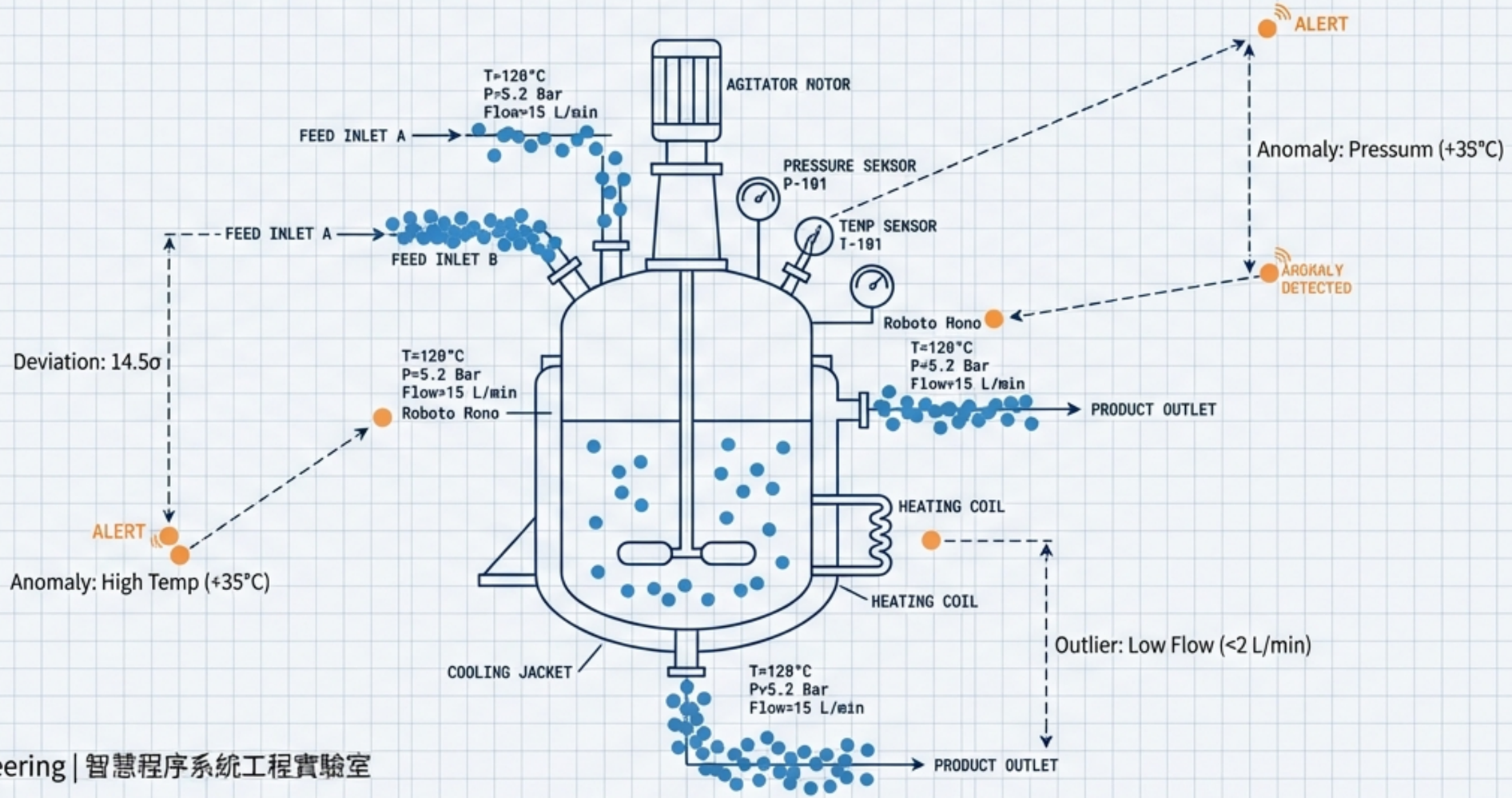
Unit 07 區域性離群因子 (LOF)

