Aprendizagem 2024
Homework I – Group 44
(ist1106827, ist107245)

Part I: Pen and paper

Consider the partially learnt decision tree from the dataset D. D is described by four input variables – one numeric with values in [0,1] and 3 categorical – and a target variable with three classes.

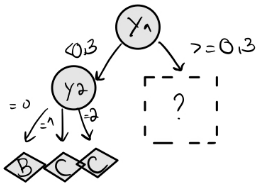
1. Complete the given decision tree using Shannon entropy (log2) and considering that: i) a minimum of 4 observations is required to split an internal node, and ii) decisions by ascending alphabetic should be placed in case of ties.

1. Calcular IG (yk) para cada feature yk

$IG(y_k) = E(class) - E(class \mid y_k)$

$E(class) = -\sum_{c \in class} p(class=c) \cdot \log_c(p(class=c))$

$E(class \mid y_k) = \sum_{j \in y_k} p(y_k = j) \cdot E(class \mid y_k = j)$



$y_1 >= 0,3$

$IG(y_2) = E(y_{out} \mid y_1 \geq 0,3) - E(y_{out} \mid y_1 \geq 0,3, y_2) = 1,558 - 1,251 = 0,307$

$E(y_{out} \mid y_1 \geq 0,3) = -(3/7 \log_2 3/7 + 2/7 \log_2 2/7 + 2/7 \log_2 2/7) = 1,558$

$E(y_{out} \mid y_1 \geq 0,3, y_2) = 4/7 \cdot (-(1/4 \log_2 1/4 + 1/4 \log_2 1/4 + 2/4 \log_2 2/4)) + 3/7 (-(2/3 \log_2 2/3 + 1/3 \log_2 1/3)) = 1,251$

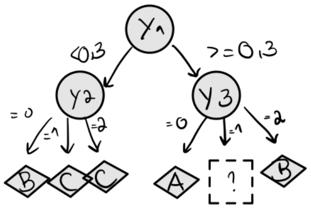
$IG(y_3) = E(y_{out} \mid y_1 >= 0.3) - E(y_{out} \mid y_1 >= 0.3, y_3) = 1,558 - 0,857 = 0,701$

$E(y_{out} \mid y_1 \geq 0,3, y_3) = 2/7 \cdot (-(\frac{2}{2} \log_2 \frac{2}{2})) + 4/7 \cdot ((\frac{1}{4} \log_2 \frac{1}{4} + \frac{1}{4} \log_2 \frac{1}{4} + \frac{2}{4} \log_2 \frac{2}{4})) + 1/7 \cdot (-(\frac{1}{1} \log_2 \frac{1}{1})) =$
$= 0,857$

$IG(y_4) = E(y_{out} \mid y_1 >= 0,3) - E(y_{out} \mid y_1 >= 0,3, y_4) = 1,558 - 0,964 = 0,594$

$E(y_{out} \mid y_1 \geq 0,3, y_4) = 4/7 \cdot (-(2/4 \log_2 2/4 + 2/4 \log_2 2/4)) + 3/7 \cdot (-(1/3 \log_2 1/3 + 2/3 \log_2 2/3)) = 0,964$

2. Escolher yk com maior IG para nó da árvore = y3



$y_1 \geq 0,3$

y3 = 0   x8 e x11 ambos A   logo A

y3 = 1   x6, x7, x9, x10 → B, 4, C, C   4 observações → split

y3 = 2   x12 → B   logo B

$y_1 \geq 0,3, \; y_3 = 1$

$IG(y_2) = E(y_{out} \mid y_1 \geq 0,3, y_3 = 1) - E(y_{out} \mid y_1 \geq 0,3, y_3 = 1, y_2) = 1,5 - 1,5 = 0$

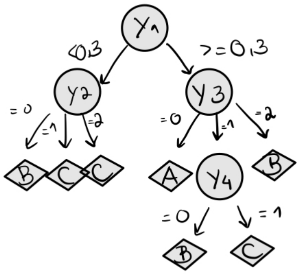
$E(y_{out} \mid y_1 \geq 0,3, y_3 = 1) = -(1/4 \log_2 1/4 + 1/4 \log_2 1/4 + 2/4 \log_2 2/4) = 1,5$

$E(y_{out} \mid y_1 \geq 0,3, y_3 = 1, y_2) = \frac{4}{4} \left[-(1/4 \log_2 1/4 + 1/4 \log_2 1/4 + 2/4 \log_2 2/4)\right] = \frac{3}{2} = 1,5$

