團隊測驗報告

報名序號：108112

團隊名稱：Metaheuristic

一、資料前處理(1/2)

將原始資料轉換為以「變數」為單位

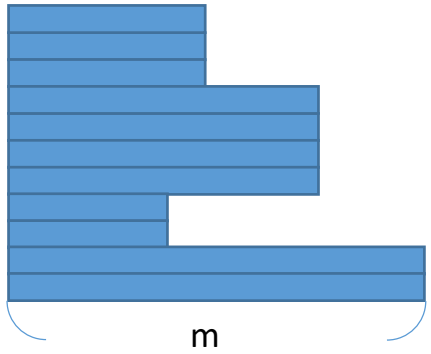
使用「線性內插法」使訓練、測試資料集之樣本變數長度轉換至訓練資料集中之最長變數的長度相同

(步驟二　如下頁所示)

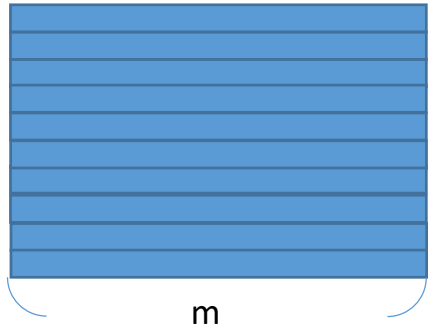
一、資料前處理(2/2)

(步驟二 之 示意圖)

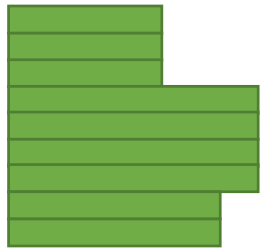
訓練資料集



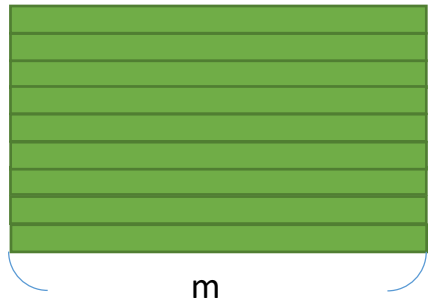
線性內插法



測試資料集



線性內插法



二、演算法和模型介紹(介紹方法細節) (1/7)

1. 主要方法：K Nearest Neighbor + Dynamic Time Warping

    1-1. K Nearest Neighbor (最近鄰居法)

    1-2. Dynamic Time Warping (動態時間校正法)

2. 演算法流程

二、演算法和模型介紹(介紹方法細節) (2/7)

1-1. K Nearest Neighbor (最近鄰居法)

- 為用於分類及迴歸的無母數統計方法
- 此演算法中，一個物件的分類係由與該物件相鄰近的鄰居，採以多數決的方式而定。
- 該物件的分類由最接近此物件之若干個鄰居所決定，其採計數量為最鄰近的K個物件，並以其中出現最多之類別為該物件最終之分類。
- 由於需採計「最鄰近」的K個鄰居，故物件之間的距離計算方式以及採計的鄰居數為此演算法最核心的兩大部分。

二、演算法和模型介紹(介紹方法細節) (3/7)

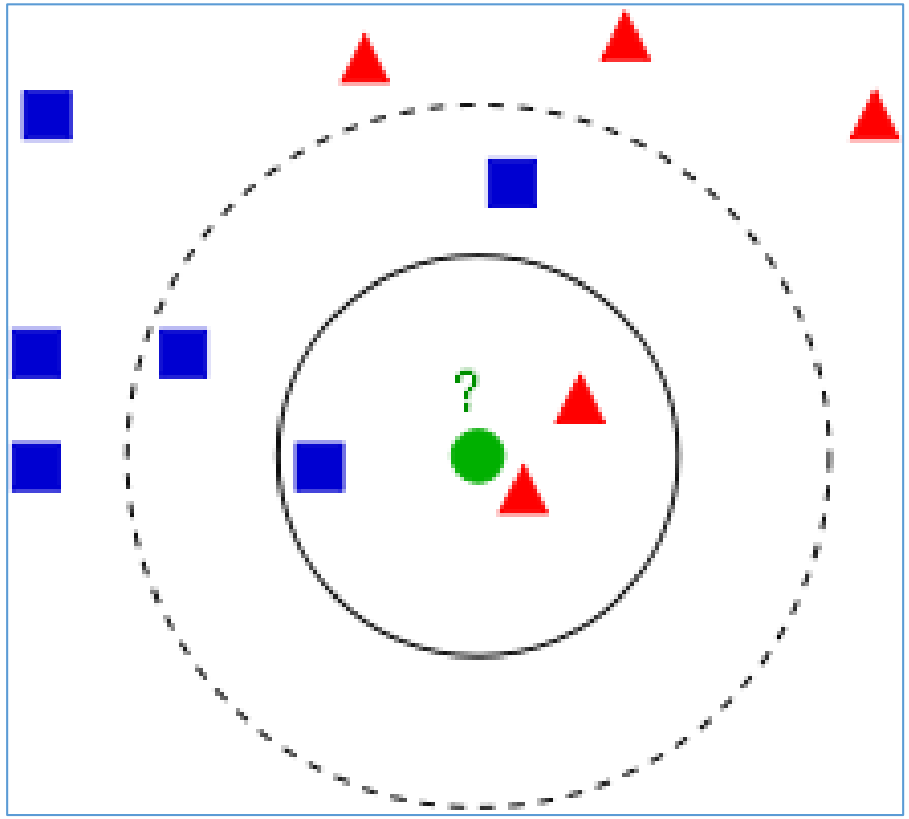
1-1. K Nearest Neighbor (續)

