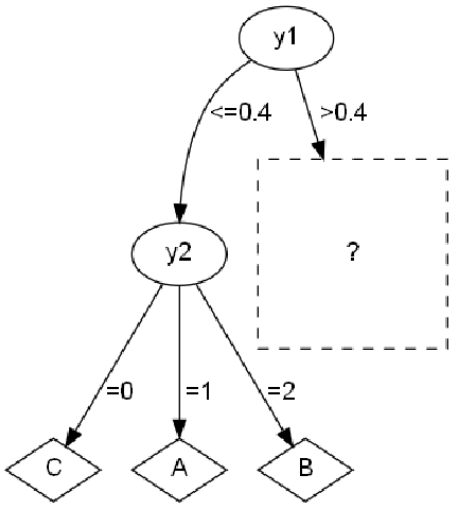
Aprendizagem 2023
Homework I – Group 28

Gonçalo Bárias (ist1103124) & Raquel Braunschweig (ist1102624)

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.

D	y_1	y_2	y_3	y_4	y_{out}
$\mathbf{x}_1$	0.24	1	1	0	A
$\mathbf{x}_2$	0.06	2	0	0	B
$\mathbf{x}_3$	0.04	0	0	0	B
$\mathbf{x}_4$	0.36	0	2	1	C
$\mathbf{x}_5$	0.32	0	0	2	C
$\mathbf{x}_6$	0.68	2	2	1	A
$\mathbf{x}_7$	0.9	0	1	2	A
$\mathbf{x}_8$	0.76	2	2	0	A
$\mathbf{x}_9$	0.46	1	1	1	B
$\mathbf{x}_{10}$	0.62	0	0	1	B
$\mathbf{x}_{11}$	0.44	1	2	2	C
$\mathbf{x}_{12}$	0.52	0	2	0	C



1. **Complete the given decision tree using Information gain with Shannon entropy (log_2). Consider that: i) a minimum of 4 observations is required to split an internal node, and ii) decisions by ascending alphabetic order should be placed in case of ties.**

   The entropy of y_{out} is given by:

$$\begin{aligned} E(y_{out}|y_1 > 0.4) = &- P(y_{out} = \text{A}, y_1 > 0.4) \log_2 \left(P(y_{out} = \text{A}, y_1 > 0.4)\right) \\ &- P(y_{out} = \text{B}, y_1 > 0.4) \log_2 \left(P(y_{out} = \text{B}, y_1 > 0.4)\right) \\ &- P(y_{out} = \text{C}, y_1 > 0.4) \log_2 \left(P(y_{out} = \text{C}, y_1 > 0.4)\right) \end{aligned} \tag{1}$$

   We can calculate $E(y_{out})$:

$$E(y_{out}|y_1 > 0.4) = -\left(\frac{3}{7}\log_2\left(\frac{3}{7}\right) + \frac{2}{7}\log_2\left(\frac{2}{7}\right) + \frac{2}{7}\log_2\left(\frac{2}{7}\right)\right) \approx 1.5567$$

   The next step is calculating $E(y_{out}|y_1 > 0.4, y_x)$, in which x will take the values of 2, 3 or 4:

$$\begin{aligned} E(y_{out}|y_1 > 0.4, y_x) = &- P(y_x = 0)E(y_{out}|y_1 > 0.4, y_x = 0) \\ &- P(y_x = 1)E(y_{out}|y_1 > 0.4, y_x = 1) \\ &- p(y_x = 2)E(y_{out}|y_1 > 0.4, y_x = 2) \end{aligned} \tag{2}$$

And the information gain of variable y_x is given by

$$IG(y_{out}|y_1 > 0.4, y_x) = E(y_{out}|y_1 > 0.4) - E(y_{out}|y_1 > 0.4, y_x) \tag{3}$$

Let's start with x = 2:

$$P(y_2 = 0, y_1 > 0.4) = \frac{3}{7}$$
$$P(y_2 = 1, y_1 > 0.4) = \frac{2}{7}$$
$$P(y_2 = 1, y_1 > 0.4) = \frac{2}{7}$$
$$E(y_{out}|y_1 > 0.4, y_2 = 0) = -\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) \approx 1.58493$$
$$E(y_{out}|y_1 > 0.4, y_2 = 1) = -\left(\frac{0}{2}\log_2\left(\frac{0}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$$
$$E(y_{out}|y_1 > 0.4, y_2 = 2) = -\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) = 0$$

Therefore, replacing these values on equation (2), gives us:

$$E(y_{out}|y_1 > 0.4, y_2) = \frac{3}{7} \times 1.58493 + \frac{2}{7} \times 1 + \frac{2}{7} \times 0 \quad \approx 0.965.$$

