ALGONQUIN COLLEGE

CST8390 - LAB
BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 7-9

LAB 6 – Decision Trees

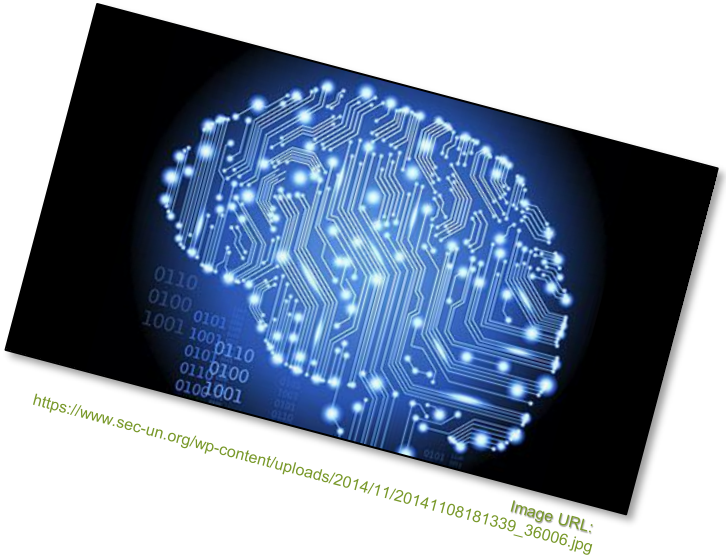
Lab 6 – Decision Trees

PART I

- Reviewing Algorithm

PART II

- Steps
- Results



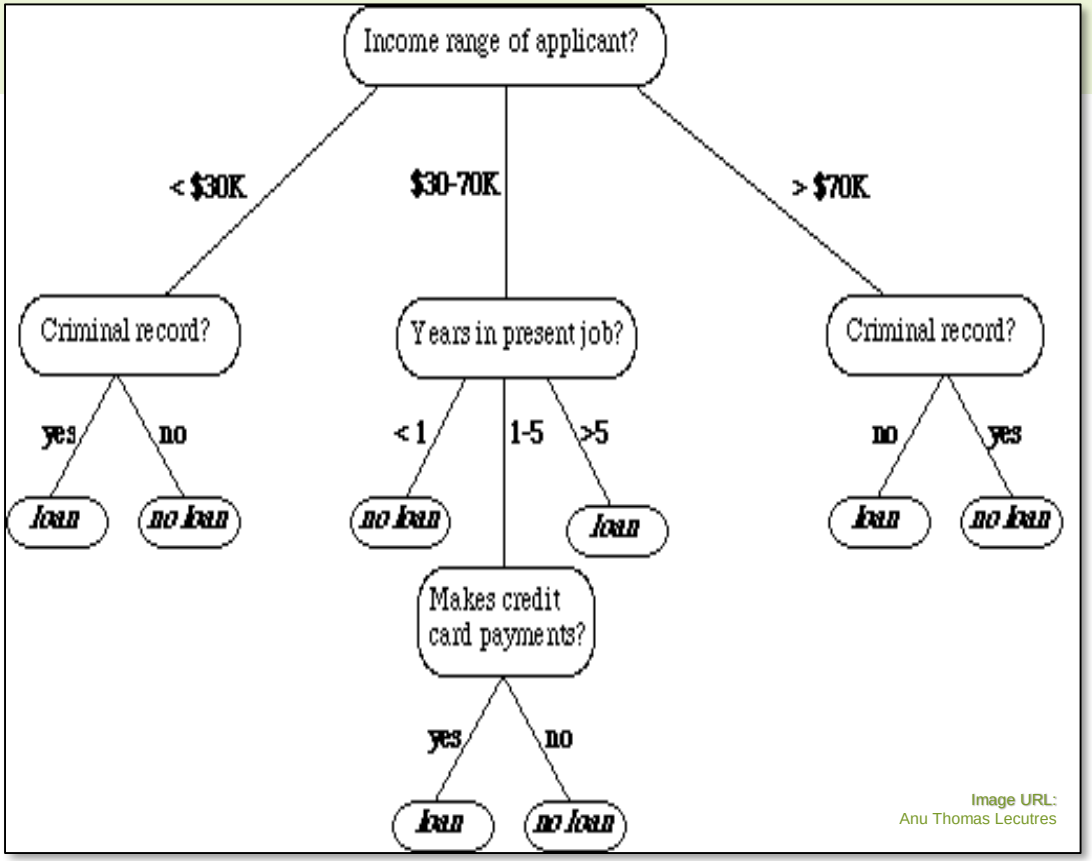
CST8390 - Lab
Business intelligence & data analytics

Lab 6 – Decision Trees

Part I – Reviewing Algorithm

Decision Trees

- **Idea**: ML algorithm used for classification.
- Decision tree is a tree where:
  - Each node is a feature (attribute)
  - Each branch is a decision (rule)
  - Each leaf represents an outcome.



