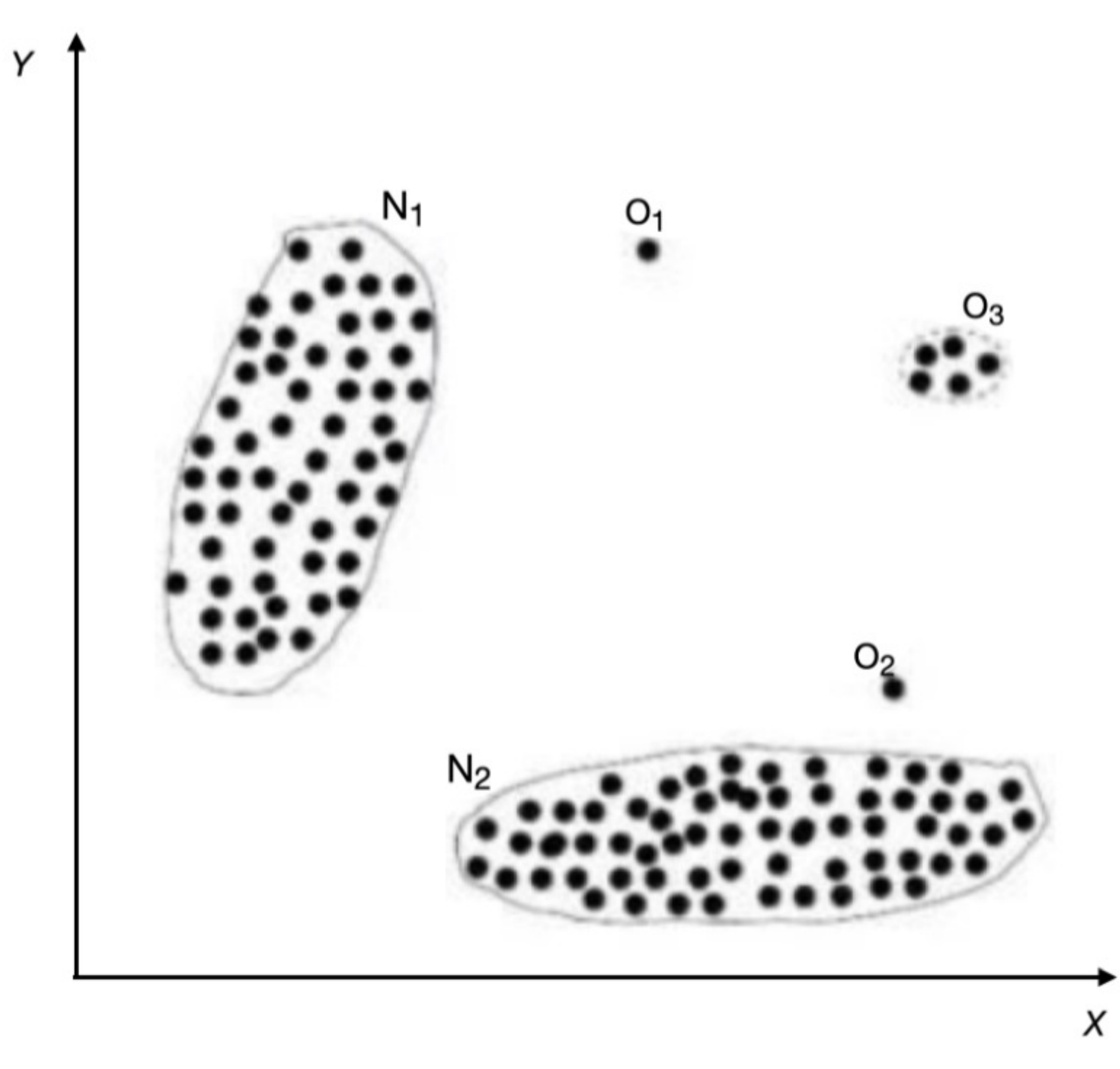
探索性資料分析(EDA)_異常值偵測

簡報閱讀 範例與作業 問題討論 學習心得(完成)

