Assignment 3 - Machine Learning
EDA132 Applied Artificial Intelligence

Johanna Petersson
adi10jpe@student.lu.se

March 7, 2016

1 Introduction

This assignment was to implement an algorithm that creates a decision tree with the help of datasets in Weka ARFF files. That tree can then be used in supervised learning agents. A single decision tree is created from one relation. A relation consists of a name, attributes that has a specified set of possible values, and examples which state the classification for different combinations of values of attributes.

2 Improvements

For this assignment, a base implementation and at least one of three defined improvements were required to be implemented. The decision fell to implement two of the given improvements.

First a pruning procedure was implemented as an improvement. Selected mostly because it seemed to be the most useful improvement to make to the program. Secondly the program was improved so that it is possible to assign real values on the attributes instead of just enumerating different values allowed.

The Weka ARFF format specifies several different types of values that each attribute can have. In the implemented solution only two types of values are allowed. These are numeric values (real and integral) and nominal values. Nominal values are an enumeration of values that the attribute may take.

3 Results

The algoritm and implementation has been run over three different example sets. The first was an example in the book Artificial Intelligence A modern Approach by Stuart Russell and Peter Norvig (3rd edition, International, was used). The definition for the structure of the data is found on page 709 and the actual data on page 711. In addition to that the files the weather.nominal and diabetes example set that was used the previous year of the course (given by another student, might be used this year as well). The data sets and the result of the sets can be found in the Appendices A, B and C.

4 Discussion

The pruning in the program was implemented according to how the textbook explained it. However, there is some uncertainty if it works as it should. For the small weather example it prunes even though it could be argued that it might be wrong to do so. And for the big diabetes set it only prunes a few things and arguably it might be supposed to prune more. Since there was no access to any example set to verify the pruning procedure there was no way to verify if it works correctly.

The improvement to represent numerical values in the tree seems to work by giving each attribute only two values for them to take. This in turn means that each node in the tree gets only two child nodes which represents bigger or smaller values respectively, seen from the split point. For the example, the split point has to be set somewhere. It was first tried to have the same split point for all attributes, but it was later changed to a point for each attribute. For the diabetes example set the split point was set to the mean values of each attribute.

5 Source Code

The implementation is split into several classes, each in its own .java file. In this section each of these files and classes are explained and some highlights of interesting points are given.

5.1 Attribute.java

This class represents an attribute with a set of values. The values are simply is the names of the values. In the case of numerical values the attribute should only have two given values, defined by a split point. The most interesting methods in this class are:

`public boolean testExample(String attributeValue, LearningExample example)` This method tests if the value is numerical and that it belongs to the right value branch defined by `attributeValue`. If it is not a numerical attribute then it just checks that the example is the current branch being evaluated.

`public void setSplitPoint(double splitPoint)` This method should only be called if it is a numerical attribute and then overrides all the previous values in the value set, be simply reseting it. This method is not well designed for production but does its job in this small example code.

`public String getKeyIfNumerical(String value)` If the attribute is numerical the value is returned. Otherwise the same value is returned.

5.2 Classification.java

This class is used when classifying an example. It contains an attribute which is the attribute used for classification. This class also takes a value which is assigned to the attribute and a boolean which indicates whether this was a positive or negative classification.

5.3 LearningExample.java

This class represents an example that is fed to the decision tree learning algorithm.

5.4 Relation.java

This class represents an entire relation, containing its examples and attributes.

5.5 DecisionTreeParser.java

This is the class that is responsible for parsing the Weka ARFF files of the program, in order to build a model with the classes in the rest of the program. It creates a `Relation` object and populates it with attributes and examples.

5.6 DecisionNode.java

This is an interface for nodes in the tree, to allow different type of nodes to be in the same collections.

5.7 AttributeNode.java

This class implements the `DecisionNode` interface and is used as a branch in the decision tree. The `Attribute` of the node determines which attribute this branch is, and the list of `LearningExample` objects are the examples that are left when we have reached this attribute in the decision tree algorithm.

