### [Isolation Forest ]

### 1. 什麼是異常檢測？

在講 Isolation Forest 之前，我們先理解一下它的應用場景。異常檢測是指從數據中找出那些「不正常」或「與眾不同」的點。例如：

* 信用卡交易中，尋找詐欺交易。
* 工廠機器監控中，檢測異常運作。
* 網絡安全中，發現入侵行為。

異常點的特徵通常是：它們跟大多數數據點的行為或分佈不一樣。Isolation Forest 就是一種專門設計來找出這些異常點的算法。

### 2. Isolation Forest 的核心想法

Isolation Forest 的核心理念很簡單：**異常點比正常點更容易被「孤立」出來**。什麼意思呢？

想像你有一堆點分佈在一張紙上：

* 正常點通常聚集在一起，像一群人聚在公園裡。
* 異常點則遠離這群點，像一個站在遠處的孤獨的人。

如果我們隨機畫一些線（分割），把這張紙分成小塊：

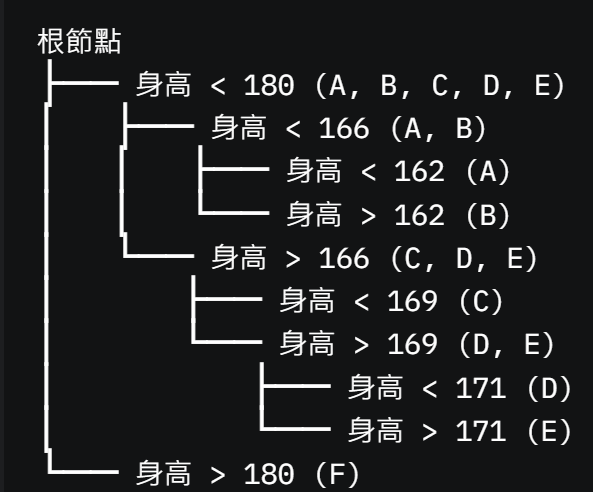
* 異常點因為遠離群體，通常很快就會被單獨分到一個小塊裡。
* 正常點因為互相靠近，需要更多次分割才能被單獨分開。