Useful Measures

- **Entropy**: Caos = Uncertainty.
  - $H(S) = \sum_{c \in C} -p(c) \log_2 p(c)$
- **Information Gain**: How better is an attribute.
  - $IG(A, S) = H(S) - \sum_{t \in T} p(t) H(t)$
- The idea is using the entropy to evaluate how important is a specific attribute.
- The process is iterative and can be done until the certainty is obtained.

Decision Trees

Decision Tree:

- http://www.saedsayad.com/decision_tree.htm
- http://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/slides/Class3-DataMiningWithWeka-2013.pdf

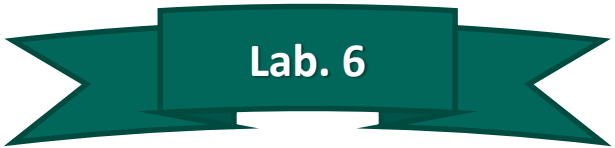
Covariance and correlation:

- http://www.dummies.com/education/math/business-statistics/how-to-measure-the-covariance-and-correlation-of-data-samples/

Image URL: https://images.idgesg.net/images/article/2018/10/analytics_binary_code_network_digital-transformation-100777428-large.jpg

Decision Trees

Demo



Step-by-step (A)

I. BASIC OPERATIONS

1. Open Diabetes dataset in text editor (from datasets that came with **Weka**). Read the information about the file. Fill in the following information (should be typed in).

a. Number of instances: _____.
b. Number of attributes: _____.
c. List of attributes (NOT abbreviation, should be typed in): _____.
d. Class labels and their relabelled values: _____.
e. Number of instances for each class label: _____.

Step-by-step (B)

2. Load the dataset in **Weka**. Take a screenshot and paste it below that shows class distribution.

3. Click on the "Choose" button on "Classify" tab and select **J48** from "trees". It is the implementation of the **C4.5 algorithm** which uses entropy to create a decision tree.

Step-by-step (C)

4. For testing the classification **accuracy**, make sure that "(Nom) Class" is selected, and cross-validation has 20 folds (Make sure that seed = 1). Click start and you should see a textual version of the **decision tree**.

a. **Copy** and **paste** the confusion matrix here.
b. Number of **leaves**: _____;
c. Size of the **tree**: _____;
d. Correctly **classified** instances: _____;

Step-by-step (D)

5. **Right click** on the result buffer and select "Visualize tree".

• From the new window, make it full screen and then **right-click** on the window and select "auto scale". It will **draw the tree** so that it's wide enough to read the text.

• You might have to **right-click** again on the screen and "Center on Top Node". You can use the mouse to pan around the tree to see all the decision splits.

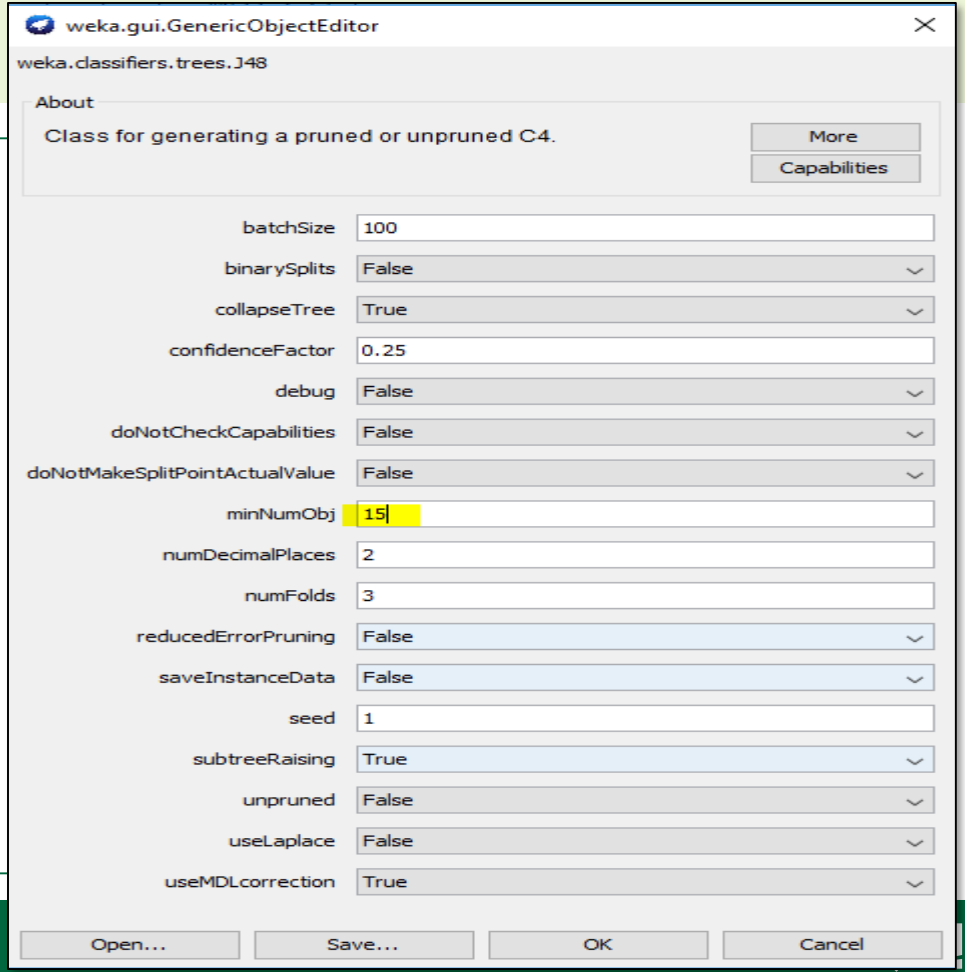
• **Right-click** again on the screen and select "Fit to Screen". Here you can see the tree all in one place, but the text might be hard to read.

