Course: Machine Learning Graded
Week 9: Graded assignment

1 Lecture 9

Consider the following dataset and answer the questions 1, 2, 3 and 4.

feature 1	feature 2	label
1	0	1
1	1	1
0	1	0
0	0	0

1. What feature will we split on at the root of our decision tree, using the Gini index measure?

   Answer: 1

   Solution:
   Gini index of the dataset will be

$$\begin{aligned} G_S &= \sum_{i=0}^{1} p_i(1-p_i) \\ &= (0.5)(0.5) + (0.5)(0.5) \\ &= 0.5 \end{aligned}$$

   If we split the data-set as per feature 1, there will be two child nodes: node0 and node1. Both nodes will be pure nodes. So, gini index of the split will be zero.

   If we split the data-set as per feature 2, there will be two child nodes: node0 and node1. Both nodes will not be pure nodes, hence will have some value for gini index (will be 0.5).

   Therefore, we split the data-set as per feature 1.

2. What will be the information gain from splitting on that feature using the Gini index measure?

   Answer: 0.5

   As per question 1, $G_i = 0.5$
   After the split, value of gini-index will be zero.
   Therefore, information gain will be 0.5.

3. What will be the cross entropy of the dataset?

   Answer: 1

   Solution:
   Gini index of the dataset will be

$$\begin{aligned} H(S) &= -\sum_{i=0}^{1} p_i \log_2(p_i) \\ &= -(0.5)\log_2(0.5) - (0.5)\log_2(0.5) \\ &= 1 \end{aligned}$$

4. What will be the information gain from splitting on that feature using the cross entropy measure?

   Answer: 1

   If we split the data-set according to feature 1, we will have pure nodes and therefore, entropy will be zero of the split.
   Therefore, value of information gain will be $1 - 0 = 0$.

5. How can we choose the right node while constructing a decision tree?

   A. A feature having high entropy.

   B. A feature having high entropy and information gain.

   C. A feature having lowest information gain.

   D. A feature having the high information gain.

   Answer: D

6. Which of the following impurity measure can be used in classification decision tree?

   A. sum of squared error

   B. Gini index

   C. Entropy

   D. Mis-classification error

   Answer: B, C, D

7. If samples of a node are equally distributed in two classes, what will be the value of gini index of that node?

   Answer: 0.5

   Solution:
   If samples of a node are equally distributed in two classes, then proportion of samples

in both classes will be 0.5 and Gini index will be given by

$$\begin{aligned} G_S &= \sum_{i=0}^{1} p_i(1-p_i) \\ &= (0.5)(0.5) + (0.5)(0.5) \\ &= 0.5 \end{aligned}$$

8. Missing values must to be filled using some strategy before applying decision tree algorithm.

   A. True

   B. False

   Answer: B

   Solution:
   Decision tree requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed.
   Therefore, missing values must **not** to be filled using some strategy before applying decision tree algorithm.

Consider the following dataset and answer the questions 9 and 10.

feature 1	feature 2	label
11	182	a
13	190	a
15	192	a
16	178	a
18	160	b
14	120	b
21	150	b

9. Which among the following split-points for feature 1 would give the best split according to the gini index measure?

   A. 13.5

   B. 15.3

   C. 16.8

   D. 18.2

   Answer: C

   Solution:
   We will find the value of gini index for all four options.

option 1:
If we split the data as per feature 1 and split value 13.5, then the child node (say node1 and node2), then data in both nodes will be as follow:
node1

feature 1	feature 2	label
11	182	a
13	190	a

$$G_{node1} = 0$$

node2

feature 1	feature 2	label
15	192	a
16	178	a
18	160	b
14	120	b
21	150	b

$$\begin{aligned} G_{node2} &= \frac{2}{5}\frac{3}{5} + \frac{2}{5}\frac{3}{5} \\ &= 0.48 \end{aligned}$$

Gini-index of the split will be

$$\begin{aligned} G_{13.5} &= \frac{2}{7}(0) + \frac{5}{7}(0.48) \\ &= 0.34 \end{aligned}$$

option 2:
If we split the data as per feature 1 and split value 15.3, then the child node (say node1 and node2), then data in both nodes will be as follow:
node1

feature 1	feature 2	label
11	182	a
13	190	a
15	192	a
14	120	b

$$G_{node1} = \frac{3}{4}\frac{1}{4} + \frac{3}{4}\frac{1}{4}$$
$$= 0.375$$

node2

feature 1	feature 2	label
16	178	a
18	160	b
21	150	b

$$G_{node2} = \frac{1}{3}\frac{2}{3} + \frac{1}{3}\frac{2}{3}$$
$$= 0.44$$

Gini-index of the split will be

$$G_{15.3} = \frac{4}{7}(0.375) + \frac{3}{7}(0.44)$$
$$= 0.214 + 0.19 = 0.40$$

option 3:
If we split the data as per feature 1 and split value 16.8, then the child node (say node1 and node2), then data in both nodes will be as follow:
node1

feature 1	feature 2	label
11	182	a
13	190	a
15	192	a
16	178	a
14	120	b

$$G_{node1} = \frac{4}{5}\frac{1}{5} + \frac{4}{5}\frac{1}{5}$$
$$= 0.32$$

node2

feature 1	feature 2	label
18	160	b
21	150	b

$$G_{node2} = 0$$

Gini-index of the split will be

$$\begin{aligned} G_{16.8} &= \frac{5}{7}(0.32) + \frac{2}{7}(0) \\ &= 0.214 + 0.19 = 0.22 \end{aligned}$$

option 4:
If we split the data as per feature 1 and split value 18.2, then the child node (say node1 and node2), then data in both nodes will be as follow:
node1

feature 1	feature 2	label
21	150	b

$$G_{node1} = 0$$

node2

feature 1	feature 2	label
11	182	a
13	190	a
15	192	a
16	178	a
18	160	b
14	120	b

$$\begin{aligned} G_{node2} &= \frac{4}{6}\frac{2}{6} + \frac{4}{6}\frac{2}{6} \\ &= 0.44 \end{aligned}$$

Gini-index of the split will be

$$\begin{aligned} G_{18.2} &= \frac{6}{7}(0.44) + \frac{1}{7}(0) \\ &= 0.38 \end{aligned}$$

Minimum gini index is for the split point 16.8, hence we will split the data as per this split point.

10. What will be the information gain from splitting on that feature 2 using the gini index measure? Write your answer correct to two decimal places.
Solution:
If we split the dataset using feature 2, split point will be 160 as this point will end up splitting pure nodes.

And gini index of split will be zero.
gini index of the dataset is given by

$$\begin{aligned} G_S &= \sum_{i=0}^{1} p_i(1 - p_i) \\ &= \frac{4}{7}\frac{3}{7} + \frac{4}{7}\frac{3}{7} \\ &= 0.49 \end{aligned}$$