Final

Lecture - 1
Math

20-06-2022
Monday

FP Growth Algorithm

① Works ~~with~~ through divide and conquer approach.

$\boxed{\text{Min sup} = 3}$ ✓

Header Table **Step-1**

{f, a, c, d, g, i, m, p, b, l, o, h, j, w, k, s, e, n

Item	Frequency
f	4
a	3
c	4
d	1
g	1
i	1
m	3
p	3
b	3
l	2
o	2
h	1
j	1
w	1
k	1
s	1
[illegible]	[illegible]
e	1
h	1

→

L

Item	Frequency
f	4
c	4
a	3
b	3
m	3
p	3

↑ Frequent pattern

(Sorted in descending order) ✓

* $\boxed{\text{f, c, a, b, m, p}}$ (order) for next step.

f, c, a, b, m, p

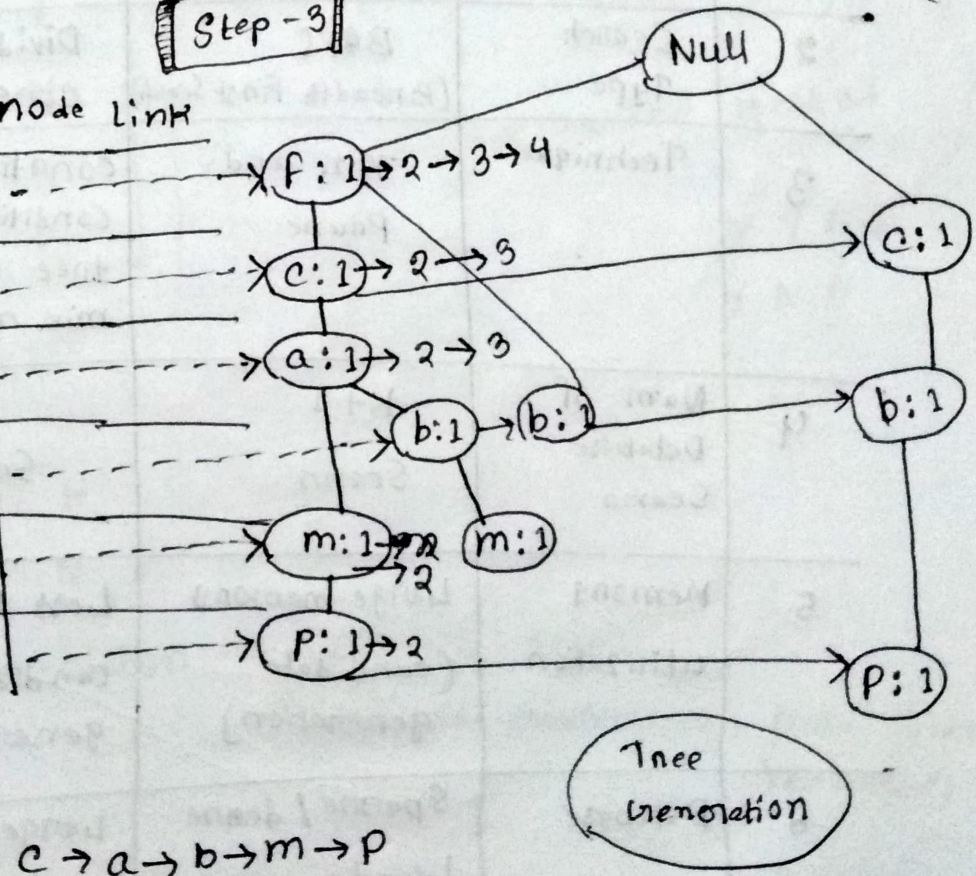
Step - 2

ID	Item	Ordered Frequent items	Item	Conditional Pattern base
T_1	f, a, c, d, g, i, m, p	f, c, a, m, p	p	{f, c, a, m : 2}, {c, b : 1}
T_2	a, b, c, f, l, m, o	f, c, a, b, m	m	{f, c, a : 2}, {f, c, a, b : 1}
T_3	b, f, h, j, o, w	f, b	b	{f, c, a : 1}, {f : 1}, {c : 1}
T_4	b, k, c, s, p	c, b, p	a	{f, c : 3}
T_5	a, f, c, e, l, p, m, n	f, c, a, m, p	c	{f : 3}

Don't need to write it for (exam).

