ML Final Project Report

Dataset: Mushroom dataset

Validation: 5-cross validation

1. Members

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 組員 | 黃梓軒 | 莫宜豪 | 林語新 | 繆鎬宇 |
| 貢獻度 | 25% | 25% | 25% | 25% |
| 使用model | KNN | Bayes model | Decision tree | SVM |
| 作業系統 | Windows | Windows | Mac OS | Windows |
| 使用語言 | Python | Python | Python | Matlab |
| 預測正確率 | 99.39759% | 77.47415% | 98% | 99.9015% |

1. Dataset

**Data Set Information:**

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like ``leaflets three, let it be'' for Poisonous Oak and Ivy.

**Attribute Information:**

1. cap-shape: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s   
2. cap-surface: fibrous=f, grooves=g, scaly=y, smooth=s   
3. cap-color: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y   
4. bruises?: bruises=t, no=f   
5. odor: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s   
6. gill-attachment: attached=a, descending=d, free=f, notched=n   
7. gill-spacing: close=c, crowded=w, distant=d   
8. gill-size: broad=b, narrow=n   
9. gill-color: black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y   
10. stalk-shape: enlarging=e, tapering=t   
11. stalk-root: bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?   
12. stalk-surface-above-ring: fibrous=f, scaly=y, silky=k, smooth=s   
13. stalk-surface-below-ring: fibrous=f, scaly=y, silky=k, smooth=s   
14. stalk-color-above-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y   
15. stalk-color-below-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y   
16. veil-type: partial=p, universal=u   
17. veil-color: brown=n, orange=o, white=w, yellow=y   
18. ring-number: none=n, one=o, two=t   
19. ring-type: cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z   
20. spore-print-color: black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y   
21. population: abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y   
22. habitat: grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

1. Model explain
2. KNN:

此model有兩個function分別為 dict 和 cross validation。dict

function為將字母a~z轉成1~26表示，cross validation function則是將本次的data 用 cross validation的方式來進行驗證，並使用confusion matrix來預測其正確率，而本次的 cross validation是使用5-cross validation來進行。

而一開始先將mushroom data以csv檔在213~233行讀入，且call dict function將所有字母data轉成0~26，若遇到?則將該筆data直接跳過，最後將feature 存在mushroom\_X，targert存在mushroom\_Y。

在程式241~252行中我們將data套進knn model中，並將距離的計算方式設為Euclidean\_distance，然後反覆去檢查knn model 的k 值為多少時，confusion matrix的正確率最高；在程式255~266行中則是將knn model的距離計算方式改為Manhattan\_distance並同樣去檢查 knn model的k值為多少時正確率最高，最後可得知本次的mushroom dataset 在knn model中k值=2使用Manhattan\_distance會有最高的預測正確率。

1. Bayes model:

此model分為四個部分，第一部分為bayes model，第二部分為讀檔，第三部分則是將資料遺失的部分以在本次資料中出現最多的資料取代，最後一部份則是進行training和predicition。

另外此model可分為兩部分進行討論:

**(一)77% classification accuracy:**

**有些條件機率的值非常低甚至等於0，然而卻在預測中佔有重要的比例，縱使經過smoothing調整，但是結果並沒有顯著的改善。**

**(二)** **Smoothing parameter k:**

**Smoothing的常數k的範圍取值對於整個預測結果有非常明顯的影**

**響，而由結果得知當k的值由高至低變化時，此model的準確率**

**也會逐漸提高。**

1. Decision tree:

在這邊我的code分成了三個部分：import/convert data, tree build, accuracy calculate。

程式一開始先將data輸入進code中，由於即將使用的library sklearn 只支援整數資料因此使用將自元換成ASCII的方式轉成整數（在這邊我將?也視為是一種data 的attribute來進行計算）。

而在建構decision tree時，在這邊使用了pre purning的方式使其樹不會大於五。最後根據統一檢驗的方式我們使用5-fold cross validation來檢測其準確度，並且將其confusion matrix印出檢驗他的準確率，發現此方式的準確度約落在0.98左右。

1. SVM:

此model主要分三個步驟。首先將8000多筆資料分成22個attribute存取在不同的vector中，接著再將對應的response variable參數丟入建立出SVM model，最後再用resubsitution validation 及 5-fold cross validation 進行驗證，並輸出confusion matrix已獲得準確度。

進行resubsitution validation進行驗證的結果，正確率是100%，而使用5-foldcross validation 驗證的正確率是約為99%。由於正確率過高，所以推測應該是overfitting造成的。

四. Result:

1. KNN:

本次作業結果:

KNN algoritm 在 K=2 時

用 Manhattan\_distance 有最好的正確率:0.993975903614

Confusion matrix:

[[3461 27]

[ 7 2149]]

1. Bayes model:

Misclassification accuracy: 22.52585%

Classification accuracy: 77.47415 %

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Prediction | |
| poisonous | edible |
| Class | poisonous | 0 | 193 |
| edible | 0 | 1431 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Prediction | |
| poisonous | edible |
| Class | poisonous | 0 | 185 |
| edible | 0 | 1440 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Prediction | |
| poisonous | edible |
| Class | poisonous | 745 | 326 |
| edible | 106 | 448 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Prediction | |
| poisonous | edible |
| Class | poisonous | 1055 | 294 |
| edible | 219 | 57 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Prediction | |
| poisonous | edible |
| Class | poisonous | 1118 | 0 |
| edible | 507 | 0 |

1. Decision tree:

K-fold cross validation (pre - pruning depth<3)

[K-fold K=5]

[101. 101. 101. …, 101. 101. 101. ]

[ [822 18]

[ 8 777]]

[112. 101. 101. …, 101. 112. 112. ]

[ [856 5]

[ 28 736]]

[101. 112. 101. …, 101. 112. 101. ]

[ [843 7]

[ 22 753]]

[101. 112. 101. …, 112. 101. 101. ]

[ [828 4]

[ 31 762]]

[112. 101. 101. …, 112. 112. 101. ]

[ [820 5]

[ 27 772]]

||||||||||||||||||||||||||

Overall Classification accuracy= 0.980920651762

4.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Prediction | |
| poisonous | edible |
| Class | poisonous | 3908 | 8 |
| edible | 0 | 4208 |

Accuracy = 0.999015