Aprendizagem 2023/24

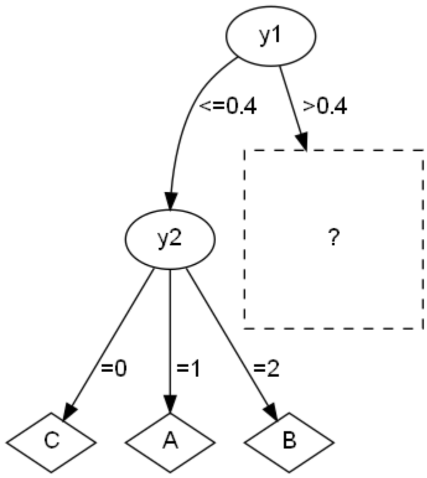
Homework I

Deadline: 29/9/2023 (Friday) 23:59 via Fenix as PDF

I. Pen-and-paper [11v]

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.

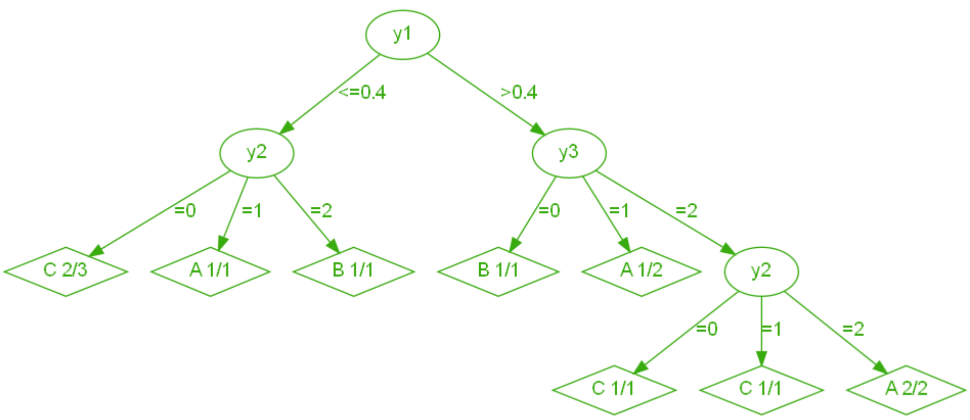
D	y_1	y_2	y_3	y_4	y_{out}
$\mathbf{x}_1$	0.24	1	1	0	A
$\mathbf{x}_2$	0.06	2	0	0	B
$\mathbf{x}_3$	0.04	0	0	0	B
$\mathbf{x}_4$	0.36	0	2	1	C
$\mathbf{x}_5$	0.32	0	0	2	C
$\mathbf{x}_6$	0.68	2	2	1	A
$\mathbf{x}_7$	0.9	0	1	2	A
$\mathbf{x}_8$	0.76	2	2	0	A
$\mathbf{x}_9$	0.46	1	1	1	B
$\mathbf{x}_{10}$	0.62	0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



[5v] 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.

$$IG(y2, z \mid y1 > 0.4) = 0.59, \;\; IG(y3, z \mid y1 > 0.4) = 0.7, \;\; IG(y4, z \mid y1 > 0.4) = 0.59$$

$$IG(y2, z \mid y1 > 0.4, y3 = 2) = 1, \;\; IG(y4, z \mid y1 > 0.4, y3 = 2) = 0.5$$



1) [2.5v] Draw the training confusion matrix for the learnt decision tree.

		true A	true B	true C
predicted	A	4	1	0
predicted	B	0	2	0
predicted	C	0	1	4

Aprendizagem 2023/24

Homework I

Deadline: 29/9/2023 (Friday) 23:59 via Fenix as PDF

2) [1.5v] Showing your calculus, identify the class with the lowest training F1 score.

$$recall(A) = 1, precision(A) = 4/5, F1(A) = 0.89$$

$$\Longrightarrow recall(B) = 0.5, precision(B) = 1, F1(B) = 2/3$$

$$recall(C) = 1, precision(C) = 4/5, F1(C) = 0.89$$

3) [1v] Considering y2 to be ordinal, assess if y1 and y2 are correlated using the Spearman coefficient.

$$Spearman(y1, y2) = Pearson(rank(y1), rank(y2)$$
$$= Pearson([3,2,1,5,4,10,12,11,7,9,6,8], [8,11,3.5,3.5,3.5,11,3,11,8,3.5,8,3.5]) = 0.08$$

4) [1v] Draw the class-conditional relative histograms of y1 using 5 equally spaced bins in [0,1]. Challenge: find the root split using the discriminant rules from these empirical distributions.