Finally, we can calculate the information gain, as per (3),

$$IG(y_{out}|y_1 > 0.4, y_2) = 1.5567 - 0.965 = 0.5917$$

Now, let's calculate for x = 3:

$$P(y_3 = 0, y_1 > 0.4) = \frac{1}{7}$$
$$P(y_3 = 1, y_1 > 0.4) = \frac{2}{7}$$
$$P(y_3 = 2, y_1 > 0.4) = \frac{4}{7}$$
$$E(y_{out}|y_1 > 0.4, y_3 = 0) = -\left(\frac{0}{1}\log_2\left(\frac{0}{1}\right) + \frac{1}{1}\log_2\left(\frac{1}{1}\right) + \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right) = 0$$
$$E(y_{out}|y_1 > 0.4, y_3 = 1) = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) = 1$$
$$E(y_{out}|y_1 > 0.4, y_3 = 2) = -\left(\frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{0}{4}\log_2\left(\frac{0}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right) = 1$$

Therefore, replacing these values on equation (2), gives us:

$$E(y_{out}|y_1 > 0.4, y_3) = \frac{1}{7} \times 0 + \frac{2}{7} \times 1 + \frac{4}{7} \times 1 \quad \approx 0.8571.$$

Finally, we can calculate the information gain, as per (3),

$$IG(y_{out}|y_1 > 0.4, y_3) = 1.5567 - 0.8571 = \mathbf{0.6996}$$

Finally, let's calculate for x = 4:

$$P(y_4 = 0, y_1 > 0.4) = \frac{2}{7}$$
$$P(y_4 = 1, y_1 > 0.4) = \frac{3}{7}$$
$$P(y_4 = 1, y_1 > 0.4) = \frac{2}{7}$$
$$E(y_{out}|y_1 > 0.4, y_4 = 0) = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{3}\right)\right) = 1$$
$$E(y_{out}|y_1 > 0.4, y_4 = 1) = -\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right) \approx 0.9183$$
$$E(y_{out}|y_1 > 0.4, y_4 = 2) = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$$

Therefore, replacing these values on equation (2), gives us:

$$E(y_{out}|y_1 > 0.4, y_4) = \frac{2}{7} \times 1 + \frac{3}{7} \times 0.9183 + \frac{2}{7} \times 1 \approx 0.965.$$

Finally, we can calculate the information gain, as per (3),

$$IG(y_{out}|y_1 > 0.4, y_4) = 1.5567 - 0.965 = 0.5917$$

Upon computing the information gains for each attribute, it is evident that y_3 yields the highest value of 0.6996. Consequently, it is selected as the next node, and since there are at least 4 observations with $y_1 > 0.4$, we split the new node.

If we fix $y_1 > 0.4$ with $y_3 = 0$ or $y_3 = 1$, we get 1 and 2 observations, respectively. This means we get two new leafs for those branches, with the $y_3 = 0$ leaf equal to class B and the $y_3 = 1$ leaf equal to A, because those are the classes that have the highest frequency for their respective conditions on the dataset. For the $y_3 = 1$ leaf there is a tie, which we just use the ascending alphabetic order to get the class. For the $y_3 = 2$ branch with the $y_1 > 0.4$ condition, we get exactly 4 observations, and so there is a split.

Below we do the required calculations to finish the decision tree:

$$E(y_{out}|y_1 > 0.4, y_3 = 2, y_2 = 0) = -\left(\frac{0}{1}\log_2\left(\frac{0}{1}\right) + \frac{0}{1}\log_2\left(\frac{0}{1}\right) + \frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) = 0$$
$$E(y_{out}|y_1 > 0.4, y_3 = 2, y_2 = 1) = -\left(\frac{0}{1}\log_2\left(\frac{0}{1}\right) + \frac{0}{1}\log_2\left(\frac{0}{1}\right) + \frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) = 0$$
$$E(y_{out}|y_1 > 0.4, y_3 = 2, y_2 = 2) = -\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) = 0$$
$$E(y_{out}|y_1 > 0.4, y_3 = 2, y_4 = 0) = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$$
$$E(y_{out}|y_1 > 0.4, y_3 = 2, y_4 = 1) = -\left(\frac{1}{1}\log_2\left(\frac{1}{1}\right) + \frac{0}{1}\log_2\left(\frac{0}{1}\right) + \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right) = 0$$
$$E(y_{out}|y_1 > 0.4, y_3 = 2, y_4 = 2) = -\left(\frac{0}{1}\log_2\left(\frac{0}{1}\right) + \frac{0}{1}\log_2\left(\frac{0}{1}\right) + \frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) = 0$$

