

HW02

Anomaly Detection

313581038 智能系統碩一 蒲品憶

1. Explain your implementation which get the best performance in detail.


我使用的是 KNN (K-Nearest Neighbors) 方法來進行異常偵測，其流程如下：

- **資料處理**：將資料集中不包含異常的資料進行分割，分為訓練集 (train) 與驗證集 (valid)，並使用 Normalizer 和 MinMaxScaler 建立前處理管線，使所有特徵值在相同尺度範圍內，避免因特徵值差異過大而影響距離計算。
- **模型建立**：透過 NearestNeighbors 模型計算每一筆驗證資料與鄰近樣本的平均距離作為異常分數 (anomaly score)，距離越大代表越異常。
- **超參數搜尋**：使用 Hyperopt 套件來對 n_neighbors (鄰居數)、algorithm、metric、leaf_size、p 等參數進行最佳化，並以驗證集的平均異常分數 (loss) 最小化 作為目標函數。
- **最佳模型評估與儲存**：當 loss 值達到當前最佳時，即儲存模型並輸出異常分數至 CSV。

此方法的特點在於使用【非監督式學習】並結合超參數優化，有效地在沒有異常標籤的情況下找出潛在異常資料。

9


313581038



0.99725

16

19h



Your Best Entry!

Your most recent submission scored 0.99725, which is an improvement of your previous score of 0.99266. Great job!

[Tweet this](#)

2. Explain the rationale for using auc score instead of F1 score for binary classification in this homework.

AUC (Area Under the ROC Curve) 與 F1 score 都是常用的二元分類指標，但在這份作業中使用 AUC 分數會更合適，原因如下：

- **異常比例不平衡**：F1 score 對正負樣本比例敏感，若異常樣本很少，容易導致 F1 分數失真；而 AUC 是基於排序的指標，不依賴閾值，較能真實反映模型的異常識別能力。
- **模型產生的是異常分數，而非分類結果**：KNN 會輸出連續的異常分數，AUC 可以評估整體排序的好壞，而 F1 score 需要先設定一個固定閾值轉換為 0/1，再進行評估。
- **AUC 具有穩定性與可解釋性**：它能直接反映模型對於區分正常與異常樣本的能力，越接近 1 表示分類效果越好。

因此，在沒有明確標籤且關心整體排序優劣的異常偵測任務中，AUC 是更合適的評估方式。

3. Discuss the difference between semi-supervised learning and unsupervised learning.

- **非監督學習 (Unsupervised Learning)**

不使用任何標籤 (label) 資料，只根據樣本之間的特徵關係進行學習。

常見應用包括：

✧ 聚類 (Clustering，如 KMeans)

✧ 降維 (如 PCA)

✧ 異常偵測 (如 Isolation Forest、Autoencoder)

- **半監督學習 (Semi-supervised Learning)**

是介於監督式與非監督式之間的方法，**僅有一小部分樣本有標籤，大部分樣本是無標籤**。模型會先利用有標籤資料進行初步學習，再結合無標籤資料進行泛化學習。適用於標籤取得困難或成本高的場景。

比較項目	非監督學習	非監督學習
使用標籤	X	V
資料需求	全部無標籤	少量有標籤 + 大量無標籤
模型訓練方式	依賴數據內部結構	利用有標籤樣本先學習，再泛化到無標籤樣本
適用情境	聚類、降維、異常偵測等	資料難以標註或標註成本高的分類問題