5.8 TerminalNode.java

This class implements the `DecisionNode` interface and only consists of a `Classification`.

5.9 DecisionTreeAlgorithm.java

This is the class that holds the algorithm for building and pruning the decision tree. It takes a `Relation` and creates a decision tree from it.

`public DecisionNode decsionTreeLearning()` This method creates a decision tree from the relation given to the constructor.

`public DecisionNode pruning(DecisionNode tree)` This method takes a decision tree and prunes it using the standard deviation algorithm explained in the book for pruning. Just like the building of the tree this method uses different types of private methods to calculate the deviation.

5.10 Program.java

This class consists of the main method of the program. This starts the program and initiates all the necessary classes and runs the decision tree algorithm for three different files. Could have been modified to take some arguments to run a specific file for example.

6 Library

A choice was made to use an external library to help get the χ^2 table needed to do a χ^2-pruning. The choice fell to the libarary *SSJ - Stochastic Simulation in Java* that was developed at Pierre L'Ecuyer's Simulation and Optimization Laboratory, located at the Department of Computer Science and Operations Research of Université de Montréal.

It is available at `http://umontreal-simul.github.io/ssj` and was downloaded as a binary release from `https://github.com/umontreal-simul/ssj/releases`. Version `3.0.0-rc1` was used.

7 Getting and running the program

The code and an executable .jar file can be found at `https://github.com/JohannaMoose/eda132` in the folder "Assignment 3 - Machine Learning". To download the code and program onto one's computer go to the URL above and click the button "Download ZIP" that can be found just above the file list. This will download a zip file to the computer with all the files; go to where the file downloaded and into the folder "Assignment 3 - Machine Learning" to see the files pertaining to this assignment.

The file `DecisionTree.jar` is runnable from the terminal on Linux/Mac computers with the command `java -jar DecisionTree.jar` (tested on Mac) if the terminal is pointed to the `dist` folder in the project. This will start the program and the analysis of the data in the nested data folder. It is most important that all files found in the dist folder are were they should be for the program to run.

Of cause, the code in the src folder should be buildable as well and should run the program without a hiccup as long as the folder structure of the src, data and lib folders stay the same, provided that Java SDK8 is installed on the computer. Nothing else is tested.

The external library references might need to be updated depending on how the compiler is set up and what program is used, they can be found in the lib folder.

A Example 18-3

The example set: The results:

```
patrons = some: yes
patrons = none: no
patrons = full
 hungry = no: no
 hungry = yes
  type = burger: yes
  type = thai
   friday/saturday = no: no
   friday/saturday = yes: yes
  type = italian: no
  type = french: yes
```

B weather.nominal

The example set: The results:

```
outlook = rainy: yes
outlook = overcast: yes
outlook = sunny: no
```