$IG(y_4) = E(y_{out} \mid y_1 \geq 0,3, y_3 = 1) - E(y_{out} \mid y_1 \geq 0,3, y_3 = 1, y_4) = 1,5 - 0,688 = 0,812$

$E(y_{out} \mid y_1 \geq 0,3, y_3 = 1, y_4) = 1/4 \left[-(0 + 1 \log_2 1 + 0)\right] + 3/4 \left[-(1/3 \log_2 1/3 + 0 + 2/3 \log_2 2/3)\right] = 0,688$

3. Repetir 2. Escolher y_k com maior IG para nó da árvore = Y_4



$\dfrac{y_1 \geq 0,3,\ y_3 = 1}{y_4 = 0 \quad x_6 \rightarrow B}$ logo B

$y_4 = 1$ $x_7, x_9, x_{10} \rightarrow A, C, C$ logo C

$y_4 = 2$ nenhuma observação

2. Draw the training confusion matrix for the learnt decision tree.

D	y_1	y_2	y_3	y_4	real y_{out}	previsto y_{out}
x_1	0,22	2	0	1	C	C
x_2	0,06	0	0	0	B	B
x_3	0,16	1	2	2	C	C
x_4	0,21	0	0	0	B	B
x_5	0,01	2	2	0	C	C
x_6	0,3	0	1	0	B	B
x_7	0,76	0	1	1	A	C
x_8	0,86	1	0	0	A	A
x_9	0,93	0	1	1	C	C
x_{10}	0,47	0	1	1	C	C
x_{11}	0,73	1	0	0	A	A
x_{12}	0,89	1	2	0	B	B

training confusion matrix

		previstos		
		A	B	C
reais	A	2	0	1
	B	0	4	0
	C	0	0	5

3. Identify which class has the lowest training F1 score.

$$F_1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

class A

$$\text{precision} = \frac{2}{2+0+0} = \frac{2}{2} = 1$$

$$\text{recall} = \frac{2}{2+0+1} = \frac{2}{3}$$

$$F_1\text{-score} = 2 \times \frac{1 \times 2/3}{1 + 2/3} = \frac{4}{5}$$

class B

$$\text{precision} = \frac{4}{0+4+0} = \frac{4}{4} = 1$$

$$\text{recall} = \frac{4}{0+4+0} = \frac{4}{4} = 1$$

$$F_1\text{-score} = 2 \times \frac{1 \times 1}{1+1} = 1$$

class C

$$\text{precision} = \frac{5}{1+0+5} = \frac{5}{6}$$

$$\text{recall} = \frac{5}{0+0+5} = \frac{5}{5} = 1$$

$$F_1\text{-score} = 2 \times \frac{5/6 \times 1}{5/6 + 1} = \frac{10}{11}$$

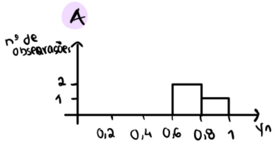
class A has the lowest training F1 score

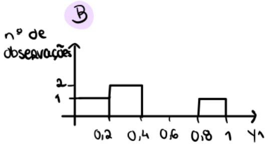
4. Draw the class-conditional relative histograms of y1 using 5 equally spaced bins in [0,1]. Find the n-ary root split using the discriminant rules from these empirical distributions.

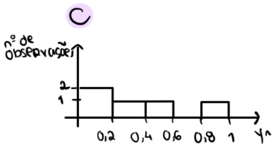
5 equally spaced bins in [0,1]
[0;0,2[→ C
[0,2;0,4[→ B
[0,4;0,6[→ C
[0,6;0,8[→ A
[0,8;1] → A

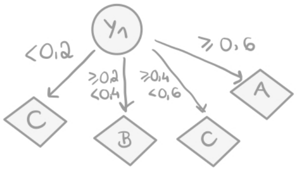
(A, B e C empatados decidimos escolher por ordem alfabética)

D	Y1	Y2	Y3	Y4	Yout
x5	0,01	2	2	0	C
x2	0,06	0	0	0	B
x3	0,16	1	2	2	C
x4	0,21	0	0	0	B
x1	0,22	2	0	1	C
x6	0,3	0	1	0	B
x10	0,47	0	1	1	C
x11	0,73	1	0	0	A
x7	0,76	0	1	1	A
x8	0,86	1	0	0	A
x12	0,89	1	2	0	B
x9	0,93	0	1	1	C









Part II: Programming

Consider the diabetes.arff data available at the homework tab, comprising 8 biological features to classify 768 patients into 2 classes (normal, diabetes).

