Decision Trees

Q2. Entropy, Gini Index and Decision Trees

a) - what is the formula for computing entropy?

- Among the probability distributions p and p' specified below, which one has a higher entropy?
- Explain why, without computing the entropies of them.

$p = (0.33, 0.33, 0.34)$ $p' = (0.8, 0.1, 0.1)$

let D is the

$H(D) = -(P_0 \log_2 P_0 + P_1 \log_2 P_1)$ Entropy $= -\sum_{i=0}^{A} P_i \log_2 P_i$

The entropy for p is high since the probability distribution for all are same.

b) What is the Gini Index of the probability distribution $(\frac{2}{3}, \frac{1}{3})$?

$$G = 1 - \sum_{i=1}^{n} P_i^2$$

$\therefore G = 1 - \left[\frac{2}{3} \times \frac{2}{3} + \frac{1}{3} \times \frac{1}{3}\right] = 1 - [0.44 + 0.111] = 0.445.$

c) Consider the following set of datapoints, each containing 3 entries of the format (X_1, X_2, Y),

i.e. in each datapoint, the first 2 entries are attributes, and the last entry is its label. The labels can be either "T" or "F".

$D = \{(1,4,T), (2,3,T), (3,0,T), (6,2,F), (7,1,F), (8,1,F)\}$

Now consider the following two possible splits:

- Split 1: $X_1 \leq 4$ and $X_1 > 4$
- Split 2: $X_2 \leq 2.5$ and $X_2 > 2.5$

For Split 2, which datapoints go to the left (under $X_2 \leq 2.5$), and which datapoints go to the right (under $X_2 > 2.5$)?

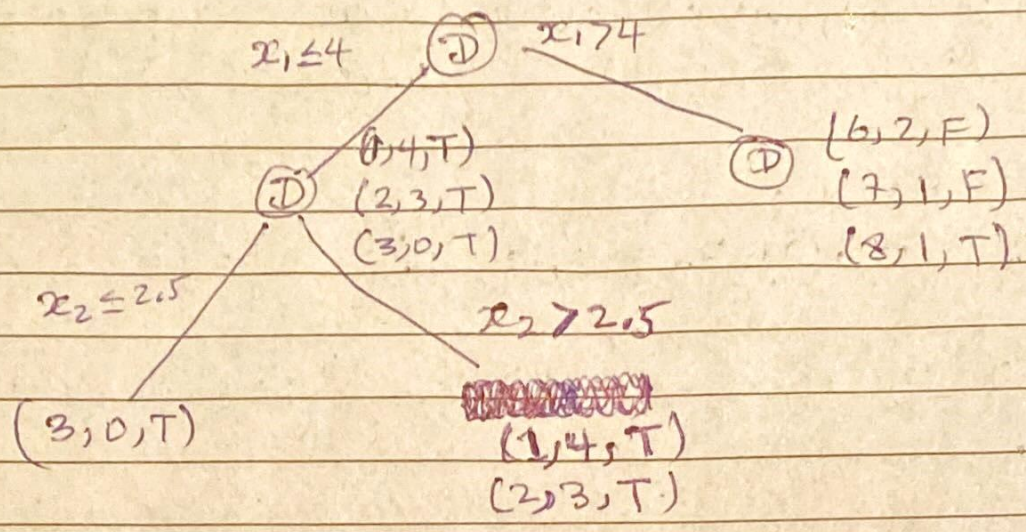
(x_1, x_2, Y)

Given datapoint

$D = \{(1,4,T), (2,3,T), (3,0,T), (6,2,F), (7,1,F), (8,1,F)\}$

split 1 : $x_1 \leq 4$ & $x_1 > 4$

split 2 : $x_2 \leq 2.5$ & $x_2 > 2.5$



For split 2 datapoints which go under $x_2 \leq 2.5$ are (3,0,T) and datapoint which are under $x_2 > 2.5$ are (1,4,T) & (2,3,T).

d) You are going to generate a decision tree for the dataset D specified in part (c).

You want to choose a split into regions R_1 and R_2 which has a small value of

$|R_1| \cdot$ (Gini index of region R_1) + $|R_2| \cdot$ (Gini index of region R_2)

Among split 1 and split 2, which one is better? Justify your answer.