From other calculations above, we know that:

$$E(y_{out}|y_1 > 0.4, y_3 = 2) = 1$$

Therefore, replacing the above values on equation (2), gives us:

$$E(y_{out}|y_1 > 0.4, y_3 = 2, y_2) = \frac{1}{4} \times 0 + \frac{1}{4} \times 0 + \frac{2}{4} \times 0 \quad = 0$$
$$E(y_{out}|y_1 > 0.4, y_3 = 2, y_4) = \frac{2}{4} \times 1 + \frac{1}{4} \times 0 + \frac{1}{4} \times 0 \quad = 0.5$$

Finally, we can calculate the information gain, as per (3),

$$IG(y_{out}|y_1 > 0.4, y_3 = 2, y_2) = 1 - 0 = \mathbf{1}$$
$$IG(y_{out}|y_1 > 0.4, y_3 = 2, y_4) = 1 - 0.5 = 0.5$$

After the computation, it is evident that y_2 yields the highest value of 1, and so it is chosen. With the conditions of $y_1 > 0.4$ and $y_3 = 2$, we can check that the conditions $y_2 = 0$, $y_2 = 1$ and $y_2 = 2$ all have less than 4 observations each. Therefore, we take the class with the highest frequency for each respective conditions on the dataset. **Finally**, we can construct the following decision tree:

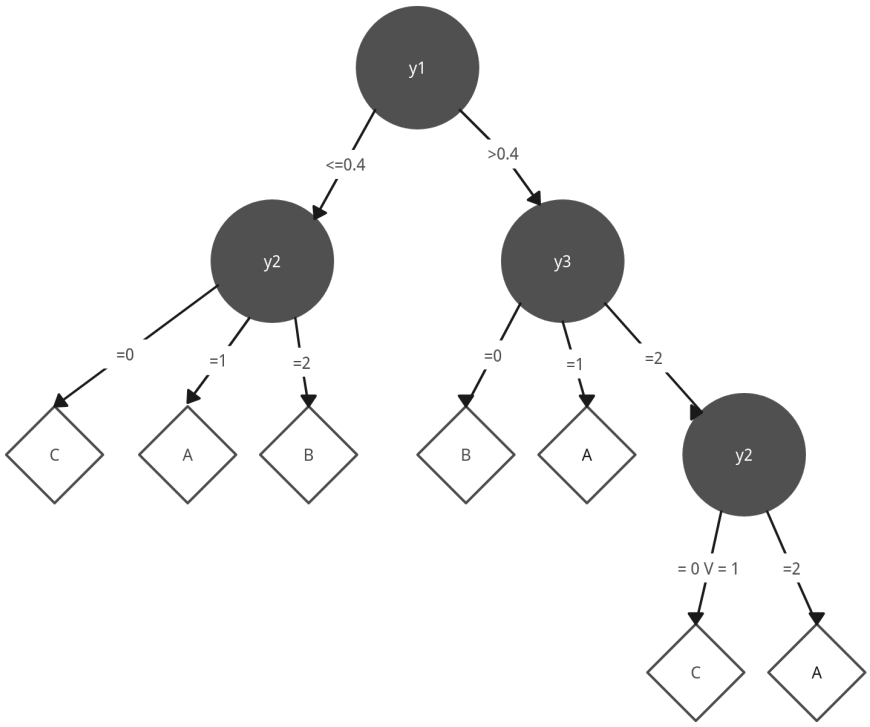


Figure 1: Decision Tree for exercise I.1

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

   Following the learnt decision tree above, we can predict the values for each observation. For each observation, we look at the value for the first variable (y_1) and follow the branch that corresponds with its value. From the node we arrive at, we do the same thing for the next variable, and we keep doing this until we reach a leaf. The class present in this leaf will be the predicted value, while the real value is the value of y_{out} for that observation. Below we present the real values along with the predicted ones:

		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
real	=	[A	B	B	C	C	A	A	A	B	B	C	C]
predicted	=	[A	B	C	C	C	A	A	A	A	B	C	C]

   Finally, we can show the count of each pair of real and predicted values in a confusion matrix (e.g. 4 pairs of AA from observations x_1, x_6, x_7 and x_8):

		Real			
		A	**B**	**C**	
Predicted	**A**	4	1	0	5
	B	0	2	0	2
	C	0	1	4	5
		4	4	4	12

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

$F1_{score}$ is given by the following equation:

$$F1_{score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

And precision and recall are given by:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{5}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{6}$$