C diabetes

The example set: The results:

```
’plas’ = >120.9
 ’mass’ = <=32.0
  ’preg’ = >3.8
   ’pedi’ = >0.5
    ’age’ = >33.2
     ’insu’ = <=79.8
      ’pres’ = <=69.1: tested_negative
      ’pres’ = >69.1
       ’skin’ = >20.5: tested_negative
       ’skin’ = <=20.5: tested_negative
     ’insu’ = >79.8
      ’pres’ = <=69.1
       ’skin’ = >20.5: tested_positive
       ’skin’ = <=20.5: tested_positive
      ’pres’ = >69.1
       ’skin’ = >20.5: tested_negative
       ’skin’ = <=20.5: tested_positive
    ’age’ = <=33.2
     ’insu’ = <=79.8: tested_positive
     ’insu’ = >79.8
      ’skin’ = >20.5: tested_positive
      ’skin’ = <=20.5: tested_negative
   ’pedi’ = <=0.5
    ’skin’ = >20.5
     ’insu’ = <=79.8: tested_negative
     ’insu’ = >79.8
      ’age’ = >33.2
       ’pres’ = <=69.1: tested_negative
       ’pres’ = >69.1: tested_positive
      ’age’ = <=33.2
       ’pres’ = <=69.1: tested_negative
       ’pres’ = >69.1: tested_negative
    ’skin’ = <=20.5
     ’age’ = >33.2
      ’insu’ = <=79.8
       ’pres’ = <=69.1: tested_negative
       ’pres’ = >69.1: tested_positive
      ’insu’ = >79.8: tested_negative
     ’age’ = <=33.2
      ’insu’ = <=79.8
       ’pres’ = <=69.1: tested_negative
       ’pres’ = >69.1: tested_negative
      ’insu’ = >79.8: tested_positive
  ’preg’ = <=3.8
   ’age’ = >33.2
    ’skin’ = >20.5
     ’insu’ = <=79.8
      ’pres’ = <=69.1: tested_negative
      ’pres’ = >69.1
       ’pedi’ = >0.5: tested_negative
       ’pedi’ = <=0.5: tested_negative
     ’insu’ = >79.8: tested_positive
    ’skin’ = <=20.5
     ’insu’ = <=79.8
      ’pedi’ = >0.5
       ’pres’ = <=69.1: tested_negative
       ’pres’ = >69.1: tested_positive
      ’pedi’ = <=0.5
       ’pres’ = <=69.1: tested_negative
       ’pres’ = >69.1: tested_negative
     ’insu’ = >79.8: tested_negative
   ’age’ = <=33.2
    ’pres’ = <=69.1
     ’insu’ = <=79.8
      ’pedi’ = >0.5
       ’skin’ = >20.5: tested_negative
       ’skin’ = <=20.5: tested_negative
      ’pedi’ = <=0.5
       ’skin’ = >20.5: tested_negative
       ’skin’ = <=20.5: tested_negative
     ’insu’ = >79.8
      ’skin’ = >20.5: tested_negative
      ’skin’ = <=20.5
       ’pedi’ = >0.5: tested_negative
       ’pedi’ = <=0.5: tested_negative
    ’pres’ = >69.1
     ’skin’ = >20.5
      ’pedi’ = >0.5
       ’insu’ = <=79.8: tested_positive
       ’insu’ = >79.8: tested_negative
      ’pedi’ = <=0.5
       ’insu’ = <=79.8: tested_negative
       ’insu’ = >79.8: tested_negative
     ’skin’ = <=20.5
      ’insu’ = <=79.8
       ’pedi’ = >0.5: tested_negative
       ’pedi’ = <=0.5: tested_negative
      ’insu’ = >79.8
```

```
      ’pedi’ = >0.5: tested_positive
      ’pedi’ = <=0.5: tested_positive
’mass’ = >32.0
 ’age’ = >33.2
  ’skin’ = >20.5
   ’preg’ = >3.8
    ’pedi’ = >0.5
     ’insu’ = <=79.8: tested_positive
     ’insu’ = >79.8
      ’pres’ = <=69.1: tested_negative
      ’pres’ = >69.1: tested_positive
    ’pedi’ = <=0.5
     ’pres’ = <=69.1: tested_positive
     ’pres’ = >69.1
      ’insu’ = <=79.8: tested_positive
