

1. For the following set of instances:

	a_1	a_2	a_3	c
7 transactions	hot	windy	dry	Yes
	mild	windy	rainy	No
	hot	windy	rainy	Yes
	cool	still	dry	Yes
	cool	still	rainy	No
	hot	still	dry	No
	mild	still	dry	Yes

Construct all of the **1-itemsets** and calculate their confidences and supports. Discuss how you would continue mining for effective Association Rules.

Association Rule mining : find "interesting" rules with high

Support (S) & Confidence (C)

$$(S \geq \tau_S, C \geq \tau_C)$$

\Rightarrow predict occurrence of an item based on the occurrences of other items in the transaction.

E.g. $\{ \text{Beer}, \text{Bread} \} \rightarrow \{ \text{Milk} \}$

itemset

co-occurrence
(not causation)

transaction
contains "A&B"

Support :

$$S(A \rightarrow B) = \frac{n(A, B)}{N}$$

total # of
transactions

Confidence :

$$C(A \rightarrow B) = \frac{n(A, B)}{n(A)}$$

frac of transactions
that contains A
also contains B.

Methods :

① Brute-force : list all the rules

\Rightarrow Computationally expensive ($2^d - 1$ possible itemsets)

* ② Two-step approach (with Apriori principle)

① Generate frequent itemsets (itemsets with support $\geq \tau_S$)

② Generate rules with high confidence (binary partitioning)

$K = 1$:

- ① 1-itemsets (all attribute / class values : items)
- # items → {hot}, {mild}, {cold}, {windy}, {still},
{dry}, {rainy}, {Yes}, {No}

Support	hot	mild	cool	windy	still	dry	rainy	YES	NO
Support	$\frac{3}{7}$	$\frac{2}{7}$	$\frac{2}{7}$	$\frac{3}{7}$	$\frac{4}{7}$	$\frac{4}{7}$	$\frac{3}{7}$	$\frac{4}{7}$	$\frac{3}{7}$

$$s(\{\text{hot}\}) = \frac{n(\{\text{hot}\})}{N} = \frac{3}{7}$$

...

$K = 2$: (Generate $k+1$ itemsets from K itemsets)

Apriori principle: ($s < t_s$)

if itemset infrequent \Rightarrow all its superset infrequent.

E.g. If $t_s = 0.5$

$$\begin{aligned} \xrightarrow{\text{infrequent}} s(\{\text{hot}\}) &= \frac{3}{7} \\ s(\{\text{Yes}\}) &= \frac{4}{7} \end{aligned} \quad \left\{ \begin{array}{l} s(\{\text{hot, Yes}\}) \leq \min(s(\{\text{hot}\}), \\ s(\{\text{Yes}\})) \\ = \frac{3}{7} \\ \Rightarrow \text{infrequent!} \end{array} \right.$$

\Rightarrow Only use freq. k -itemsets to generate $k+1$ itemsets.

freq 1-itemsets:

{still}, {dry}, {Yes}

\Rightarrow 2-itemset: (potentially interesting)

{still, dry}, {still, Yes}, {dry, Yes}

...

(find support of the itemsets \Rightarrow freq 2-itemsets)

\Rightarrow 3-itemset:

{still, dry, Yes}

Generate rules: (binary partitioning freq itemsets)

$K=1$: {still}, {dry}, {Yes}

E.g. For {still}: (defective!)

$$\{ \text{still} \} \rightarrow \emptyset \quad \text{OR} \quad \emptyset \rightarrow \{ \text{still} \}$$

\Rightarrow Can't calculate conf for these rules.

$K=2$: (freq 2-itemsets)

E.g. for {still, dry}

$$\{ \text{still} \} \rightarrow \{ \text{dry} \} \quad \text{OR} \quad \{ \text{dry} \} \rightarrow \{ \text{still} \}$$

check if $\text{conf} \geq \text{Ic}$

For $K > 2$:

(right of \rightarrow)

Start with one item in consequent to $K-1$ items.

