CS 422: Data Mining

Department of Computer Science
Illinois Institute of Technology
Vijay K. Gurbani, Ph.D.

Fall 2021: Homework 4 (10 points)

Due date: Sun, October 03 2021, 11:59:59 PM Chicago Time

Please read all of the parts of the homework carefully before attempting any question. If you detect any ambiguities in the instructions, please let me know right away instead of waiting until after the homework has been graded.

Problem 1.1: Decision Tree on the Iris Dataset

Using the built-in Iris dataset in R [1], you will create a Decision Tree model. To create the model, you will use the rpart and rpart.plot packages. If these packages are not installed on your system, please install them first.

To build a decision tree model, first read the rpart and rpart.plot help file in R. You can get the help file by typing "?rpart" and "?rpart.plot" in the R REPL. Examples on how to build the tree and plot it are given towards the end section of the help file displayed in RStudio.

Use all of the data to train the decision tree.

Note that the first parameter to build a decision tree model is a formula, this is the same type of formula you used to build linear regression models. Pay close attention to the "method" parameter and set it appropriately to train the tree as a classification tree.

Once you have trained the decision tree, plot the tree using the following command (look at the rpart.plot help file to understand the parameters below):

```
> rpart.plot(model, extra=104, fallen.leaves=T, type=4, main="Iris Dataset Decision Tree")
```

Answer the following questions based on your reading of the help file and the plot of the tree. **(All questions carry equal points).**
(You must write these answers in a word processor file and submit the file as a PDF. In addition to the PDF file, you must also submit the R code in a .Rmd file and the resulting .nb.html file as usual.)

(a) How many levels are there in the decision tree?

(b) What is the default class label associated with each vertex?

Your output should look like the following:

```
Level N, Vertex 1: Default class label is <X>
Level N, Vertex 2: Default class label is <Y>
...
```

(c) Starting from the root note, what is the name of the first attribute used for a decision, and what are the split points? Your answer should be of the form:

```
Level N, split on attribute: <attribute name>.
```

```
        Split points: < X.X left subtree, >= X.X right subtree
...
```

(d) Each vertex has three lines.

(i) At each vertex, what do the three numbers in the middle line signify?

(ii) At each vertex, what does the last line signify?

[1] The Iris data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features are measured from each sample: the length and the width of the sepals and petals, in centimeters.