      ’insu’ = >79.8: tested_positive
   ’preg’ = <=3.8
    ’pres’ = <=69.1: tested_positive
    ’pres’ = >69.1
     ’insu’ = <=79.8
      ’pedi’ = >0.5: tested_positive
      ’pedi’ = <=0.5: tested_positive
     ’insu’ = >79.8
      ’pedi’ = >0.5: tested_positive
      ’pedi’ = <=0.5: tested_positive
  ’skin’ = <=20.5
   ’insu’ = <=79.8
    ’preg’ = >3.8
     ’pres’ = <=69.1: tested_positive
     ’pres’ = >69.1
      ’pedi’ = >0.5: tested_positive
      ’pedi’ = <=0.5: tested_positive
    ’preg’ = <=3.8
     ’pedi’ = >0.5: tested_negative
     ’pedi’ = <=0.5
      ’pres’ = <=69.1: tested_positive
      ’pres’ = >69.1: tested_positive
   ’insu’ = >79.8: tested_negative
 ’age’ = <=33.2
  ’pedi’ = >0.5
   ’skin’ = >20.5
    ’preg’ = >3.8
     ’pres’ = <=69.1: tested_negative
     ’pres’ = >69.1
      ’insu’ = <=79.8: tested_positive
      ’insu’ = >79.8: tested_positive
    ’preg’ = <=3.8
     ’pres’ = <=69.1
      ’insu’ = <=79.8: tested_positive
      ’insu’ = >79.8: tested_positive
      ’pres’ = >69.1
       ’insu’ = <=79.8: tested_positive
       ’insu’ = >79.8: tested_positive
    ’skin’ = <=20.5: tested_positive
   ’pedi’ = <=0.5
    ’pres’ = <=69.1
     ’preg’ = >3.8
      ’skin’ = >20.5: tested_negative
      ’skin’ = <=20.5
       ’insu’ = <=79.8: tested_negative
       ’insu’ = >79.8: tested_negative
     ’preg’ = <=3.8
      ’insu’ = <=79.8
       ’skin’ = >20.5: tested_negative
       ’skin’ = <=20.5: tested_positive
      ’insu’ = >79.8
       ’skin’ = >20.5: tested_positive
       ’skin’ = <=20.5: tested_negative
    ’pres’ = >69.1
     ’preg’ = >3.8
      ’skin’ = >20.5
       ’insu’ = <=79.8: tested_negative
       ’insu’ = >79.8: tested_positive
      ’skin’ = <=20.5
       ’insu’ = <=79.8: tested_positive
       ’insu’ = >79.8: tested_negative
     ’preg’ = <=3.8
      ’insu’ = <=79.8
       ’skin’ = >20.5: tested_negative
       ’skin’ = <=20.5: tested_negative
      ’insu’ = >79.8
       ’skin’ = >20.5: tested_negative
       ’skin’ = <=20.5: tested_negative
’plas’ = <=120.9
 ’preg’ = >3.8
  ’pedi’ = >0.5
   ’insu’ = <=79.8
    ’skin’ = >20.5
     ’age’ = >33.2
      ’mass’ = <=32.0
       ’pres’ = <=69.1: tested_negative
       ’pres’ = >69.1: tested_negative
      ’mass’ = >32.0
       ’pres’ = <=69.1: tested_positive
       ’pres’ = >69.1: tested_positive
     ’age’ = <=33.2: tested_negative
    ’skin’ = <=20.5
     ’mass’ = <=32.0
      ’age’ = >33.2
       ’pres’ = <=69.1: tested_negative
     ’pres’ = >69.1: tested_negative
    ’age’ = <=33.2
     ’pres’ = <=69.1: tested_negative
     ’pres’ = >69.1: tested_positive
   ’mass’ = >32.0: tested_negative
 ’insu’ = >79.8
  ’pres’ = <=69.1
   ’mass’ = <=32.0
    ’skin’ = >20.5
     ’age’ = >33.2: tested_negative
     ’age’ = <=33.2: tested_negative
    ’skin’ = <=20.5
     ’age’ = >33.2: tested_positive
     ’age’ = <=33.2: tested_positive
   ’mass’ = >32.0: tested_positive
  ’pres’ = >69.1: tested_positive
’pedi’ = <=0.5
 ’insu’ = <=79.8
  ’age’ = >33.2
   ’pres’ = <=69.1
    ’mass’ = <=32.0
     ’skin’ = >20.5: tested_negative
