TÉCNICO LISBOA

Aprendizagem 2021/22

Homework II – Group 057

I. Pen-and-paper

1)

$$z = \begin{pmatrix} 1 \\ 3 \\ 2 \\ 0 \\ 6 \\ 4 \\ 5 \\ 7 \end{pmatrix} \qquad \Phi_j = \begin{pmatrix} 1 & \sqrt{2} & 2 & (\sqrt{2})^3 \\ 1 & \sqrt{27} & 27 & (\sqrt{27})^3 \\ 1 & \sqrt{20} & 20 & (\sqrt{20})^3 \\ 1 & \sqrt{14} & 14 & (\sqrt{14})^3 \\ 1 & \sqrt{53} & 53 & (\sqrt{53})^3 \\ 1 & \sqrt{3} & 3 & (\sqrt{3})^3 \\ 1 & \sqrt{8} & 8 & (\sqrt{8})^3 \\ 1 & \sqrt{85} & 85 & (\sqrt{85})^3 \end{pmatrix}$$

$$w_j = (\Phi_j^T \Phi_j)^{-1} \Phi_j^T z$$

$$= \left(\begin{pmatrix} 1 & \sqrt{2} & 2 & (\sqrt{2})^3 \\ 1 & \sqrt{27} & 27 & (\sqrt{27})^3 \\ 1 & \sqrt{20} & 20 & (\sqrt{20})^3 \\ 1 & \sqrt{14} & 14 & (\sqrt{14})^3 \\ 1 & \sqrt{53} & 53 & (\sqrt{53})^3 \\ 1 & \sqrt{3} & 3 & (\sqrt{3})^3 \\ 1 & \sqrt{8} & 8 & (\sqrt{8})^3 \\ 1 & \sqrt{85} & 85 & (\sqrt{85})^3 \end{pmatrix}^T \begin{pmatrix} 1 & \sqrt{2} & 2 & (\sqrt{2})^3 \\ 1 & \sqrt{27} & 27 & (\sqrt{27})^3 \\ 1 & \sqrt{20} & 20 & (\sqrt{20})^3 \\ 1 & \sqrt{14} & 14 & (\sqrt{14})^3 \\ 1 & \sqrt{53} & 53 & (\sqrt{53})^3 \\ 1 & \sqrt{3} & 3 & (\sqrt{3})^3 \\ 1 & \sqrt{8} & 8 & (\sqrt{8})^3 \\ 1 & \sqrt{85} & 85 & (\sqrt{85})^3 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & \sqrt{2} & 2 & (\sqrt{2})^3 \\ 1 & \sqrt{27} & 27 & (\sqrt{27})^3 \\ 1 & \sqrt{20} & 20 & (\sqrt{20})^3 \\ 1 & \sqrt{14} & 14 & (\sqrt{14})^3 \\ 1 & \sqrt{53} & 53 & (\sqrt{53})^3 \\ 1 & \sqrt{3} & 3 & (\sqrt{3})^3 \\ 1 & \sqrt{8} & 8 & (\sqrt{8})^3 \\ 1 & \sqrt{85} & 85 & (\sqrt{85})^3 \end{pmatrix}^T \begin{pmatrix} 1 \\ 3 \\ 2 \\ 0 \\ 6 \\ 4 \\ 5 \\ 7 \end{pmatrix} =$$

$$= \begin{pmatrix} 8,196 & -6,231 & 1,305 & -0,079 \\ -6,231 & 5,078 & -1,104 & 0,069 \\ 1,305 & -1,104 & 0,247 & -0,016 \\ -0,079 & 0,069 & -0,016 & 0,001 \end{pmatrix} \begin{pmatrix} 1 & \sqrt{2} & 2 & (\sqrt{2})^3 \\ 1 & \sqrt{27} & 27 & (\sqrt{27})^3 \\ 1 & \sqrt{20} & 20 & (\sqrt{20})^3 \\ 1 & \sqrt{14} & 14 & (\sqrt{14})^3 \\ 1 & \sqrt{53} & 53 & (\sqrt{53})^3 \\ 1 & \sqrt{3} & 3 & (\sqrt{3})^3 \\ 1 & \sqrt{8} & 8 & (\sqrt{8})^3 \\ 1 & \sqrt{85} & 85 & (\sqrt{85})^3 \end{pmatrix}^T \begin{pmatrix} 1 \\ 3 \\ 2 \\ 0 \\ 6 \\ 4 \\ 5 \\ 7 \end{pmatrix} = \begin{pmatrix} 1,769 & -0,081 & -0,669 & -1,007 & 1,379 & 0,905 & -0,785 & -0,511 \\ -1,044 & -0,032 & 0,531 & 0,904 & -1,307 & -0,392 & 0,850 & 0,510 \\ 0,193 & 0,044 & -0,091 & -0,187 & 0,323 & 0,052 & -0,195 & -0,139 \\ -0,01 & -0,004 & 0,004 & 0,011 & -0,021 & -0,002 & 0,012 & 0,011 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 2 \\ 0 \\ 6 \\ 4 \\ 5 \\ 7 \end{pmatrix}$$

$$= \begin{pmatrix} 4,584 \\ -1,687 \\ 0,338 \\ -0,013 \end{pmatrix} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{pmatrix}$$

$$f(x, w) = 4,584 - 1,687 x_1 + 0,338 x_2 - 0,013 x_3$$

2)

$$\hat{x}_9 = 4,584 - 1,687 \times 2 + 0,338 \times 0 - 0,013 \times 0 = 1,21$$

$$\hat{x}_{10} = 4,584 - 1,687 \times 1 + 0,338 \times 2 - 0,013 \times 1 = 3,56$$

