Universidade da Beira Interior
Departamento de Informática
Inteligência Artificial

Practical exercises 7

Ano letivo 2018-19

Exercises

1. This problem is about classification of flowers into 3 classes: versicolor, viginica and setosa. There are 4 atributes or features: sepal length, sepal width, petal length, petal width.

   Use the scikit-learn package to create a decision tree that can classify the data in the Iris dataset (see example on the link above).

   The structure that contains the data has the following fields: `iris['data']` is the set of input vectors with flower attributes; `iris['target']` is an array with the output labels; `iris['target_names']` has the names that correspond to the numeric output lables. There is also a `iris['DESCR']` that contains a description of the problem. Make sure you read it.

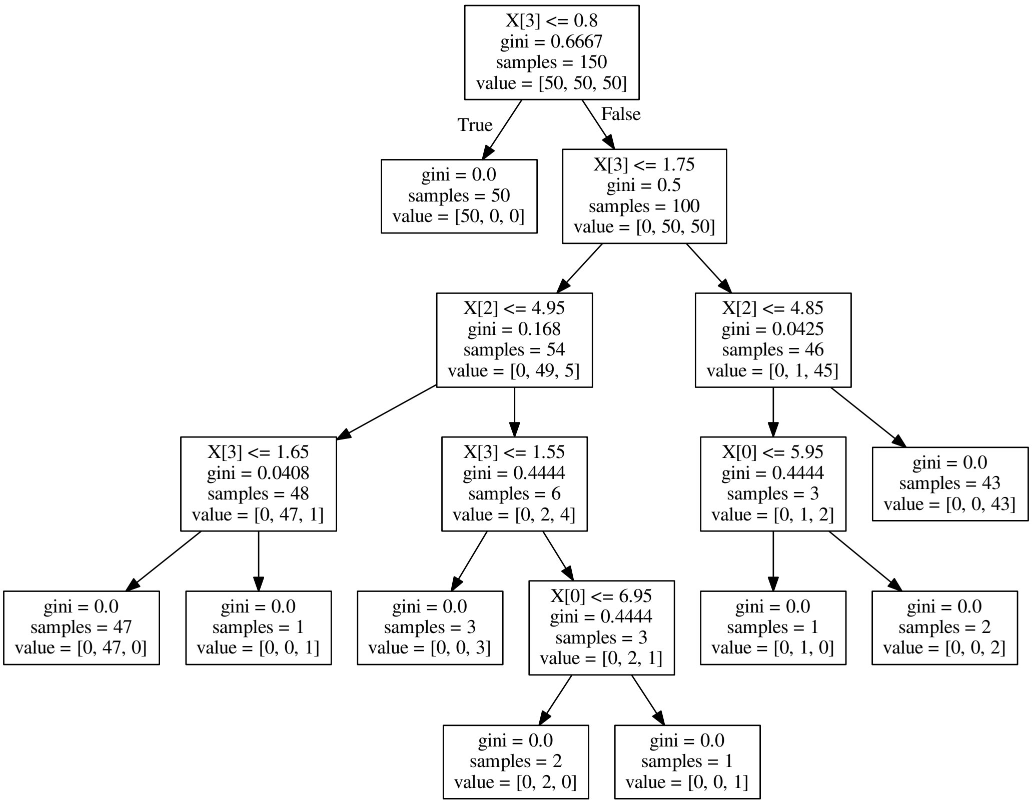
   You should obtain a decision tree similar to the one below.

   (a) Make sure you understand the meaning of all the values in the tree nodes. The value 'gini' represents the Gini impurity. It is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. High values are 'bad'. The minimum (zero) is reached when all cases in the node fall into a single target category.

   (b) What does $X[3]$ represent?

   (c) What is the meaning of $value = [0, 0, 3]$?

   (d) What is the class your model assigns to a flower that has the following features: [2.1, 2.7, 3.9, 1.2]? Use the decision tree to get the result, not your inspection of the tree.



2. Consider the data set in the file `CTG.csv`. For details on it see here. You can load the data into your program and separate the inputs (`X`) from the outputs (`Y`) using the following code:

```
import numpy as np
data = np.loadtxt(’CTG.csv’, delimiter=’,’)
X = data[:,0:21]
Y = data[:,21]
```

This data set is much larger than the previous one. We are going to use it to build a learning curve.

   (a) Create a test set using the first 126 points. The remaining of the data set will be used for training.

   (b) Build training sets of increasing size with 100, 200, 500, 1000, 2000 data points and observe the behaviour of the test error as the training set size increases.

   (c) Create a learning curve by plotting the values obtained in the previous questions.

3. We want to estimate the generalization capability by obtaining test error estimates using the 10-fold cross validation method. Compare the error you obtain using this procedure with the learning curve of the previous exercise. What do you conclude?
   Sugestion: try using `ShuffleSplit` from the `sklearn.model_selection` package.

4. This is a regression problem. Read the data from the file with:

   ```
   import numpy as np
   data = np.loadtxt('regression.txt')
   ```

   The input data are the weights in Kg of 1 year old babies and the targets (output values) are the height of the babies in cm.

   (a) Create a regression tree and estimate the height of a baby that weights 11.5Kg.

   (b) Plot the values in the training set along with the input and prediction obtained in the previous question (use a different color to distinguish from the training set data points).