TITLE OF PROJ. OR STUDY PROJ. OR STUDY NO.

SUBJECT CISC-483-683 WORKS

Assignment 4, Fall 2019 Answers COMPUTER DATE

1. Consider the subtree of Location at the node COST (Amenities = some, Location = south).

Cost = low has 1 error out of 3

Cost = med has 3 errors out of 4

Cost = high has 0 errors out of 1

weighted error is $\frac{3}{8}(\frac{1}{3}) + \frac{4}{8}(\frac{3}{4}) + \frac{1}{8}(0) = 4/8 = \frac{1}{2}$

(You can also do this as just $\frac{1}{8} + \frac{3}{8} + \frac{0}{8}$; with reduced error pruning, the answers are the same. This is the same as counting the errors and dividing by the number of instances — ie. 4/8)

If we remove the test at COST, then the answer at Location = south will be determined by the majority class ~~of~~ the training instances that reached the COST node, which is (5 YES, 3 NO) so the answer would be YES.

Thus the error rate at COST on the test data is 3/8.

So prune COST.

Now look at the node labelled CONDITION. (Amenities = some, LOCATION = west).

CONDITION = excellent, error is 0 out of 0

CONDITION = good, error is 1 out of 2

CONDITION = OK, error is 1 out of 3

CONDITION = poor, error is 1 out of 4

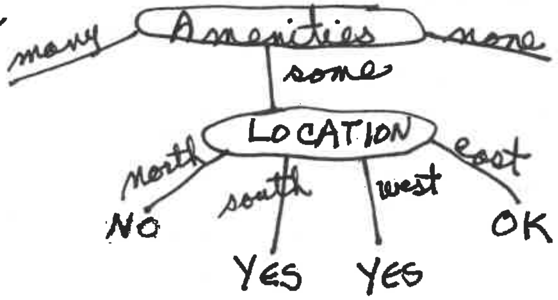
weighted error is $0/9(0) + 2/9(\frac{1}{2}) + 3/9(\frac{1}{3}) + 4/9(\frac{1}{4}) = 1/3$

If we remove the test at CONDITION, then the answer at CONDITION (according to the training data) will be YES.

The error rate for LOCATION = west will be $3/9 = 1/3$

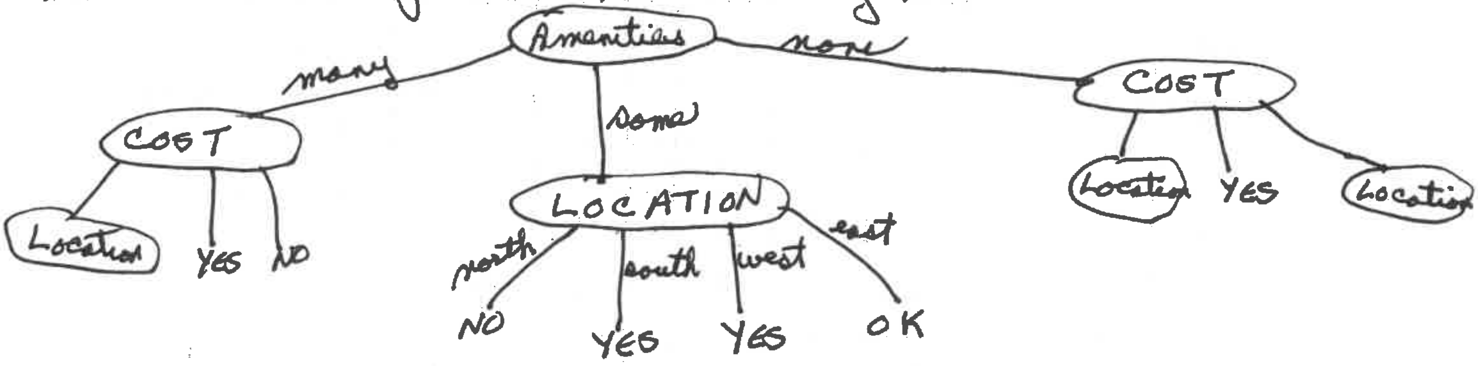
So prune.

