20T1: COMP9417 Machine Learning and Data Mining

Lectures: Tree Learning
Topic: Questions from lectures

Introduction

Some questions and exercises from the course lectures covering aspects of supervised tree learning for classification and regression.

Expressiveness of decision trees

Question 1 Give decision trees to represent the following Boolean functions, where the variables `A`, `B`, `C` and `D` have values `t` or `f`, and the class value is either `True` or `False`:

a) $A \wedge \neg B$

b) $A \vee [B \wedge C]$

c) $A \ \ XOR \ \ B$

d) $[A \wedge B] \vee [C \wedge D]$

Can you observe any effect of the increasing complexity of the functions on the form of their expression as decision trees ?

Decision tree learning

Question 2 Here is small dataset for a two-class prediction task. There are 4 attributes, and the class is in the rightmost column. Look at the examples. Can you guess which attribute(s) will be most predictive of the class ?

species	rebel	age	ability	homeworld
pearl	yes	6000	regeneration	no
bismuth	yes	8000	regeneration	no
pearl	no	6000	weapon-summoning	no
garnet	yes	5000	regeneration	no
amethyst	no	6000	shapeshifting	no
amethyst	yes	5000	shapeshifting	no
garnet	yes	6000	weapon-summoning	no
diamond	no	6000	regeneration	yes
diamond	no	8000	regeneration	yes
amethyst	no	5000	shapeshifting	yes
pearl	no	8000	shapeshifting	yes
jasper	no	6000	weapon-summoning	yes

You probably guessed that attributes 3 and 4 were not very predictive of the class, which is true. However, you might be surprised to learn that attribute "species" has higher information gain than attribute "rebel". Why is this ? Refer to slides 37-38 on "Attributes with Many Values" in the lecture notes.

Suppose you are told the following: for attribute "species" the Information Gain is 0.52 and *Split Information* is 2.46, whereas for attribute "rebel" the Information Gain is 0.48 and *Split Information* is 0.98.

Which attribute would the decision-tree learning algorithm select as the split when using the *Gain Ratio* criterion instead of Information Gain ? Is Gain Ratio a better criterion than Information Gain in this case ?

Question 3 Assume we learn a decision tree to predict class Y given attributes A, B and C from the following training set, with no pruning.

A	B	C	Y
0	0	0	0
0	0	1	0
0	0	1	0
0	1	0	0
0	1	1	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	0	1
1	1	1	0
1	1	1	1

What would be the training set error for this dataset ? Express your answer as the number of examples out of twelve that would be misclassified.

Question 4 One nice feature of decision tree learners is that they can learn trees to do *multi-class* classification, i.e., where the problem is to learn to classify each instance into exactly one of $k > 2$ classes.

Suppose a decision tree is to be learned on an arbitrary set of data where each instance has a discrete class value in one of $k > 2$ classes. What is the maximum training set error, expressed as a fraction, that any dataset could have ?