N.KAUSIK
CS21M037

Page No.
Date

Tutorial 9

1. a) $P_{NO} = 5/14 \qquad P_{Yes} = 9/14$

$$\text{Entropy} = -P_{NO}\log P_{NO} - P_{Yes}\log P_{Yes}$$

$$= 0.94$$

b) When we split with Age,

$$P_{Young} = \frac{5}{14} \qquad P_{mid} = \frac{4}{14} \qquad P_{Senior} = \frac{5}{14}$$

$$\text{Avg Entropy} = \frac{5}{14}\left(-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}\right) + \frac{4}{14}\left(-0 - \frac{4}{4}\log\frac{4}{4}\right) + \frac{5}{14}\left(-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}\right)$$

$$= 0.694$$

c) Info Gain for Age $= 0.94 - 0.694 = 0.246$

d) $IG_{Age} = 0.246$

When we split with income,

$$IG_{Income} = 0.94 - \frac{5}{14}\left(-\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4}\right) - \frac{4}{14}\left(-\frac{2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6}\right) - \frac{5}{14}\left(-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}\right)$$

$$= 0.029$$

Similarly,

$IG_{student} = 0.151$

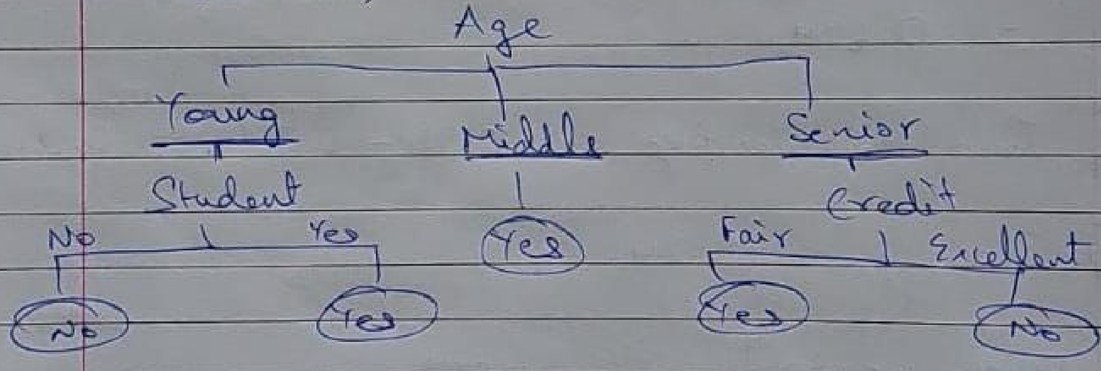
$IG_{Credit} = 0.048$

Hence max IG is Age
We will use root
node as Age.

e) For Age = Middle Aged, Buys Computer is always Yes. Hence we will not grow the tree further for it.

But for other 2, we will.
Hence # data points = 5 + 5 = 10

f) For Age = Middle, Buys Comp = Yes
For Age = Young,
If Student is Yes, Buys is Yes and if No, it is also No.
For Age = Senior,
If Credit is Fair, Buys is Yes and if Credit is Excellent, Buys is No.
∴ Tree is,



2. All are parallelisable except AdaBoost at train time.

3. 3) In given method, if some critical attributes are not taken for a tree, it gives a bad classifier. Hence overall performance reduces.

4. 2) and 4)

5. 2) Since weights are limited, outliers cant keep increasing their weights and cause the classifier to become bad. Hence Robust to Outliers

3) If limit is too less, classifier cannot become better to fit all points and sometimes even the correct points.