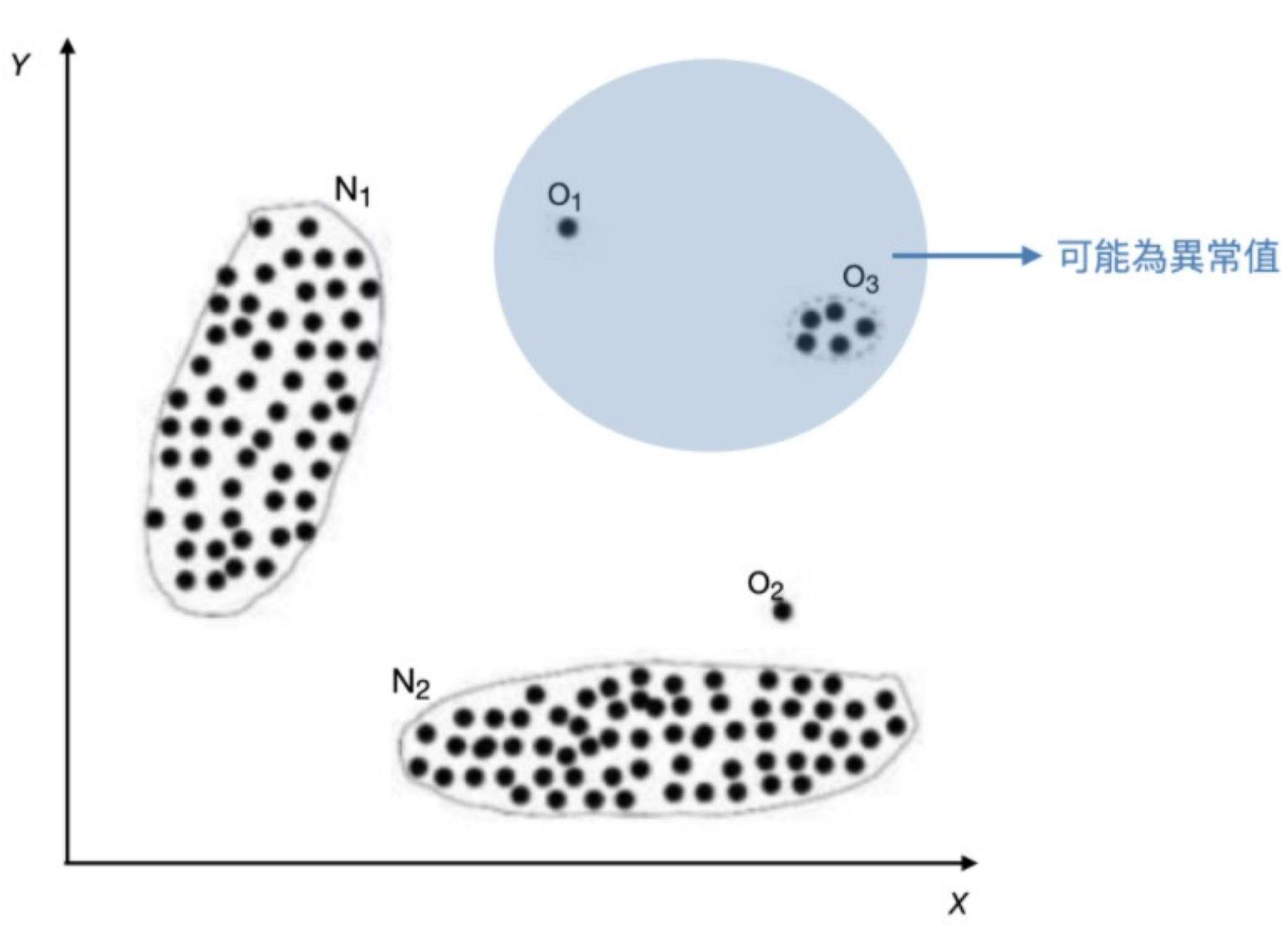
重要知識點

- 異常值的定義
- 掌握辨識出異常值的方法
- 判斷異常值出現是否該刪除或保留

下圖中有哪邊你覺得可能是異常？



你想的和我想的的一樣？



什麼是異常值？

- 異常值就是偏離樣本整體數據的值，所以異常值偵測的原理是採用一種相對的概念。
- 異常檢測 (anomaly detection) 對不符合預期模式，異常也被稱為離群值、新奇、噪聲、偏差和例外。

