Prof. Corso, jcorso@buffalo.edu, SUNY Buffalo
Prepared by David Johnson

Solutions

Problem 1: Recall (2pts) (Answer in one sentence only.)

Name one reason one might use a decision tree or forest rather than another type of classifier.

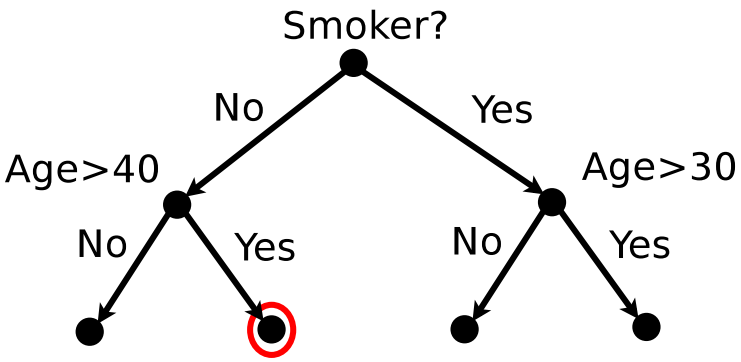
There are several valid answers to this question.

- *A learned decision tree model lends itself well to interpretation, and has much more intuitive meaning than most classifiers.*
- *Decision trees can handle cases where instances contain a mix of numerical and categorical (or real and discrete) data.*
- *Decision trees are reasonably robust to noisy or missing training data.*
- *Decision trees can easily handle cases where a disjunctive hypothesis is required (that is, where a single class of object actually consists of several sub-groups with very distinct feature representations—e.g. digital and analog clocks)*

Problem 2: Work (8 Points) (Show all derivations/work and explain.)

Given the data set and decision tree below, calculate the variance and misclassification impurity at the indicated node.

Class label	Feature	Feature
Has cancer?	Age	Smoker?
No	55	No
No	48	No
No	25	Yes
No	18	No
Yes	33	Yes
Yes	74	No
Yes	65	Yes



Begin by evaluating the first split. Smokers go to the right, while non-smokers go to the left. Since the node we are interested in is a child of the left-side node, we only need to worry about instances that go to the left on the first split (i.e. non-smokers).

There are 4 non-smokers, aged 55, 48, 18 and 74. The next split is between people younger than 40 and people 40 and older. The 55, 48 and 74 year olds thus go right, into the node with which we are concerned.

This leaves 3 people at the selected node, 1 with and 2 without cancer, so the node is $\frac{2}{3}$ *No Cancer and* $\frac{1}{3}$ *Cancer. Thus:*

Variance impurity $= \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{9}$

Misclassification impurity $= 1 - \max(\frac{1}{3}, \frac{2}{3}) = 1 - \frac{2}{3} = \frac{1}{3}$

Half credit was given for answers that got one of the two impurity measures wrong. I also deducted 2 points for using Gini impurity instead of variance impurity—they are not quite *the same, even for the two class case (specifically,* $imp_G = 2 \cdot imp_V$*).*