應對多密度化工數據的異常檢測方案

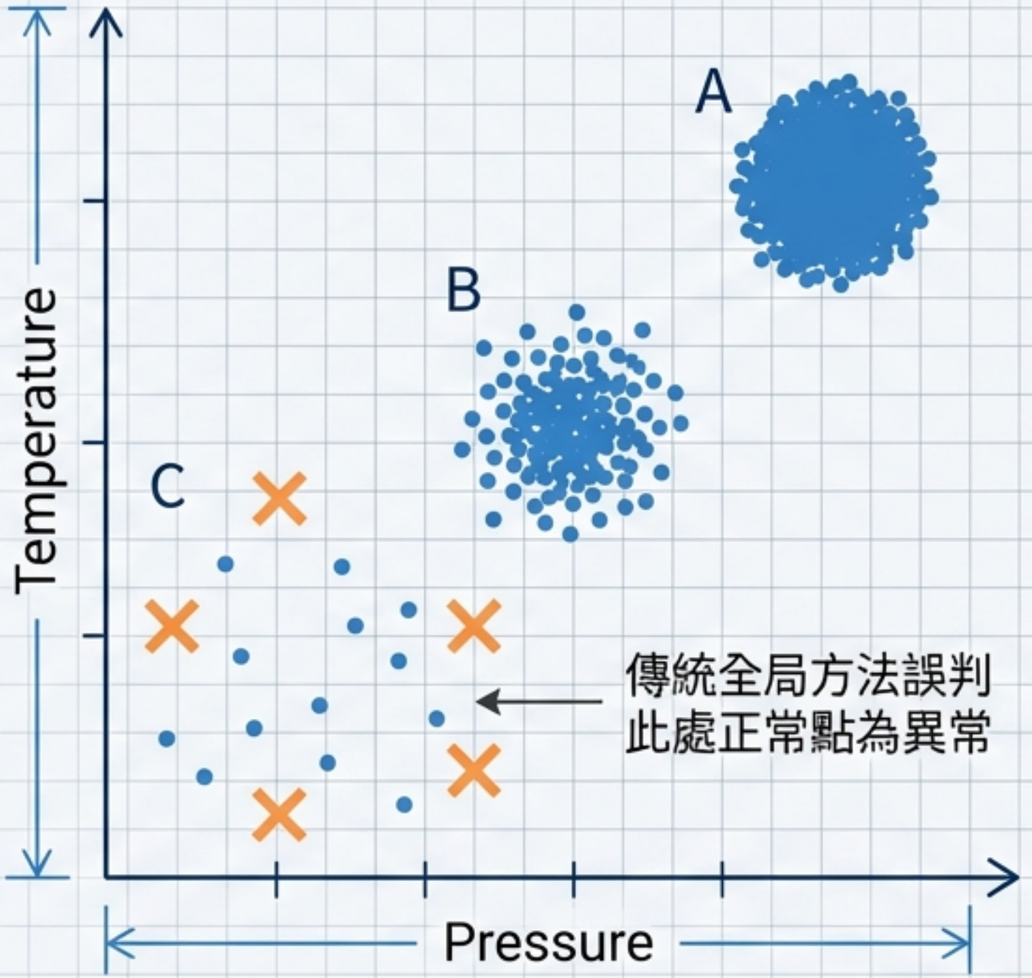


AI in Chemical Engineering | 智慧程序系統工程實驗室
Updated: 2026-01-28

工業現場的挑戰：多模式製程數據 (Multi-Modal Processes)

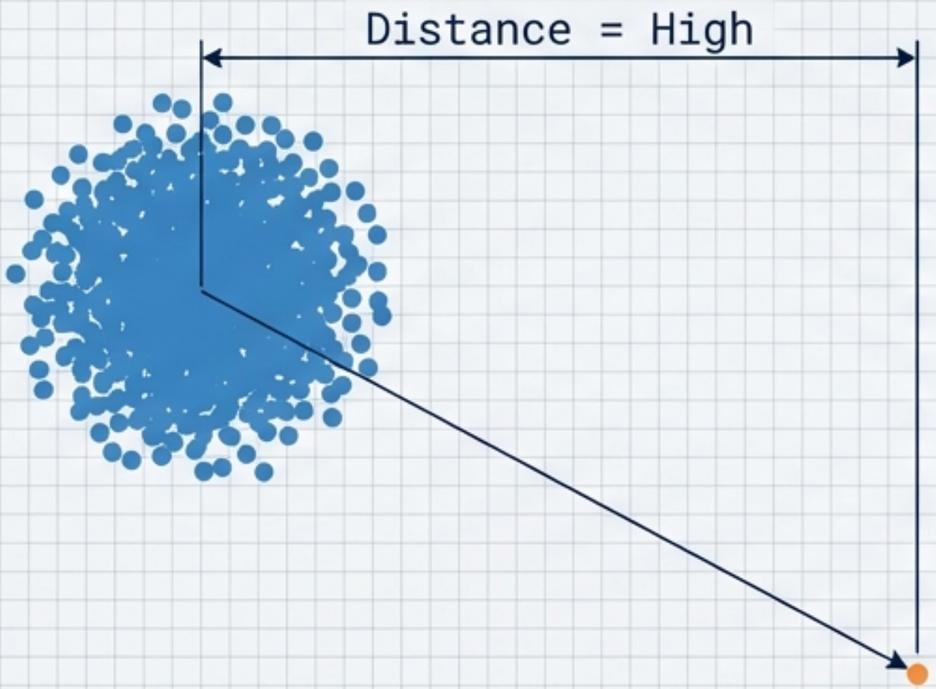
操作配方 (Operating Recipes)

- 1. Recipe A: 高溫高壓 (80-90°C)
  → [數據極度密集]
- 2. Recipe B: 中溫中壓 (60-70°C)
- 3. Recipe C: 低溫低壓 (40-50°C)
  → [數據相對稀疏]

