1 Attribute Type

Answer: interval

2 Missing values

Answer:

- Eliminate data objects with missing values
- Estimate missing values
- Ignore missing values during data analysis

3 Jaccard coefficient

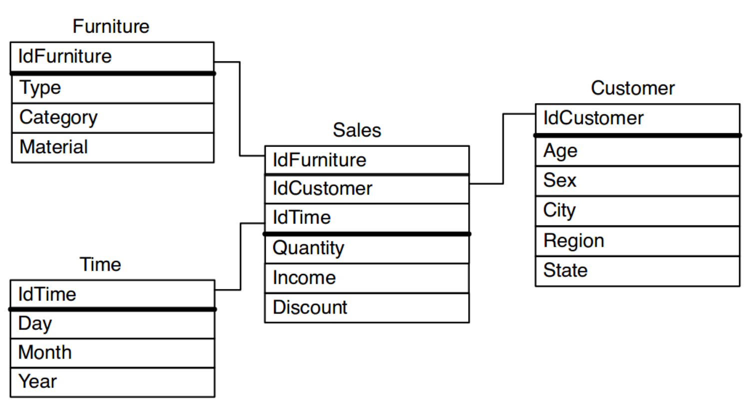
Answer: 0.5

4 CityBlock distance

Answer: 5

5 Modeling

Answer:



6 OLAP

Answer:

- {Month, ItemName, City}: Yes, after roll-up from ItemName to Brand

- {Month, Brand, Country}: No, cannot retrieve city in the query
- {Year, Brand, City}: No, cannot retrieve month and city in the query
- {ItemName, City} where year = 2016: No, cannot retrieve the specified month in the query

7 Apriori Algorithm

Answer:

1) Applying Apriori

Pass(k)	Candidate k-itemsets and their support	Frequent k-itemsets
k=1	{H}(4), {B}(2), {K}(2), {C}(3), {I}(4)	{H}, {B}, {K}, {C}, {I}
k=2	{H, B}(2), {H, K}(1), {H, C}(2), {H, I}(2), {B, K}(1), {B, C}(0), {B, I}(0), {K, C}(0), {K, I}(1), {C, I}(3)	{H, B}, {H, C}, {H, I}, {C, I}
k=3	{H, C, I}(2)	{H, C, I}
k=4	{}	

2)

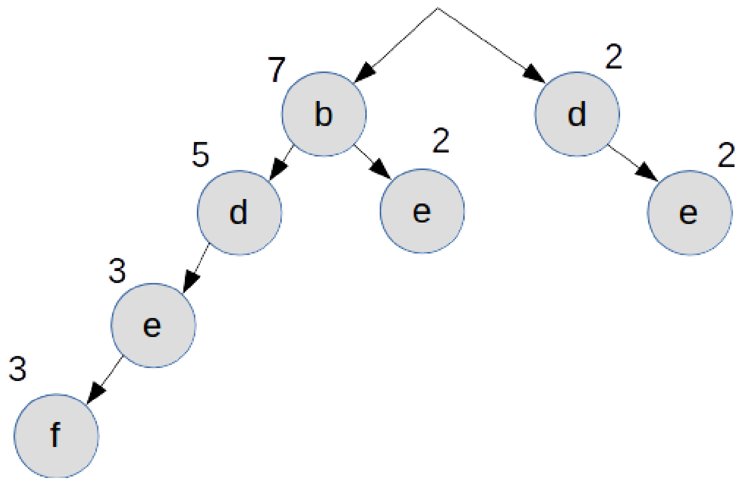
$\{H\} \rightarrow \{C, I\}$ (confidence=2/4=0.5)
$\{C\} \rightarrow \{H, I\}$ (confidence=2/3=0.66)
$\{I\} \rightarrow \{H, C\}$ (confidence=2/4=0.5)
$\{H, C\} \rightarrow \{I\}$ (confidence=2/2=1)
$\{H, I\} \rightarrow \{C\}$ (confidence=2/2=1)
$\{C, I\} \rightarrow \{H\}$ (confidence=2/3=0.66)

Therefore, the four qualified association rules are $\{C\} \rightarrow \{H, I\}$, $\{H, C\} \rightarrow \{I\}$, $\{H, I\} \rightarrow \{C\}$, and $\{C, I\} \rightarrow \{H\}$.

