Answers to the Exam in Data Mining 2006-12-15

1. **Classification**

   (a) See any textbook on the subject.

   (b) $O(qn)$

   (c)
   - How to split nodes (binary split, multiway split)
   - How to evaluate how good splits are (GINI-meassure, entropy)
   - Stopping conditions.

   (d)
   - **Pros:**
     - Fast classification, O(depth of tree).
     - Easy to interpret.
     - Inexpensive to construct.
     - ...
   - **Cons:**
     - Difficult to construct the optimal decision tree.
     - Works poorly for some data since we are splitting on one characteristic at the time, which leads to rectangular classification borders.
     - ...

---

2. **Evaluation meassures in rule-based classification**

   (a) i. The accuracies of the rules are 80% (for R1), 75% (for R2), and 52.6% (for R3), respectively. Therefore R1 is the best candidate and R3 is the worst candidate according to rule accuracy.

   ii. Assume the initial rule is $\emptyset \rightarrow +$. This rule covers $p_0 = 100$ positive examples and $n_0 = 400$ negative examples.

   - R1 covers $p_1 = 4$ positive examples and $n_1 = 1$ negative example. Therefore, the FOIL's information gain for this rule is

$$4 * \left(log_2 \left(\frac{4}{5} \right) - log_2 \left(\frac{100}{500} \right) \right) = 8$$

- R2 covers $p_1 = 30$ positive examples and $n_1 = 10$ negative example. Therefore, the FOIL's information gain for this rule is

$$30 * \left(log_2 \left(\frac{30}{40} \right) - log_2 \left(\frac{100}{500} \right) \right) = 57.2$$

- R3 covers $p_1 = 100$ positive examples and $n_1 = 90$ negative examples. Therefore, the FOIL's information gain for this rule is

$$100 * \left(log_2 \left(\frac{100}{190} \right) - log_2 \left(\frac{100}{500} \right) \right) = 139.6$$

R3 is the best candidate and R1 is the worst candidate.

(b) Rule accuracy is only concerned with the accuracy of the rule when it is applied. FOIL's information gain also takes into account how often the rule can be applied, and how much better it is than the default rule.

---

3. **Association patterns evaluation and sequential patterns**

(a) Support:

- Symmetric
- Not invariant under inversion
- Not invariant under row-column scaling
- Not invariant under null-addition

Confidence:

- Asymmetric
- Not invariant under inversion
- Not invariant under row-column scaling
- Invariant under null-addition

(b) i. <{a} {b} {c} {d}>
<{a} {b e} {c}>
<{a} {e} {c d}>
<{b} {c} {d} {e}>
<{b e} {c d}>

ii. <{a} {b e} {c}>

---

4. **Clustering**

(a) Core Points: C, E, F, H, I, L

(b) Border Points: A, B, G, K, O, M

(c) Directly Density Reachable from I: E, F, H

(d) Directly Density Reachable from M: None (M is not a core point)

(e) (Row, Column) = (5, 3) P then becomes a border point for both clusters.

(f) (6, 4), (5, 5), (6,5). P becomes a core point, reachable from both clusters.

---

5. **Association rules**

(a) MinSuppCount = ceil(0.3 * 8) = 3
MinConf = 0.77

C1	**L1**	**C2**	**L2**	**C3**	**C'3**	**L3**
{a}	{a}	{a b}	{a b}	{a b c}	{a b c}	{a b c}
{b}	{b}	{a c}	{a c}	{d e f}		
{c}	{c}	{a d}	{b c}			
{d}	{d}	{a e}	{d e}			
{e}	{e}	{a f}	{d f}			
{f}	{f}	{b c}				
		{b d}				
		{b e}				
		{b f}				
		{c d}				
		{c e}				
		{c f}				
		{d e}				
		{d f}				
		{e f}				

Final rules
{a c}→{b}
{a}→{b}
{b}→{a}
{b}→{c}
{c}→{b}
{e}→{d}
{f}→{d}

(b) **Low support and high confidence** $I_1 \cup I_2$ is seldom bought, but when I_1 is bought we know that there is a high probability that I_2 is also bought. The high confidence tells us that I_1 is relatively uncommon, and if I_2 is also uncommon we have a strong rule, however seldom applicable.
Example: {Expensive beer} → {plastic bag} is a pretty uninteresting rule, but {Ipod} → {Special Ipod headphones} is more interesting.

High support and low confidence $I_1 \cup I_2$ is relatively often bought together, but since I_1 is bought even more often, we cannot say for sure that somebody

interested in I_1 will also be interested in I_2. These rules are only interesting if the other rules also have a low confidence.
Example: {plastic bag} → {Cheap beer}.

(c) Given facts and relations between support of subsets:

$$\frac{S_{\{1234\}}}{S_{\{12\}}} \geq C_{\texttt{min}} \quad S_{\{123\}} \leq S_{\{12\}} \leq S_{\{1\}}$$
$$\frac{S_{\{1234\}}}{S_{\{34\}}} < C_{\texttt{min}} \quad S_{\{234\}} \leq S_{\{34\}} \leq S_{\{3\}}$$

i. Will definitely appear in the final set

ii. Might appear in the final set

iii. Might appear in the final set

iv. Will definitely not appear in the final set