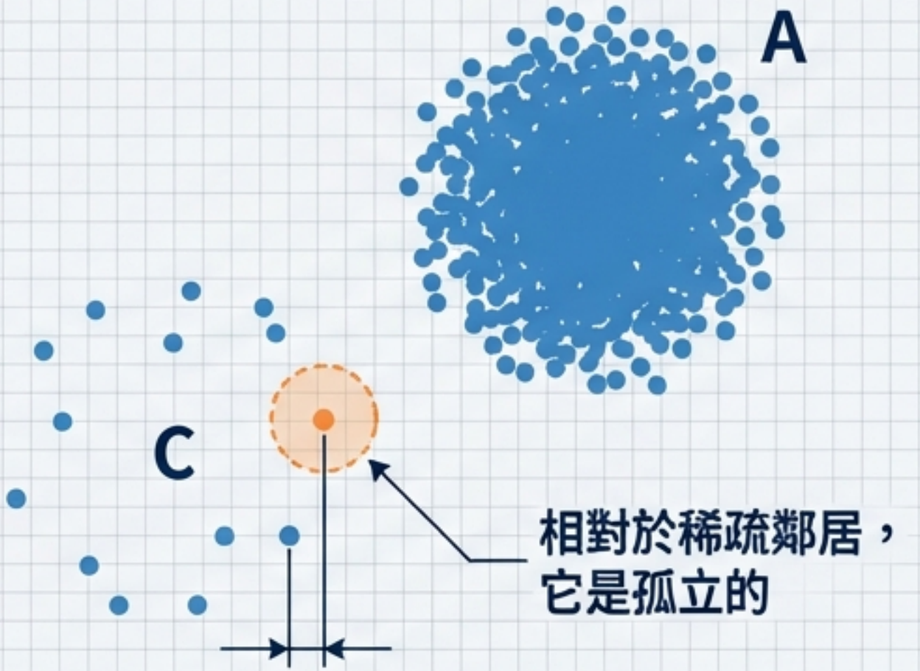


LOF 核心理念：異常是相對於鄰居而言

全局觀點 (Global View)

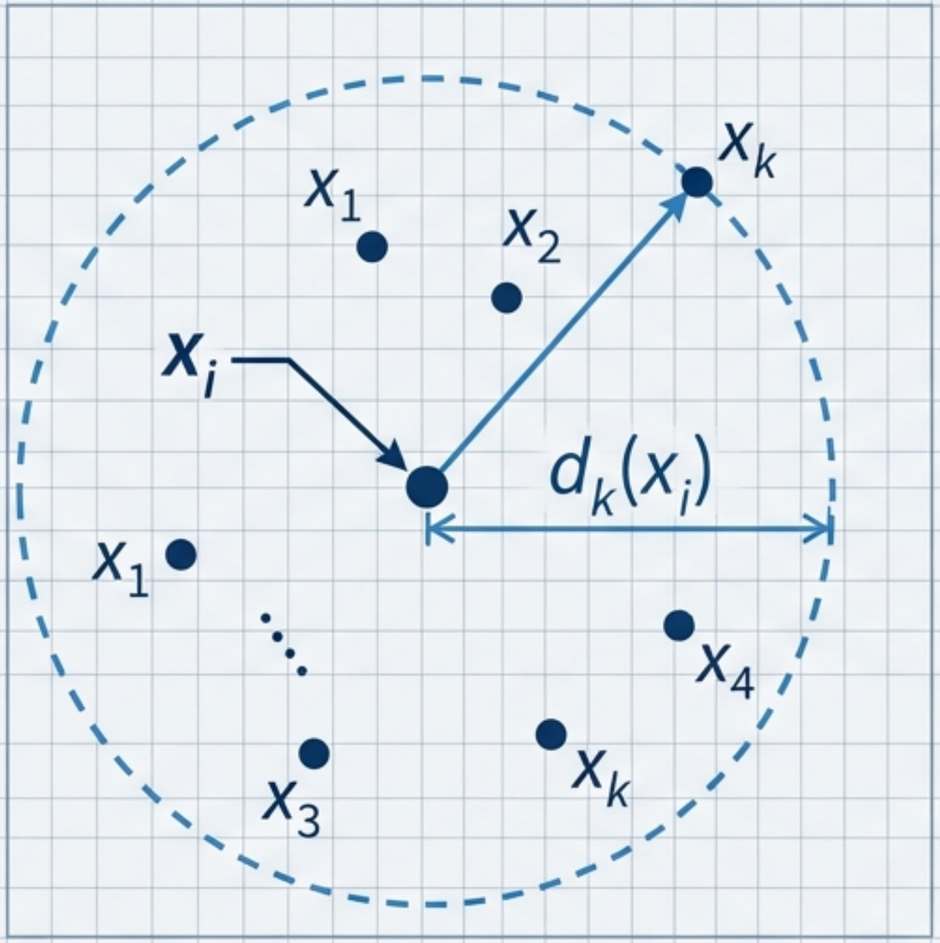


局部觀點 (Local View)



「異常點的局部密度明顯低於其鄰近點的局部密度。」

機制詳解 (1)：定義檢查範圍 (k-Distance)