8 FP-growth

Answer:

1)



2)

Item	Conditional sub-database	Conditional FP-tree	Frequent Item sets
f	{{b,d,e}:3}	$\langle b:3, d:3, e:3 \rangle$	{f}:3, {b,f}:3, {d,f}:3, {e,f}:3, {b,d,f}:3, {b,e,f}:3, {d,e,f}:3, {b,d,e,f}:3
e	{{b,d}:3, {b}:2, {d}:2}	$\langle b:5, d:3 \rangle$, $\langle d:2 \rangle$	{e}:7, {d,e}:5, {b,e}:5
ed	{{b}:3}	b:3	{b,d,e}:3
(eb	empty	empty)	
d	{{b}:5}	b:5	{d}:7, {b,d}:5
b	empty	empty	{b}:7

9 K-means clustering

Step 1a: compute squared Euclidean distances between data points and mean vectors

	$m_1 = (2, 0)$	$m_2 = (3, 4)$
$x^{(1)} = (0, 3)$	(0-2)2+(3-0)2=13	(0-3)2+(3-4)2=10
$x^{(2)} = (1, 4)$	(1-2)2+(4-0)2=17	(1-3)2+(4-4)2=4
$x^{(3)} = (3, 1)$	(3-2)2+(1-0)2=2	(3-3)2+(1-4)2=9
$x^{(4)} = (4, 2)$	(4-2)2+(2-0)2=8	(4-3)2+(2-4)2=5
$x^{(5)} = (5, 1)$	(5-2)2+(1-0)2=10	(5-3)2+(1-4)2=13

Pick the smallest in each row, which results that $x^{(1)}$, $x^{(2)}$, and $x^{(4)}$ belong to Cluster 2, and $x^{(3)}$, and $x^{(5)}$ belong to Cluster 1

Step 1b: update mean vectors

$m_1 = (x^{(3)} + x^{(5)})/2 = ((3, 1) + (5, 1))/2 = (4, 1)$

$m_2 = (x^{(1)} + x^{(2)} + x^{(4)})/3 = ((0, 3) + (1, 4) + (4, 2))/3 = (5/3, 3)$ or (1.67, 3)

Step 2a: compute squared Euclidean distances between data points and mean vectors

	$m_1 = (4,1)$	$m_2 = (5/3,3)$
$x^{(1)} = (0,3)$	(0-4)2+(3-1)2=20	(0-5/3)2+(3-3)2=2+7/9 (or 2.78)
$x^{(2)} = (1,4)$	(1-4)2+(4-1)2=18	(1-5/3)2+(4-3)2=1+4/9 (or 1.44)
$x^{(3)} = (3,1)$	(3-4)2+(1-1)2=1	(3-5/3)2+(1-3)2=5+7/9 (or 5.78)
$x^{(4)} = (4,2)$	(4-4)2+(2-1)2=1	(4-5/3)2+(2-3)2=6+4/9 (or 6.44)
$x(5) = (5,1)$	(5-4)2+(1-1)2=1	(5-5/3)2+(1-3)2=15+1/9 (or 15.11)

Pick the smallest in each row, which results that $x^{(1)}$ and $x^{(2)}$ belong to Cluster 2, and $x^{(3)}$, $x^{(4)}$, and $x^{(5)}$ belong to Cluster 1

Step 2b: update mean vectors

$m_1 = (x(3)+x(4)+x(5))/3 = ((3,1)+(4,2)+(5,1))/3 = (4,4/3)$ or (4,1.33)

$m_2 = (x(1)+x(2))/2 = ((0,3)+(1,4))/2 = (0.5,3.5)$

Step 3a: compute squared Euclidean distances between data points and mean vectors