Step - 3

ID	S.C	Node Link
f	4	
c	4	
a	3	
b	3	
m	3	
p	3	



f → c → a → b → m → p

Ans:

Decision Tree

Ex: 8.1 (Dataset - Table 8.1) Book page - 338

Soln: Formula: $\text{Entrophy}(t) = -\sum_{j} P(j|t) \log_2 P(j|t)$

↳ percentage of distribution

Decision column → class: buys_computer $\quad S = [9+, 5-]$

$$\left.\begin{array}{l} \text{Yes} - 9 \\ \text{No} - 5 \end{array}\right\} 14$$

Information gain

$$\text{Info}(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$= -0.64 \times (-0.64) - 0.36 \times (-1.48)$$

$$= 0.94 \text{ bits}$$

Main Entropy

Now, calculating the expected information Fte requirment for each attributes:

For "age":

Youth = 5

middle_aged = 4

Senior = 5

$$\text{Info}_{age}(D) = \frac{5}{14} \times \left(-\frac{2}{5}\right) \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5}) +$$

$$\frac{4}{14} \times \left(-\frac{4}{4} \log_2 \frac{4}{4}\right) + \frac{5}{14} \left(-\frac{3}{5} \log_2\right.$$

$$\left.\frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5}\right)$$

$$= 0.36 \times (0.053 + 0.44) + 0.28 \times 0 +$$

$$0.36 \times (0.44 + 0.53)$$

$$= 0.69 \text{ bits}$$

~~Now~~ Now, $\text{Gain}(age) = \text{Info}(D) - \text{Info}_{age}(D)$

$$= 0.94 - 0.69$$

$$= \boxed{0.25 \text{ bits}}$$ (overall information gain with respect to age)

For "Income":

High = 4, medium = 6, low = 4

$$\text{Info}_{Income}(D) = \frac{4}{14} \times \left(-\frac{2}{4} \log_2 \frac{2}{4}\right) - \frac{2}{4} \log_2 \frac{2}{4}) +$$

$$\frac{6}{14} \times \left(-\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6}\right) +$$

$$\frac{4}{14} \times \left(-\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4}\right)$$

$$= 0.28 \times (0.5 + 0.5) + 0.43 (0.39 + 0.53)$$

$$+ 0.28 (0.31 + 0.5)$$

$$= 0.902 \text{bits}$$

Now, $\text{Gain}(age) = \text{Info}(D) - \text{info}_{Income}(D)$

$$= 0.94 - 0.902$$

$$= \cancel{0.038}$$

$$= 0.03 \text{ bits}$$

For "Student":

yes = 7, no = 7

$$Info_{student}(D) = \frac{7}{14} \times \left(-\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}\right) +$$

$$\frac{7}{14} \times \left(-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}\right)$$

$$= 0.5(0.19 + 0.40) + 0.5(0.52 + 0.46)$$

$$= 0.785 \text{ bits}$$

Now, $Gain(student) = Info(D) - Info_{student}(D)$

$$= 0.94 - 0.785$$

$$= 0.15 \text{ bits}$$

For "credit-rating":

fair = 8, excellent = 6

$$Info_{credit-rating}(D) = \frac{8}{14}\left(-\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}\right) +$$

$$\frac{6}{14}\left(-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right)$$

$$= 0.57(0.31 + 0.5) + 0.43(0.5 + 0.5)$$

$$= 0.89 \text{ bits}$$

Now, $Gain(credit\text{-}rating) = Info(D) - Info_{credit\text{-}\cancel{card}\ rating}(D)$

$$= 0.94 - 0.89$$

$$= 0.05 \text{ bits}$$

We can see "age" has the highest information gain among the attributes, it is selected as the splitting attribute.

Decision Tree

age

youth

income	student	credit_rating	class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

senior

income	student	credit_rating	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

middle-aged

income	student	credit_rating	class
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

Gain ration

Formula: 1) $\text{Split Info}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$

2) $\text{Gain Ratio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}_A(D)}$

$\boxed{\text{Ex: 8.2}}$ computation of gain ratio for the attribute "income".

$$\text{Split Info}_{income}(D) = -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right)$$

$$= 0.52 + 0.52 + 0.52$$

$$= 1.56$$

From "ex: 8.1" we found Gain(income) = 0.03.

