉及概率測度及大數定律等，因此不同於現有的統計方法，從本質上看，它避開了從歸納到演繹的傳統過程，實現了高效的從訓練樣本到預測樣本的「轉導推理」，因此大大簡化了通常的分類和迴歸等問題，故也選用此法**

* + 1. **原理**



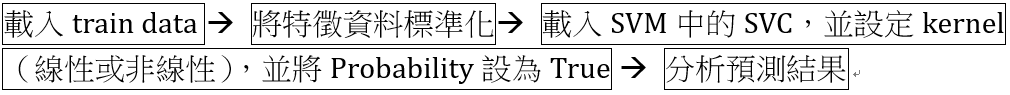






**虛線上的點X1,X2 其實就是所謂的支援向量(Support vector)，我們主要是利用支援向量來算出Margin，並最大化Margin(最小化Cost Function)**

* + 1. **流程**



* + 1. **優缺點**

**優點:**

1. 切出來的線很漂亮，擁有最大margin的特性
2. 可以很容易透過更換Kernel，做出非線性的線（非線性的決策邊界）

**缺點:**

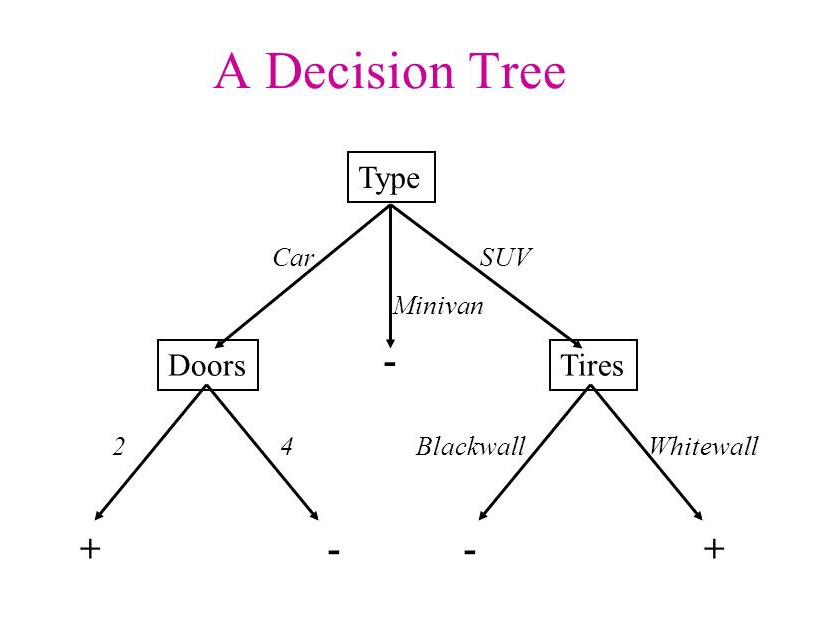
1. 效能較不佳，由於時間複雜度為O(n²)當有超過一萬筆資料時，運算速度會慢上許多
   1. **隨機森林模型(Random Forest)(RF)**
      1. **目的**

random forest classifier的時間複雜度較低，為O(n\*log(n)\*d\*k)，也更適合處理高維度的資料。在分析上能顯示各個特徵的重要性，還能利用oob error快速評估模型的表現，非常適合用於本次的實驗