	$m_1 = (4,4/3)$	$m_2 = (1/2,7/2)$
$x^{(1)} = (0,3)$	(0-4)2+(3-4/3)2=18+7/9 (or 18.78)	(0-0.5)2+(3-3.5)2=0.5
$x^{(2)} = (1,4)$	(1-4)2+(4-4/3)2=16+1/9 (or 16.11)	(1-0.5)2+(4-3.5)2=0.5
$x^{(3)} = (3,1)$	(3-4)2+(1-4/3)2=1+1/9 (or 1.11)	(3-0.5)2+(1-3.5)2=12.5
$x^{(4)} = (4,2)$	(4-4)2+(2-4/3)2=4/9 (or 0.44)	(4-0.5)2+(2-3.5)2=14.5
$x^{(5)} = (5,1)$	(5-4)2+(1-4/3)2=1+1/9 (or 1.11)	(5-0.5)2+(1-3.5)2=26.5

Pick the smallest in each row, which results that $x^{(1)}$ and $x^{(2)}$ belong to Cluster 2, and $x^{(3)}$, $x^{(4)}$, and $x^{(5)}$ belong to Cluster 1.

Since there is no change in the cluster assignment, the algorithm ends and outputs

$m_1 = (4,4/3)$ or (4,1.33)

$m_2 = (0.5,3.5)$

Cluster($x^{(1)}$)=2

Cluster($x^{(2)}$)=2

Cluster($x^{(3)}$)=1

Cluster($x^{(4)}$)=1

Cluster($x^{(5)}$)=1

10 DBSCAN pros and cons

Answer:

- Data is high-dimensional
- Data has varying densities

11 Cross-validation

Purpose of cross validation. One of the following answers or the like is acceptable:

- Cross-validation can be used to evaluate the performance of a supervised model (e.g. a classifier or regressor)
- Cross-validation is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set.
- Cross-validation can be used to tune hyper-parameters in algorithms or models

Cross validation procedure. It requires the following parts in the answer

- has explained training/validation/test sets
- has correctly show K-fold rotation steps

12 Decision tree

Answer:

1)

$$\text{Entropy}(p) = -\frac{3}{8} \cdot \log_2 \frac{3}{8} - \frac{5}{8} \cdot \log_2 \frac{5}{8} \approx 0.9544$$

If using "NotHeavy" for root splitting,

$$\sum_{i=1}^{k} \frac{n_i}{n} \text{ Entropy } (i) = \frac{5}{8}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{3}{8}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) \approx 0.9512$$

$$\text{GAIN}_{\text{NotHeavy}} = \text{Entropy}(p) - \sum_{i=1}^{k} \frac{n_i}{n} \text{ Entropy } (i) = 0.0032$$

If using "Smelly" for root splitting,

$$\sum_{i=1}^{k} \frac{n_i}{n} \text{ Entropy } (i) = \frac{3}{8}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) + \frac{5}{8}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) \approx 0.9512$$

$$\text{GAIN}_{\text{Smelly}} = \text{Entropy}(p) - \sum_{i=1}^{k} \frac{n_i}{n} \text{ Entropy } (i) = 0.0032$$

If using "Spotted" for root splitting,

$$\sum_{i=1}^{k} \frac{n_i}{n} \text{ Entropy } (i) = \frac{3}{8}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) + \frac{5}{8}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) \approx 0.9512$$

$$\text{GAIN}_{\text{Spotted}} = \text{Entropy}(p) - \sum_{i=1}^{k} \frac{n_i}{n} \text{ Entropy } (i) = 0.0032$$

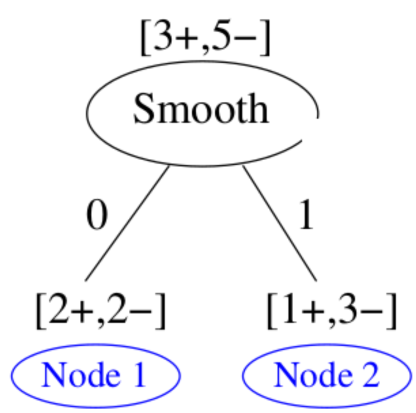
If using "Smooth" for root splitting,

$$\sum_{i=1}^{k} \frac{n_i}{n} \text{ Entropy } (i) = \frac{4}{8}\left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}\right) + \frac{4}{8}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) \approx 0.9056$$

