Liam Fruzyna

COSC 4610

Assignment 2

1) Consider the given training examples for a binary classification problem.

a) What is the entropy of the entire training data?

$-([\frac{5}{10}log_2(\frac{5}{10})] + [\frac{5}{10}log_2(\frac{5}{10})] = 1$

b) What are the information gains of a1 and a2, report which one gives a better split.

a1: $(\frac{5}{10})(-([\frac{3}{5}log_2(\frac{3}{5})] + [\frac{2}{5}log_2(\frac{2}{5})])) + (\frac{5}{10})(-([\frac{2}{5}log_2(\frac{2}{5})] + [\frac{3}{5}log_2(\frac{3}{5})])) = \frac{1}{2}0.971 + \frac{1}{2}0.971 = 0.971$

a2: $(\frac{5}{10})(-([\frac{2}{5}log_2(\frac{2}{5})] + [\frac{3}{5}log_2(\frac{3}{5})])) + (\frac{5}{10})(-([\frac{3}{5}log_2(\frac{3}{5})] + [\frac{2}{5}log_2(\frac{2}{5})])) = \frac{1}{2}0.971 + \frac{1}{2}0.971 = 0.971$

Both a1 and a2 have the same information gain (0.971) so neither gives a better split.

c) Compute the Gini index for a1 and a2, report which one gives a better split.

a1: $1 - (\frac{5}{10}^2 + \frac{5}{10}^2) = 1 - \frac{1}{2} = \frac{1}{2}$

a2: $1 - (\frac{5}{10}^2 + \frac{5}{10}^2) = 1 - \frac{1}{2} = \frac{1}{2}$

Both a1 and a2 have the same gini index (1/2) so neither gives a better split.

d) Compute the information gain for a3 for each potential split point, report which split point is the best for a3.

Split between 1 and 3

$(\frac{1}{10})(-([\frac{1}{1}log_2\frac{1}{1})] + [\frac{0}{1}log_2(\frac{0}{1})])) + (\frac{9}{10})(-([\frac{4}{9}log_2(\frac{4}{9})] + [\frac{5}{9}log_2(\frac{5}{9})])) = 0.892$

Split between 3 and 4

$(\frac{2}{10})(-([\frac{2}{2}log_2(\frac{2}{2})] + [\frac{0}{2}log_2(\frac{0}{2})])) + (\frac{8}{10})(-([\frac{3}{8}log_2(\frac{3}{8})] + [\frac{5}{8}log_2(\frac{5}{8})])) = 0.764$

Split between 4 and 5

$(\frac{4}{10})(-([\frac{2}{4}log_2(\frac{2}{4})] + [\frac{2}{4}log_2(\frac{2}{4})])) + (\frac{6}{10})(-([\frac{3}{6}log_2(\frac{3}{6})] + [\frac{3}{6}log_2(\frac{3}{6})])) = 1.000$

Split between 5 and 7

$(\frac{5}{10})(-([\frac{3}{5}log_2(\frac{3}{5})] + [\frac{2}{5}log_2(\frac{2}{5})])) + (\frac{5}{10})(-([\frac{2}{5}log_2(\frac{2}{5})] + [\frac{3}{5}log_2(\frac{3}{5})])) = 0.971$

Split between 7 and 8

$(\frac{6}{10})(-([\frac{4}{6}log_2(\frac{4}{6})] + [\frac{2}{6}log_2(\frac{2}{6})])) + (\frac{4}{10})(-([\frac{1}{5}log_2(\frac{1}{5})] + [\frac{3}{5}log_2(\frac{3}{5})])) = 0.875$

Split between 8 and 10

$(\frac{8}{10})(-([\frac{4}{8}log_2(\frac{4}{8})] + [\frac{4}{8}log_2(\frac{4}{8})])) + (\frac{2}{10})(-([\frac{1}{2}log_2(\frac{1}{2})] + [\frac{1}{2}log_2(\frac{1}{2})])) = 1.000$

Split between 10 and 12

$$(\frac{9}{10})(-([\frac{4}{9}log_2(\frac{4}{9})] + [\frac{5}{9}log_2(\frac{5}{9})])) + (\frac{1}{10})(-([\frac{1}{1}log_2(\frac{1}{1})] + [\frac{0}{1}log_2(\frac{0}{1})])) = 0.892$$

The best split point would be between 3 and 4 because it has the lowest information gain.

2) Implement a decision-tree algorithm for breast cancer diagnosis with a given data set.

a) If you found any missing values, how did you deal with them in your code?

There were missing values in the 'Bare Nuclei' column. I dealt with them by computing the average of the existing values and rounding down because values in the columns are integers between 1 and 10.

b) Show the accuracy of the classifier with 2 different measures, entropy and gini index.

After a few tests the gini measured decision tree tended to result in a slightly lower accuracy score. For example the entropy was 0.931 while the gini was 0.926.