Gini index for R_1

$x \leq 4 \Rightarrow$ 3T and 0F

$1 - \left[\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right] = 0$

$x > 4 \Rightarrow$ 1T and 2F

$1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right]$

$1 - 0.555 = 0.445$

weighted entropy $= \left(\frac{3}{6} \times 0 + \frac{3}{6} \times 0.445\right) = \frac{0 + 0.4445}{2} = 0.225$

Exam paper: CS5100J : Data Analysis

Decision Trees

Q2. d) For Region 2

$x \leq 2.5 \Rightarrow$ 1T & 0F $\quad$ $x > 2.5 \Rightarrow$ 2T and 0F

$= 1 - [(1)^2 + (0)^2] = 0$ $\quad$ $1 - [(1)^2 + (0)^2] = 0$

weighted $= \frac{1}{3} + 0 + \frac{2}{3} + 0 = 0$

Split 1 classifies the difference same as of split 2 through entropy. split 1 is more better than split 2.

e) For any decision tree T of a classification problem, let its "penalized Gini" be

$$\alpha|T| + \sum_{R} (\text{Gini index of region } R)$$

where,

- $\alpha \geq 0$
- $|T|$ is the number of leaves of the tree, and
- R runs over all the regions/leaves of the tree.

What technique are we using via the introduction of the term $\alpha|T|$?

Explain the motivation for introducing this term. (5 sentences)

We need to get better α and prune for best setoff

Technique used for $\alpha|T|$ is penalise RSS and Pruning Trees.

$$\alpha|T| + \sum_{m=1}^{|T|} RSS(R_m)$$

- $|T|$ = number of leaves $\quad \alpha = 0$, penalised RSS = 0

$y = \alpha + RSS$ • we get critical α and penalised RSS

Q1.d) In a leaf of a decision tree

Exam paper: CS5100R Data Analysis

Decision Tree)

Q2. Prediction (Decision tree).

b) Compute the probability to assign the test object to each of two classes according to a Decision Tree model - based on the partition of the feature space defined by

(1) $R_1 = \{x \text{ such that } x_2 \leq \xi\}$, $R_2 = \{x \text{ such that } x_2 > \xi\}$, $\xi = 7$

and training data set D. Justify all steps in your cal (Max 4 lines) the test object (4,4)

(x_1, x_2)	Y
• (3, 8)	1
- (5, 6)	0
- (5, 5)	0
- (4, 0)	0
• (8, 9)	0
- (9, 6)	0
- (1, 2)	1
(7, 0)	0
- (9, 5)	0
- (6, 2)	0

$R_1 \Rightarrow x_2 \leq 7$ $\quad x_2 \leq 7$ $\quad x_2 > 7$

R_1 $\quad$ R_2

$R_1 \Rightarrow$ (

$R_2 \Rightarrow (3,8), (8,9)$

c) The Gini Impurity of a region R is defined as

$$G(R) = \sum_{i=0}^{1} P_i (1 - P_i) = 1 - \sum_{i=0}^{1} P_i^2$$

Where; • $P_i = Z_R^{-1} \sum_{(x,y)} 1_{y=i} 1_{x \in R}$,

• $Z_R = \sum_{(x,y)} 1_{x \in R}$,

i.e. P_i is the probability of an object to have label $y=i$, given it's belongs to region R.

Compute the information gain associated with (1)

i.e.

$$Z_{R_1 \cup R_2} G(R_1 \cup R_2) - (Z_{R_1} G(R_1) + Z_{R_2} G(R_2))$$

Where, R_1, R_2, and Z_R are defined as above.

Exam paper: CS5100 - Final ; Data Analysis

Q1. Decision Trees

- Decision trees are grown by splitting the feature space in increasingly pure regions.
- There exist several methods to measure the impurity of a set of labelled observations, e.g. $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$ by looking at the object labels, y_n.
- Two examples of impurity measures are
  - the Gini index and
  - the entropy.

a) Consider the case where $y_n \in \{0, 1\}$ and $x_n \in \{i\}_{i=1}^{10}$ for all $n = 1, \ldots, N$ and

- the task

Decision Trees

- Entropy - Measures the purity of split.
  - rangs from 1 to 0, [0, 1]
  - when, E=0, pure split (pure subset)
  - when, E=1, worest split (impure subset)

Formula: $Entropy = -\sum_{i=0}^{n} P_i \log_2 P_i$

- Information Gain

$$IG(S, f) = H(S) - \sum_{V \in Val} \frac{|S_V|}{|S|} H(S_V)$$

where • S_V the subset after splitting • S the subset
- $H(S_V)$ Entropy after splitting
- $H(S)$ Entropy

- The higher the IG the better

- Gini Index (Impurity)

$$GI = 1 - \sum_{i=1}^{n} P_i^2$$

- ranges between 0.5 and 0.