Isolation Forest 利用這個特性，通過隨機分割數據來「孤立」每個點，並觀察哪些點特別容易被孤立——這些就是異常點。

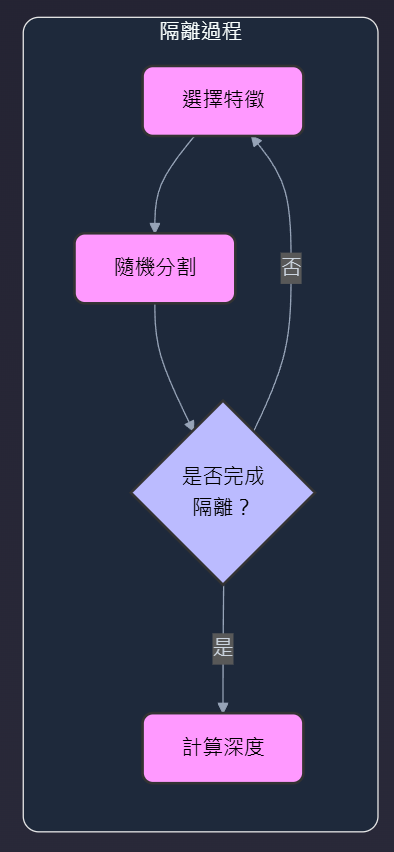
### 3. 怎麼理解這個算法？

舉例:請找出異常點

* A: 160
* B: 165
* C: 168
* D: 170
* E: 172
* F: **200**



#### 構建多棵樹

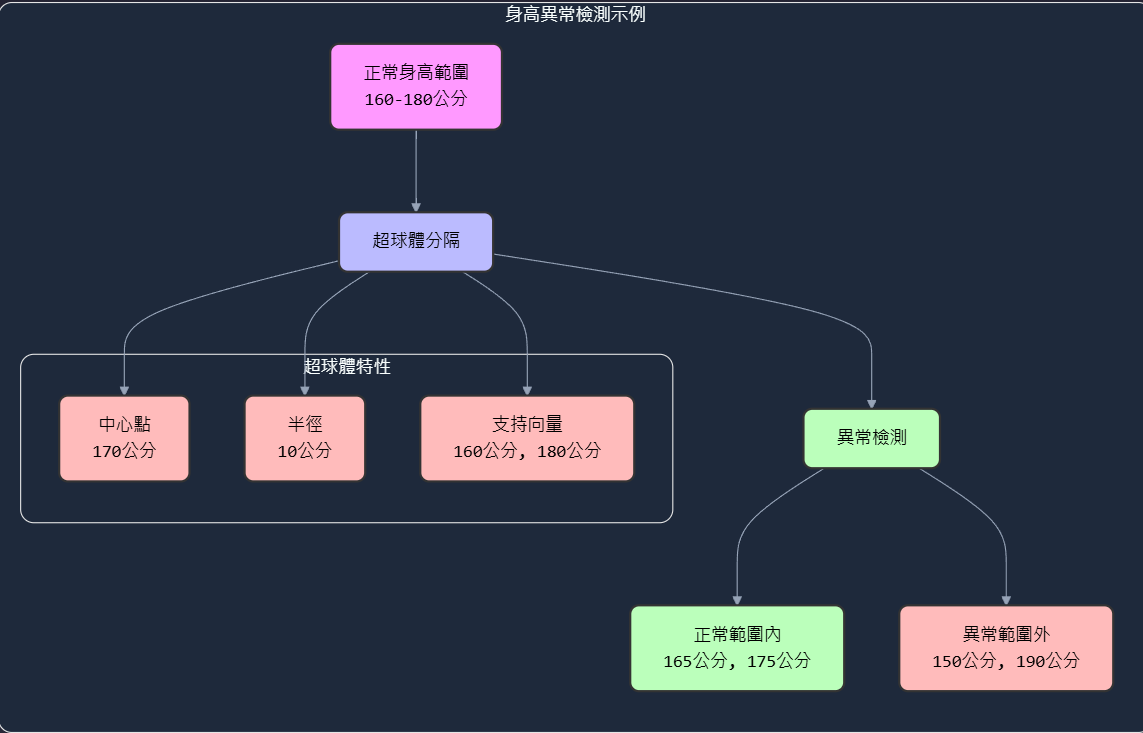
* 一棵樹可能不夠穩定，因為分割是隨機的。所以 Isolation Forest 會構建多棵孤立樹（比如 100 棵），形成一個「森林」。
* 對每棵樹計算某個點的路徑長度，然後取平均值。
* 
* 