1. ANOVA is a statistical test that can be used to assess the discriminative power of a single input variable. Using f_classif from sklearn, identify the input variables with the worst and best discriminative power. Plot their class-conditional probability density functions.

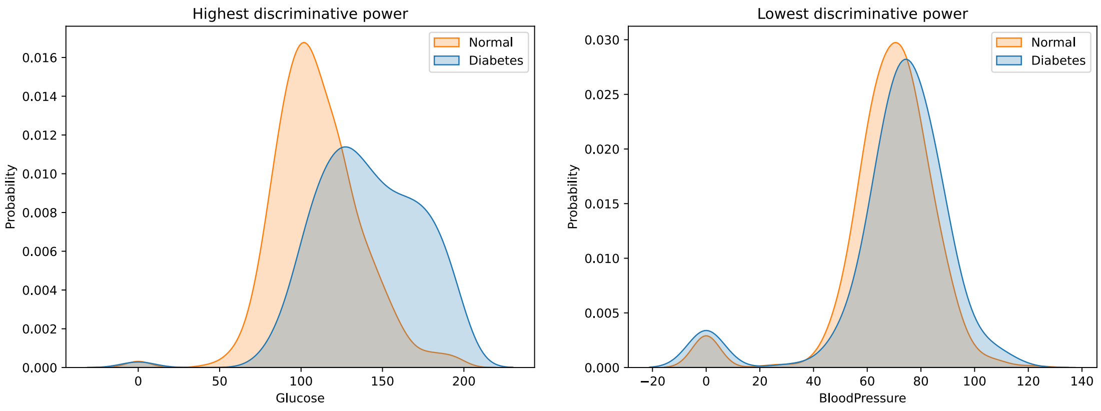


Figure 1: Class-conditional probability density functions of the input variables with the best and worst discriminative power

2. Using a stratified 80-20 training-testing split with a fixed seed (random_state=1), assess in a single plot both the training and testing accuracies of a decision tree with minimum sample split in 2, 5,10, 20, 30, 50, 100 and the remaining parameters as default.
   [optional] Note that split thresholding of numeric variables in decision trees is non- deterministic in sklearn, hence you may opt to average the results using 10 runs per parameterization.

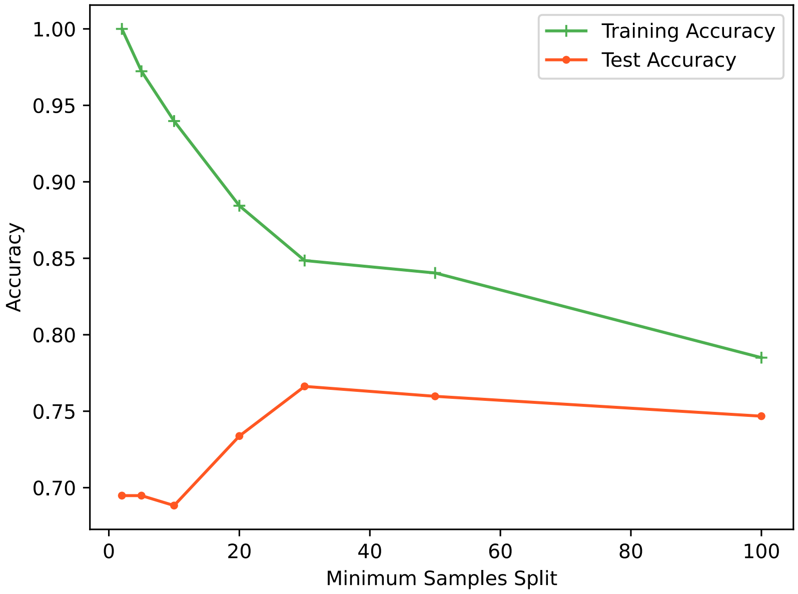


Figure 2: Accuracy of trained decision tree with with minimum sample split, applied to both a test and training sets.

3. Critically analyze these results, including the generalization capacity across settings.

   In the results above, you can observe that the training accuracy decreases as the minimum samples split increases. However, the test accuracy initially improves and then plateaus, indicating that the model begins to generalize better as the minimum samples split increases. Eventually, though, the performance on the test set stabilizes while the training accuracy continues to drop, which may suggest that the model is becoming too simple and losing some capacity to learn more nuanced patterns in the data.