     ’skin’ = <=20.5: tested_negative
    ’mass’ = >32.0
     ’skin’ = >20.5: tested_negative
     ’skin’ = <=20.5: tested_positive
   ’pres’ = >69.1
    ’mass’ = <=32.0
     ’skin’ = >20.5: tested_negative
     ’skin’ = <=20.5: tested_negative
    ’mass’ = >32.0
     ’skin’ = >20.5: tested_negative
     ’skin’ = <=20.5: tested_negative
  ’age’ = <=33.2
   ’skin’ = >20.5
    ’mass’ = <=32.0
     ’pres’ = <=69.1: tested_negative
     ’pres’ = >69.1: tested_negative
    ’mass’ = >32.0
     ’pres’ = <=69.1: tested_negative
     ’pres’ = >69.1: tested_negative
   ’skin’ = <=20.5
    ’pres’ = <=69.1
     ’mass’ = <=32.0: tested_negative
     ’mass’ = >32.0: tested_negative
    ’pres’ = >69.1
     ’mass’ = <=32.0: tested_negative
     ’mass’ = >32.0: tested_negative
 ’insu’ = >79.8
  ’age’ = >33.2
```

```
    ’mass’ = <=32.0: tested_negative
    ’mass’ = >32.0
     ’pres’ = <=69.1: tested_negative
     ’pres’ = >69.1
      ’skin’ = >20.5: tested_negative
      ’skin’ = <=20.5: tested_negative
   ’age’ = <=33.2: tested_negative
’preg’ = <=3.8
 ’mass’ = <=32.0
  ’age’ = >33.2
   ’pres’ = <=69.1: tested_negative
   ’pres’ = >69.1
    ’skin’ = >20.5: tested_negative
    ’skin’ = <=20.5
     ’insu’ = <=79.8
      ’pedi’ = >0.5: tested_positive
      ’pedi’ = <=0.5: tested_positive
     ’insu’ = >79.8: tested_negative
  ’age’ = <=33.2
   ’skin’ = >20.5
    ’pedi’ = >0.5: tested_negative
    ’pedi’ = <=0.5
     ’pres’ = <=69.1
      ’insu’ = <=79.8: tested_negative
      ’insu’ = >79.8: tested_negative
     ’pres’ = >69.1
      ’insu’ = <=79.8: tested_negative
      ’insu’ = >79.8: tested_negative
   ’skin’ = <=20.5
    ’pedi’ = >0.5
     ’pres’ = <=69.1
      ’insu’ = <=79.8: tested_negative
      ’insu’ = >79.8: tested_negative
     ’pres’ = >69.1: tested_negative
    ’pedi’ = <=0.5: tested_negative
 ’mass’ = >32.0
  ’pedi’ = >0.5
   ’skin’ = >20.5
    ’age’ = >33.2
     ’insu’ = <=79.8
      ’pres’ = <=69.1: tested_positive
      ’pres’ = >69.1: tested_positive
     ’insu’ = >79.8: tested_negative
    ’age’ = <=33.2
     ’insu’ = <=79.8
      ’pres’ = <=69.1: tested_negative
      ’pres’ = >69.1: tested_negative
     ’insu’ = >79.8
      ’pres’ = <=69.1: tested_negative
      ’pres’ = >69.1: tested_negative
 ’skin’ = <=20.5
  ’age’ = >33.2: tested_negative
  ’age’ = <=33.2
   ’insu’ = <=79.8
    ’pres’ = <=69.1: tested_negative
    ’pres’ = >69.1: tested_negative
   ’insu’ = >79.8
    ’pres’ = <=69.1: tested_positive
    ’pres’ = >69.1: tested_positive
’pedi’ = <=0.5
 ’age’ = >33.2
  ’insu’ = <=79.8
   ’skin’ = >20.5: tested_negative
   ’skin’ = <=20.5
    ’pres’ = <=69.1: tested_negative
    ’pres’ = >69.1: tested_negative
  ’insu’ = >79.8: tested_positive
 ’age’ = <=33.2
  ’skin’ = >20.5
   ’pres’ = <=69.1
    ’insu’ = <=79.8: tested_negative
    ’insu’ = >79.8: tested_negative
   ’pres’ = >69.1
    ’insu’ = <=79.8: tested_negative
    ’insu’ = >79.8: tested_negative
  ’skin’ = <=20.5
   ’insu’ = <=79.8
    ’pres’ = <=69.1: tested_negative
    ’pres’ = >69.1: tested_negative
   ’insu’ = >79.8: tested_negative
```