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{2} (\hat{x}_t - x_t)^2}{2}} = \sqrt{\frac{(1,21-2)^2 + (3,56-4)^2}{2}} \simeq 0,6394$$

TÉCNICO LISBOA

Aprendizagem 2021/22

Homework II – Group 057

3)

{1,2,4,7} → 0
{0,3,5,9} → 1

x_i	y_3 binarization	t_i
x_1	1	N
x_2	1	N
x_3	0	N
x_4	1	N
x_5	0	P
x_6	0	P
x_7	0	P
x_8	1	P

$$H(t) = -\left(\frac{4}{8}\log\frac{4}{8} + \frac{4}{8}\log\frac{4}{8}\right) = 1$$

$$IG(t \mid y_i) = H(t) - H(t \mid y_i)$$

$$H(t \mid y_1) = \frac{4}{8}H(t \mid y_1=1) + \frac{2}{8}H(t \mid y_1=0) + \frac{2}{8}H(t \mid y_1=2) =$$

$$= \frac{4}{8}\left(-\left(\underbrace{\frac{3}{4}\log\frac{3}{4}}_{-0{,}3113} + \underbrace{\frac{1}{4}\log\frac{1}{4}}_{-0{,}5}\right)\right) + \frac{2}{8}\left(-\left(\underbrace{\frac{1}{2}\log\frac{1}{2}}_{-0{,}5} + \underbrace{\frac{1}{2}\log\frac{1}{2}}_{-0{,}5}\right)\right) + \frac{2}{8}\left(-\left(\cancel{\frac{2}{2}\log\frac{2}{2}}\right)\right) =$$

$$= 0{,}40565 + 0{,}25 = 0{,}65565$$

- $IG_{y_1} = 1 - 0{,}65565 = 0{,}34435$ → maior ganho de info ⇒ raiz da árvore

$$H(t \mid y_2) = \frac{2}{8}H(t \mid y_2=0) + \frac{3}{8}H(t \mid y_2=1) + \frac{3}{8}H(t \mid y_2=2) =$$

$$= \frac{2}{8}\left(-\left(\cancel{\frac{2}{2}\log\frac{2}{2}}\right)\right) + \frac{3}{8}\left(-\left(\underbrace{\frac{2}{3}\log\frac{2}{3}}_{-0{,}38998} + \underbrace{\frac{1}{3}\log\frac{1}{3}}_{-0{,}52832}\right)\right) + \frac{3}{8}\left(-\left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right)\right)$$

$$= 0{,}688725$$

- $IG_{y_2} = 1 - 0{,}688725 = 0{,}311275$

$$H(t \mid y_3) = \frac{4}{8}H(t \mid y_3=0) + \frac{4}{8}H(t \mid y_3=1) = \frac{4}{8}\left(-\left(\underbrace{\frac{1}{4}\log\frac{1}{4}}_{-0{,}5} + \underbrace{\frac{3}{4}\log\frac{3}{4}}_{-0{,}3113}\right)\right) + \frac{4}{8}\left(-\left(\frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4}\right)\right)$$

$$= 0{,}8113$$

- $IG_{y_3} = 1 - 0{,}8113 = 0{,}1887$

$\{y_1=0\}$ - conditional dataset:

	y_2	y_3	t
x_3	2	0	N
x_8	2	1	P

↓

$IG_{y_2}=0$

$\{y_1=0,\ y_3=0\}$ - conditional dataset:
apenas x_3 → N

$\{y_1=0,\ y_3=1\}$ - conditional dataset:
apenas x_8 → P

$\{y_1=1\}$ - conditional dataset:

	y_2	y_3	t
x_1	1	1	N
x_2	1	1	N
x_4	2	1	N
x_6	1	0	P

↑
maior IG

$\{y_1=1,\ y_3=0\}$ - conditional dataset:
apenas x_6 → P

$\{y_1=1,\ y_3=1\}$ - conditional dataset:

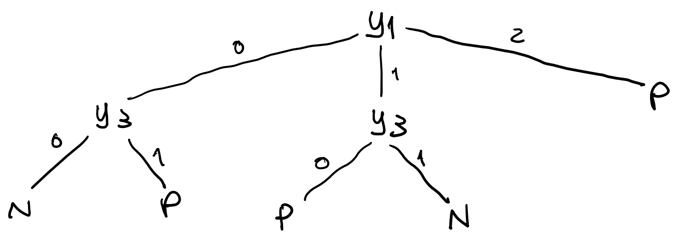
	y_2	t
x_1	1	N
x_2	1	N
x_4	2	N

todos N

TÉCNICO LISBOA

Aprendizagem 2021/22

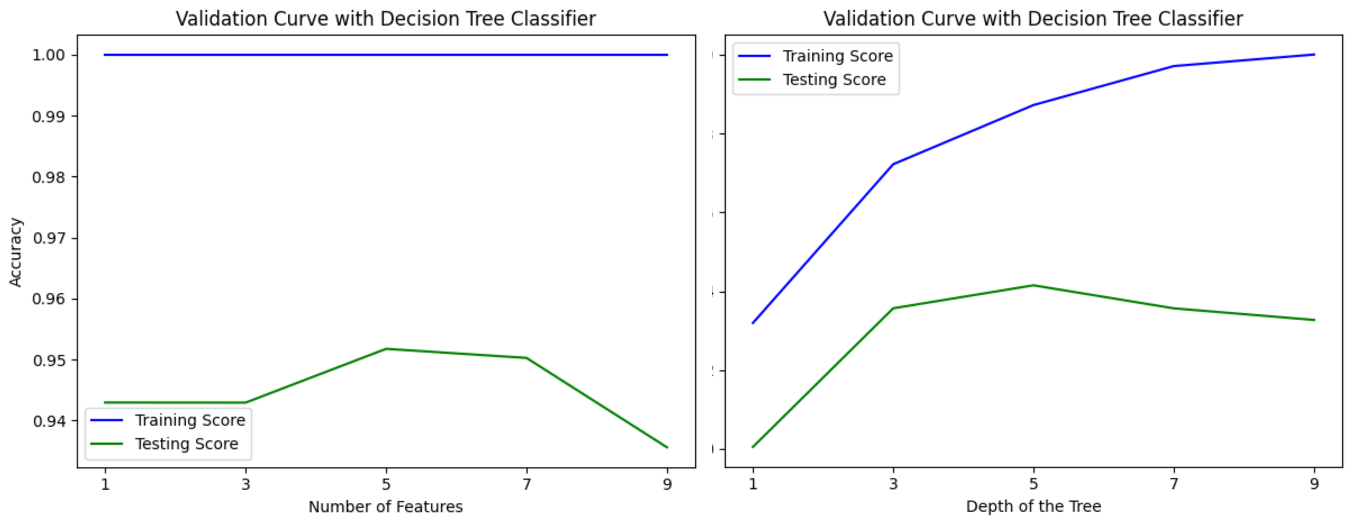
Homework II – Group 057



4) x_9 previsto = P x_9 observado = N
x_{10} previsto = N x_{10} observado = P
Concluímos assim que, nos x de teste, a árvore de decisão tem uma accuracy de 0%.

II. Programming and critical analysis

5)



6) No primeiro gráfico, há um aumento da accuracy na curva de testing até 5 features. A partir deste valor, a curva começa a descer, o que sugere que as restantes features não fornecem informação que melhore a árvore. Isto acontece, pois, o modelo está "sobreajustado" (overfitting) ao conjunto de treino, ou seja, o modelo perde as suas capacidades de generalização e comete erros em sets de teste, apesar de ter uma boa performance no training set.
No segundo gráfico, observamos que aumentar a tree depth até 5 resulta numa melhoria tanto nos train sets como nos test sets. No entanto, para valores superiores a 5, a accuracy no testing set diminui (contrariamente ao que acontece no training). Isto acontece, novamente, devido a overfitting.

7) Selecionamos para o valor da profundidade da árvore 5, porque este valor é o que fornece melhor accuracy na curva de teste.

III. APPENDIX

```
from scipy.io import arff
import pandas as pd
import numpy as np
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.model_selection import validation_curve
import matplotlib.pyplot as plt

data = arff.loadarff('breast.w_modified.arff')
df = pd.DataFrame(data[0])

X = df.iloc[:, 0:9]
y = df.iloc[:, -1]
y = y.astype('string')

X_train, X_test, y_train, y_test = train_test_split(X, y)

param_range = np.arange(1, 10, 2)

train_score, test_score = validation_curve(DecisionTreeClassifier(), X, y,
                                       param_name = "max_features",
                                       param_range = param_range,
                                       scoring = "accuracy")
# change param_name to "max_depth" for ex.5.ii

# mean and standard deviation of training score
mean_train_score = np.mean(train_score, axis = 1)
std_train_score = np.std(train_score, axis = 1)

# mean and standard deviation of testing score
mean_test_score = np.mean(test_score, axis = 1)
std_test_score = np.std(test_score, axis = 1)

# mean accuracy scores for training and testing scores
plt.plot(param_range, mean_train_score,
     label = "Training Score", color = 'b')
plt.plot(param_range, mean_test_score,
   label = "Testing Score", color = 'g')

# creating the plot
plt.xticks(param_range, param_range)
plt.title("Validation Curve with Decision Tree Classifier")
plt.xlabel("Depth of the Tree")
plt.ylabel("Accuracy")
plt.tight_layout()
plt.legend(loc = 'best')
plt.show()
plt.savefig("plot.png")  #savefig, don't show
```

END