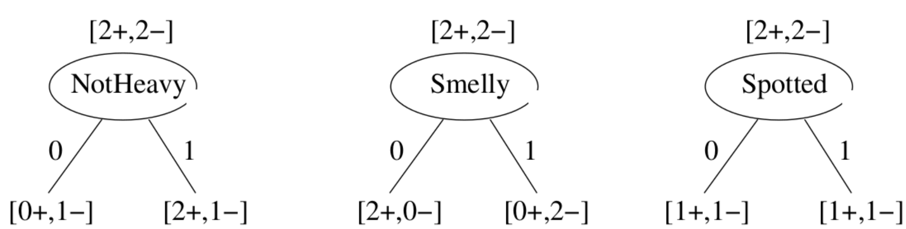
$$\text{GAIN}_{\text{Smooth}} = \text{Entropy}(p) - \sum_{i=1}^{k} \frac{n_i}{n} \text{ Entropy } (i) = 0.0488$$

So we should use "Smooth" for the root splitting because its GAIN is the largest.

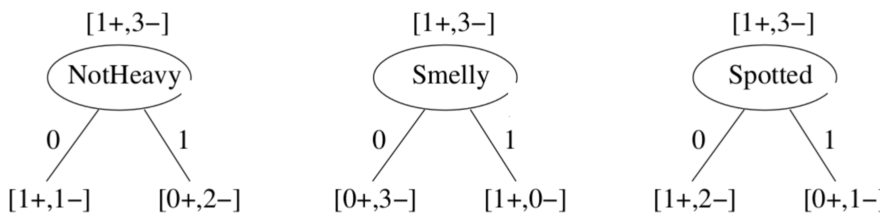
2)



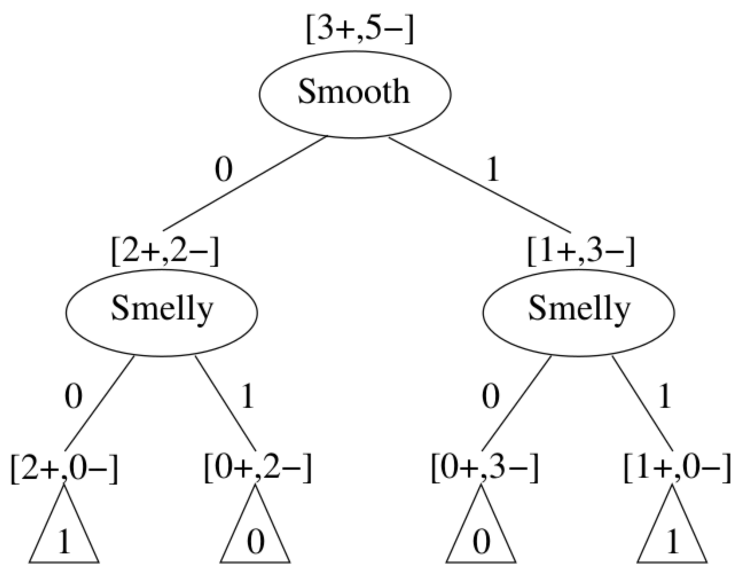
Node 1: Smooth = 0



Node 2: Smooth = 1



It can be seen that after splitting with "Smooth" and then "Smelly", all training samples have been classified. The Entropy after the two-level decision becomes zero (i.e. giving the maximum GAIN). The GAINs using other features are smaller. Therefore the resulting decision tree is



3)

- For U: Smooth = 1, Smelly = 1 ⇒ Edible = 1
- For V: Smooth = 1, Smelly = 1 ⇒ Edible = 1
- For W: Smooth = 0, Smelly = 1 ⇒ Edible = 0