定義 Definition

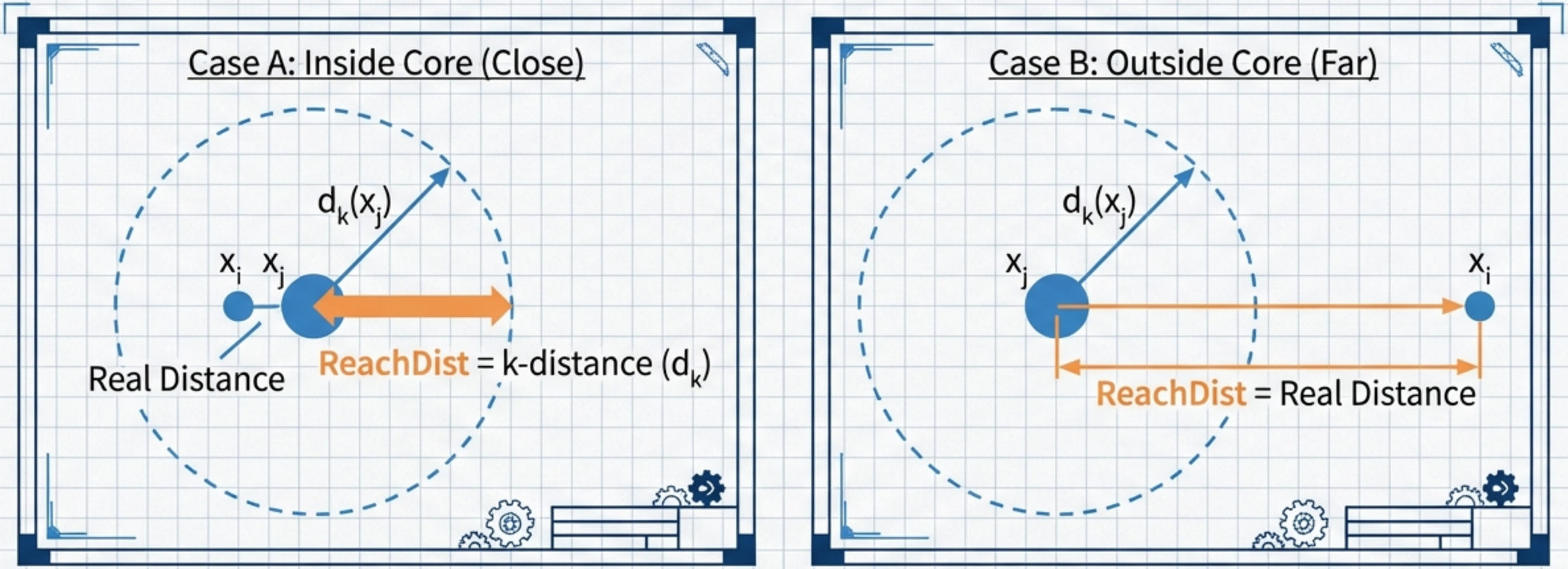
k-distance: 第 k 個最近鄰居的距離
k-neighborhood (N_k)：圓圈內的所有點

工程類比：為每個感測器讀值定義動態的檢查半徑

$$d_k(x_i) = \text{distance}(x_i, x_i^{(k)})$$

機制詳解 (2)：可達距離 (Reachability Distance)

目的：消除統計噪聲，確保密度估計的穩定性 (Stability).



$$\text{reach-dist}_k(x_i, x_j) = \max\{d_k(x_j), \text{distance}(x_i, x_j)\}$$

機制詳解 (3)：LRD 與 LOF 得分計算

Step 1: 局部可達密度 (LRD)

計算鄰居平均可達距離的倒數

$$\text{Density} \propto \frac{1}{\text{Distance}}$$

Step 2: 局部離群因子 (LOF)

比較：鄰居的平均 LRD / 我的 LRD

$$\text{Ratio} = \frac{\text{Neighbors' Density}}{\text{My Density}}$$

Step 3: 判讀結果 (Interpretation)

Ratio ≈ 1 ：正常 (Normal)
Ratio > 1.5：警告 (Warning)
Ratio > 2.0：**異常 (Anomaly)**

LOF Score

Python 實作：Scikit-Learn 工具箱

```python
from sklearn.neighbors import LocalOutlierFactor

# 初始化模型：針對多密度化工數據
lof = LocalOutlierFactor(
    n_neighbors=15,  # 關鍵參數 k（檢查範圍）
    contamination='auto',# 自動決定異常比例
    novelty=False        # False: 檢測歷史數據 (Outlier Detection)
)                        # True:  檢測新數據 (Novelty Detection)

# 訓練並取得結果 (-1 為異常，1 為正常)
y_pred = lof.fit_predict(X_train)
lof_scores = -lof.negative_outlier_factor_  # 轉為正數得分
```