- 以右圖為例：

1. 綠色圓形為測試樣本。此例中希望判斷其分類為藍色正方形或紅色三角形。

2. 若 K = 3 (即圖中「實線」圓圈)，可見此圈中包含最近之3個訓練樣本。其中，紅色三角形出現次數大於藍色正方形。故由此結果可知，當K = 3 時測試樣本應被分與紅色三角形同類別。

3. 若 K = 5 (即圖中「虛線」圓圈)，可見此圈之範圍中包含最鄰近之5個訓練樣本。其中，藍色正方形出現次數大於紅色三角形。故由此結果可知當 K = 5 之下測試樣本應被分與藍色正方形相同之類別。



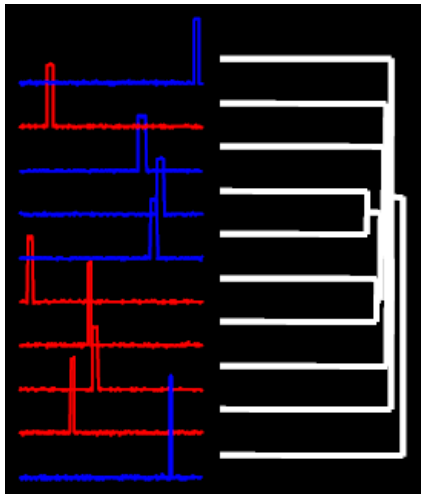
(https://zh.wikipedia.org/wiki/%E6%9C%80%E8%BF%91%E9%84%B0%E5%B1%85%E6%B3%95)

二、演算法和模型介紹(介紹方法細節) (4/7)

1-2. Dynamic Time Warping (動態時間校正法)

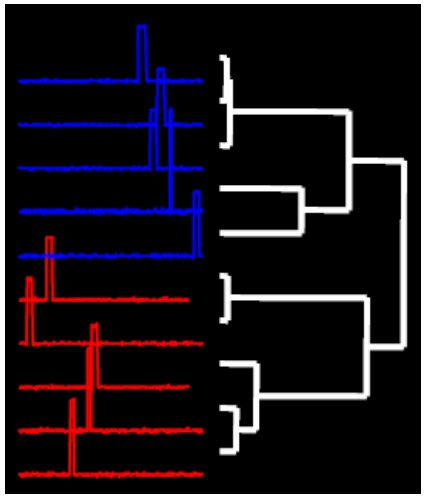
1-2-1. 目的與用途

- DTW為計算兩組時間序列之間之相似度的演算法
- 透過算得之相似度，便可進行分群、分類等流程



(https://www.cs.unm.edu/~mueen/DTW.pdf)

借助 DTW



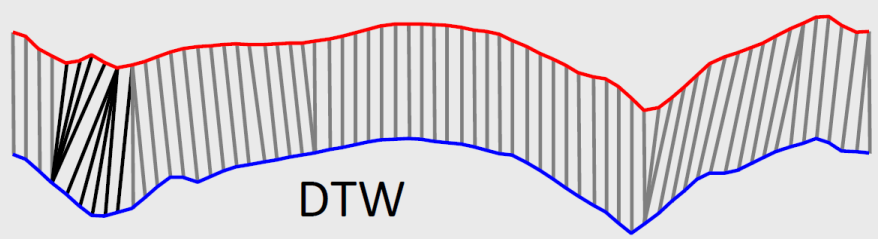
(https://www.cs.unm.edu/~mueen/DTW.pdf)

二、演算法和模型介紹(介紹方法細節) (5/7)

1-2-2. 優點與特色

- 所計算之時間序列，其時間長度可為「不相同」。
  (此亦為此法有別於歐氏距離(Euclidean distance)之特色)