\therefore Moving an item from antecedent to consequent + conf.

$$\begin{aligned} \text{conf } (\{A, B\} \rightarrow \{C\}) &= \frac{n(A, B, C)}{n(A, B)} \xleftarrow{\text{same}} \\ &\quad \xleftarrow{\text{smaller than}} n(A) \\ \text{conf } (\{A\} \rightarrow \{B, C\}) &= \frac{n(A, B, C)}{n(A)} \xleftarrow{\text{same}} \end{aligned}$$

(stop when $\text{conf} < \text{Ic}$)

2. What does "correlation does not imply causation" mean? Why is it important to keep this adage in mind, when working in the field of Data Mining?

"Correlation does not imply causation" is an important adage which highlights the fact some events are reliably seen together, even though there isn't a causal relationship between the events.

E.g.

Farmers have breakfast \Rightarrow sunrise

People have sports drink \Rightarrow injury

Data Mining :

look for patterns about the data.

(useful) patterns which are:

(i) Valid : actually attested in the data

(ii) Non-trivial : isn't immediately self-evident from data
(e.g. instances are composed of attributes)

(iii) Previously unknown: something we didn't already know
about (many non-trivial patterns are
already well-understood by experts)

A large num of possible patterns (e.g. ARM)

\Rightarrow likely to find some which are entirely statistical
quirks , do not actually imply causal relationship.

3. Review the concepts of **Recommendation Systems**:

- (a) What is Content-based Recommendation?
- (b) What is Collaborative Filtering?

(a) Content-based

Making recommendation based on the content of items.

(Comparing them to items that users have seen or enjoyed)

(b) Collaborative Filtering

Recommendations based on different users' preferences.

(need a large num of users who have submitted ratings
of various items ; difficult to start to solve)

4. Consider the following rating table between five users and six items:

ID	Item A	Item B	Item C	Item D	Item E	Item F
User 1	5	6	7	4	3	?
User 2	4	?	3	?	5	4
User 3	?	2	4	1	1	?
User 4	7	4	3	7	?	4
User 5	1	?	3	2	2	7

- (a) Predict the value of the unknown rating for User 4 using User-based Collaborative Filtering.
(i.e. Find the Pearson correlation between users and adjust User 4's mean score).

Find similarity between users (pearson correlation)

$$P(X, Y) = \frac{\sum_{i \in X \cap Y} (r_{xi} - \mu_x)(r_{yi} - \mu_y)}{\sqrt{\sum_{i \in X \cap Y} (r_{xi} - \mu_x)^2 \sum_{i \in X \cap Y} (r_{yi} - \mu_y)^2}}$$

↑ rated by both X & Y

(like cos similarity with mean subtracted)

① Calculate users' average (μ)

$$\mu_1 = \frac{5+6+7+4+3}{5} = 5$$

$$\mu_2 = 4 \quad \mu_3 = 2$$

$$\mu_4 = 5 \quad \mu_5 = 3$$

② Find Similarity

$$P(1,4) = \frac{\sum_{i \in U_1 \cap U_4} (r_{1i} - \mu_1)(r_{4i} - \mu_4)}{\sqrt{\sum_{i \in U_1 \cap U_4} (r_{1i} - \mu_1)^2} \sqrt{\sum_{i \in U_1 \cap U_4} (r_{4i} - \mu_4)^2}}$$

users

$$= \frac{(5-5)(7-5) + (6-5)(4-5) + \dots}{\sqrt{(5-5)^2 + (6-5)^2 + \dots} \sqrt{(7-5)^2 + (4-5)^2 + \dots}}$$

$$= -0.793$$

$\Rightarrow U_1$ & U_4 negatively correlated

If U_1 likes an item (higher than μ_1),

U_4 might dislike an item (lower than μ_4).