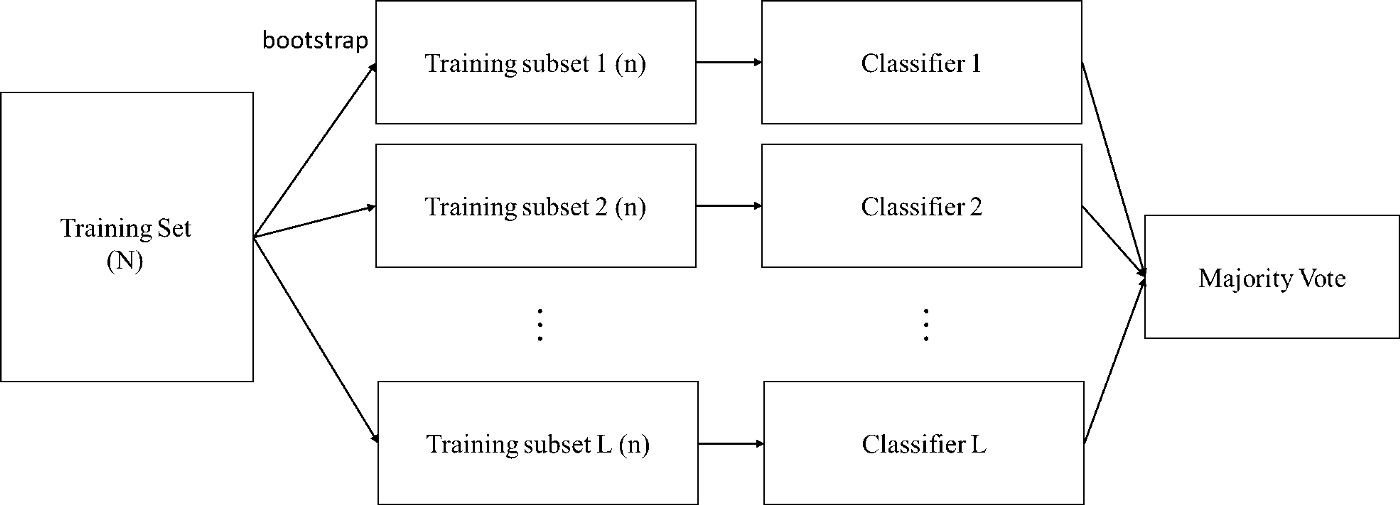
* + 1. **原理**

在講Random forest之前，必須要先介紹Decision tree和Bagging這兩個概念

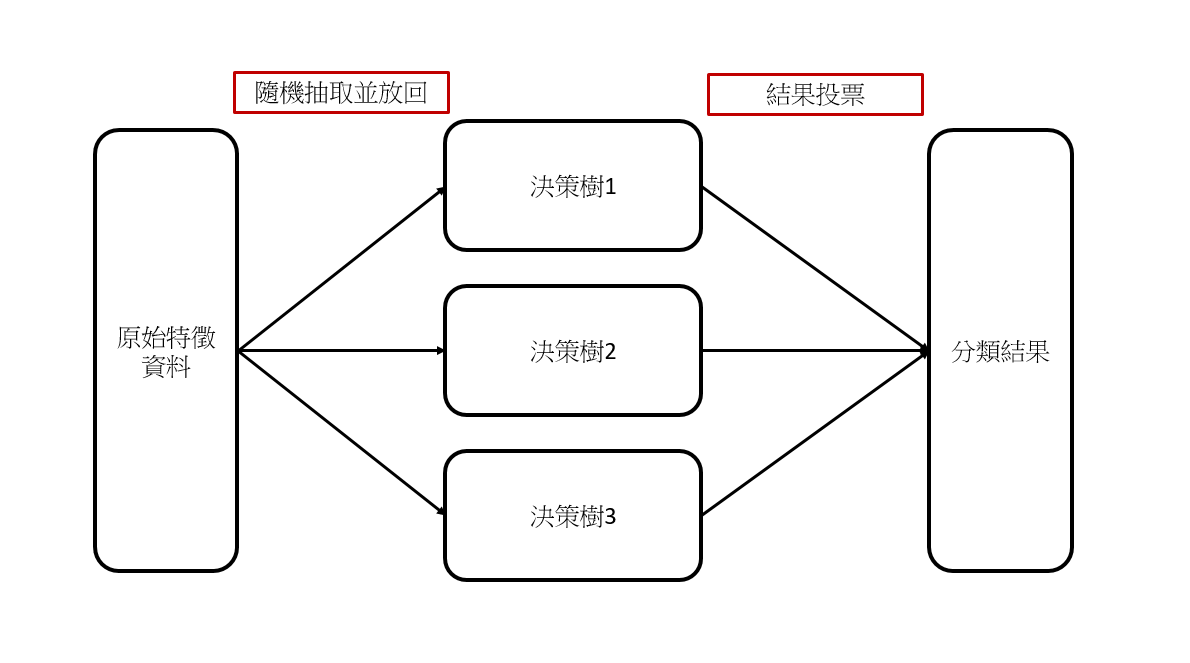
Decision tree：透過一層一層的決策，逐步篩選出符合的結果，如下圖所示



