# Data Mining: Mining Association Rules Examples

## Course Enrollments

**Itemset.** $I = \{$ CSC365, CSC366, CSC402, CSC405, CSC416, CSC454, CSC456, CSC464, CSC465, CSC471, CSC474, CSC480$\}$.

| Column | Course Number | Course |
|---|---|---|
| 1 | CSC365 | Intro Databases |
| 2 | CSC366 | Database Design, Modeling, Implementation |
| 3 | CSC402 | Software Requirements |
| 4 | CSC405 | Software Construction |
| 5 | CSC416 | Autonomous Mobile Robotics |
| 6 | CSC454 | Implementation of OS |
| 7 | CSC456 | Intro Computer Security |
| 8 | CSC464 | Intro Networks |
| 9 | CSC465 | Advanced Networks |
| 10 | CSC471 | Intro Graphics |
| 11 | CSC474 | Computer Animation |
| 12 | CSC480 | Artificial Intelligence |

**Market Baskets.** The market baskets in our dataset consist of the Computer Science electives selected by individual students. Consider the list of 20 market baskets in Figure 1

This list can be represented as a **full binary matrix** as shown in Figure 2.

**Example 1.** Consider the itemset $T = \{\mathsf{CSC365}, \mathsf{CSC366}, \mathsf{CSC416}\}$. Support set of $T$ in the dataset is $Sup(T) = \{s_5, s_{18}\}$. Therefore,

$$support(T) = \frac{|Sup(T)|}{20} = \frac{2}{20} = 0.1.$$

| | |
|---|---|
| $s_1$ | CSC365, CSC366, CSC402, CSC405, CSC464 |
| $s_2$ | CSC402, CSC405, CSC454, CSC456, CSC480 |
| $s_3$ | CSC365, CSC416, CSC454, CSC464 |
| $s_4$ | CSC365, CSC366, CSC471, CSC474 |
| $s_5$ | CSC365, CSC366, CSC416, CSC471, CSC474, CSC480 |
| $s_6$ | CSC402, CSC405, CSC480 |
| $s_7$ | CSC416, CSC454, CSC456, CSC464, CSC465, CSC480 |
| $s_8$ | CSC456, CSC464, CSC465 |
| $s_9$ | CSC471, CSC474 |
| $s_{10}$ | CSC365, CSC456, CSC464, CSC471, CSC480 |
| $s_{11}$ | CSC416, CSC456, CSC464, CSC480 |
| $s_{12}$ | CSC365, CSC366, CSC402, CSC480 |
| $s_{13}$ | CSC365, CSC402, CSC405, CSC464 |
| $s_{14}$ | CSC402, CSC471, CSC480 |
| $s_{15}$ | CSC365, CSC366, CSC456, CSC464, CSC465 |
| $s_{16}$ | CSC471, CSC474, CSC480 |
| $s_{17}$ | CSC454, CSC471 |
| $s_{18}$ | CSC365, CSC366, CSC416, CSC480 |
| $s_{19}$ | CSC402, CSC405, CSC471, CSC474 |
| $s_{20}$ | CSC454, CSC480 |

Figure 1: Student Enrollment Dataset: Market Baskets

| Item | 365 | 366 | 402 | 405 | 416 | 454 | 456 | 464 | 465 | 471 | 474 | 480 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $s_1$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| $s_2$ | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| $s_3$ | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| $s_4$ | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| $s_5$ | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| $s_6$ | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| $s_7$ | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 |
| $s_8$ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| $s_9$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| $s_{10}$ | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| $s_{11}$ | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| $s_{12}$ | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| $s_{13}$ | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| $s_{14}$ | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| $s_{15}$ | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| $s_{16}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| $s_{17}$ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| $s_{18}$ | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| $s_{19}$ | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| $s_{20}$ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| Count: | 9 | 6 | 7 | 5 | 5 | 6 | 6 | 6 | 3 | 8 | 5 | 12 |
| Support: | 0.45 | 0.3 | 0.35 | 0.25 | 0.25 | 0.3 | 0.3 | 0.3 | 0.15 | 0.4 | 0.25 | 0.6 |

Figure 2: Student Enrollment Dataset: Full Binary Vectors

**Example 2.** Consider an association rule $R_1 = \mathsf{CSC402} \longrightarrow \mathsf{CSC405}$.

The support set for $R_1$ is $Sup(R_1) = \{s_1, s_2, s_6, s_{13}, s_{19}\}$. The support of $R_1$ is

$$support(R_1) = \frac{|Sup(R_1)|}{20} = \frac{5}{20} = 0.25.$$

The support set for $\{\mathsf{CSC402}\}$ is $\{s_1, s_2, s_6, s_{12}, s_{13}, s_{14}, s_{19}\}$. The confidence of the rule $R_1$ is then

$$confidence(R_1) = \frac{support(R_1)}{support(\{\mathsf{CSC402}\})} = \frac{0.25}{0.35} = \frac{5}{7} = 0.714.$$

## Apriori Algorithm

**minConf.** Consider the value of minimal support, minSup = 0.25.

**Goal.** We trace the work of the Apriori algorithm in discovery of frequent itemsets with support of at least minSup (0.25).