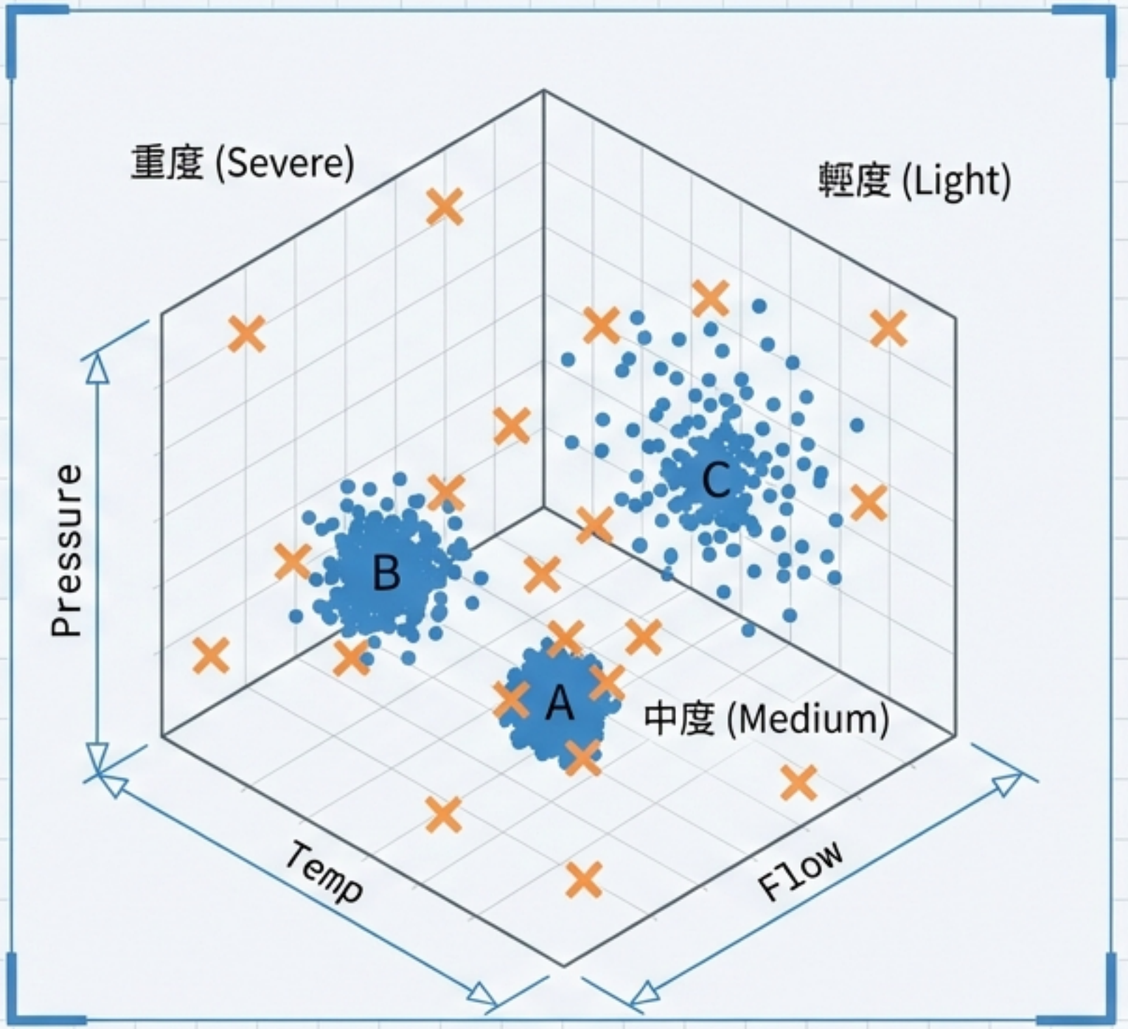
決定敏感度的核心參數

模式切換開關

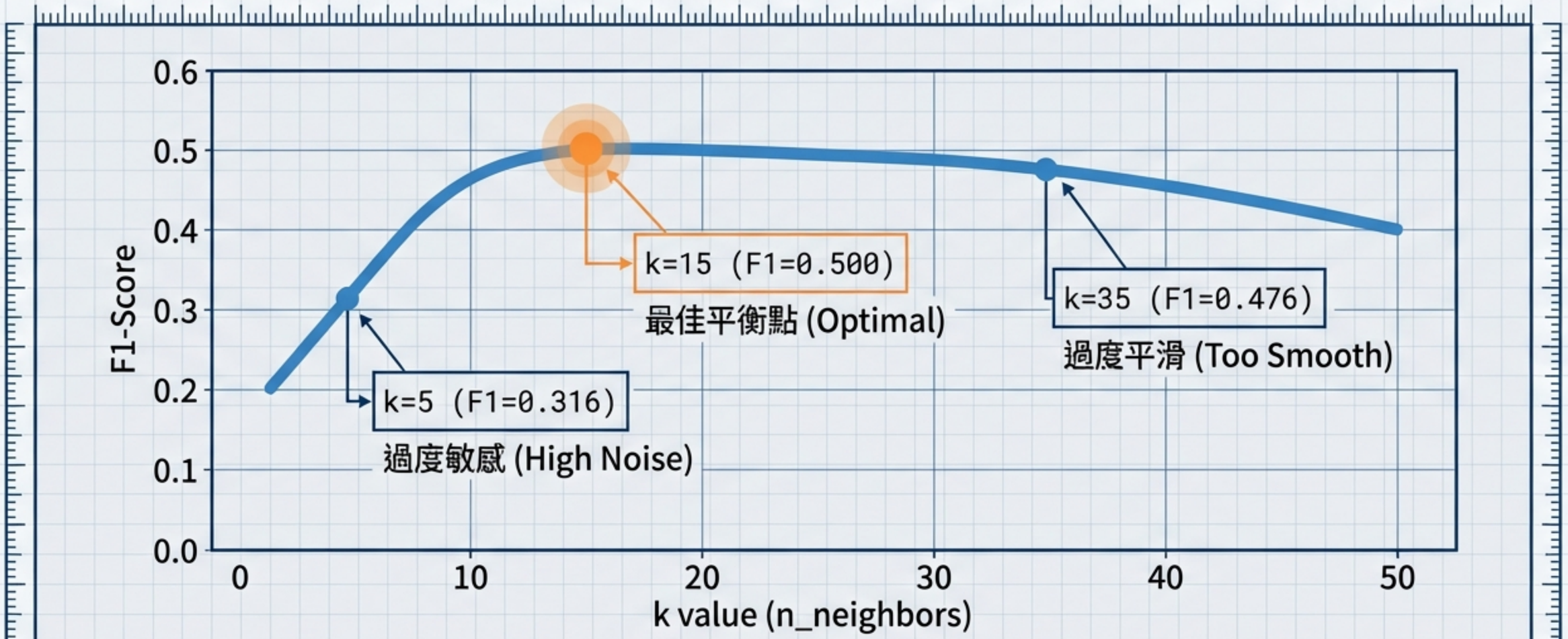
實驗設計：批次反應器模擬 (Batch Reactor Simulation)