發現異常值該刪除嗎？

對異常資料進行處理前，需要先辨別出到底哪些是真正的資料異常。

從資料異常的狀態看分為兩種：
- 一種是**「偽異常」**，這些異常是由於業務特定運營動作產生的，其實是正常反映業務狀態，而不是資料本身的異常規律。
- 一種是**「真異常」**，這些異常並不是由於特定的業務動作引起的，而是客觀地反映了資料本身分佈異常的分佈個案。

大多數資料挖掘或資料工作中，異常值都會在資料的預處理過程中被認為是噪音而剔除，以避免其對總體資料評估和分析挖掘的影響。但在以下幾種情況下，我們無須對異常值做拋棄處理。

為什麼會出現異常值？

數據輸入錯誤：人工在數據收集、記錄、輸入造成的錯誤，可能會成為數據中的異常值。例如在人工記錄時將 10 記錄成 10000。

測量誤差：當你使用錯誤的測量儀器測量時，通常會出現異常值。

故意離群：這個通常在進行問卷調研時問題設計不合理或過於敏感出現的。例如在調查用戶年收入時，可能會有很多用戶故意報低或報高。

抽樣錯誤：例如調查普通員工收入時，錯誤的抽取了高層的員工作為樣本的一部分，這會使數據集中出現異常值。

自然異常值：異常值出現的原因不是人工造成的。例如在做用戶價值分析時，通常會發現前10%的用戶消費金額遠遠高於其他用戶，這時候這部分用戶可以單獨取出做分析。

因為出現異常值的原因不同，了解背後的原因才能決定處理方式
刪除/以非離群值的資料統計值取代/分群處理

以下幾種情況下，我們無須對異常值做拋棄處理

1. **異常值正常反映了真實的結果**
   - 例如：由業務部門的特定動作導致的資料分佈異常，如果拋棄異常值將導致無法正確反饋業務結果。
2. **異常檢測模型**
   - 異常檢測模型是針對整體樣本中的異常資料進行分析和挖掘，以便找到其中的異常個案和規律，這種資料應用圍繞異常值展開，因此異常值不能做拋棄處理。
3. **包容異常值的資料建模**
   - 如果資料演算法和模型對異常值不敏感，那麼即使不處理異常值也不會對模型本身造成負面影響，ex：決策樹。

異常值的判別方法 1 - 簡單統計分析

簡單統計分析
- 對屬性值進行一個描述性的統計（規定範圍），從而檢視哪些值是不合理的（範圍以外的值）
- 適用範圍：儀器量測出來的數值，超過儀器的規格。

新增異常值的判別方法 2 - 3σ 原則

3σ 原則 (3倍標準差)