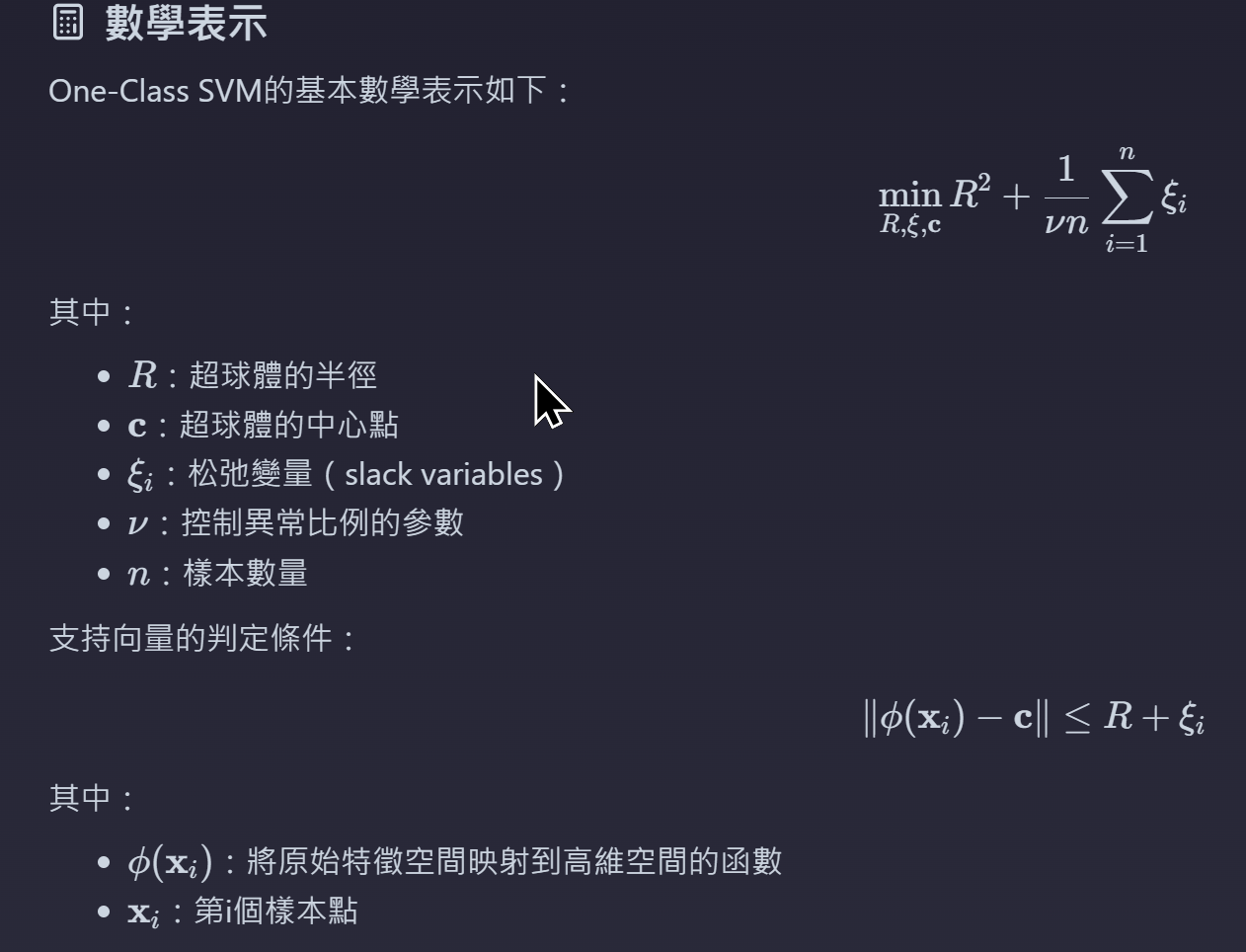
#### 異常分數

* 根據平均路徑長度，計算每個點的「異常分數」（Anomaly Score）。
* 公式是這樣的（簡化版）：
  + 分數接近 1：表示路徑很短，很可能是異常點。
  + 分數接近 0：表示路徑很長，很可能是正常點。
  + 分數在 0.5 附近：不確定，需要更多分析。

[One-Class SVM ]

One-Class SVM的核心思想是：在高維空間中找到一個最小的超球體（Hypersphere），使得大部分正常樣本都被包含在這個超球體內，而異常樣本則位於超球體的外部。這與傳統的SVM不同，傳統SVM是尋找一個超平面來分隔不同類別的樣本。





[Local Outlier Factor ]