數據生成參數

- 配方數量: 3 (Dense, Medium, Sparse)
- 正常數據變異性: +30-40% (模擬真實噪聲)
- 異常設計 (Hierarchical Anomalies):
  - 輕度 (Light): 40%
  - 中度 (Medium): 40%
  - 重度 (Severe): 20%

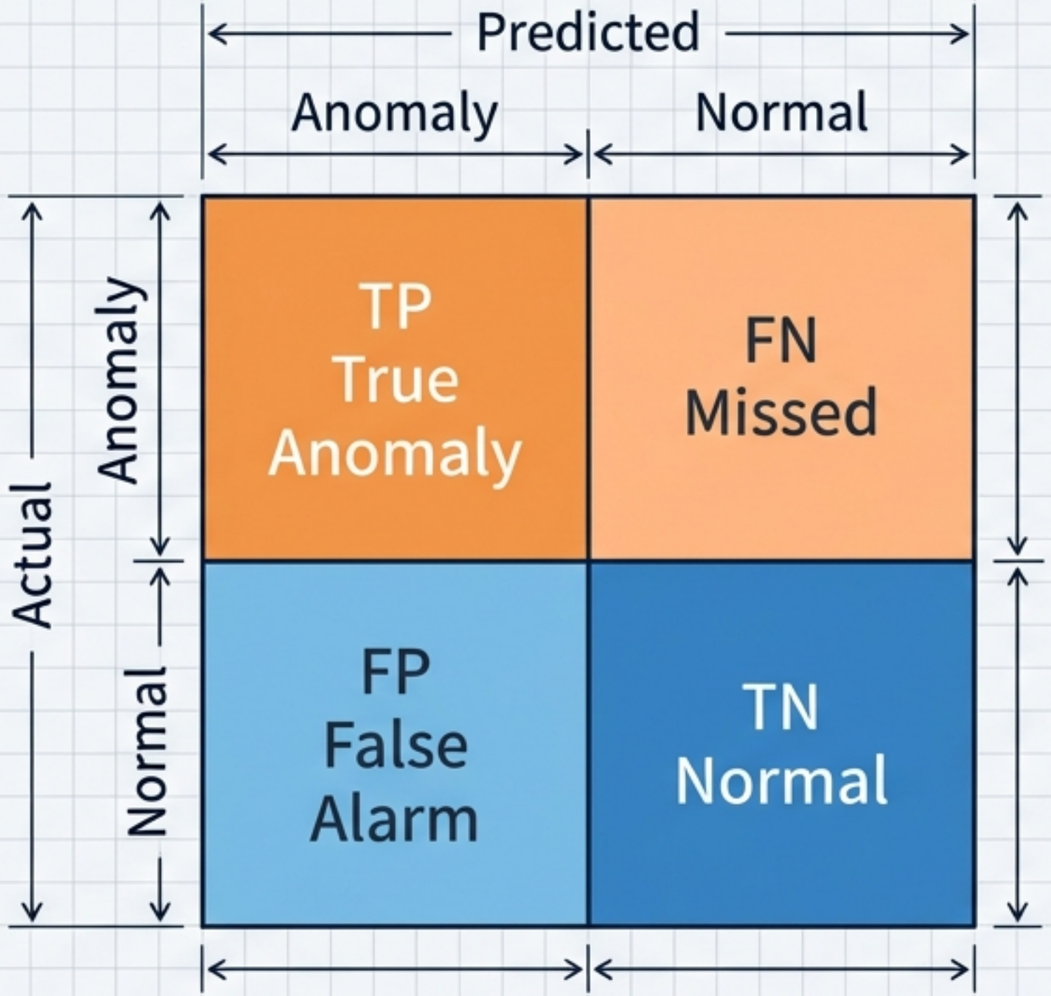


超參數分析：尋找最佳的 'k' (n_neighbors)



結論：對於中等規模數據 (n < 500)，k=10 ~ 20 通常是最佳範圍。

效能評估：真實數據的考驗



Model Metrics (k=15)

