# Data Mining (Extra Credits)

(CSE 634/590)

Problem From Text Book

Abhishek Sharma (106932443)

Apurva Maharishi (106676884)

# Given Data (From Text Book)

- T100 - {M,O,N,K,E,Y}
- T200 - {D,O,N,K,E,Y}
- T300 - {M,A,K,E}
- T400 - {M,U,C,K,Y}
- T500 - {C,O,K,I,E}

- Min_sup = 60% = Minimum Support Count(%)
- Min_conf = 80% = Minimum Confidence(%)
- Let the Classes be E (Item 11) and Y (Item 10).

# Transactional Database

| | M Item 1 | D Item 2 | C Item 3 | O Item 4 | A Item 5 | U Item 6 | N Item 7 | K Item 8 | I Item 9 | Y Item 10 Class | E Item 11 Class |
|---|---|---|---|---|---|---|---|---|---|---|---|
| T100 | + | - | - | + | - | - | + | + | - | + | + |
| T200 | - | + | - | + | - | - | + | + | - | + | + |
| T300 | + | - | - | - | + | - | - | + | - | - | + |
| T400 | + | - | + | - | - | + | - | + | - | + | - |
| T500 | - | - | + | + | - | - | - | + | + | - | + |

# Minimum Support Count

- Total No. of Transaction = 5
- Minimum Support Count = 60% of 5
  = 3

# Frequent one Itemsets

| One item Itemsets | |
|---|---|
| M | 3 |
| D | 1 |
| C | 2 |
| O | 3 |
| A | 1 |
| U | 1 |
| N | 2 |
| K | 5 |
| E | 4 |
| Y | 3 |
| I | 1 |

Minimum Support Count = 3

| Frequent One Itemsets | |
|---|---|
| M | 3 |
| O | 3 |
| K | 5 |
| E | 4 |
| Y | 3 |

# Frequent 2 Itemsets

| Candidate 2-Itemsets | |
|---|---|
| {M,O} | 1 |
| {M,K} | 3 |
| {M,E} | 2 |
| {M,Y} | 2 |
| {O,K} | 3 |
| {O,E} | 3 |
| {O,Y} | 2 |
| {K,E} | 4 |
| {K,Y} | 3 |
| {E,Y} | 2 |

Use Apriori Principle for Pruning

| Candidate 2-Itemsets After Pruning | |
|---|---|
| {M,O} | 1 |
| {M,K} | 3 |
| {M,E} | 2 |
| {M,Y} | 2 |
| {O,K} | 3 |
| {O,E} | 3 |
| {O,Y} | 2 |
| {K,E} | 4 |
| {K,Y} | 3 |
| {E,Y} | 2 |

# Continued..

| Candidate 2-Itemsets After Pruning | |
|---|---|
| {M,O} | 1 |
| {M,K} | 3 |
| {M,E} | 2 |
| {M,Y} | 2 |
| {O,K} | 3 |
| {O,E} | 3 |
| {O,Y} | 2 |
| {K,E} | 4 |
| {K,Y} | 3 |
| {E,Y} | 2 |

Minimum Support Count = 3

| Frequent 2-Itemsets | |
|---|---|
| {M,K} | 3 |
| {O,K} | 3 |
| {O,E} | 3 |
| {K,E} | 4 |
| {K,Y} | 3 |

# Frequent 3 Itemsets

| Candidate 3-Itemsets | |
|---|---|
| {O,K,E} | 3 |
| {K,E,Y} | 2 |

Use Apriori Principle for Pruning

| Candidate 3-Itemsets After Pruning | |
|---|---|
| {O,K,E} | 3 |

Minimum Support Count = 3

-> There are no 4 Itemsets

| Frequent 3-Itemsets | |
|---|---|
| {O,K,E} | 3 |

# All Frequent Itemsets

- M - 3
- O - 3
- K - 5
- E - 4
- Y - 3
- {M,K} - 3
- {O,K} - 3
- {O,E} - 3
- {K,E} - 4
- {K,Y} - 3
- {O,K,E} - 3

# Association rules

- C(X→Y) = P(Y|X)
  = sup_count(XUY) /sup_count(X)

  We have included only 2 and 3 frequent item sets, because One Itemsets will not help us in making the association rules.

- Let the Classes be Y (Item 10) and E (Item 11)
- So, we are interested in finding the Rules of the form A -> Y (Item 10) and A -> E (Item 11)

# Association Rules by Classification

| Rule_No | Rule | Confidence | Confidence(%) |
|---|---|---|---|
| R1 | O → E | 3/4 | 75% |
| R2 | K → E | 4/5 | 80% |
| R3 | K → Y | 3/5 | 60% |
| R4 | {O,K} → E | 3/3 | 100% |

- Since our classes are E and Y, so with confidence of 80% the rule R3 cannot be included. So we reduce the confidence to 60% to include R3.

# Selected Rules

- Rule_No Rule [Actual Support Count, Actual Confidence]
- R1 O → E [60%, 75%]
- R2 K → E [80%, 80%]
- R3 K → Y [60%, 60%]
- R4 {O,K} → E [60%, 100%]

# Test Data

| | M Item 1 | D Item 2 | C Item 3 | O Item 4 | A Item 5 | U Item 6 | N Item 7 | K Item 8 | I Item 9 | Y Item 10 Class | E Item 11 Class |
|---|---|---|---|---|---|---|---|---|---|---|---|
| T100 | + | - | - | - | + | - | - | + | - | + | - |
| T200 | + | - | - | + | - | - | + | + | + | - | + |
| T300 | - | - | + | + | + | - | - | - | - | + | - |
| T400 | + | - | + | + | - | + | - | - | - | - | + |
| T500 | - | + | + | - | - | - | - | + | + | - | + |

# Continued..

- T100 satisfies the rule:
  - K->Y  [Success]
- T200 satisfies the rule:
  - {O,K}->E  [Success]
- T300 satisfies the rule:
  - {C,O,A} ->Y [Failure]

# Continued..

- T400 satisfies the rule:
  - O -> E [Success]
- T500 satisfies the rule:
  - K -> E [Success]
- Predictive Accuracy = 80%
- Error Rate = 20%

Thank You