Therefore, $\text{GainRatio(income)} = \frac{0.03}{1.56} = 0.019$ Ans:

For "age":

$$\text{SplitInfo}_{\cancel{income}\ age}(D) = -\frac{5}{14} \times \log_2 \frac{5}{14} - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{5}{14} \times \log_2 \frac{5}{14}$$

$$= 0.53 + 0.52 + 0.53$$

$$= 1.58$$

From "ex: 8.1" we found Gain (age) = 0.25

Therefore, $\text{GainRatio(age)} = \frac{0.25}{1.58}$

$$= 0.16$$

For "Student":

$$\text{Split Info}_{Student}(D) = -\frac{7}{14} \times \log_2\left(\frac{7}{14}\right) - \frac{7}{14} \times \log_2\left(\frac{7}{14}\right)$$

$$= 0.5 + 0.5$$

$$= 1$$

Previously we found ~~gain~~ Gain(student) = 0.15

Therefore, $\text{GainRatio}(student) = \frac{0.15}{1} = 0.15$

For "credit_rating":

$$\text{Split Info}_{credit\_rating}(D) = -\frac{8}{14} \times \log_2\left(\frac{8}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right)$$

$$= 0.46 + 0.52$$

$$= 0.98$$

Previously we found Gain(credit_rating) = 0.05.

Therefore, $\text{GainRatio}(credit\_rating) = \frac{0.05}{0.98}$

$$= 0.05$$

Ans.

Gini Index

Formulas:

1) $Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$

2) $Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$

3) $\Delta Gini(A) = Gini(D) - Gini_A(D)$

$\boxed{\text{Ex: 8.3}}$

Impurity of D:

$$Gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2$$
$$= 0.459$$

Calculating Gini Index for each attribute.

For "income":

Considering each of the possible splitting subsets. ~~consi~~
Consider the subset {low, medium}. This would result in 10 tuples in partition D_1 satisfying the condition "income" ∈ {low, medium}." The remaining four tuples of D would be assigned to partition D_2.

The Gini Index value computed based on this partitioning is:

$Gini_{income \in \{low, medium\}}(D)$

$$= \frac{10}{14} Gini(D_1) + \frac{4}{14} Gini(D_2)$$

$$= \frac{10}{14}\left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right)$$

$$= 0.443$$

$$= Gini_{income \in \{high\}}(D)$$

Ans:

Subset $\{low, high\}$:

$Gini_{income \{low, high\}}(D)$

$= \frac{8}{14} Gini(D_1) + \frac{6}{14} Gini(D_2)$

~~$= \frac{8}{14}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{3}{4}\right)^2\right) +$~~

~~$\frac{6}{14}(1 -$~~

$= \frac{8}{14}\left(1 - \left(\frac{5}{8}\right)^2 - \left(\frac{3}{8}\right)^2\right) + \frac{6}{14}\left(1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2\right)$

$= 0.4588$

$= Gini_{income \{medium\}}(D)$

$\{low, medium\}$ ✓
$\{low, high\}$ ✓
~~$\{high, medium\}$~~
$\{medium, high\}$

Subset $\{medium, high\}$:

$Gini_{income \{medium, high\}}(D)$

$= \frac{10}{14} Gini(D_1) + \frac{4}{14} Gini(D_2)$

$= \frac{10}{14}\left(1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2\right)$

$= 0.450$

$= Gini_{income \{low\}}(D)$

From the calculation we can see the best binary split for attribute "income" is on $\{low, medium\}$ (or $\{high\}$). Because it minimizes the Gini index.