Precision: 0.500
50% Reliability

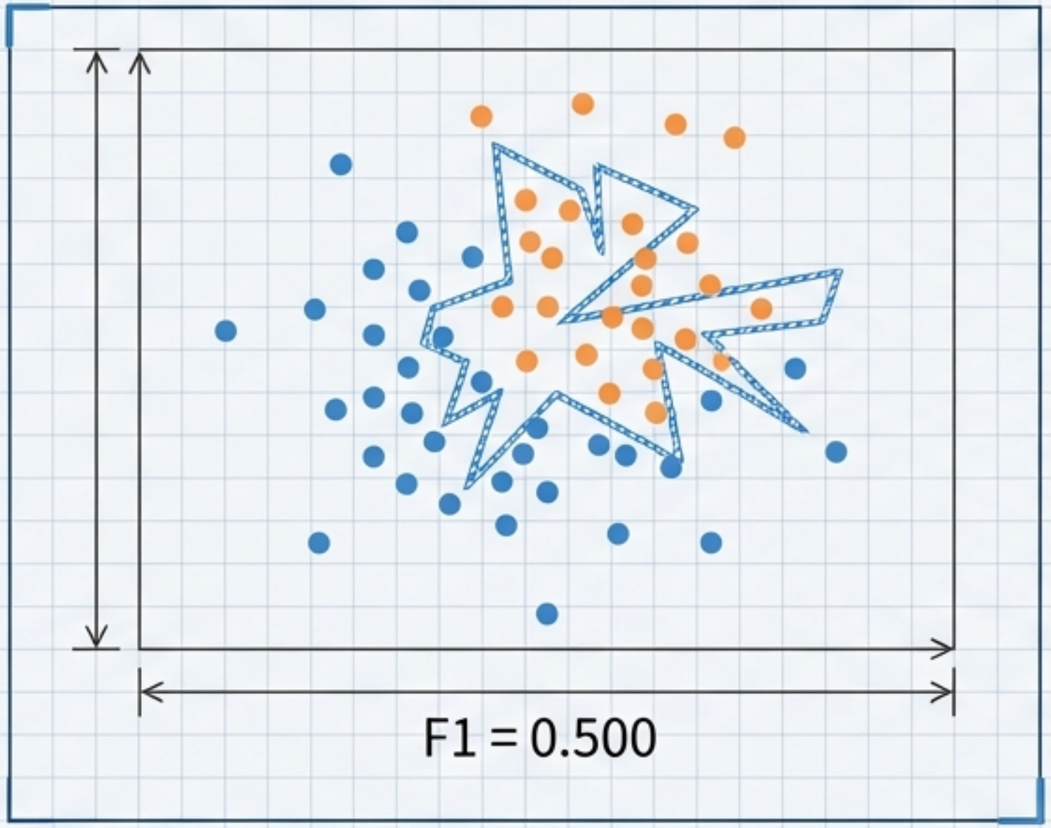
Recall: 0.500
Caught 50% anomalies

AUC: 0.835
Good Separability

Why not AUC 1.0?
真實工業數據存在模糊邊界與輕微異常。
這是合理的現實結果 (Realistic Outcome).

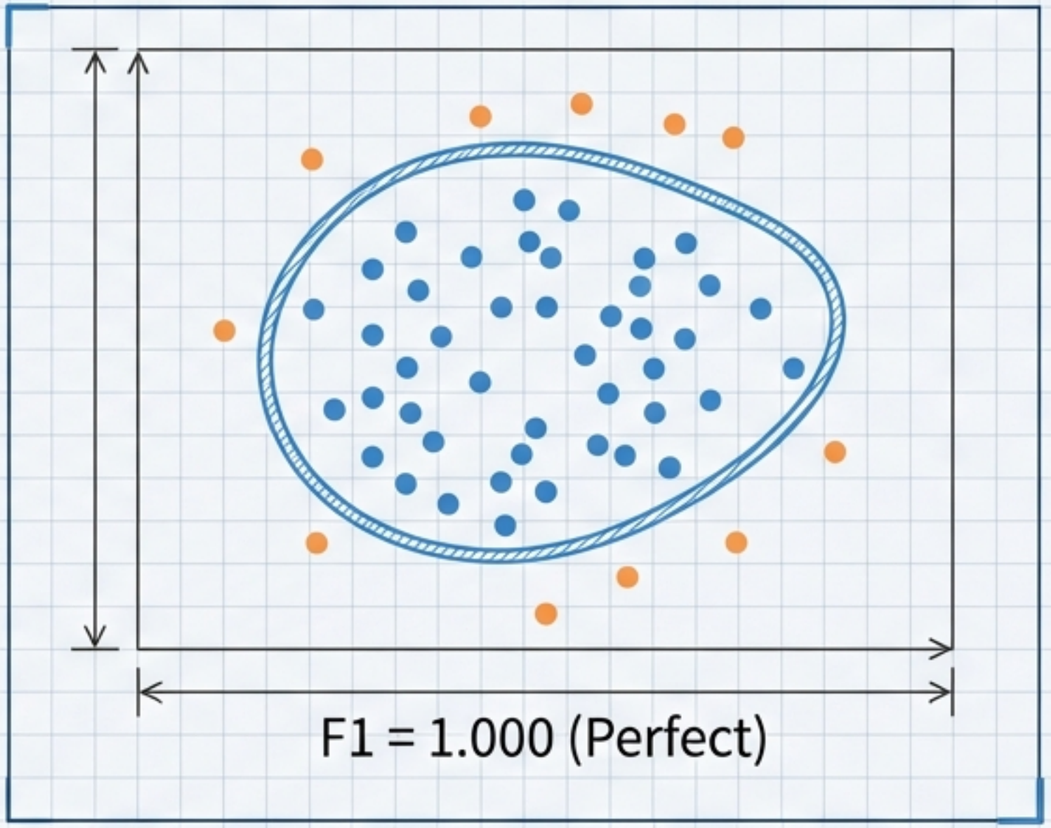
進階應用：新穎性檢測 (Novelty Detection)

Outlier Detection (Mixed Data)



Noto Sans TC

Novelty Detection (Clean History)