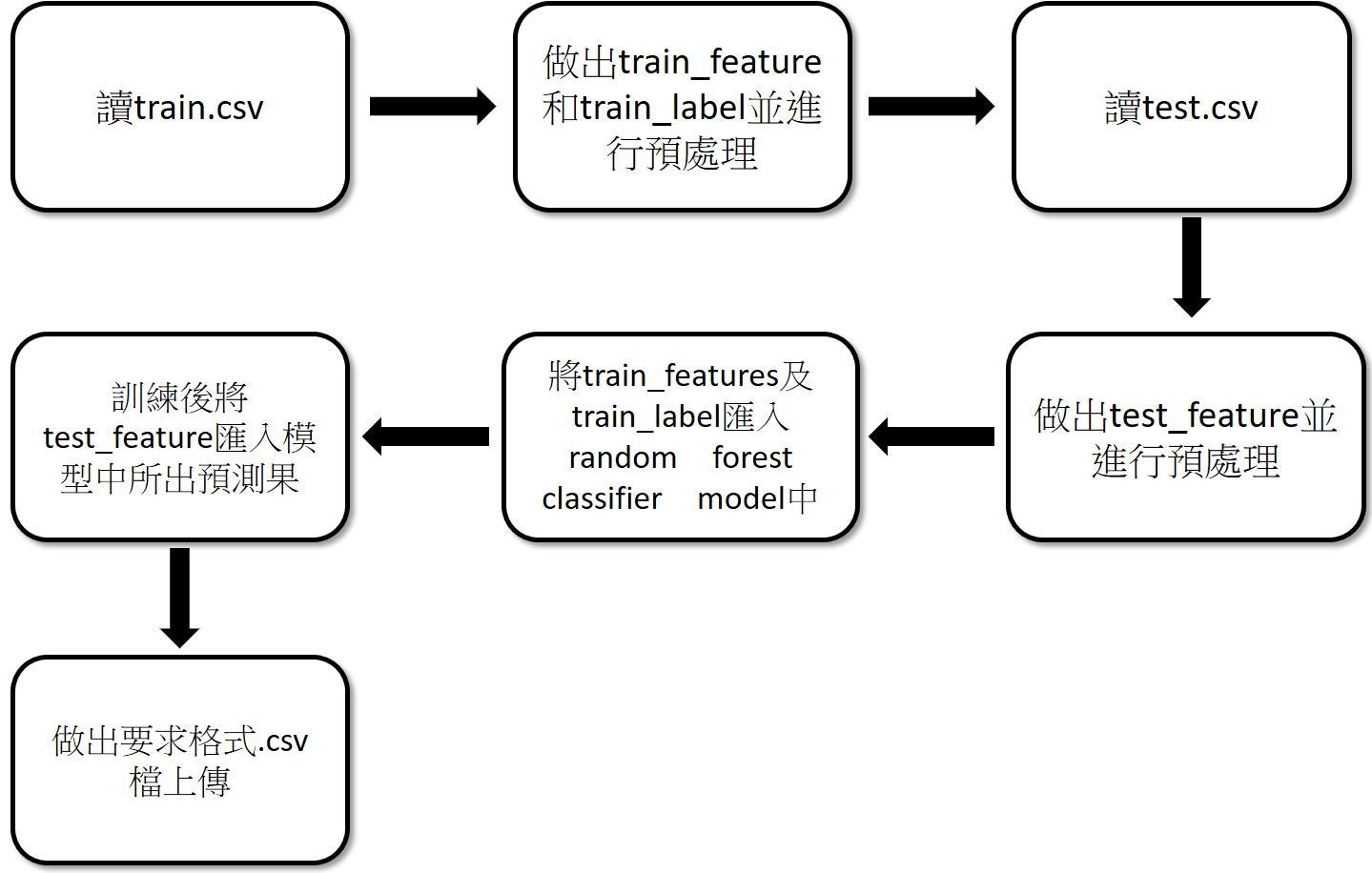
Bagging：從訓練資料中隨機抽取，取出後放回(bootstrap)，利用抽出之樣本訓練多個分類器，每個分類器的權重一致，最後用投票方式(Majority vote)得到最終結果，如下圖所示



Random forest：是一個結合bagging及decision trees的演算法，如其名字所述，由多個decision tree所組成，每顆決策樹獨立運算出結果，並透過投票得到最後的分類結果，如下圖所示



* + 1. **流程**



* + 1. **優缺點**

**優點**

1. 時間複雜度低，適合處理大量高維度的資料
2. 附有feature importance以及oob error等有利分析的功能
3. 訓練速度快
4. 能夠平衡失衡資料集的誤差
5. 對於缺失值以及離群值的敏感度低
6. 能夠避免overfitting
7. 能解決回歸與分類兩種問題