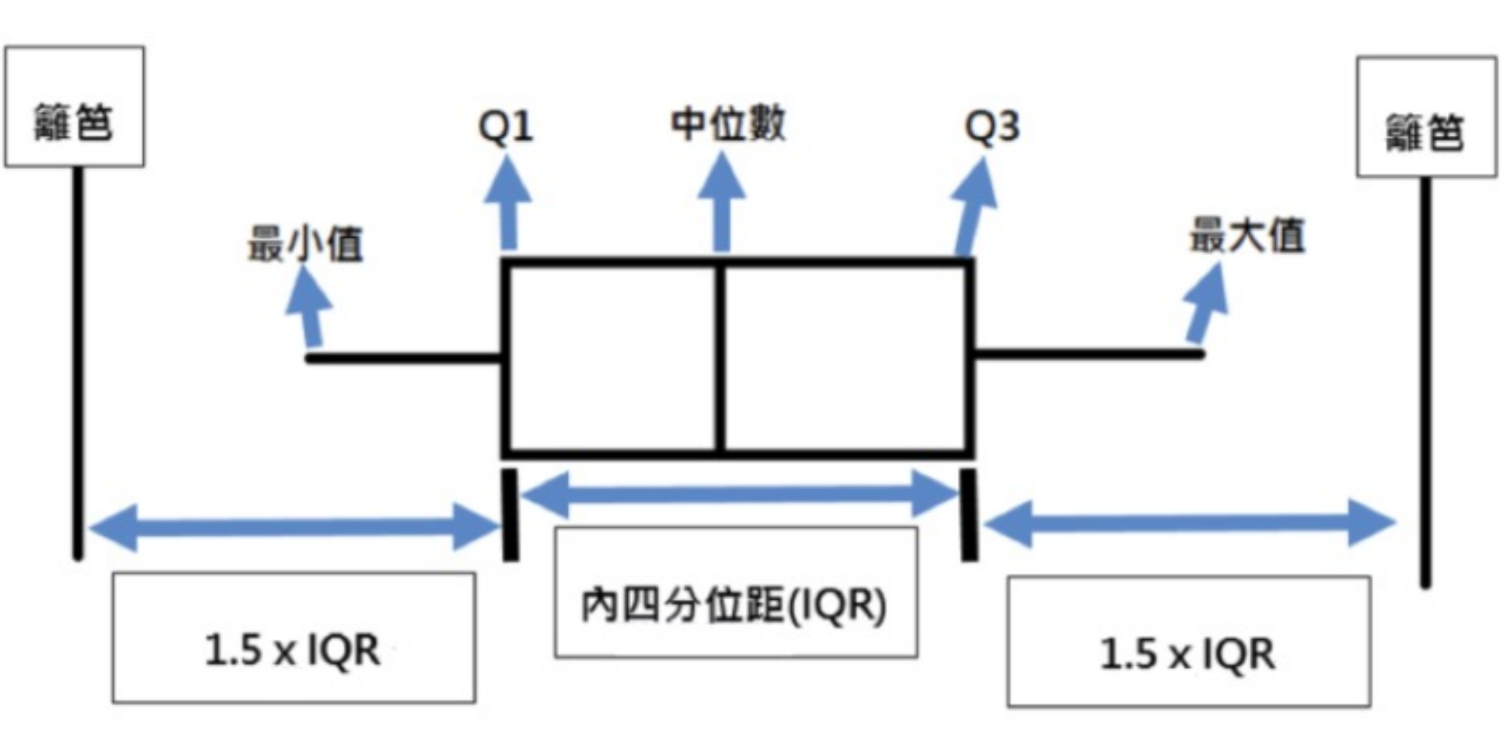
若資料服從正態分佈：根據正態分佈的定義可知，距離平均值 3σ 之外的概率為 $P(|x-\mu|>3\sigma) \leqq 0.003$

這屬於極小概率事件，在預設情況下我們可以認定，距離超過平均值 3σ 的樣本是不存在的。

因此，當樣本距離平均值大於 3σ，認為該樣本為異常值。

異常值的判別方法 3 - 盒鬚圖判別法

- 透過數據，算出Q1 (第一四分位數)、Q3 (第三四分衛數)
- IQR = Q3-Q1
- 最大值與最小值，為籬笆內的最大最小值。
- 超出籬笆外的定義為離群值/異常值。
- 一般而言 d 取 1.5，根據不同資料特性 調整 d 的大小。



異常值的處理方法

刪除異常值：
- 如果異常值是由於數據輸入錯誤、數據處理錯誤或異常值數目很少，我們可以刪除它們。

數據轉換：
- 轉換數據也可以剔除異常值，例如對數據取對數可以減少極端值的變化。

聚類：
- 我們也可以用決策樹直接處理帶有異常值的數據（決策樹基本不會受到異常值和缺失值的影響），或是對不同的觀測值分配權重。

替換：
- **在替換前需判斷為真異常還是偽異常，如果是真異常**，類似替換缺失值，我們也可以替換異常值。我們可以使用均值、中位數、眾數替換方法。

分離對待：
- 如果異常值的數目比較多，在統計模型中我們應該對它們分別處理。一個處理方法是異常值一組，正常值一組，然後分別建立模型，最後對結果進行合併。

知識點回顧

- 異常值通常是一種相對的概念，偏離大部分資料的樣態。
- 簡易偵測異常值的方法有三種，分別為簡單統計分析、3σ原則與盒鬚圖判別法。
- 異常值出現是否該刪除或保留，要根據應用性質與領域知識而定。

回到今天的範例

怎麼運用 python 簡易偵測異常值

Python 語法

- 透過統計量的觀察，看有無異常值

```
df_train['Age'].describe()
```

- 進行 3 倍標準差原則的計算，從而檢視哪些值是可疑的異常值

```
out_index=outliers_z_score(df_train['Age'],3)
```