   Looking at the gap between the training and test accuracy curves, we can assess the generalization capacity. A small gap with both accuracies being relatively close to each other usually signifies good generalization. In this case, the most beneficial minimum samples split appears to be around 20-50, where the test accuracy is highest and the gap between the training and test accuracy is minimized.

4. To deploy the predictor, a healthcare provider opted to learn a single decision tree (random_state=1) using all available data and ensuring that the maximum depth would be 3 in order to avoid overfitting risks.

   i. Plot the decision tree.

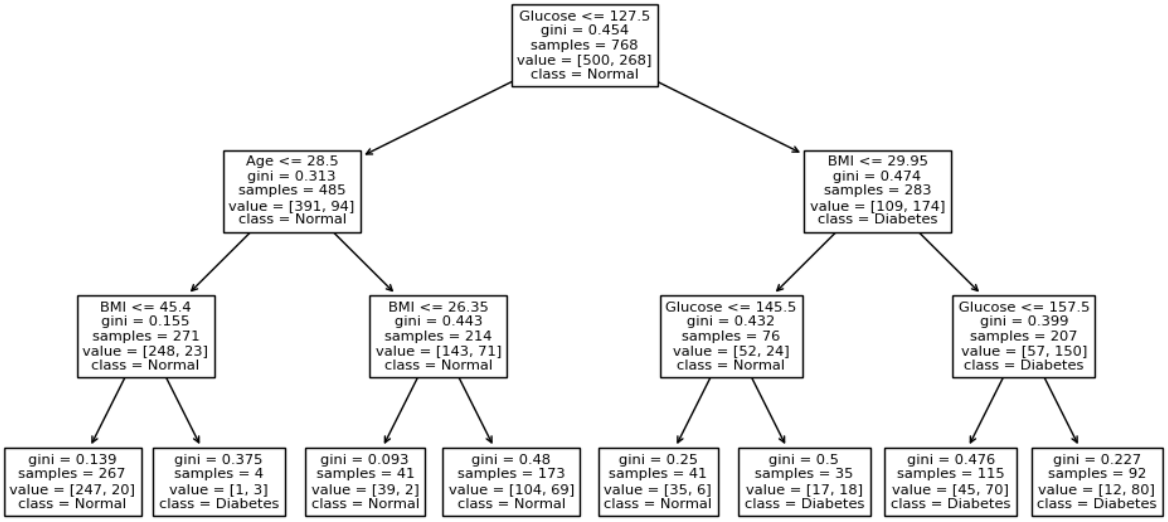


Figure 3: Decision tree classifier trained on the diabetes dataset, with a maximum depth of 3

   ii. Explain what characterizes diabetes by identifying the conditional associations together with their posterior probabilities.

   The decision tree uncovers key conditional associations that distinguish diabetes from normal cases based on glucose levels, age, and BMI. Here are the posterior probabilities for diabetes at various leaf nodes:

   Leaf node for $Glucose \leq 127.5$ and $Age \leq 28.5$ and $BMI \leq 45.4$: $\frac{3}{4} \approx 75\%$

   Leaf node for $Glucose \geq 127.5$ and $BMI \leq 29.95$ and $Glucose \geq 145.5$: $\frac{18}{35} \approx 51\%$

   Leaf node for $Glucose \geq 127.5$ and $BMI \geq 29.95$ and $Glucose \leq 157.5$: $\frac{70}{115} \approx 61\%$

   Leaf node for $Glucose \geq 127.5$ and $BMI \geq 29.95$ and $Glucose \geq 157.5$: $\frac{80}{92} \approx 87\%$

   The most important predictor is glucose level: patients with glucose levels below 127.5 are predominantly classified as ”Normal”, indicating that lower glucose levels are associated with a lower likelihood of diabetes. Conversely, when glucose exceeds 127.5, there is a significant increase in diabetes cases. For patients with high glucose, BMI becomes a crucial factor. Individuals with a BMI above 29.95 and glucose levels higher than 157.5 have a high probability of being diagnosed with diabetes, suggesting that both high BMI and glucose are strong indicators of the disease. On the other hand, lower BMI and glucose levels are associated with a higher chance of being ”Normal”. These associations help predict whether a patient is likely to have diabetes, with glucose levels serving as the primary determinant, followed by BMI.

TÉCNICO
LISBOA