**缺點**

1. 對於資料數量少，或是低維度的資料，分類結果較差
2. 在某些noise較大的分類或迴歸問題上會過擬合
3. 相對於Decision tree，需要更長的時間以及更多的儲存空間作運算
   1. **深度神經網路(Deep Neural Network)(DNN)**
      1. **目的**

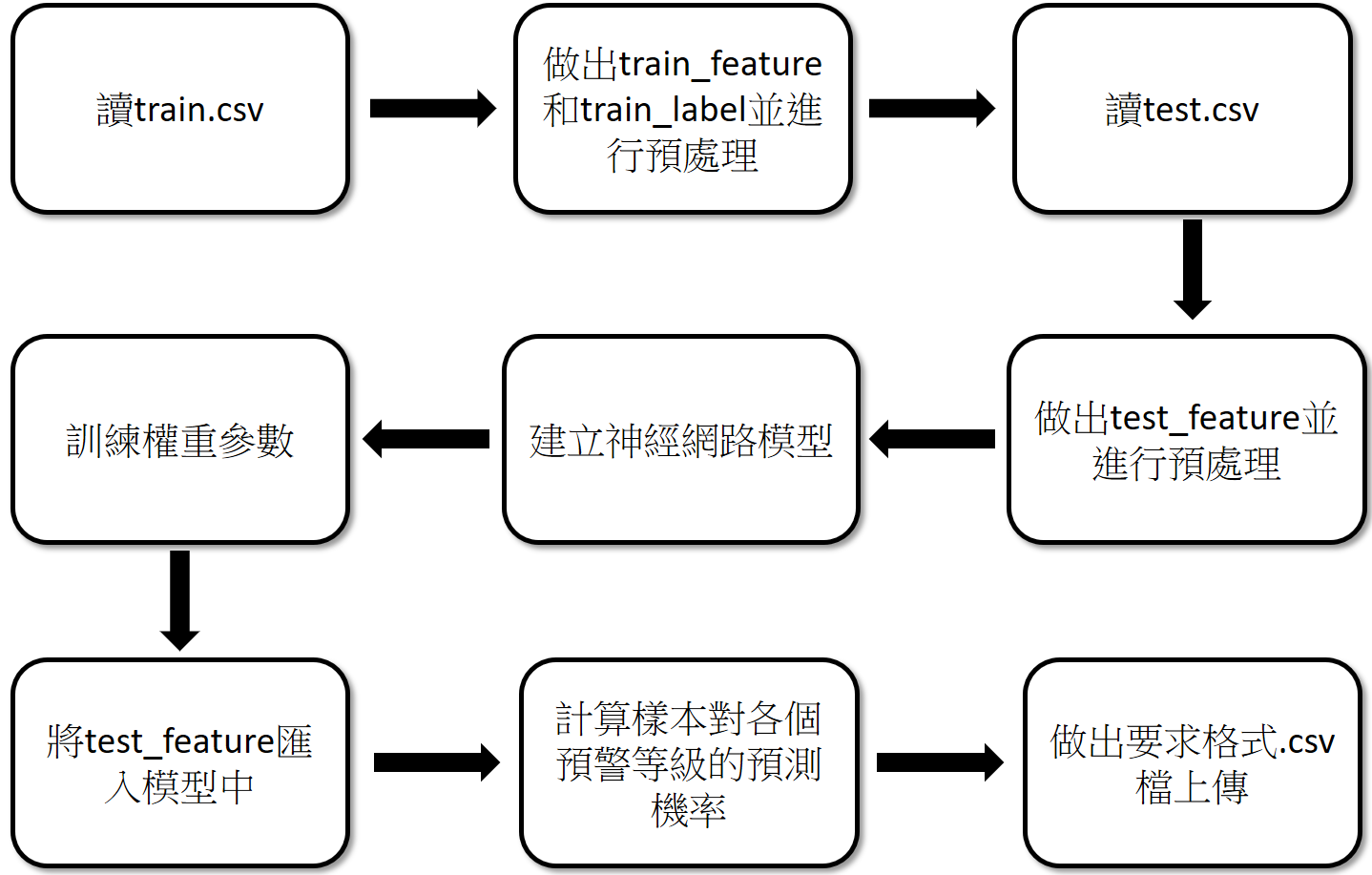
**在機器學習的世界中，神經網路就像是人類的大腦神經結構，而神經元就像是大腦的神經細胞，是神經網路中最基礎的結構，在它們互相結合下，可以建構龐大的運作網路，能大大的提高預測準確率，也非常適合處理分類問題，故也選用此方法**