- 透過「選擇性一對多對應(one-to-many mapping with tolerance)」增加算得之準確率。
  (下圖黑線處即為「選擇性排除(allows some points to be unmapped)」)



(https://www.cs.unm.edu/~mueen/DTW.pdf)

- DTW 雖為距離量度(distance measure)，但已非常接近公制單位(metric)
  (公制單位(metric)之優點在於，其結果不受路徑所擾)
  (意即 DTW 所算得之結果為穩定、收斂，僅含有極少的隨機可能)

- DTW 不須預先將資料進行特徵提取(feature extraction)再輸入演算法中
  故可避免分析方面對於領域知識(domain knowledge)之需求

二、演算法和模型介紹(介紹方法細節) (6/7)

1-2-3. 使用原因

1. 資料維度不同

- 每個資料夾中的各txt檔，其維度(列數、行數)並不完全相同
- 因此造成資料在對應上造成無法完全對其的情況
- 希望藉由 DTW 之「選擇性排除配對」的特性，以將不同長度的變數相對應並計算倆倆之間的相似度

2. 資料內容單調(一)

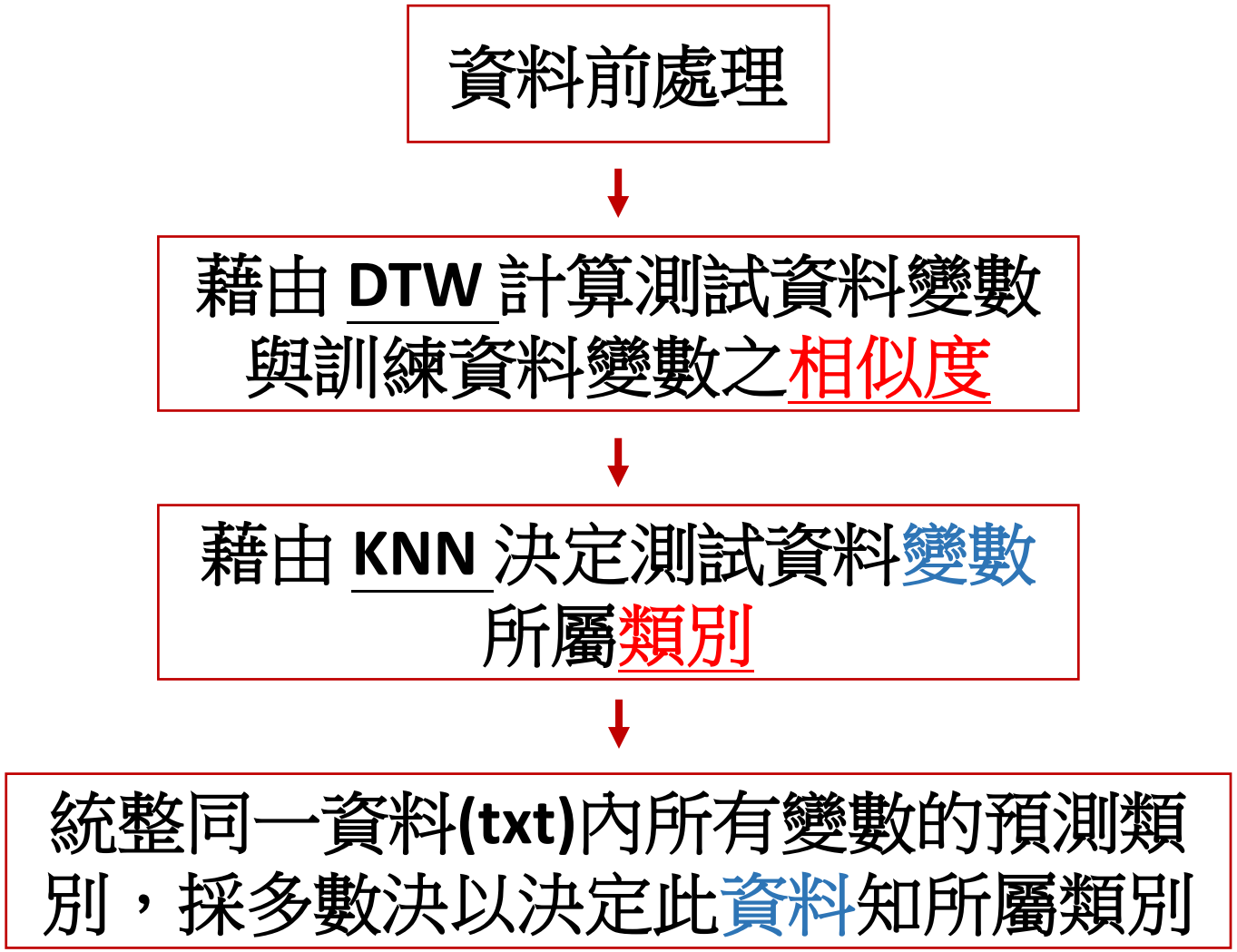
- 每個變數皆為「溫度」之描述，因此變數內容偏單調
- 此類資料在分析時，容易因採用的演算法或模型選用的不同，而在每次運算、執行時產生不同的結果
- 希望藉由 DTW 的「穩定、收斂性」，以維持資料分析結果之固定呈現