Therefore, **let's start by calculating the precision for A, B and C** by replacing the values on (5):

$$Precision_A = \frac{4}{4+1} = \frac{4}{5}$$

$$Precision_B = \frac{2}{2+0} = 1$$

$$Precision_C = \frac{4}{4+1} = \frac{4}{5}$$

Now, it's time to calculate the recalls for A, B and C, using the equation on (6):

$$Recall_A = \frac{4}{4+0} = 1$$

$$Recall_B = \frac{2}{2+2} = \frac{1}{2}$$

$$Recall_C = \frac{4}{4+0} = 1$$

Finally, let's calculate the $F1_{score}$, using the equation (4):

$$F1_{score}A = 2 \cdot \frac{\frac{4}{5} \cdot 1}{\frac{4}{5} + 1} \approx 0.8889$$

$$F1_{score}B = 2 \cdot \frac{1 \cdot \frac{1}{2}}{1 + \frac{1}{2}} \approx 0.6667$$

$$F1_{score}C = 2 \cdot \frac{\frac{4}{5} \cdot 1}{\frac{4}{5} + 1} \approx 0,8889$$

The class with the lowest training F1 score is B, with a score of 0.6667.

4. **Considering y_2 to be ordinal, assess if y_1 and y_2 are correlated using the Spearman coefficient.**

   To calculate the Spearman coefficient when there's rank, we have to use the following equation:

$$\text{Spearman}(y_x, y_y) = \text{Pearson}(y'_x, y'_y) = \frac{\text{cov}(y'_x, y'_y)}{\sigma_{y'_x}\sigma_{y'_y}} = \frac{\sum_{i=1}^{n}(y'_{x_i} - \bar{y'_x})(y'_{y_i} - \bar{y'_y})}{\sqrt{\sum_{i=1}^{n}(y'_{x_i} - \bar{y'_x})^2}\sqrt{\sum_{i=1}^{n}(y'_{y_i} - \bar{y'_y})^2}} \tag{7}$$

   Firstly, **let's order y_1 and y_2** so we can calulate the ranks and y'_1 and y'_2:

$$\begin{aligned}
ordered_y_1 &= [0.04, 0.06, 0.24, 0.32, 0.36, 0.44, 0.46, 0.52, 0.62, 0.68, 0.76, 0.9] \\
ranks_y_1 &= [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12] \\
y'_1 &= [3, 2, 1, 5, 4, 10, 12, 11, 7, 9, 6, 8] \\
ordered_y_2 &= [0, 0, 0, 0, 0, 0, 1, 1, 1, 2, 2, 2] \\
ranks_y_2 &= [3.5, 3.5, 3.5, 3.5, 3.5, 3.5, 8, 8, 8, 11, 11, 11] \\
y'_2 &= [8, 11, 3.5, 3.5, 3.5, 11, 3.5, 11, 8, 3.5, 8, 3.5]
\end{aligned}$$

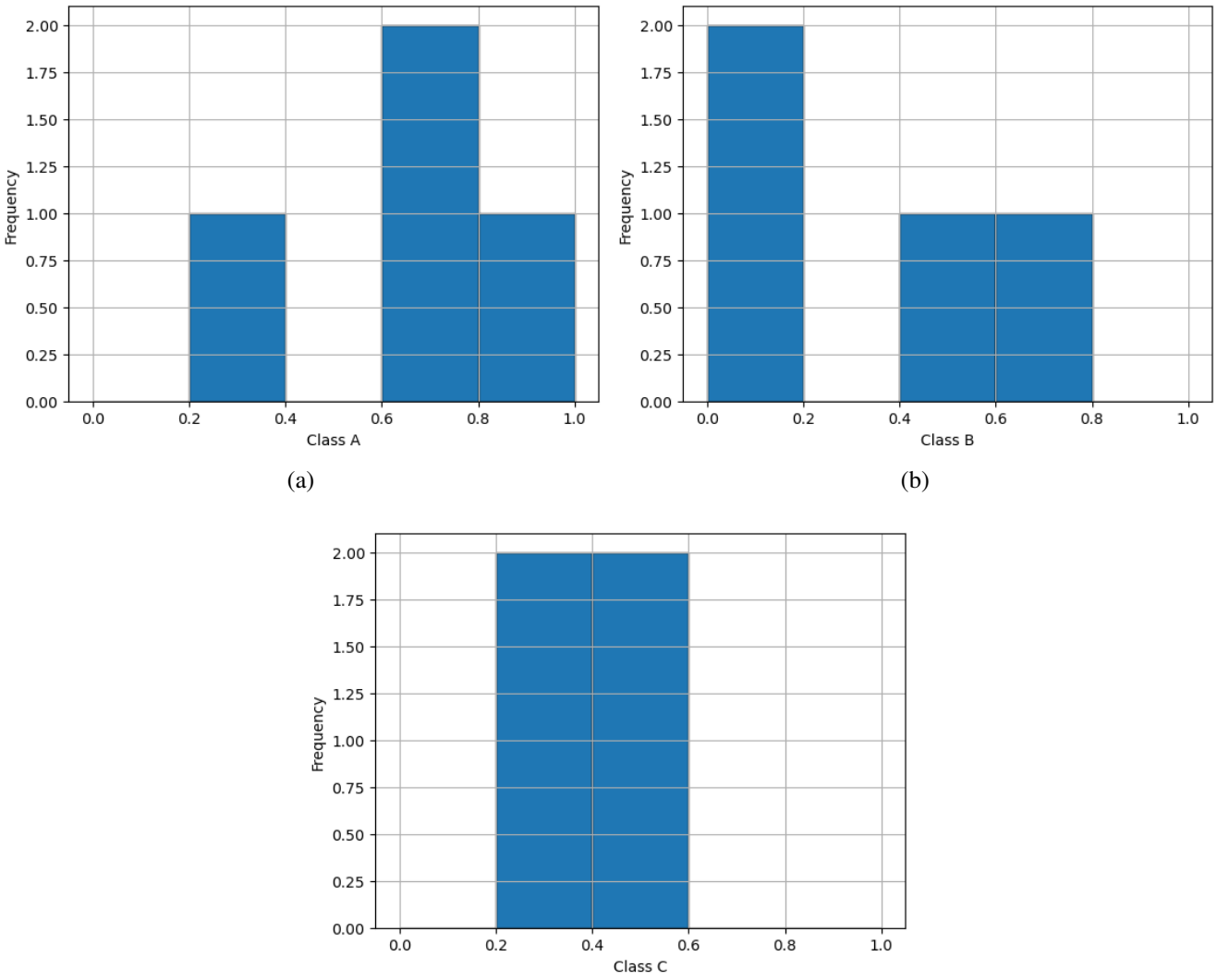
   Now, we have all we need to calculate **the Spearman coefficient** using the expression at (7). Here is the result:

$$Spearman(y_1, y_2) = Pearson(y'_1, y'_2) \approx 0.07966$$

   Since the rank correlation (under Spearman coefficient) obtained a value relatively close to 0, we can conclude that the variables y_1 and y_2 are almost non correlated.

5. **Draw the class-conditional relative histograms of y_1 using 5 equally spaced bins in $[0, 1]$. Find the root split using the discriminant rules from these empirical distributions.**

   The following histograms were created by observing the values of y_1 and fixing the class in y_{out}. Since there are three classes A, B and C, we get three histograms. In each histogram we have 5 equally spaced bins in $[0, 1]$, meaning that each bin has a length of 0.2. For each class, when a value of y_1 falls into a bin, we increase the value of that bin by one.



(a) (b)

(c)

Figure 2: Class-conditional relative histograms of y_1 with 5 equally spaced bins in $[0, 1]$

In order to find the root split using the discriminant rules from the empirical distributions, we made a decision tree with only a root split on y_1. Each branch will have the values of one bin, meaning there will be 5 branches (e.g. the first branch is for values of y_1 that belong to $[0, 0.2[$). The leaf for each branch corresponds to the class that has the highest value for that specific bin in its histogram.

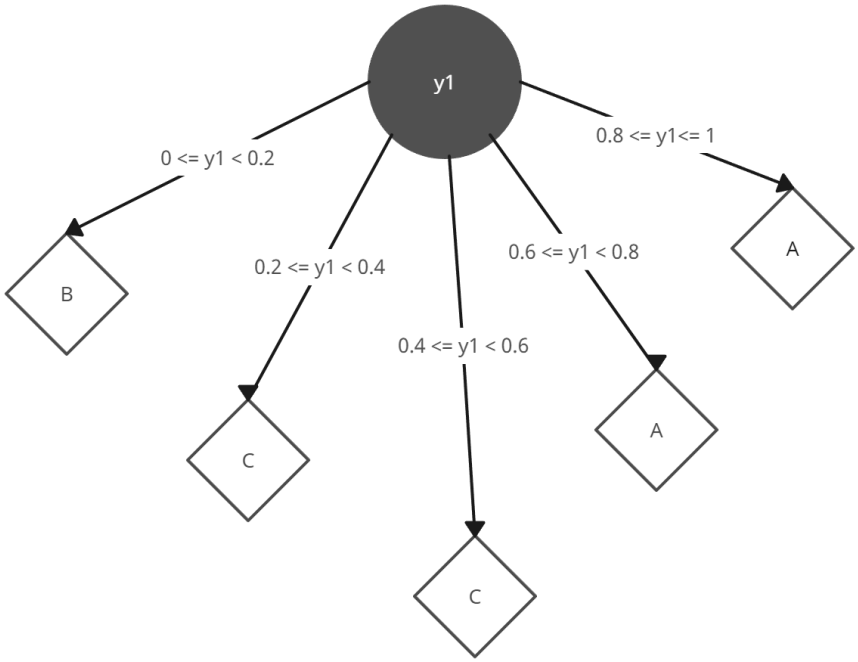


Figure 3: Decision Tree using the discriminant rules from the empirical distributions

Part II: Programming

Consider the `column_diagnosis.arff` data available at the homework tab, comprising 6 biomechanical features to classify 310 orthopaedic patients into 3 classes (`normal`, `disk hernia`, `spondilolysthesis`).