* + 1. **原理**

1. **建立輸入層、多層隱藏層、輸出層和每個層的神經元，每個神經元都有一個輸出，其輸出稱為激勵值(介於0~1)，每一層神經元中激勵值的操作結果會影響下一層的激勵值，一層一層之間激勵值的傳遞最後即可得到輸出判斷結果**
2. **前一層的每個神經元的輸出分別乘其對應的「權重值」最後相加 🡺 減下一層神經元的「偏置值」🡺 帶入激勵函式將其轉為介於0~1的「激勵值」**
3. **數學形式:**

**0 ≤ 激勵值 = f(a1w1+a2w2+……+anwn - b) ≤ 1**

* + 1. **流程**



* + 1. **優缺點**

**優點**

1. 可以建構非線性的模型
2. 有良好的推廣性，對於未知的輸入亦可得到正確的輸出
3. 可以接受不同種類的變數作為輸入，適應性強
4. 可應用的領域相當廣泛

**缺點**

1. 以迭代方式更新鍵結值與閥值，計算量大，相當耗費電腦資源
2. 訓練的過程中無法得知需要多少神經元個數，太多或太少的神經元均會影響系統的準確性，因此往往需以試誤的方式得到適當的神經元個數
3. **初步實驗結果**

除了上面詳細說明的方法之外，我們還試了很多其他ML的方法，下圖一的實驗結果為對train data做5-fold cross validation後的估計結果，圖二為Gradient Boosting Classifier(GBC)在AIdea public leaderboard的評估結果

|  |
| --- |
| 圖一 |
| 圖二 |

1. **分析**

* 到目前為止結果都不是很理想，可能是因為有效train data數只有319筆的關係，導致DNN這種DL方法無法發揮，因此表現是最差的
* ML方法雖然普遍表現較好，但是應該還有很大的進步空間
* Gradient Boosting Classifier(GBC)的表現最好，之後應該會由這個架構去做後續改良

1. **未來工作**

未來工作主要分為改良feature和改良model兩大方向

* 1. **改良feature**

根據我們過去的經驗用AI方法來處理問題時，feature是非常重要的，同一個train data如果處理成不同feature，最後的performance甚至可能會差到十幾%，所以feature要如何改良應該會成為關鍵

* + 1. **缺失值處理: 以同一個地點的其他3個season的眾數來填補缺失值**

我們目前是用-1來取代缺失值，這個做法可能對於model的學習來說不太好，而且其實缺失值還蠻多的，佔了快25%，應該要更妥善處理，所以我們後來有想到一個可能的解決方案，那就是以同一個地點的其他3個season的眾數來填補缺失值，因為經過觀察，發現缺失值都是海岸地形種類，從直觀上來想，海岸地形因為會影響人類活動，理論上來說應該也會和海廢等級有關，因此海岸地形這個feature應該要用到，所以缺失值勢必要用某個值去填補，不能直接去掉。此外，我們還發現缺失值會發生在每一個地點的第4個season，而且同一個地點的其他3個season的海岸地形種類十分接近，代表第4個season很有可能也是同一個海岸地形，這其實蠻合理的，畢竟同一個地點的海岸地形通常不會因為season的不同而變化，不過同一個地點在不同season時，某些海岸地形會略有不同，所以我們最後決定以同一個地點的其他3個season的眾數來填補缺失值



* + 1. **將可能會誤導model的feature去掉**

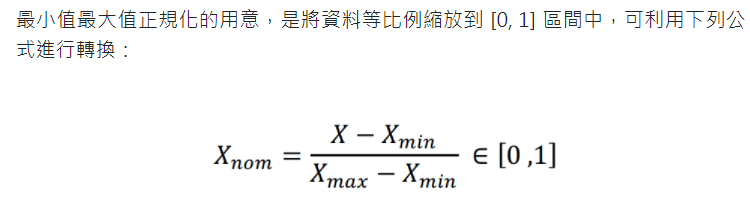
根據觀察，同一個地點的海廢等級通常不會差異太大，所以地點可能會是一個重要因素，以經緯度來表示地點應該是最適當的，但是資料中除了經緯度還給了Station、County、Location、縣市、海岸段、Region，這些feature我們目前是先轉成代號再丟到model裡去train，但是這些代號可能會誤導model，舉例來說，有可能2個地點的Station的代號差異很大，但是其實這2個地點並沒有差很遠，而且這樣也不好做data augmention，所以可能會試著把除了經緯度以外代表地點的feature去掉

