COMP3411/9814 Artificial Intelligence
Term 1, 2020

Assignment 2 – Machine Learning

Due: Tuesday 21 April, 10pm

Marks: 30% of final assessment for COMP3411/9814 Artificial Intelligence

Question 1: Decision trees

Consider the decision tree learning algorithm of Figure 7.7 and the data of Figure 7.1 from Poole & Mackworth [1], also presented below. Suppose, for this question, the stopping criterion is that all of the examples have the same classification. The tree of Figure 7.6 was built by selecting a feature that gives the maximum information gain. This question considers what happens when a different feature is selected.

Example	*Author*	*Thread*	*Length*	*Where_read*	*User_action*
e_1	*known*	*new*	*long*	*home*	*skips*
e_2	*unknown*	*new*	*short*	*work*	*reads*
e_3	*unknown*	*followup*	*long*	*work*	*skips*
e_4	*known*	*followup*	*long*	*home*	*skips*
e_5	*known*	*new*	*short*	*home*	*reads*
e_6	*known*	*followup*	*long*	*work*	*skips*
e_7	*unknown*	*followup*	*short*	*work*	*skips*
e_8	*unknown*	*new*	*short*	*work*	*reads*
e_9	*known*	*followup*	*long*	*home*	*skips*
e_{10}	*known*	*new*	*long*	*work*	*skips*
e_{11}	*unknown*	*followup*	*short*	*home*	*skips*
e_{12}	*known*	*new*	*long*	*work*	*skips*
e_{13}	*known*	*followup*	*short*	*home*	*reads*
e_{14}	*known*	*new*	*short*	*work*	*reads*
e_{15}	*known*	*new*	*short*	*home*	*reads*
e_{16}	*known*	*followup*	*short*	*work*	*reads*
e_{17}	*known*	*new*	*short*	*home*	*reads*
e_{18}	*unknown*	*new*	*short*	*work*	*reads*
e_{19}	*unknown*	*new*	*long*	*work*	?
e_{20}	*unknown*	*followup*	*short*	*home*	?

Figure 7.1: Examples of a user's preferences

a) Suppose you change the algorithm to always select the first element of the list of features. What tree is found when the features are in the order [*Author, Thread, Length, WhereRead*]? Does this tree represent a different function than that found with the maximum information gain split? Explain.

b) What tree is found when the features are in the order [*WhereRead, Thread, Length, Author*]? Does this tree represent a different function than that found with the maximum information gain split or the one given for the preceding part? Explain.

c) Is there a tree that correctly classifies the training examples but represents a different function than those found by the preceding algorithms? If so, give it. If not, explain why.

Question 2: Decision trees

The goal is to take out-of-the-box models and apply them to a given dataset. The task is to analyse the data and build a model to predict whether income exceeds $50K/yr based on census data (also known as "Census Income" dataset).
Use the data set **Adult Data Set** from the Machine Learning repository [2].
Use the supervised learning methods discussed in the lectures, Decision Trees.

Do not code these methods: instead use the implementations from scikit-learn. Read the scikit-learn documentation on Decision Trees [3], and the linked pages describing the parameters of the methods.

This question will help you master the workflow of model building. For example, you'll get to practice how to use the critical steps:

- Importing data
- Cleaning data
- Splitting it into train/test or cross-validation sets
- Pre-processing
- Transformations
- Feature engineering

Use the sklearn documentation pages for instructions. You should need the classification algorithms.

There are also available Tutorials:

- Sklearn – official tutorial for the sklearn package
- Predicting wine quality with scikit-learn – Step-by-step tutorial for training a machine learning model

The data is available here: http://archive.ics.uci.edu/ml/

machine-learning-databases/adult/

Preferences

1. Poole &Mackworth, Artificial Intelligence: Foundations of Computational Agents, Chapter 7, Supervised Machine Learning)

2. http://archive.ics.uci.edu/ml/datasets/Adult).

3. https://scikit-learn.org/stable/modules/tree.html

4. https://scikit-learn.org/stable/modules/naive_bayes.html

Submission

This assignment must be submitted electronically.
Put your zID and your name at the top of every page of your submission!

give cs3411 assign2 ...

The give script will accept *.pdf *.txt *.doc *.rtf
Late submissions will incur a penalty of 10% per day, applied to the maximum mark.

Group submissions will not be allowed. By all means, discuss the assignment with your fellow students. But you must write (or type) your answers individually. **Do NOT copy anyone else's assignment, or send your assignment to any other student.**