通過比較每個數據點的局部密度與其鄰近點的密度來判斷是否為異常值。

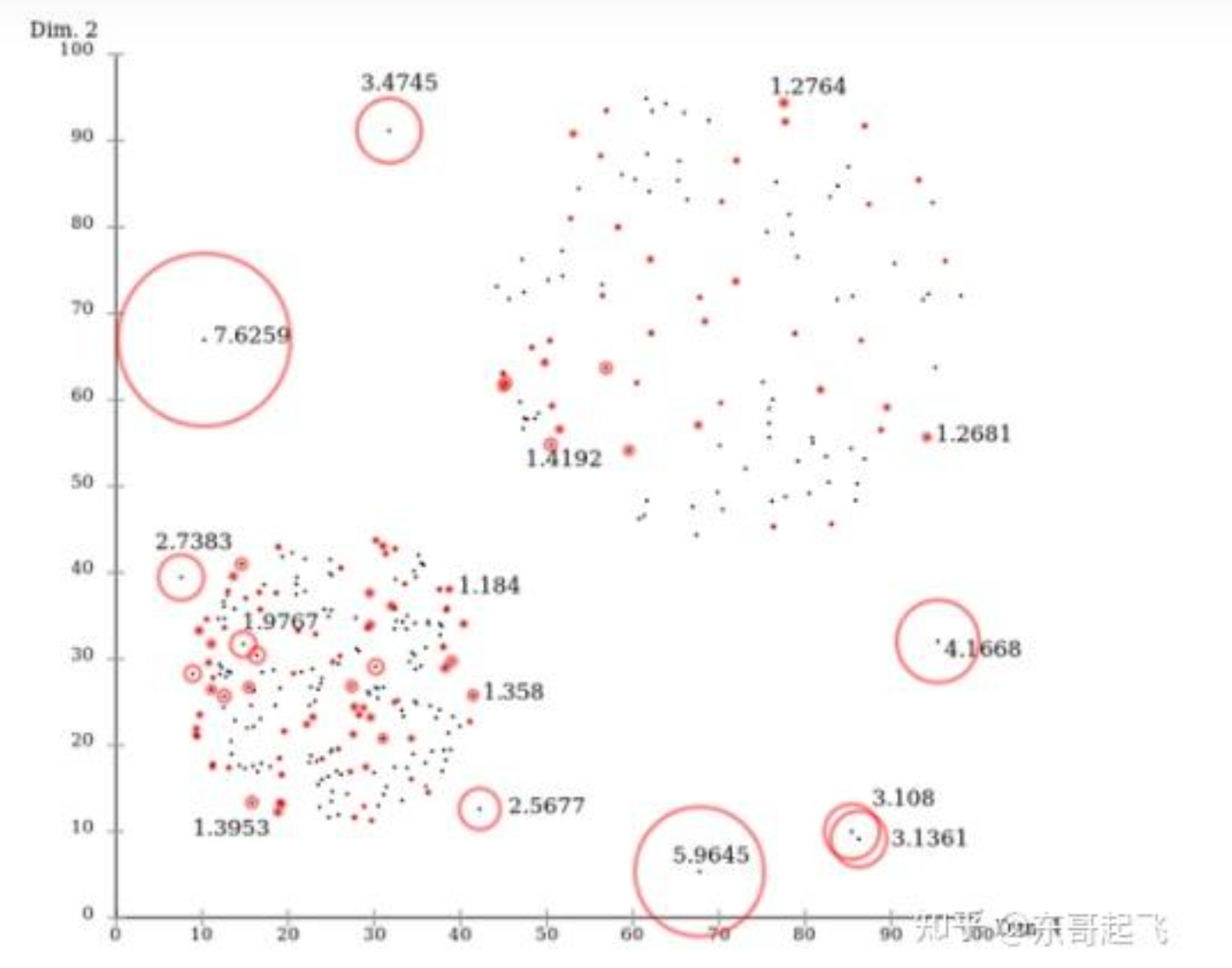
**密度計算步驟**：

* + k鄰近點（k=5）：選擇每個點周圍的5個最近的鄰居
  + 局部可達密度（LRD）：計算每個點在其鄰域內的密度
  + LOF值計算：比較每個點的密度與其鄰近點的平均密度

**異常檢測標準：**

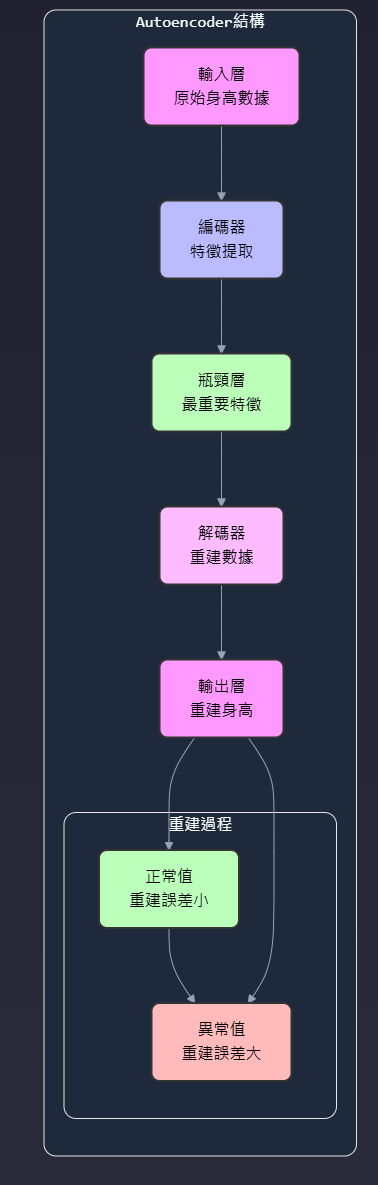
* + LOF值≈1：表示該點的密度與鄰近點相似，為正常值
  + LOF值>1：表示該點的密度明顯低於鄰近點，可能為異常值

LOF值越大越異常,趨近於一表示正常



[Autoencoder ]

通過一個瓶頸結構來強迫網絡學習數據的最重要特徵。想象一下，你要把一本書的內容用一句話總結，必須保留最關鍵的信息。



例子：

1. **正常範圍內的學生**：
   * 輸入：{身高: 170, 體重: 60, BMI: 20.5, 年級: 2, 性別: 男}
   * 通過編碼器：提取身高-體重-BMI的關聯模式
   * 通過瓶頸層：保留核心特徵（如BMI與身高的關係）
   * 重建：{身高: 169.8, 體重: 59.5, BMI: 20.4, 年級: 2, 性別: 男}
   * 重建誤差：所有特徵都在正常範圍內
2. **異常範圍的學生**：
   * 輸入：{身高: 150, 體重: 80, BMI: 35.5, 年級: 2, 性別: 男}
   * 通過編碼器：難以提取正常的模式
   * 通過瓶頸層：無法保留合理的特徵組合
   * 重建：{身高: 165, 體重: 65, BMI: 23.8, 年級: 2, 性別: 男}
   * 重建誤差：BMI值差異過大，表示異常