TÉCNICO LISBOA

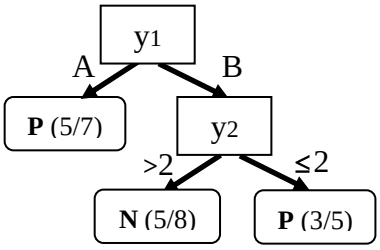
Aprendizagem 2022/23

**Homework I**

Deadline: 7/10/2022 (Friday) 23:59 via Fenix as PDF

## I. Pen-and-paper [12v]

Given the following decision tree learnt from 20 observations using Shannon entropy, with leaf annotations `(#correct/#total)`




1) [2.5v] Draw the training confusion matrix.

   TP=5+3, TN=5, FP=2+2, FN=3

2) [2v] Identify the training F1 after a post-pruning of the given tree under a maximum depth of 1.

   Right branch: 7 N and 6 P, hence classification is N (7/13).

   TP=5, TN=7, FP=2, FN=6. Hence F1=5/9

3) [1.5v] Identify two different reasons as to why the left tree path was not further decomposed.

   Possibilities: $y_2$ does not discriminate the target class on A-conditional data; unsatisfaction of the necessary minimum number of instances to split (parameter in sklearn); post-pruning to avoid overfitting risks, etc.

4) [3.5v] Compute the information gain of variable y1.

$$IG(y1) = E(z) - E(z|y1) = 0.993 - 0.949 = 0.044$$

$$E(z) = -\frac{11}{20} log \frac{11}{20} - \frac{9}{20} log \frac{9}{20} = 0.993$$

$$E(z|y1) = -\frac{7}{20} \times \left(\frac{5}{7} log \frac{5}{7} + \frac{2}{7} log \frac{2}{7}\right) - \frac{13}{20} \times \left(\frac{7}{13} log \frac{7}{13} + \frac{6}{13} log \frac{6}{13}\right) = 0.949$$

## II. Programming [8v]

Considering the `pd_speech.arff` data available at the homework tab.

1) [6v] Using a stratified 70-30 training-testing split with a fixed seed (`random_state=1`), assess in a single plot the training and testing accuracies of a decision tree with no depth limits (and remaining default behavior) for a varying number of selected features in {10,40,100,250,700}. Feature selection should be performed based on the discriminative power of input variables according to mutual information criterion.

```
import pandas as pd, numpy as np
from scipy.io.arff import loadarff

# load and prepare data
data = loadarff('pd_speech.arff')
df = pd.DataFrame(data[0])
df['class'] = df['class'].str.decode('utf-8')
X, y = df.drop('class', axis=1), df['class']
```

TÉCNICO LISBOA

Aprendizagem 2022/23

**Homework I**

Deadline: 7/10/2022 (Friday) 23:59 via Fenix as PDF

```
from sklearn import metrics, tree
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import SelectKBest, mutual_info_classif

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=1)
predictor = tree.DecisionTreeClassifier()

# either raw function or SelectKBest are acceptable
feature_scores = mutual_info_classif(X_train, y_train)
sorted_features = np.argsort(feature_scores)
```

```
# iterate per m best features
for m in [10,40,100,250,700]:
    train_m_accs, test_m_accs = [], []

    # option with raw mutual_info_classif
    top_features = sorted_features[-m:]
    X_m_train, X_m_test = X_train.iloc[:,top_features], X_test.iloc[:,top_features]

    # option with SelectKBest
    '''selector = SelectKBest(mutual_info_classif, k=m)
       selector.fit(X_train, y_train)
       X_m_train, X_m_test = selector.transform(X_train), selector.transform(X_test)'''

    predictor.fit(X_m_train, y_train)
    train_accs.append(round(metrics.accuracy_score(y_train, predictor.predict(X_m_train)),2))
    test_accs.append(round(metrics.accuracy_score(y_test, predictor.predict(X_m_test)),2))

print("Train accuracies:",train_accs,"\nTest accuracies:",test_accs)
```

`Train accuracies: [1.0, 1.0, 1.0, 1.0, 1.0]`
`Test accuracies: [0.78, 0.81, 0.82, 0.85, 0.85]` (estimates can vary due to non-determinism)

Line and scatter plots accepted.

2) [2v] Why training accuracy is persistently 1? Critically analyze the gathered results.

Notes: overfitting to training data (100% training accuracy while significantly lower testing accuracy) attributed to the fact that the learnt trees have no limited depth. Arguably, the top 250 features yield nearly as much discriminative info as the original feature space. Additional comments on the impact of data dimensionality on testing accuracies acceptable in accordance with the gathered accuracy estimates.

END