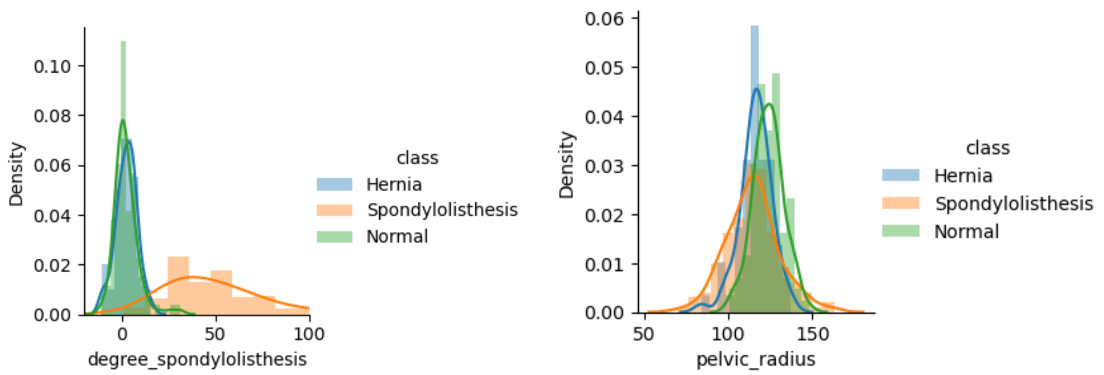
Comparing the class-conditional empirical distributions (side-by-side or in a single plot), we can assess the intervals in y1 where a given class yields higher probability to occur as a proxy to minimize the entropy:

$$\theta \leq 0.2\ (\{\mathbf{x_2}, \mathbf{x_3}\});\ 0.2 < \theta \leq 0.6\ (\{\mathbf{x_1}, \mathbf{x_4}, \mathbf{x_5}, \mathbf{x_9}, \mathbf{x_{11}}, \mathbf{x_{12}}\});\ \theta > 0.6\ (\{\mathbf{x_6}, \mathbf{x_7}, \mathbf{x_8}, \mathbf{x_{10}}\})$$

II. Programming [9v]

Considering 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) [1v] 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.




```
from sklearn.feature_selection import f_classif
scores, pvalues = f_classif(X, y)
col_min = data.columns[np.array(scores).argmin()]
col_max = data.columns[np.array(scores).argmax()]
```

2) [4v] 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 the split thresholding of numeric variables is non-deterministic in `sklearn`, hence we recommend averaging the results for 10 runs per parameterization.

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework I

Deadline: 29/9/2023 (Friday) 23:59 via Fenix as PDF

```
Depths:           [1,    2,    3,    4,    5,    6,    8,    10 ]
Train accuracies: [0.78, 0.84, 0.85, 0.9,  0.93, 0.97, 1.0,  1.0]
Test accuracies:  [0.75, 0.78, 0.78, 0.85, 0.84, 0.85, 0.8,  0.8]
```

```python
from sklearn import metrics, tree
from sklearn.model_selection import train_test_split

train_accs, test_accs = [], []
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, stratify=y, random_state=0)

for depth in [1,2,3,4,5,6,8,10]:
    train_acc, test_acc = [], []
    for i in range(10):
        predictor = tree.DecisionTreeClassifier(max_depth=depth, random_state=3)
        predictor.fit(X_train, y_train)
        train_acc.append(metrics.accuracy_score(y_train, predictor.predict(X_train)))
        test_acc.append(metrics.accuracy_score(y_test, predictor.predict(X_test)))
    train_accs.append(round(np.average(train_acc),2))
    test_accs.append(round(np.average(test_acc),2))
```

3) [2v] Critically analyze these results, including the generalization capacity across settings.

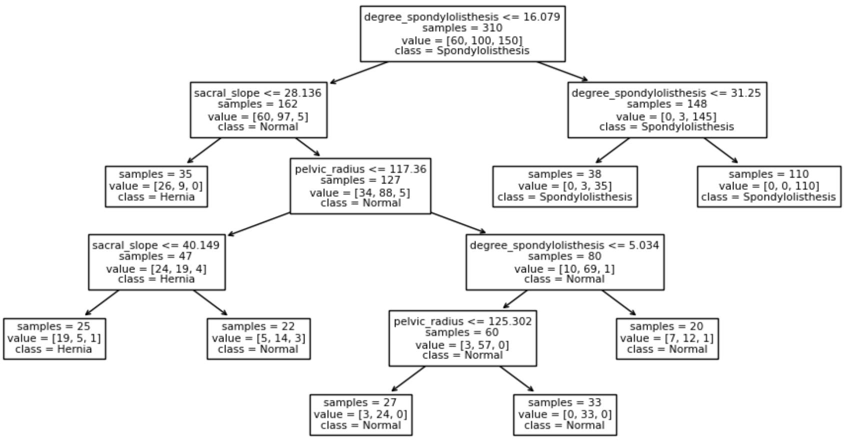
There seems to be considerable underfitting and overfitting risks for specific learning setting:

i) higher depths are associated with 100% training accuracy, decreased testing accuracy, and heightened training-testing differences (overfitting risk);
ii) too restricted depths are associated with suboptimal testing accuracy (underfitting risk).

A depth limit in {4,5,6} seems to trade-off these risks, yielding an optimal testing accuracy.

4) [2v] To deploy the predictor, a healthcare opted to learn a single decision tree (`random_state=0`) using *all* available data and ensuring that each leave has a minimum of 20 individuals in order to avoid overfitting risks.

i. Plot the decision tree.
ii. Characterize a hernia condition by identifying the hernia-conditional associations.



Hernia-conditional associations: individuals with a spondylolisthesis degree below 16.079 **and either** a sacral slope below 28.136 (yielding a posterior probability of 26/35=0.74 to have an hernia) **or** a pelvic radius below 117.36 and sacral slope below 40.149 (yielding a probability of 19/25=0.76).

END