* + 1. **feature的特徵值為代號**

這些train data的feature除了經緯度以外，所有特徵值都是代號，這種類型的特徵值可能和一般的數值型特徵值處理方式略不同，無法和圖像pixel一樣直接丟到model去學

* + 1. **改變normalize的方式**

原本是用z-score standardization，但是z-score standardization會將分布會強行轉成標準高斯分布，原本的分布可能會跑掉，所以當原本的分布很不對稱時會出問題，所以未來可以試試看用Min-Max Normalization，公式如下圖所示



* + 1. **找出和海廢等級比較有關係的feature**

海廢等級代表海岸的汙染程度，海廢等級越高就代表海岸越髒，因此海廢等級和垃圾量有關，而垃圾量和人類活動有關，而人類活動可能和地點、季節、地形(ex: 沙灘)、垃圾類型這些feature有關，所以我們可以去做這個feature和海廢等級的關係圖、correlation或是p-value來找出關鍵feature並增加權重

* + 1. **做data augmentation**

等做好feature selection找到關鍵性的feature後，就可以去做data augmentation來增加train data數，以避免過擬合問題

* 1. **改良model**
     1. **用LSTM**

或許可以換成用LSTM之類考慮時序關係的model架構

* + 1. **ensemble learning**

如果未來可以用多個不同的model得到不錯的結果，就可以拿這些model去做ensemble learning，也就是將所以model的prob array取平均來得到最後的prob array和預測結果

-References

1. PATTERN RECOGNITION AND MACHINE LEARNING
2. PYTHON機器學習與深度學習特訓班
3. Github
4. [資料分析&機器學習] 第4.1講 : Kaggle競賽-鐵達尼號生存預測‑(前16%排名)

<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%AC4-1%E8%AC%9B-kaggle%E7%AB%B6%E8%B3%BD-%E9%90%B5%E9%81%94%E5%B0%BC%E8%99%9F%E7%94%9F%E5%AD%98%E9%A0%90%E6%B8%AC-%E5%89%8D16-%E6%8E%92%E5%90%8D-a8842fea7077>

1. 我如何分析客戶流失預測？Kaggle比賽思路分享

<https://medium.com/finformation%E7%95%B6%E7%A8%8B%E5%BC%8F%E9%81%87%E4%B8%8A%E8%B2%A1%E5%8B%99%E9%87%91%E8%9E%8D/%E6%88%91%E5%A6%82%E4%BD%95%E5%88%86%E6%9E%90%E5%AE%A2%E6%88%B6%E6%B5%81%E5%A4%B1%E9%A0%90%E6%B8%AC-kaggle%E6%AF%94%E8%B3%BD%E6%80%9D%E8%B7%AF%E5%88%86%E4%BA%AB-daecd888a91>

1. [資料分析&機器學習] 第3.4講：支援向量機(Support Vector Machine)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%AC3-4%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>

1. [機器學習 ML NOTES]Kaggle比賽心得(2%經歷)

<https://medium.com/@super135799/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-ml-notes-kaggle%E6%AF%94%E8%B3%BD%E5%BF%83%E5%BE%97-2-%E7%B6%93%E6%AD%B7-7e8667cf1dc6>

1. Feature Engineering 特徵工程中常見的方法

<https://vinta.ws/code/feature-engineering.html>

1. Basic feature analysis (Date+Categorical+Revenue)

<https://www.kaggle.com/super13579/basic-feature-analysis-date-categorical-revenue>

1. R筆記–Ensemble Learning(集成學習)

<https://rpubs.com/skydome20/R-Note16-Ensemble_Learning>

1. 機器學習模型的時間複雜度

<https://kknews.cc/zh-tw/code/zyv254a.html>

1. 機器學習: Ensemble learning之Bagging、Boosting和AdaBoost

<https://medium.com/@chih.sheng.huang821/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-ensemble-learning%E4%B9%8Bbagging-boosting%E5%92%8Cadaboost-af031229ebc3>

1. 隨機森林 RF 算法的原理（一）

<https://www.twblogs.net/a/5c8a02b2bd9eee35cd6a97fc>

1. 隨機森林(RANDOM FOREST)的底層概念、操作細節，與推薦相關資源

<http://notebookpage1005.blogspot.com/2018/03/random-forest.html>

1. 隨機森林（Random forest,RF）的生成方法以及優缺點

<https://www.itread01.com/content/1547100921.html>