Now with both the COST and CONDITION nodes pruned, we have



LOCATION = north, we have 1 error out of 1
LOCATION = south, we have 3 errors out of 8
LOCATION = west, we have 3 errors out of 9
LOCATION = east, we have 1 error out of 4
weighted error is $\frac{1}{22}(1)+\frac{8}{22}(3/8)+\frac{9}{22}\left(\frac{3}{9}\right)+\frac{4}{22}\left(\frac{1}{4}\right) = \frac{8}{22}$

If we remove the LOCATION node (ie., remove the test at LOCATION), according to the training data, the answer will be YES (it could also be NO since we have the same number of YES as NO).
Error rate if we prune is 9/22. (If you had selected NO as the answer, the error rate would be even worse.)
So we do not prune. The resulting tree is



b) COST = low, (Amenities = some, Location = south)
error rate on training data is $\frac{15}{31}$
COST = med, (Amenities = some, Location = south)
error rate on training data is $15/38$
COST = high, (Amenities = some, Location = south)
error rate on training data is $3/11$
Pessimistic error: use z value of .84 from table (entry for .60/2)
COST = low, pessimistic error = .5588
COST = med, pessimistic error = .4627
COST = high, pessimistic error = .3966
weighted pessimistic error = $\frac{31}{80}(.5588)+\frac{38}{80}(.4627)+\frac{11}{80}(.3966)$
$\approx .49$

If we prune the COST node, the answer will be NO since the training set has 16+15+3 yes's and 15+23+8 NO's that reach here (ie., Amenities = some, Location = south)
error rate for these instances on training data is $34/80$
pessimistic error is .472
So prune since pessimistic error if prune is smaller.

G-46 REV. 9/96

ENGINEERING DUPONT COMPUTATION SHEET

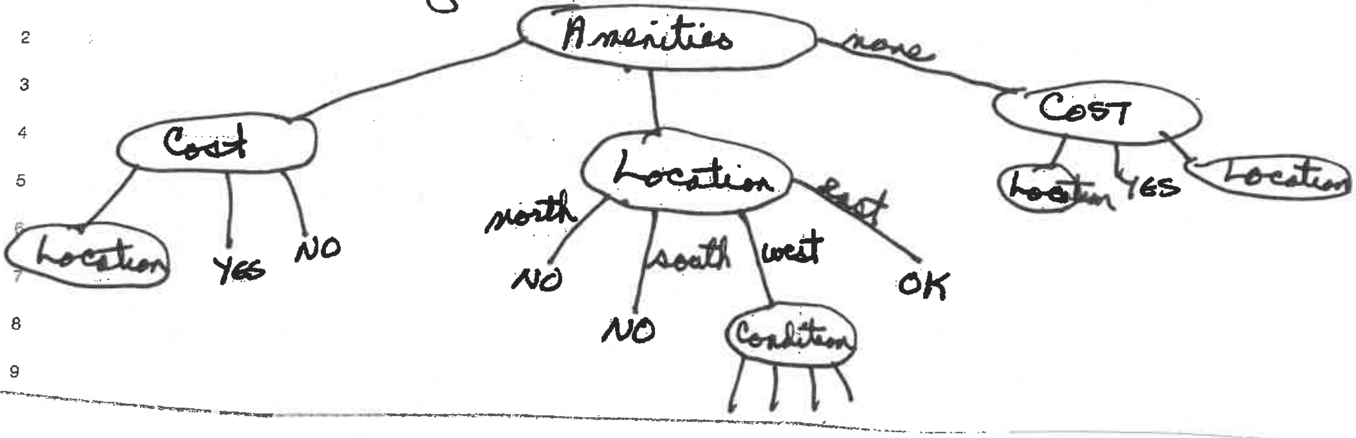
SHEET NO. 3

TITLE OF PROJ. OR STUDY ____ PROJ. OR STUDY NO. ____

SUBJECT ____ WORKS ____

COMPUTER ____ DATE ____ 20 ____

So the resulting tree is



2a) success rate = $\frac{2500}{3000} = 5/6 \approx .833$

confidence is 90% so $z = 1.645$

(you could use $z = 1.64$
or $z = 1.65$
I used the midpoint between the two)

$$\text{lower bound} = \frac{\frac{5}{6} + \frac{(1.645)^2}{6000} - 1.645\sqrt{\frac{5/6}{3000} - \frac{\left(\frac{5}{6}\right)^2}{3000} + \frac{(1.645)^2}{4(3000)^2}}}{1 + \frac{(1.645)^2}{3000}}$$

$\approx .822$

b) $$\text{upper bound} = \frac{\frac{5}{6} + \frac{(1.645)^2}{6000} + 1.645\sqrt{\frac{5/6}{3000} - \frac{\left(\frac{5}{6}\right)^2}{3000} + \frac{(1.645)^2}{4(3000)^2}}}{1 + \frac{(1.645)^2}{3000}}$$

$\approx .844$

3. unpruned tree has many more nodes and leaves than the pruned tree

the pruned tree has a higher success rate than the unpruned tree

the difference in accuracy is because the pruned tree is more general than the unpruned tree (which is more closely fitted to the training data) and thus the pruned tree is likely to be more successful on new test data