Roboto Mono

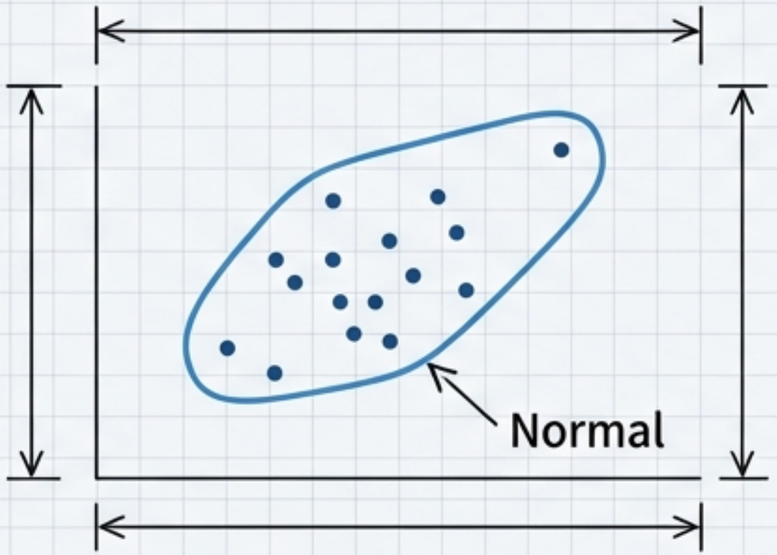
Training on Golden Batch only

應用場景：當擁有純淨的歷史數據時，切換 `novelty=True` 可實現最佳即時監控。

演算法對決：LOF vs. Isolation Forest

LOF (Local Outlier Factor)

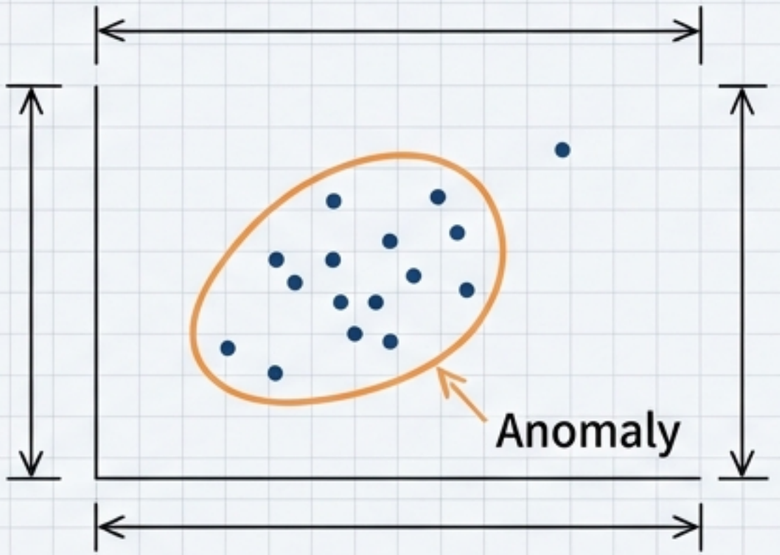
F1 Score: 0.500 ✅



適應局部密度：正確識別稀疏配方為正常。

Isolation Forest

F1 Score: 0.400 ⚠️



全局隔離：誤判稀疏配方為異常 (False Alarm)。

Isolation Forest 難以處理多密度 (Multi-Density) 數據集。

策略總結：LOF 的優勢與代價

優勢 Advantages

- ✓ 處理多密度數據 (Multi-Density Handling)
- ✓ 無需假設分佈 (No Gaussian Assumption)
- ✓ 提供量化異常分數 (Quantifiable Score)

劣勢 Disadvantages

- ⚠ 計算成本較高 $O(n^2)$ (Computationally Expensive)
- ⚠ 對參數 k 敏感 (Sensitive to Hyperparameters)
- ⚠ 高維度表現下降 (Curse of Dimensionality)

最佳適用：複雜、中等規模的製造業數據集 (Complex, Medium-sized Manufacturing Data)。

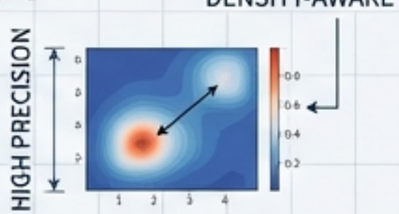
實務部署指南 (Deployment Guide)

↳ Code is read much more often than it is written. — 請保持代碼與模型的可維護性。

結論與下一步

核心價值

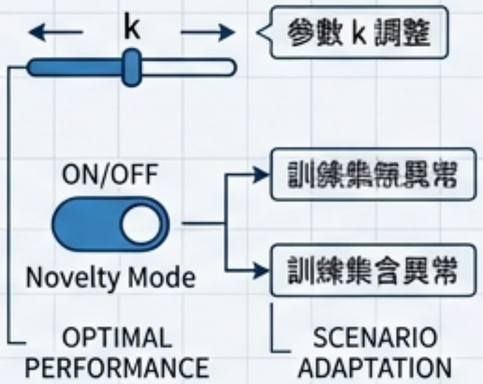
LOF 是應對「非均勻密度」工業數據的最佳檢查員。



關鍵洞察

從「全局距離」轉向「局部密度」視角，大幅降低誤報。

實作建議



注意參數 k 的調整，並針對不同場景選擇 Novelty 模式。

下一步 (Next Step):
開啟 Unit07_LOF.ipynb
開始實作練習