National University of Computer and Emerging Sciences, Lahore Campus

Course:	Artificial Intelligence		Course Code:	CS401
Program:	BS(Computer Science)		Semester:	Fall 2016
Duration:	20 min		Total Marks:	10
Paper Date:	21-03-17		Weight	2%
Section:			Page(s):	
Exam:	Quiz 3		Reg. No	

Instruction/Notes:

Given the training data what attribute will be chosen as root node? Using Gini as impurity measure.

Training Data

Training Data				
Color	Size	Act	Age	Class(Inflated)
PURPLE	SMALL	DIP	CHILD	F
PURPLE	SMALL	DIP	CHILD	F
PURPLE	LARGE	DIP	CHILD	F
PURPLE	LARGE	DIP	CHILD	F
YELLOW	SMALL	STRETCH	ADULT	T
YELLOW	SMALL	STRETCH	CHILD	T
YELLOW	SMALL	DIP	ADULT	T
YELLOW	LARGE	STRETCH	ADULT	T
YELLOW	LARGE	STRETCH	CHILD	T

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2, \qquad (4.4)$$

Where C is number of classes, and P(i|t) is fraction of records that belong to class i at node t.

Question 2 Classify the following test instance using kNN and training data given in above table. Use K=3

PURPLE	LARGE	STRETCH	CHILD	T

All the attributes are symmetric (in case of binary)

$$Gini(color) = 1 - \left(\frac{4}{9}\right)^2 - \left(\frac{5}{9}\right)^2 = x$$

$$Gini(C_1) = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0$$

$$Gini(C_2) = 1 - \left(\frac{0}{5}\right)^2 - \left(\frac{5}{5}\right)^2 = 0$$

$$Gain(color) = x - \frac{5}{9}(0) - \frac{4}{9}(0) = 0$$

$$Gini(S_1) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = y$$

$$Gini(S_2) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = z$$

$$Gain(Size) = x - \frac{5}{9}(y) - \frac{4}{9}(z)$$

$$Gini(A_1) = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = v$$

$$Gini(A_2) = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$$

$$Gain(Age) = x - \frac{6}{9}(v) - \frac{3}{9}(0)$$

$$Gini(AC_1) = 1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = w$$

$$Gini(AC_2) = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$$

$$Gain(Act) = x - \frac{5}{9}(w) - \frac{4}{9}(0)$$

as x y w z v all are > 0

so Gain of color is highest so it will be chosen as root node.

Training Data

Training Data

Color	Size	Act	Age	Class(Inflated)
PURPLE	SMALL	DIP	CHILD	F
PURPLE	SMALL	DIP	CHILD	F
PURPLE	LARGE	DIP	CHILD	F
PURPLE	LARGE	DIP	CHILD	F
YELLOW	SMALL	STRETCH	ADULT	T
YELLOW	SMALL	STRETCH	CHILD	T
YELLOW	SMALL	DIP	ADULT	T
YELLOW	LARGE	STRETCH	ADULT	T
YELLOW	LARGE	STRETCH	CHILD	T

Similarity × 4

2
2
3 ✓ F
3 ✓ F
1
2
0
2
3 ✓ T

$$Gini(t) = 1 - \sum^{c-1}[p(i|t)]^2. \qquad (4.4)$$

Q2

Similarity = SMC measure

with given test instance

most frequent in k nearest neighbors

so class of test instance is F