For "age":

{youth, middle-age}
{youth, senior}
{middle-age, senior}

$\left.\begin{array}{l}\{youth, middle\text{-}age\} \\ \{youth, senior\} \\ \{middle\text{-}age, senior\}\end{array}\right\} \rightarrow$ Subsets

Subset {~~yot~~ youth, middle-age}

$Gini_{age} \in \{youth, middle\text{-}age\} (D)$

$= \frac{\cancel{10}\,9}{14} Gini(D_1) + \frac{5}{14} Gini(D_2)$

$= \frac{9}{14}\left(1-\left(\frac{6}{9}\right)^2-\left(\frac{3}{9}\right)^2\right) + \frac{5}{14}\left(1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2\right)$

$= 0.457$

$= \cancel{0.297}$

$= Gini_{age} \{senior\} (D)$

$= \cancel{0.367}$

Subset {youth, senior}

$Gini_{age} \in \{youth, senior\} (D)$

$= \frac{10}{14} Gini(D_1) + \frac{4}{14} Gini(D_2)$

$= \frac{10}{14}\left(1-\left(\frac{5}{10}\right)^2-\left(\frac{5}{10}\right)^2\right) + \frac{4}{14}\left(1-\left(\frac{4}{4}\right)^2\right)$

$= 0.357$

$= Gini_{age} \in \{middle\text{-}age\} (D)$

Subset {middle age, senior}

$Gini_{age} \in \{\text{middle age, senior}\}(D)$

$= \frac{9}{14} Gini(D_1) + \frac{5}{14} Gini(D_2)$

$= \cancel{\frac{9}{14}\left(1-\left(\frac{4}{4}\right)^2\right) + \frac{5}{14}\left(1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2\right)}$

$= \frac{9}{14}\left(1-\left(\frac{7}{9}\right)^2-\left(\frac{2}{9}\right)^2\right) + \frac{5}{14}\left(1-\left(\frac{2}{5}\right)^2-\left(\frac{3}{5}\right)^2\right)$

$= 0.3936$

$= Gini_{age} \in \{\text{youth}\}(D)$

Let's consider "Student":

binary attribute

$Gini_{student}(D) = \frac{7}{14}\left(1-\left(\frac{6}{7}\right)^2-\left(\frac{1}{7}\right)^2\right) + \frac{7}{14}\left(1-\left(\frac{3}{7}\right)^2-\left(\frac{4}{7}\right)^2\right)$

$= 0.3673$

Similarly, $gini_{credit\ rating}(D) = 0.4285$

Reduction of impurity:

~~age → 0.459 − 0.357 = 0.1019~~

~~income → 0.458 − 0.443~~

~~age → 0.457 − 0.357 = 0.1~~

~~income → 0.458 − 0.443 = 0.02~~

~~student →~~

age → 0.459 − 0.357
= 0.102

income → 0.459 − 0.443
= 0.016

student → 0.459 − 0.367
= 0.092

credit rating →
0.459 − 0.428
= 0.031

Ans

K-means clustering

↱ Division - Cluster
- Supervised Algo - Divide data into different category.
- Unsupervised Algo - Divide data into different category.
↳ Division - classification / class
- But there is functional difference → working procedure difference.

Example:

Q: Num of obj = 6. Num of clusters = 2

NO	X	Y
1	1	1
2	2	3
3	1	2
4	3	3
5	2	2
6	3	1

} choosen points

Soln:

Step-1: choose random k points and set as cluster centers.

$c_1 = (2,2)$ $\quad$ $c_2 = (3,3)$

Step-2: calculating the distance between ~~objects~~ Objects into cluster centroids by using Euclidean Distance.

$$D_i = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$\boxed{C_1:}$ For $(2,2)$ w