**Step 1. Itemsets of size 1.** First, we discover frequent itemsets of size 1.

$$F_1 = \{\{\mathsf{CSC365}\}, \{\mathsf{CSC366}\}, \{\mathsf{CSC402}\}, \{\mathsf{CSC405}\}, \{\mathsf{CSC416}\}, \{\mathsf{CSC454}\}, \{\mathsf{CSC456}\}, \{\mathsf{CSC464}\}, \{\mathsf{CSC471}\}, \{\mathsf{CSC474}\}, \{\mathsf{CSC480}\}\}.$$

**Note:** $support(\{\mathsf{CSC465}\}) = 0.15 < \mathsf{minSup}$, so CSC465 is excluded from consideration. All other columns have support of 0.25 or higher and they are included.

**Step 2.1. Itemsets of size 2. Join Step.** On this step, we construct the list of all *pairs* of items from $C_1$.

**Note:** The **join step** for size 2 itemsets is *trivial*: it involves computing cartesian product of $C_1$.

$$C_2 = F_1 \times F_1.$$

**Step 2.2. Itemsets of size 2. Pruning Step.** For itemsets of size 2, the pruning step of the cadidateGen() function is trivial. Nothing is pruned, $C_2$ remains intact.

**Step 2.3. Itemsets of size 2. Support computation.** Step 2.1 generated $\frac{11 \cdot 10}{2} =$ 55 possible pairings. We now need to prune this set, by excluding from it all pairs of courses that have low support. We can construct the following Support table for our dataset:

| | 365 | 366 | 402 | 405 | 416 | 454 | 456 | 464 | 471 | 474 | 480 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 365 | — | **6** | 2 | 1 | 2 | 0 | 2 | 2 | 3 | 2 | **5** |
| 366 | | — | 2 | 1 | 2 | 0 | 1 | 2 | 2 | 2 | 3 |
| 402 | | | — | **5** | 0 | 1 | 1 | 1 | 2 | 1 | **5** |
| 405 | | | | — | 0 | 1 | 1 | 1 | 1 | 1 | 3 |
| 416 | | | | | — | 2 | 3 | 2 | 1 | 1 | 4 |
| 454 | | | | | | — | 3 | 2 | 1 | 0 | 3 |
| 456 | | | | | | | — | 4 | 1 | 0 | 4 |
| 464 | | | | | | | | — | 1 | 0 | 3 |
| 471 | | | | | | | | | — | **5** | 4 |
| 474 | | | | | | | | | | — | 2 |
| 480 | | | | | | | | | | | — |

From the table above, the following pairs of courses exceed minsup:

| Itemset | Baskets | Frequency | support |
|---|---|---|---|
| {CSC365, CSC366} | $\{s_1, s_4, s_5, s_{12}, s_{15}, s_{18}\}$ | 6 | 0.3 |
| {CSC365, CSC480} | $\{s_5, s_{10}, s_{12}, s_{13}, s_{18}\}$ | 5 | 0.25 |
| {CSC402, CSC405} | $\{s_1, s_2, s_6, s_{13}, s_{19}\}$ | 5 | 0.25 |
| {CSC402, CSC480} | $\{s_2, s_6, s_{12}, s_{13}, s_{14}\}$ | 5 | 0.25 |
| {CSC471, CSC474} | $\{s_4, s_5, s_9, s_{16}, s_{19}\}$ | 5 | 0.25 |

So, $F_2 = \{$ {CSC365, CSC366}, {CSC365, CSC480}, {CSC402, CSC405}, {CSC402, CSC480}, {CSC471, CSC474} $\}$.

