Aprendizagem 2021/22

Homework II – Group 114

I. Pen-and-paper

1) Answer 1

① $\phi_j(x) = \|x\|_2^j \qquad f(x,w) = \sum_{j=0}^{3} w_j \cdot \phi_j(x)$

$$w = (\phi^T \cdot \phi)^{-1} \cdot \phi^T \cdot z$$

$$\phi = \begin{bmatrix} 1 & \sqrt{2} & 2 & (\sqrt{2})^3 \\ 1 & \sqrt{27} & 27 & (\sqrt{27})^3 \\ 1 & \sqrt{20} & 20 & (\sqrt{20})^3 \\ 1 & \sqrt{14} & 14 & (\sqrt{14})^3 \\ 1 & \sqrt{53} & 53 & (\sqrt{53})^3 \\ 1 & \sqrt{3} & 3 & (\sqrt{3})^3 \\ 1 & \sqrt{8} & 8 & (\sqrt{8})^3 \\ 1 & \sqrt{85} & 85 & (\sqrt{85})^3 \end{bmatrix}$$

$\phi_1(x_1) = (\sqrt{1^2+1^2+0^2})^1 = \sqrt{2}$

$\phi_1(x_2) = (\sqrt{1^2+1^2+5^2})^1 = \sqrt{27}$

$\phi_1(x_3) = (\sqrt{0^2+2^2+4^2})^1 = \sqrt{20}$

$\phi_1(x_4) = (\sqrt{1^2+2^2+3^2})^1 = \sqrt{14}$

$\phi_1(x_5) = (\sqrt{2^2+0^2+7^2})^1 = \sqrt{53}$

$\phi_1(x_6) = (\sqrt{1^2+1^2+1^2})^1 = \sqrt{3}$

$\phi_1(x_7) = (\sqrt{2^2+0^2+2^2})^1 = \sqrt{8}$

$\phi_1(x_8) = (\sqrt{0^2+2^2+9^2})^1 = \sqrt{85}$

$$\phi^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \sqrt{2} & \sqrt{27} & \sqrt{20} & \sqrt{14} & \sqrt{53} & \sqrt{3} & \sqrt{8} & \sqrt{85} \\ 2 & 27 & 20 & 14 & 53 & 3 & 8 & 85 \\ (\sqrt{2})^3 & (\sqrt{27})^3 & (\sqrt{20})^3 & (\sqrt{14})^3 & (\sqrt{53})^3 & (\sqrt{3})^3 & (\sqrt{8})^3 & (\sqrt{85})^3 \end{bmatrix}$$

$$z = \begin{bmatrix} 1 \\ 3 \\ 2 \\ 0 \\ 6 \\ 4 \\ 5 \\ 7 \end{bmatrix}$$

Utilizamos ferramentas computacionais para o cálculo de matrizes e chegamos a

$$w = \begin{bmatrix} 4{,}594 \\ -1{,}696 \\ 0{,}340 \\ -0{,}013 \end{bmatrix}$$

$$f(x,w) = 4{,}594 + (-1{,}696)\cdot\|x\| + 0{,}340\cdot\|x\|^2 - 0{,}013\,\|x\|^3$$

2) Answer 2

②

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\text{Predicted}_i - \text{Observed}_i)^2}{n}}$$

$f(x_9, w) = 4{,}594 - 1{,}696\cdot\sqrt{2} + 0{,}340\cdot 2 - 0{,}013(\sqrt{2})^3 =$

$(\|x_9\| = \sqrt{2}) \quad = \cancel{2{,}458}\ 2{,}839$

$f(x_{10}, w) = 4{,}594 - 1{,}696\cdot\sqrt{6} + 0{,}340\cdot 6 - 0{,}013(\sqrt{6})^3 =$

$(\|x_{10}\| = \sqrt{6}) \quad = 2{,}289$

~~RMSE = (2-2,839)+(~~

$$RMSE = \sqrt{\frac{(2{,}839-2)^2 + (2{,}289-4)^2}{2}} = 1{,}347$$

3) Answer 3

③ Considerando a binarização e classificações "P" e "N" transformamos os dados para o seguinte formato

	y_1	y_2	y_3	output
x_1	1	1	0	N
x_2	1	1	1	N
x_3	0	2	1	N
x_4	1	2	0	N
x_5	2	0	1	P
x_6	1	1	0	P
x_7	2	0	0	P
x_8	0	2	1	P

Para descobrirmos a "root" da árvore vamos calcular a entropia e ganho de informação dos atributos:

$E(y_1=0) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$

$E(y_1=1) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) = 0{,}811$

$E(y_1=2) = -\frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0$

$E(y_2=0) = -\frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0$

$E(y_2=1) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0{,}918$

$E(y_2=2) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0{,}918$

$E(y_3=0) = -\frac{4}{8}\log_2\left(\frac{4}{8}\right) - \frac{4}{8}\log_2\left(\frac{4}{8}\right) = 1$