1. **Apply `f_classif` from `sklearn` to assess the discriminative power of the input variables. Identify the input variable with the highest and lowest discriminative power. Plot the class-conditional probability density functions of these two input variables.**

```python
import numpy as np, pandas as pd, seaborn as sns, matplotlib.pyplot as plt
from scipy.io.arff import loadarff
from sklearn.feature_selection import f_classif

# Read the ARFF file and prepare data
data = loadarff("./data/column_diagnosis.arff")
df = pd.DataFrame(data[0])
df["class"] = df["class"].str.decode("utf-8")
X, y = df.drop("class", axis=1), df["class"]

# Apply f_classif
f_scores, _ = f_classif(X, y)

# Obtains the variables with the highest and lowest discriminative power.
h_disc_power_var = X.columns[np.argmax(f_scores)]
l_disc_power_var = X.columns[np.argmin(f_scores)]

plt.figure(figsize=(8, 6))

```

```
# Plot for the highest discriminative power variable
for class_label in np.unique(y):
    class_data = X.loc[y == class_label, h_disc_power_var]
    sns.kdeplot(
        class_data,
        label=f"Class {class_label} - {h_disc_power_var}",
        linewidth=2,
    )

# Plot for the lowest discriminative power variable
for class_label in np.unique(y):
    class_data = X.loc[y == class_label, l_disc_power_var]
    sns.kdeplot(
        class_data,
        label=f"Class {class_label} - {l_disc_power_var}",
        linestyle="--",
        linewidth=2,
    )

plt.xlabel("Variables")
plt.ylabel("Density")

plt.legend()
plt.grid(True)
plt.show()
```

As you can see in the graph ahead, the highest discriminative power variable is *degree_spondilolysthesis* and the lowest discriminative power variable is *pelvic_radius*.

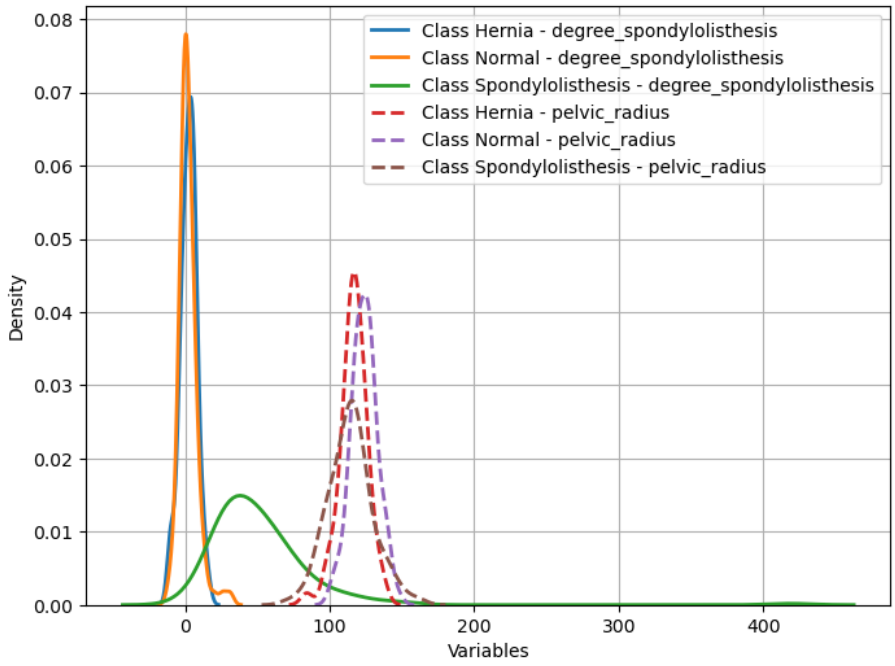


Figure 4: Class-conditional probability density functions of the highest and lowest discriminative power variables.

