# 機器學習作業 Lab02 報告

**資訊工程學系 黃右萱(0416323)**

## 題敘：

使用 White Wine Quality Data Set 資料庫訓練 K-Nearest-Neighbor Classification Model，並以 K-fold Cross Validation、Resubstitution Validation 進行驗證。

## 做法：

利用 Manhattan Norm, Euclidean Norm, Cosine Similarity, Minkowski Norm 做為判斷，搜尋資料中與 query data 最接近的前 k 筆(k = pow(dataSize,0.4) )資料綜合判斷得到 prediction，值得注意的是，對於 Cosine Similairty，我們採用一個等效的作法：先將 Feature Data 進行 L2-Normalization，再採用 Euclidean Norm 進行訓練。

詢問時，搜尋方式採用 Linear Search (Brute)、KD-Tree Search、Ball-Tree Search 與前述訓練算法交叉運用，最後採用 norm-based 加權對若干最鄰近點進行綜合判斷。

評估採用 Resubstitution 和 10-Fold 兩種方法交叉比對，並記錄 Query Time 的總和進行參照。

## Library:

1. Scikit-Leanrn 建構 Classification Model
2. Numpy 做矩陣運算
3. Math 做數學運算
4. Time 程序內計時

## 結果分析：**[完整結果資料在 result.txt 檔案中]**

總體上，各方法 K - Fold 的結果約為 65%，Resubstitution 約為 100%。

時間效率上而言，Linear Search 在 Manhattan Norm, Euclidean Norm, Cosine Similarity, Minkowski Norm 耗時分別約為[0.38, 0.5, 0.51, 0.51]，而 KD-Tree 與 Ball-Tree 的時間分別為[0.08, 0.06, 0.07, 0.07]、[0.18, 0.19, 0.16, 0.17]，相比之下 Linear Search 時間效率較其餘兩者搜尋方法差。

對於 K-Fold 交叉檢驗，Euclidean 之 Validation 約為 64.3%，而其他方法約為 64.9%，可見 Euclidean 訓練效果相對遜色。

## 備註：

1. 語言：Python 2.7.6
2. 環境：Ubuntu 14.04.5
3. K-fold: 採用 10-fold