$E(y_3=1) = -\frac{4}{8}\log_2\left(\frac{4}{8}\right) - \frac{4}{8}\log_2\left(\frac{4}{8}\right) = 1$

$E(\text{total})$ (do output) $= -\frac{4}{8}\log_2\left(\frac{4}{8}\right) - \frac{4}{8}\log_2\left(\frac{4}{8}\right) = 1$

$Gain(\text{Output}, y_1) = E(\text{total}) - \sum [\,p(\text{output}|y_1)\cdot E(\text{output}|y_1)]$

$= 1 - \left(\frac{2}{8}\cdot 1 + \frac{4}{8}\cdot 0{,}811 + \frac{2}{8}\cdot 0\right) = 0{,}3445$

$Gain(\text{Output}, y_2) = 1 - \left(\frac{2}{8}\cdot 0 + \frac{3}{8}\cdot 0{,}918 + \frac{3}{8}\cdot 0{,}918\right) =$

$= 0{,}3115$

$Gain(\text{Output}, y_3) = 1 - \left(\frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 1\right) = 0$

y_1 é o atributo com maior ganho de informação, então será a root da árvore:

Vamos repetir o processo tirando y_1 (omitindo valores que não estão no set de treino)

$E(y_2=2 \mid y_1=0) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$

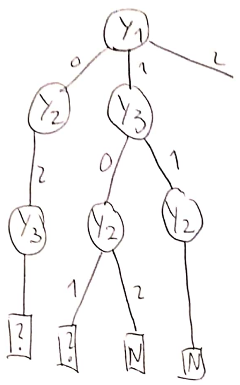
$E(y_3=1 \mid y_1=0) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$

~~Como os valores do set~~

$Gain(y_1=0, y_2) = 1 - \left(\frac{2}{2}\cdot 1\right) = 0$

$Gain(y_1=0, y_3) = 1 - \left(\frac{2}{2}\cdot 1\right) = 0$

Como y_2 e y_3 têm o mesmo ganho, podemos optar por qualquer um destes atributos, vamos optar por y_2 de forma aleatória.

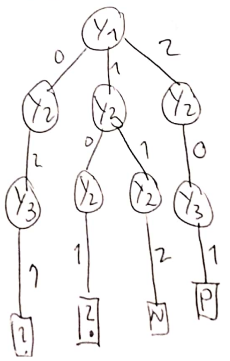


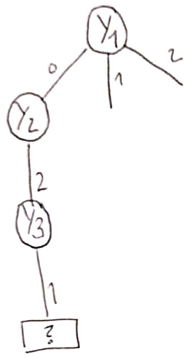
Vamos calcular para $Y_1=2$:

$E(Y_2=0|Y_1=2) = -\frac{1}{1}\log_2(\frac{1}{1}) = 0$

$E(Y_3=1|Y_1=2) = -\frac{1}{1}\log_2(\frac{1}{1}) = 0$

O Ganho também é igual entre os atributos, desta vez escolhemos Y_2:





~~Como~~, (0,1,2) é indefinido, pois, tem classificações N e P

Vamos calcular agora para $Y_1=1$:

$E(Y_3=0|Y_1=1) = -\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3}) = 0,918$

$E(Y_3=1|Y_1=1) = 0 - \frac{1}{1}\log_2(\frac{1}{1}) = 0$

$E(Y_2=1|Y_1=1) = -\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3}) = 0,918$

$E(Y_2=2|Y_1=1) = 0 - \frac{1}{1}\log_2(\frac{1}{1}) = 0$

$Gain(Y_1=1, Y_2) = 1-(\cancel{\frac{3}{4} 0,918} + \frac{3}{4}\cdot 0,918 + \frac{1}{4}\cdot 0)$

$Gain(Y_1=1, Y_3) = 1-(\frac{3}{4}\cdot 0,918 + \frac{1}{4}\cdot 0)$

Mais uma vez Y_2 e Y_3 têm o mesmo ganho. Desta vez selecionamos Y_3:

4) Answer 4

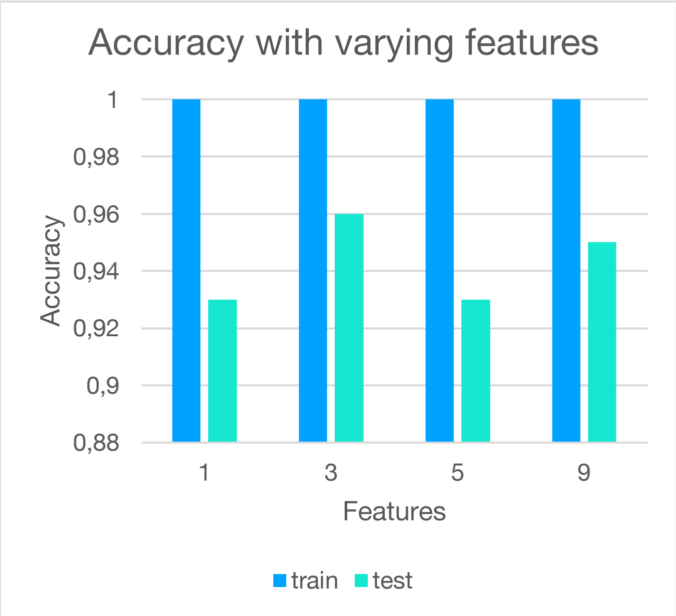
4 Assumindo os valores não representados na árvore como indefinidos quando computamos (2,0,0) obtemos indefinido e quando computamos (1,2,1) = (1,2,0) obtemos "N". ~~Portanto~~ Em termos de "accuracy" significa 0%