- 盒鬚圖判別法

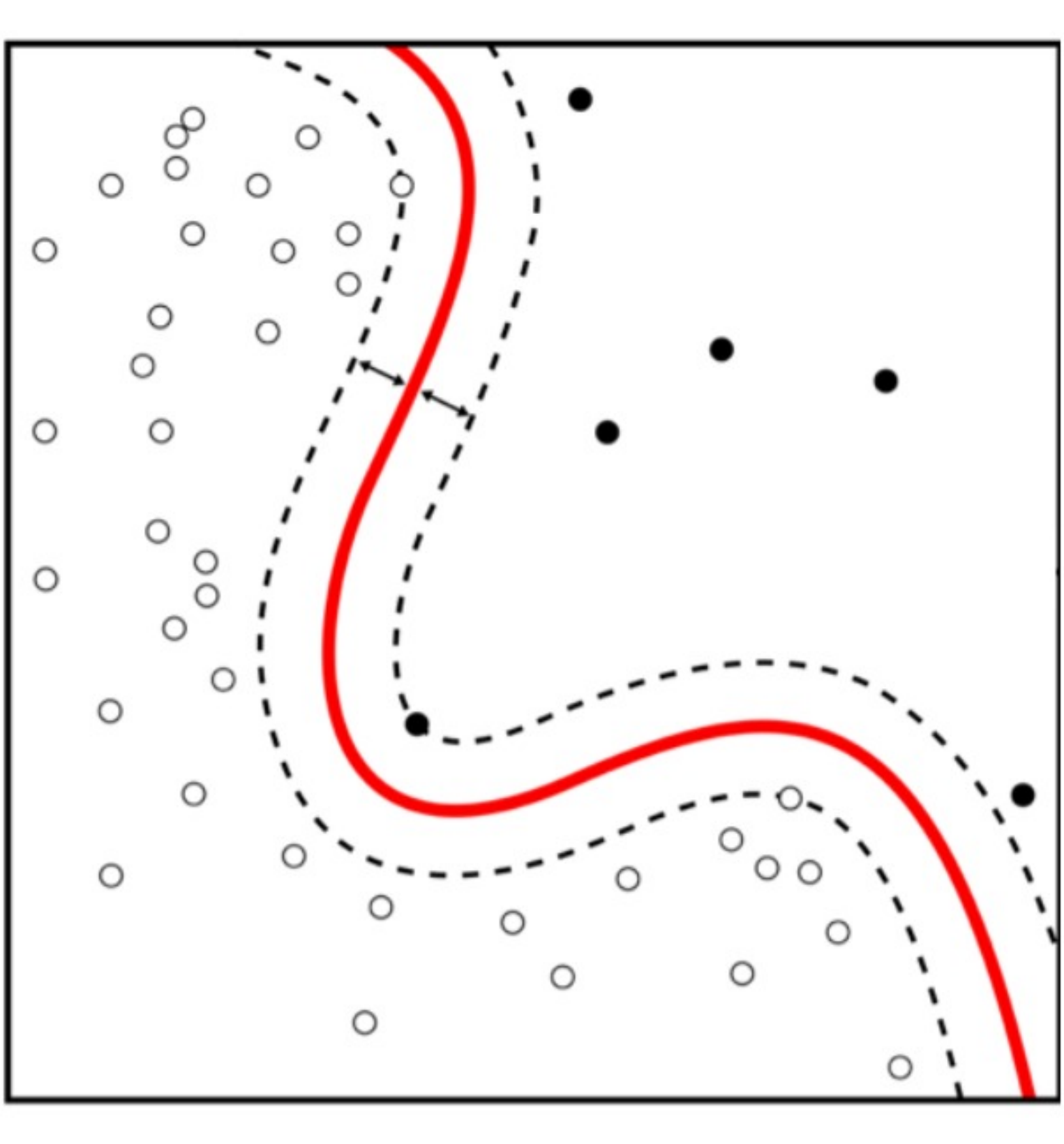
```
out_index2=outliers_iqr(df_train['Age'],1.5)
```

延伸閱讀

進階的異常偵測
網站：異常檢測

透過演算法來辨識異常
- **無監督異常檢測**
- **監督式異常檢測**
- **半監督式異常檢測**

偵測異常，在每一個領域都很重要，也很容易發生，所以還有很多進階的異常偵測模型持續發展中。



下一步：閱讀範例與完成作業