**Step 3.1. Itemsets of size 3. Join Step.** On this step, we join all pairs of sets from $F_2$ trying to form candidate frequent itemsets of size 3. We are able to join the following pairs of sets:

| First itemset | Second itemset | Join | ID |
|---|---|---|---|
| {<u>CSC365</u>, CSC366} | {<u>CSC365</u>, CSC480} | {CSC365, CSC366, CSC480} | $c_1$ |
| {<u>CSC402</u>, CSC405} | {<u>CSC402</u>, CSC480} | {CSC402, CSC405, CSC480} | $c_2$ |
| {CSC365, <u>CSC480</u>} | {CSC402, <u>CSC480</u>} | {CSC365, CSC402, CSC480} | $c_3$ |

$C_3 = \{$ {CSC365, CSC366, CSC480}, {CSC402, CSC405, CSC480}, {CSC365, CSC402, CSC480} $\}$.

**Step 3.2. Itemsets of size 3. Pruning Step.** For {CSC365, CSC366, CSC480}:

{CSC365, CSC366} $\in F_2$
{CSC365, CSC480} $\in F_2$
{CSC366, CSC480} $\notin F_2$

For {CSC402, CSC405, CSC480}:

{CSC402, CSC405} $\in F_2$
{CSC402, CSC480} $\in F_2$
{CSC405, CSC480} $\notin F_2$

For {CSC365, CSC402, CSC480}:

{CSC365, CSC480} $\in F_2$
{CSC402, CSC480} $\in F_2$
{CSC365, CSC402} $\notin F_2$

Therefore, all three elements of $C_3$ are **not frequent itemsets** and the **Apriori Algoritm** can stop there and return $F = F_1 \cup F_2$ as the set of all frequent itemsets.

**Takehome Problem**

Run **Apriori Algorithm** by hand with minSup = 0.2.

## Generation of Association Rules

**Frequent Itemsets.** In previous section, we discovered that Student Enrollment dataset has 11 (eleven) frequent itemsets of size 1 (all singleton sets except for {CSC465}) and five frequent itemsets of size 2:

{{CSC365, CSC366}, {CSC365, CSC480}, {CSC402, CSC405}, {CSC402, CSC480}, {CSC471, CSC474}}.

This gives rise to 10 candidate association rules with a single item on the right side. For each of them, we compute confidence.

| ID | Rule | Frequent Itemset Support | Left side support | Confidence |
|---|---|---|---|---|
| $R_1$ | CSC365 $\longrightarrow$ CSC366 | $\frac{6}{20}$ | $\frac{9}{20}$ | $\frac{2}{3} = 0.667$ |
| $R_2$ | CSC366 $\longrightarrow$ CSC365 | $\frac{6}{20}$ | $\frac{6}{20}$ | 1 |
| $R_3$ | CSC365 $\longrightarrow$ CSC480 | $\frac{5}{20}$ | $\frac{9}{20}$ | $\frac{5}{9} = 0.555$ |
| $R_4$ | CSC480 $\longrightarrow$ CSC365 | $\frac{5}{20}$ | $\frac{12}{20}$ | $\frac{5}{12} = 0.41667$ |
| $R_5$ | CSC402 $\longrightarrow$ CSC405 | $\frac{5}{20}$ | $\frac{7}{20}$ | $\frac{5}{7} = 0.714$ |
| $R_6$ | CSC405 $\longrightarrow$ CSC402 | $\frac{5}{20}$ | $\frac{5}{20}$ | 1 |
| $R_7$ | CSC402 $\longrightarrow$ CSC480 | $\frac{5}{20}$ | $\frac{7}{20}$ | $\frac{5}{7} = 0.714$ |
| $R_8$ | CSC480 $\longrightarrow$ CSC402 | $\frac{5}{20}$ | $\frac{12}{20}$ | $\frac{5}{12} = 0.41667$ |
| $R_9$ | CSC471 $\longrightarrow$ CSC474 | $\frac{5}{20}$ | $\frac{8}{20}$ | $\frac{5}{8} = 0.625$ |
| $R_{10}$ | CSC474 $\longrightarrow$ CSC471 | $\frac{5}{20}$ | $\frac{5}{20}$ | 1 |

Depending on the values of minConf, we will report the following:

- minConf = 1. We report rules $R_2, R_6$ and $R_{10}$.
- $0.668 \leq$ minConf $< 1$. We report rules $R_2, R_6$ and $R_{10}$ from above, plus $R_5$ and $R_7$.
- minConf $> 0.5$. In addition to the rules above, we report $R_1, R_3$ and $R_9$.

**Takehome Problem**

After discovering all frequent itemsets with support of at least 0.2, report all association rules in the dataset for minConf levels of 1, 0.75, 0.666 and 0.5.