2. **Using a stratified 70-30 training-testing split with a fixed seed (`random_state=0`), assess in**

a single plot both the training and testing accuracies of a decision tree with depth limits in $\{1, 2, 3, 4, 5, 6, 8, 10\}$ 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.**

```python
import pandas as pd, matplotlib.pyplot as plt
from scipy.io.arff import loadarff
from sklearn import metrics, tree
from sklearn.model_selection import train_test_split

# Read the ARFF file and prepare data
data = loadarff("./data/column_diagnosis.arff")
df = pd.DataFrame(data[0])
df["class"] = df["class"].str.decode("utf-8")
X, y = df.drop("class", axis=1), df["class"]

DEPTH_LIMIT = [1, 2, 3, 4, 5, 6, 8, 10]
training_accuracy, test_accuracy = [], []

# Split the dataset into a testing set (30%) and a training set (70%)
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, stratify=y, random_state=0
)

for depth_limit in DEPTH_LIMIT:
    # Create and fit the decision tree classifier
    predictor = tree.DecisionTreeClassifier(
        max_depth=depth_limit, random_state=0
    )
    predictor.fit(X_train, y_train)

    # Use the decision tree to predict the outcome of the given observations
    y_train_pred = predictor.predict(X_train)
    y_test_pred = predictor.predict(X_test)

    # Get the accuracy of each test
    train_acc = metrics.accuracy_score(y_train, y_train_pred)
    training_accuracy.append(train_acc)
    test_acc = metrics.accuracy_score(y_test, y_test_pred)
    test_accuracy.append(test_acc)

plt.plot(
    DEPTH_LIMIT, training_accuracy,
    label="Training Accuracy", marker="+", color="#f8766d"
)
plt.plot(
    DEPTH_LIMIT, test_accuracy,
    label="Test Accuracy", marker=".", color="#00bfc4"
)

plt.xlabel("Depth Limit")
plt.ylabel("Accuracy")

plt.legend()
plt.grid(True)
```

```
plt.show()
```

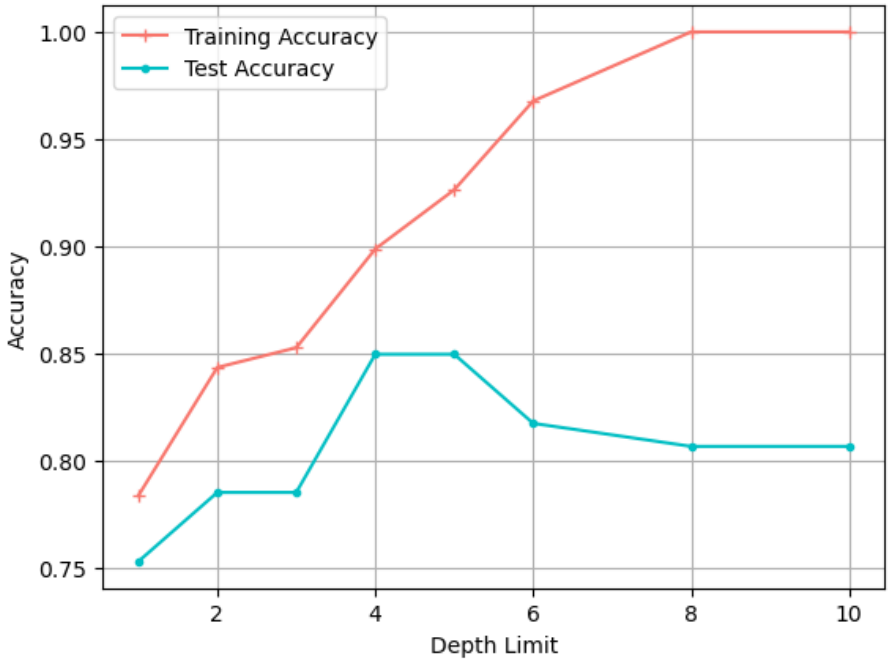


Figure 5: Accuracy of the trained decision tree, applied to both a test and training sets, for varying depth limits.

3. **Comment on the results, including the generalization capacity across settings.**

   The graphic shows that as the depth limit of the decision tree increases, the training accuracy rises steadily until reaching 100%. This observation suggests that deeper trees can better fit the training data, capturing complex patterns and achieving higher accuracy when evaluated on the dataset.

   The testing accuracy initially improves as the depth limit increases, indicating improved generalization. However, beyond a certain depth limit (around 4 or 5 in this case), the testing accuracy starts to decline. Therefore, we conclude there is a loss in generalization capacity, which causes overfitting, implying that, in this case, overly complex decision trees can fit noise in the training data, performing poorly on new, unseen data.

   The optimal depth limit appears to be around 4 or 5, which seems to be the point where the accuracy is maximized for the testing data, striking a balance between model complexity and generalization to new data, this way we can avoid both underfitting (too simple) and overfitting (too complex).

