Machine Learning HW1

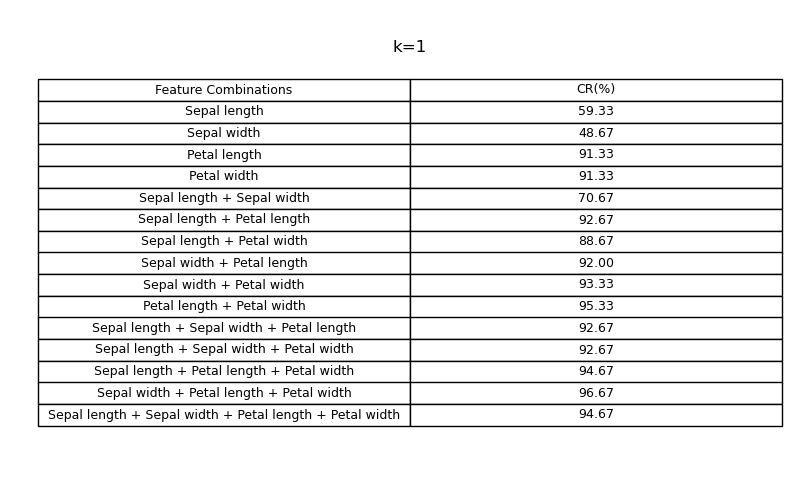
312512005 黃名諄

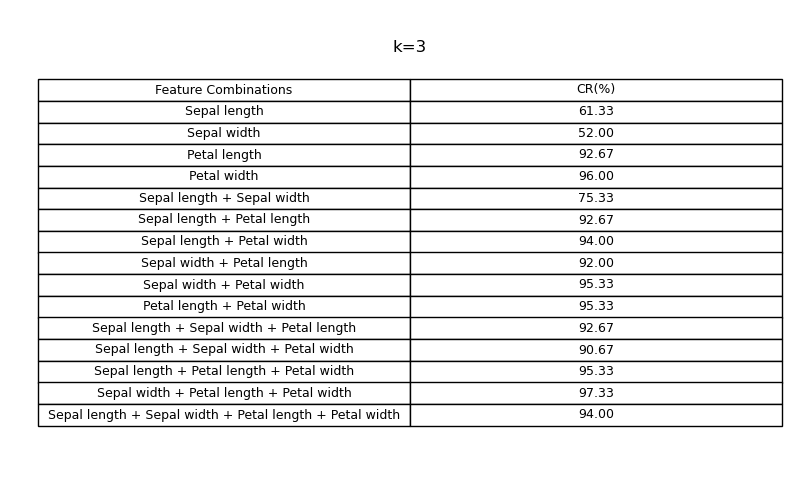
1. Scatter plot :

|  |  |
| --- | --- |
|  |  |
| Sepal length v.s. Sepal width | Sepal length v.s. Petal length |
|  |  |
| Sepal length v.s. Petal width | Sepal width v.s. Petal length |
|  |  |
| Sepal width v.s. Petal width | Petal length v.s. Petal width |

1. Classification Rate(CR):

* K=1



* K=3

1. 結果討論:
   * 散佈圖使以兩兩特徵組合繪製，因此可對應到CR table中使用兩種特徵去做knn分類的情況來做比較。
   * 在k=1及k=3使用兩種特徵的情況下，Sepal length + Sepal width 的分類率都是6種中最低的，只有70.67 %(k=1)及75.33 %(k=3)，在分佈圖Sepal length v.s. Sepal width中也能看出，Versicolor 和Virginica的分佈是混雜在一起的，只有Setosa是明顯分開的，因此用此兩種特徵去做knn分類，只能有效分出Setosa，對於其他兩類則難以區分，造成分類率低的問題
   * 其他5種組合下，分類率都不錯，從散佈圖來看，也可發現3種類別的劃分區域較Sepal length + Sepal width來的明顯，和較高的分類率互相吻合，而當中又以Setosa分布區域最獨立，可預期此品種的分類預測會最正確，Versicolor 和Virginica的邊界還是有一些混雜，因此這兩類可預期會有預測錯誤發生，而這也是分類率無法達到100%的主要原因
   * 另一方面，因k=3使用了更多點去做分類上的預測，等於多了幾個能提供分類依據的點，比起k=1多了些許驗證性，其分類率可能較高的結果也是可預期的，在此6種情況中，k=3的分類率都大於等於k=1的情況。
     + 在k=1或 k=3中，可發現單獨使用Sepal length 和Sepal width的分類率都非常低，因此此兩特徵各自在3品種中可能較沒有明顯差異能用來分類。
   * 所有特徵組合中，不論在k=1或 k=3中，都可發現用

Sepal width + Petal length +Petal width的情況能有最高的分類率，可得到結論，使用此3種特徵可能能夠最有效的區分這3個品種。

* + - 在實作過程中，有發現在單獨使用Sepal length 和Sepal width特徵時，點間距離會常有一樣的情況(也對應到低分類率的事實)，這種情況下，因浮點數精度問題，選最近點的結果依據不同的算法可能會有些許不同，和同學討論後，我認為這是導致大家的CR有些許細微差異的原因。我的邏輯是先依距離從小到大排列(使用argsort)，遇到相同距離時也會依照原順序排列(照類別1、2、3的順序)，再根據k值選前k個項。

1. 心得:

這次作業讓我能實際去撰寫knn分類器，比起過去大學部的課上都只有使用開源套件來操作，此次實作能讓我更深刻的將上課學到的邏輯概念和自己的程式碼連結起來，knn的邏輯在上課時聽起來不難，透過作業能更了解實作上架構的考量及不同情況或參數下分類能力的表現，也透過作業更熟悉像numpy、pyplot等函式功能使用，對於未來作業肯定有所幫助。