3. 資料內容單調(二)

- 由變數內容單調方面可知，在進行特徵提取(feature extraction)時，除非對於該領域知識有所涉入否則將很難進行提取
- 希望藉由 DTW 之「不須經由特徵提取」的特性，利用變數隨者時間變化呈現的線段形狀當作主要特徵。

二、演算法和模型介紹(介紹方法細節) (7/7)

2. 演算法流程



三、演算法模擬結果 (1/2)

使用交叉驗證方法以模擬在不同的訓練及測試樣本比例下找出能夠提供最佳準確率之參數組合

	NN	Window_size	K_fold	median	mean	sd
1	5	0	50	1	1.0000000	0.00000000
2	7	0	50	1	1.0000000	0.00000000
3	3	10	50	1	1.0000000	0.00000000
4	5	10	50	1	1.0000000	0.00000000
5	7	10	50	1	1.0000000	0.00000000
6	5	0	45	1	1.0000000	0.00000000
7	7	0	45	1	1.0000000	0.00000000
8	5	10	45	1	1.0000000	0.00000000
9	7	10	45	1	1.0000000	0.00000000
10	3	0	50	1	0.9960000	0.02828427
11	3	0	45	1	0.9931217	0.03235350
12	7	0	25	1	0.9929231	0.02471933
13	3	10	45	1	0.9916667	0.04128614
14	1	0	50	1	0.9876667	0.05003514
15	1	10	50	1	0.9861667	0.04819914
16	5	0	25	1	0.9843939	0.03708641
17	5	10	25	1	0.9810101	0.05985221
18	7	10	25	1	0.9800909	0.04108380
19	1	10	45	1	0.9768254	0.06762873
20	1	0	45	1	0.9760582	0.05676066

	NN	Window_size	K_fold	median	mean	sd
1	3	0	10	0.7332016	0.7342981	0.06645854
2	5	10	10	0.7228261	0.7308206	0.08231067
3	3	10	10	0.7083333	0.7231790	0.05915883
4	7	10	10	0.6950758	0.7192254	0.09312648
5	7	0	10	0.6956522	0.7148670	0.10732324
6	1	50	10	0.7113095	0.7141328	0.09462096
7	5	0	10	0.7036364	0.7133440	0.07461367
8	1	0	10	0.6660870	0.6665940	0.08748782
9	1	10	10	0.6306818	0.6630840	0.10209254
10	3	50	10	0.5998024	0.6191723	0.07812386
11	5	50	10	0.5780632	0.5557628	0.09833633
12	7	50	10	0.5108696	0.5318109	0.06396436
13	1	50	5	0.3555556	0.3749517	0.07364058
14	1	10	5	0.3404255	0.3438915	0.04836195
15	1	0	5	0.3333333	0.3389614	0.03930186
16	5	10	5	0.3555556	0.3354774	0.06251249
17	3	0	5	0.3404255	0.3334107	0.06257476
18	3	10	5	0.3333333	0.3253382	0.05355054
19	5	0	5	0.3181818	0.3219697	0.02968770
20	7	0	5	0.3043478	0.3127845	0.04705619
21	7	10	5	0.2888889	0.3093237	0.05103185
22	3	50	5	0.2954545	0.2872400	0.03986073
23	5	50	5	0.2444444	0.2482064	0.03945675
24	7	50	5	0.2222222	0.2263889	0.03460018

三、演算法模擬結果 (2/2)

〔結論〕

1. K_fold越大，代表訓練資料與測試資料比例越高。
   在訓練樣本越多的情形下，測試結果越佳。
   此次競賽中訓練與測試資料比例約為7:1，應考慮此限制並進一步選擇最佳參數組合。

2. 繼上述條件之下，透過模擬結果最終選定三組最佳參數組合如下表所示：

組合編號	NN	Window_size	K_Fold
A	3	0	10
B	5	10	10
C	1	50	10

將測試資料經過上述三種參數下的模型進行預測，再依預測結果採多數決以決定最終結果。