Exam IV syllabus

Chap 22: Classification Evaluation
  → metrics, ROC curve
  → cross validation, confidence intervals, t-test

Chap 13: K means, Kernel Kmeans, EM

Chap 15: Density based
  → Kernel density estimation
  → DBSCAN, Dencity

Chap 16: Spectral Clustering
  → matrix
  → Ratio, Normalized cut

Chap 17: Clustering Validation
  Internal, External
  Relative: SC, gap statistic

Chap 8 & 9: Stream mining, Assoc rule, summaries

Pattern Mining vs Clustering

<table>
<thead>
<tr>
<th>Pattern Mining</th>
<th>Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>unsupervised</td>
<td>unsupervised</td>
</tr>
<tr>
<td>local</td>
<td>group all points, consider all attribute ( (P^n) )</td>
</tr>
</tbody>
</table>
  → subset of the points
  → subset of attribute

Items sets: subset patterns
Sequences: subsequence patterns
Graphs: subgraph patterns

Amazon: Cooccurrence patterns (case analysis)

I: set of items (attributes)
T: set of transactions (points)
The task is to find all frequent itemsets with minimum support \( \Theta \). The support of an itemset \( X \), \( \text{supp}(X) \), is the number of transactions that contain \( X \) as a subset:

\[
\text{supp}(X) = \left| \{ X \in \text{trans} \mid X \text{ contains } X \} \right|
\]

An itemset \( X \subseteq I \) has a support of \( \text{supp}(X) \) if \( X \) is frequent. The frequent itemset \( X \) can be represented by the indicator vector \( t(X) \). For example, if \( X = \{A, B, C\} \), then

\[
t(X) = [1, 0, 0, 0, 1]
\]

We say that \( X \) is frequent if \( \text{supp}(X) \geq \Theta \), where \( \Theta \) is called the minimum support threshold.

**Apriori Algorithm**

Given a dataset \( D \) with items \( \{A, B, C, D, E\} \) and transactions, the Apriori algorithm is used to find frequent itemsets. The minimum support \( \text{minsup} \) is set to 3 or 50\%.

\[
\text{minsup} = 3
\]

\[
I = \{ A, B, C, D, E \}
\]

All possible subsets of \( I \) are considered as potential frequent itemsets.
\[
\begin{align*}
\text{all possible subsets } &= \text{ power set} \\
|\mathcal{P}(S)| &= 2^{|S|} = 32 \\
I &\approx 10^5 \\
\end{align*}
\]

**Eclat Algorithm**

\[
D = \begin{bmatrix}
A & B & C & D & E \\
2 & 2 & 3 & 6 & 6 \\
5 & 4 & 8 & 9 & 5 \\
\end{bmatrix}
\]

Invert the data

18 frequent itemsets excluding \(\emptyset\)
\[
\frac{(x, y)}{(x, y)} = \frac{P(x)}{P(y)\gamma} = \frac{p(x|y)}{p(y|x)}
\]

Given \( x \) and \( y \) such that \( \gamma \), solve for \( x \) and \( y \) such that

\[
\frac{(x, y)}{(x, y)} = \frac{P(x)}{P(y)\gamma} = \frac{p(x|y)}{p(y|x)}
\]

This is a complex equation involving conditional probabilities.

**Task:** Find all possible solutions, given the constraints.

---

**Diagram:**

- Red and green paths intersecting at various points, indicating different sub-problem solutions.
- Nodes labeled with letters and numbers, connecting through arcs.

- Diagram shows relationships and dependencies between elements, possibly representing a network or graph theory problem.
Combinational space (power set) is exponentially large.

\[ F = \text{set of all frequent itemsets} \]

\[ M = \text{set of all maximal frequent itemsets} \]

\[ \{ \text{ABDE}_3, \text{CE}_3, \text{BC}_n \} \rightarrow \text{lossy compression} \]

A\#B has the frequent
\[ s_p(A\#B) \geq 3 \]

\[ C = \text{set of all closed frequent itemsets} \]

\[ \rightarrow \text{There is no frequent superset with the same weight.} \]
There is no frequent superset with the same support.

\[ F \subseteq C \subseteq M \]

\[ C = \{ \emptyset, B, B \cup C, D \cup C, C, A, \emptyset, A \cup B \cup C \} \]

\[ \text{sup}(C) = 4 \]

There is no frequent subset with the same support.

\[ \{ \emptyset, A, C, D \cup C, A \cup D \}, \{ A \emptyset, C \}, D \cup C \} \]

Also a compressed set.

Associated Rule

\[ F \subseteq C \subseteq M \]

Summary sets

\[ F \hookrightarrow \text{Apriori} \quad \text{Exam IV} \]

\[ \text{Eclat} \]

Unique top element closed set

Smallest sets that have the same itemset as the closed itemset.

Abduction graph