Q2. a) Entropy $= -\sum_{i=0}^{n} P_i \log_2 P_i$

The p has higher entropy since the probability distributions for all are the same.

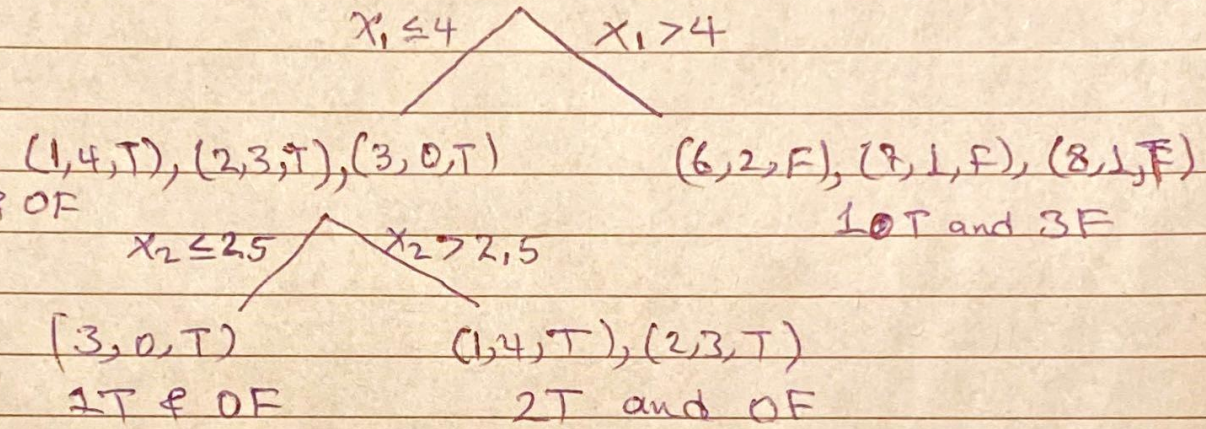
b) $(\frac{2}{3}, \frac{1}{3})$ Gini Index $= 1 - \sum_{i=1}^{n} P_i^2$

$\therefore GI = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 1 - \left[\left(\frac{4}{9} + \frac{1}{9}\right)\right] = 1 - \frac{5}{9} = 0.444$

c) $D = \{(1,4,T), (2,3,T), (3,0,T), (6,2,F), (7,1,F), (8,1,F)\}$

Split 1: $X_1 \leq 4$ & $X_1 > 4$ format (X_1, X_2, Y)

split 2: $X_2 \leq 2.5$ & $X_2 > 2.5$



d) Gini Index $= 1 - \sum_{i=1}^{n} P_i^2$

- For R_1, when $X_1 \leq 4$, 3T & 0F and when $X_1 > 4$, 1T & 3F

$GI_{X_1 \leq 4} = 1 - \left[\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right] = 0$ $GI_{X_1 > 4} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 1 - 1 = 0$

$= 1 - 1 = 0$ $= 1 - \frac{5}{9} = 0.444$

Weighted entropy $= |R_1| \cdot (GI \text{ of } R_1)$

$= \left(\frac{3}{6} \times 0 + \frac{3}{6} \times 0.444\right) = \frac{0 + 0.444}{2} = 0.225$

- For R_2, when $X_2 \leq 2.5$, 1T & 0F and when $X_2 > 2.5$ 2T & 0F

$GI_{X_2 \leq 2.5} = 1 - \left[\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right] = 0$ $GI_{X_2 > 2.5} = 1 - \left[\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2\right]$

$= 1 - 1 = 0$

Weighted entropy $= |R_2| \cdot (GI \text{ of } R_2) = \frac{1}{3} \times 0 + \frac{2}{3} \times 0 = 0$

$\therefore$ split 1 is better