• Have this window open for your lab demonstration. Also, take a **screenshot** and paste it here.

Step-by-step (E)

6. Set minNumObj to 15 in the settings window of the classifier, as shown

(This means that don't continue splitting if the nodes get very small. Default value is 2):



Step-by-step (F)

Run the classifier with this setting and fill in the following information:

a. **Copy** and **paste** the confusion matrix here.
b. Number of **leaves**: _____;
c. Size of the **tree**: _____;
d. Correctly **classified** instances: _____;

7. Take a **screenshot** of the tree and paste it here (from "Visualize tree").

8. Now, turn off pruning by setting unpruned property to (also, se minNumObj to **5**, seed = **1**) in the settings window of the Trueclassifier, as shown (this means that we are not reducing the size of the tree even if it is not giving much value for the task):

Step-by-step (G)

Run the classifier with this setting and fill in the following information:

a. **Copy** and **paste** the confusion matrix here.
b. Number of **leaves**: _____;
c. Size of the **tree**: _____;
d. Correctly **classified** instances: _____;

9. Run the classifier again with unpruned property to **True** and minNumObj to **15**, and fill in the answers for the questions below:

a. **Copy** and **paste** the confusion matrix here.
b. Number of **leaves**: _____;
c. Size of the **tree**: _____;
d. Correctly **classified** instances: _____;

Step-by-step (H)

10. Take a screenshot of the tree and paste it here (from "Visualize tree").

11. Decision trees have a problem with **overfitting**.

- One way to correct **overfitting** is with using random forests.
- This uses many decision trees, each built with **random subset** of the data.
- When a new item is going to be classified, the trees all vote when classifying each data item, with the majority deciding the final answer.
- The probability of an outlier being selected to be in several of the trees is highly unlikely so they will have less impact on the final classification.

Step-by-step (I)

To run the random forest algorithm, click the "Choose" button and select "Random Forest". Select Run the algorithm and paste the confusion matrix here:

a. Details of **random forest**: _____ with _____ iterations
b. **Time** taken to build model: _____.
c. **Correctly** classified instances: _____.

Show your answers to the lab professor when you are done.

Step-by-step (J)

REMEMBER:

- You should be ready with all your results in the result pane, and should show trees for steps 5, 7 and 10.

FOR YOUR ANALYSIS:

*** Option 1:** Explain with your own words what is a Decision Tree and where to use it.

*** Option 2:** Explain how to decide what is the strategy to decide how is the better parameter to use in a root node.

Image URL: https://www.marketingdirecto.com/wp-content/uploads/2018/01/ciencia-datos.jpg

Note: Due Date: Week 9 in corresponding lab sessions.

Open questions…

- Before we finish, do you have any doubt / question?

Image URL: https://cdn-az.allevents.in/events3/banners/26b150363d5757da8578fa6a1481585a368f12d4247adfe95837e6ea6c5ab2af-rimg-w1200-h549-gmir.jpg?v=1569691155

See you…

- Remember:
  - Labs require practice and it is ok committing errors and learning with them.
  - Do not forget to show your results…
  - Any questions, let me know…

sousap@algonquincollege.com

Image URL: https://thumbs.gfycat.com/MaleFrigidBull-size_restricted.gif

Thank you for your attention!