$$P(2,4) = \dots = 0.667$$

Pearson Correlation



ID	Item A	Item B	Item C	Item D	Item E	Item F	Mean	$P(4)$
User 1	5	6	7	4	3	?	5	-0.793
User 2	4	?	3	?	5	4	4	0.667
User 3	?	2	4	1	1	?	2	-0.894
User 4	7	4	3	7	?	4	5	N/A
User 5	1	?	3	2	2	7	3	-0.605

To predict rating of U_4 for Item E: (two options)

① Use users with the most positive scores

② Use users with the largest abs-valued scores

In this example :

choose to use all 4 users

$$\hat{r}_{uj} = \mu_u + \frac{\sum_v P(u,v) \cdot (r_{vj} - \mu_v)}{\sum_v |P(u,v)|}$$

↑
user associated
bias (high / low for
all items)

deviation

$$\begin{aligned}\hat{r}_{uj} &= \mu_u + \frac{P(1,4). (r_{1e} - \mu_1) + P(2,4). (r_{2e} - \mu_2) + P(3,4). (r_{3e} - \mu_3) + P(5,4). (r_{5e} - \mu_5)}{|P(1,4)| + |P(2,4)| + |P(3,4)| + |P(5,4)|} \\ &\approx 5 + \frac{(0.793)(3-5) + (0.667)(5-4) + (-0.894)(1-2) + (-0.605)(2-3)}{|(-0.793)| + |0.667| + |(-0.894)| + |(-0.605)|} \\ &= 5 + \frac{3.752}{2.959} \approx 6.268\end{aligned}$$

- (b) Predict the value of the unknown rating for User 4 using Item-based Collaborative Filtering. (i.e. Find the correlation between items (using "Adjusted Cosine Similarity") and take a weighted average of User 4's scores).

Adjusted Cosine :

$$\begin{aligned}AC(M, N) &= \frac{\sum_{i \in M \cap N} (r_{im} - \mu_i) (r_{in} - \mu_i)}{\sqrt{\sum_{i \in M \cap N} (r_{im} - \mu_i)^2} \sqrt{\sum_{i \in M \cap N} (r_{in} - \mu_i)^2}} \\ &= \frac{\sum_i Sim \cdot Sis}{\sqrt{\sum_i Sim^2} \sqrt{\sum_i Sis^2}} \quad (\text{where } Sim = r_{im} - \mu_i)\end{aligned}$$

(centered by user means : μ_i)

Compare with Pearson Correlation :

$$P(M, N) = \frac{\sum_{i \in M \cap N} (r_{im} - \mu_m) (r_{in} - \mu_n)}{\sqrt{\sum_{i \in M \cap N} (r_{im} - \mu_m)^2} \sqrt{\sum_{i \in M \cap N} (r_{in} - \mu_n)^2}}$$

(centered by item means : μ_m & μ_n)

ID	Item A	Item B	Item C	Item D	Item E	Item F	Mean	P(4)
User 1	5	6	7	4	3	?	5	-0.793
User 2	4	?	3	?	5	4	4	0.667
User 3	?	2	4	1	1	?	2	-0.894
User 4	7	4	3	7	?	4	5	N/A
User 5	1	?	3	2	2	7	3	-0.605
Mean	4.25	4	4	3.5	2.75	5		
AC(e)	0.408	-0.894	-0.882	0.943	N/A	-0.707		
P(e)	0.259	0.71	-0.076	0.990	N/A	-0.707		

Use positive correlated items only. (neg values will ruin the average)

$$\hat{r}_{uj} = \frac{\sum_i AC(i,j) \cdot r_{uj}}{\sum_i |AC(i,j)|}$$

user associated bias
 already included
 in r_{uj}
 \Rightarrow no need $\mu + \dots$
 (weighted avg of ratings)

$$\begin{aligned}\hat{r}_{4e} &= \frac{AC(a,e) \cdot r_{4a} + AC(d,e) \cdot r_{4d}}{|AC(a,e)| + |AC(d,e)|} \\ &\approx \frac{(0.408)(7) + (0.943)(7)}{(0.408) + (0.943)} = \frac{9.457}{1.351} = 7\end{aligned}$$