4. **To deploy the predictor, a healthcare team opted to learn a single decision tree (`random_state=0`) using *all* available data as training data, and further ensuring that each leaf has a minimum of 20 individuals in order to avoid overfitting risks.**

   (a) **Plot the decision tree.**

```
import matplotlib.pyplot as plt, pandas as pd, numpy as np
from scipy.io.arff import loadarff
from sklearn.tree import DecisionTreeClassifier, plot_tree
```

```
4
5  # Read the ARFF file and prepare data
6  data = loadarff("./data/column_diagnosis.arff")
7  df = pd.DataFrame(data[0])
8  df["class"] = df["class"].str.decode("utf-8")
9  X, y = df.drop("class", axis=1), df["class"]
10
11 # Create and train the decision tree classifier
12 clf = DecisionTreeClassifier(random_state=0, min_samples_leaf=20)
13 clf.fit(X, y)
14
15 # Set style and plot the decision tree
16 plt.figure(figsize=(15, 10))
17 plot_tree(clf, filled=True, feature_names=list(X.columns),
18           class_names=list(np.unique(y)), rounded=True, fontsize=12)
19 plt.show()
```

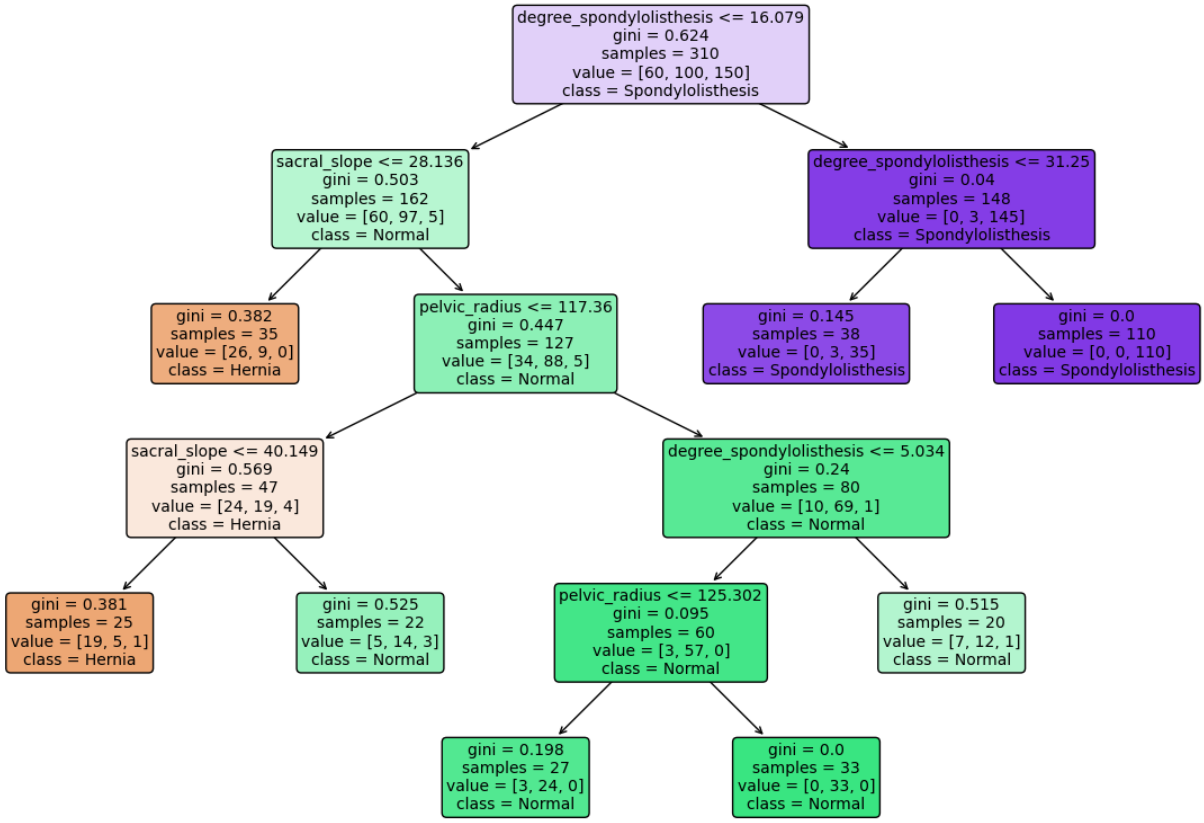


Figure 6: Decision Tree for exercise II.4a

(b) **Characterize a hernia condition by identifying the hernia-conditional associations.**

The hernia condition can be characterized by:

i. degree_spondilolysthesis $\leq$ 16.079 and sacral_slope $\leq$ 28.136

ii. degree_spondilolysthesis $\leq$ 16.079, sacral_slope $>$ 28.136, pelvic_radius $\leq$ 117.36 and sacral_slope $\leq$ 40.149