1. $D_1 = \{(1,1),(2,2)\}$
$= \sqrt{(2-1)^2+(2-1)^2}$
$= \sqrt{1+1}$
$= 1.41$

2. $D_1 = \{(2,3),(2,2)\}$
$= \sqrt{(2-2)^2+(2-3)^2}$
$= \sqrt{0+1}$
$= 1$

3. $D_1 = \{(1,2),(2,2)\}$
$= \sqrt{(2-1)^2+(2-2)^2}$
$= \sqrt{1+0}$
$= 1$

4. $D_1 = \{(3,3),(2,2)\}$
$= \sqrt{(2-3)^2+(2-3)^2}$
$= \sqrt{1+1}$
$= 1.41$

5. $D_1 = \{(2,2),(2,2)\}$
$= \sqrt{(2-2)^2+(2-2)^2}$
$= 0$

6. $D_1 = \{(3,1),(2,2)\}$
$= \sqrt{(2-3)^2+(2-1)^2}$
$= 1.41$

$\boxed{C_2:}$ For $(3,3)$ w

1. $D_2 = \{(1,1),(3,3)\}$
$= \sqrt{(3-1)^2+(3-1)^2}$
$= \sqrt{8} = 2.82$

2. $D_2 = \{(2,3),(3,3)\}$
$= \sqrt{(3-2)^2+(3-3)^2}$
$= 1$

3. $D_2 = \{(1,2),(3,3)\}$
$= \sqrt{(3-1)^2+(3-2)^2}$
$= \sqrt{5} = 2.23$

4. $D_2 = \{(3,3),(3,3)\}$
$= \sqrt{(3-3)^2+(3-3)^2}$
$= 0$

5. $D_2 = \{(2,2),(3,3)\}$
$= \sqrt{(3-2)^2+(3-2)^2}$
$= 1.41$

6. $D_2 = \{(3,1),(3,3)\}$
$= \sqrt{(3-3)^2+(3-1)^2}$
$= 2$

$C_1 = \{(1,1),(1,2),(2,2),(3,1)\}$

$C_2 = \{(2,3),(3,3)\}$

Note: For same value we can take as per our wish.
$(2,2)$

Step-3: Recalculating the position of the centroid

$$\text{Mean} = \left(\frac{x_1 + x_2 + \cdots + x_n}{n}, \frac{y_1 + y_2 + \cdots + y_n}{n} \right)$$

$$C_1 = \left(\frac{1+1+2+3}{4}, \frac{1+2+2+1}{4} \right)$$

$$\text{New } C_1 = (1.75, 1.5)$$

$$C_2 = \left(\frac{2+3}{2}, \frac{3+3}{2} \right) = (2.5, 3)$$

$\Rightarrow$ New $C_2 = (2.5, 3)$

Step-4: Go back to step 2, unless the centroids are not changing.

C_1: For $(1.75, 1.5)$

$$D_1 = \{(1,1)\ (1.75, 1.5)\}$$

$$= \sqrt{(1.75-1)^2 + (1.5-1)^2}$$

$$= 0.90$$

$$D_1 = \{(2,3), (1.75, 1.5)\}$$

$$= \sqrt{(1.75-2)^2 + (1.5-3)^2}$$

$$= 1.52$$

$$D_1 = \{(1,2), (1.75, 1.5)\}$$

$$= \sqrt{(1.75-1)^2 + (1.5-2)^2}$$

$$= 0.90$$

C_2: For $(2.5, 3)$

$$D_2 = \{(\cancel{2}1, 1), (2.5, 3)\}$$

$$= \sqrt{\cancel{(2.5-1)^2 + (3-1)^2}} = \sqrt{(2.5-1)^2 + (3-1)^2}$$

$$= 2.5$$

$\Rightarrow$

$$D_2 = \{(2,3), (2.5, 3)\}$$

$$= \sqrt{(2.5-2)^2 + (3-3)^2}$$

$$= 0.5$$

$$D_2 = \{(1,2), (2.5, 3)\}$$

$$= \sqrt{(2.5-1)^2 + (3-2)^2}$$

$$= 1.80$$

$D_1 = \{(3,3),(1.75,1.5)\}$

$= \sqrt{(1.75-3)^2+(1.5-3)^2}$

$= 1.95$

$D_1 = \{(2,2),(1.75,1.5)\}$

$= \sqrt{(1.75-2)^2+(1.5-2)^2}$

$= 0.55$

$D_1 = \{(3,1),(1.75,1.5)\}$

$= \sqrt{(1.75-3)^2+(1.5-1)^2}$

$= 1.34$

$D_2 = \{(3,3),(2.5,3)\}$

$= \sqrt{(2.5-3)^2+(3-3)^2}$

$= 0.5$

$D_2 = \{(2,2),(2.5,3)\}$

$= \sqrt{(2.5-2)^2+(3-2)^2}$

$= 1.11$

$D_2 = \{(3,1),(2.5,3)\}$

$= \sqrt{(2.5-3)^2+(3-1)^2}$

$= 2.06$

$\therefore \quad C_1 = \{(1,1),(1,2),(2,2),(3,1)\}$

$C_2 = \{(2,3),(3,3)\}\cancel{(3,1)}$

So, we can see the result of the two iteration are same i.e. clustering outcome are same.