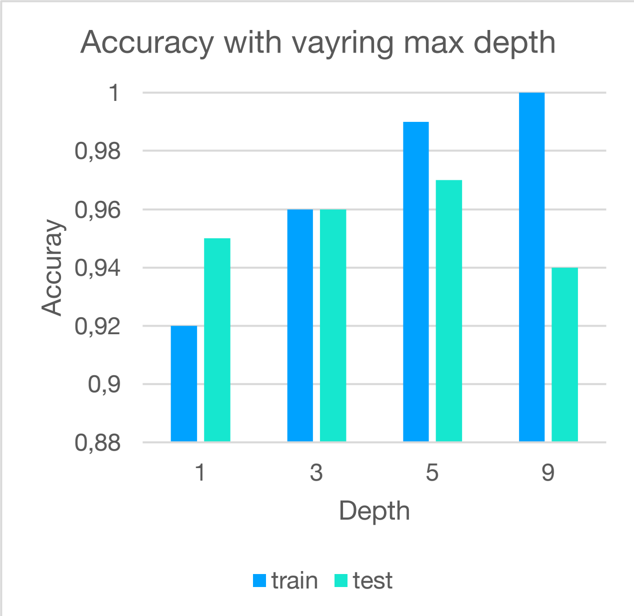
Aprendizagem 2021/22
Homework II – Group 114

II. Programming and critical analysis

5) Answer 5

a. number of selected features in {1,3,5,9} b .maximum tree depth in {1,3,5,9}





6) Answer 6

A correlação observada demonstra que há features que dependem de outras features, estando por vezes associadas/ligadas. Por exemplo, se escolhermos 3 features vamos ter um score muito melhor que se escolhermos 5, devendo-se ao facto de encontrar facilmente 3 features que dependam simultaneamente umas das outras, não sendo o mesmo para 5 features.

Quanto mais profunda for a árvore, mais "ramos" vão surgir, aumentando o erro, e torna a árvore menos eficaz. No caso em que surgem mais folhas pode levar a casos em que seja detetado múltiplas classificações erradas ou indefinidas.

7) Answer 7

A depth = 3 é a escolhida com base na accuracy do conjunto de treino e de teste, visto que diminui a probabilidade de overfitting e consequentemente aumenta a accuracy . Uma escolha do depth com valores de extremo (1 ou 9) aumenta o erro no conjunto de teste e de treino. Ainda assim a amostra de dados não justifica a escolha de um depth máximo.

III. APPENDIX

```
import pandas as pd
import numpy as np
from scipy.io.arff.arffread import print_attribute
from sklearn import tree
from scipy.io.arff import loadarff
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn import metrics
import matplotlib.pyplot as plt
import seaborn as sns

if __name__ == '__main__':
    raw_data = loadarff('breast.w.arff')
    df_data = pd.DataFrame(raw_data[0])
    classe = df_data.pop('Class')
    df_data = df_data.astype(int)
    Y = classe.str.decode('utf-8')
for v in range(0,683):
    if Y.loc(axis=0)[v] == 'benign':
        Y.loc(axis=0)[v] = 1
    else:
        Y.loc(axis=0)[v] = 0

Y = np.array(Y)
X_train, X_test, y_train, y_test = train_test_split(df_data,Y, test_size=0.30)

#Exercise 5 i
for i in [1,3,5,9]:
    tre = DecisionTreeClassifier(criterion = 'entropy',max_features=i)
    y_train = y_train.astype('int')
    y_test = y_test.astype('int')
    tre.fit(X_train,y_train)
    p=tre.score(X_train, y_train).round(2)
    j=tre.score(X_test, y_test).round(2)
    print("Train : %0.3f  Test : %.3f Dif : %0.3f" % (p,j, p-j))

#Exercise 5 ii
for i in [1,3,5,9]:
    tre = DecisionTreeClassifier(max_depth=i)
    y_train = y_train.astype('int')
    y_test = y_test.astype('int')
    tre.fit(X_train,y_train)
    p=tre.score(X_train, y_train).round(2)
    j=tre.score(X_test, y_test).round(2)
    print("Train : %0.3f  Test : %.3f Dif : %0.3f" % (p,j, p-j))
```

END