problem 1.

(a) Six items in the dataset, so, the possible rules are $3^6 - 2^{6+1} + 1 = 602$.

(b) 4. Since the longest transaction has 4 items.

(c) $C_6^3 = \frac{6 \times 5 \times 4}{3 \times 2} = 20$

(d)

{milk}	$\frac{5}{10}$	{milk, Diaper}	$\frac{4}{10}$
{Beer}	$\frac{4}{10}$	{milk. Bread}	$\frac{3}{10}$
{Diapers}	$\frac{7}{10}$	{milk, Butter}	$\frac{3}{10}$
{Bread}	$\frac{5}{10}$	{Diaper, Bread}	$\frac{3}{10}$
{Butter}	$\frac{5}{10}$	{Diaper, Butter}	$\frac{2}{10}$
{Cookies}	$\frac{4}{10}$	{Bread, Butter}	$\frac{5}{10}$

so {Bread, Butter} has the largest support $= \frac{1}{2}$.

(e) {Bread, Butter}. since the support of {Butter} {Bread} and {Bread, Butter} are all the same. So the confidence all equals to 1.

{Bread} → {Butter} = {Butter} → {Bread}.

problem 2.

$$\boxed{\begin{array}{l}\{\text{video game}\} : \{vg\} \\ \{\text{movie}\} : \{m\}\end{array}}$$

we know: $c(\{m\} \rightarrow \{vg\}) < \frac{\sigma\{vg\}}{N}$.

so: $$\frac{\sigma(\{m, vg\})}{\sigma(\{m\})} < \frac{\sigma(\{vg\})}{N}. \quad ①$$

$$\sigma(\{vg\}) = \sigma(\{\bar{m}, vg\}) + \sigma(\{m, vg\}) \quad ②$$

$$N = \sigma(\{\bar{m}\}) + \sigma(\{m\}) \quad ③$$

$$\Rightarrow \frac{\sigma(\{m,vg\})}{\sigma(\{m\})} < \frac{\sigma(\{\bar{m},vg\}) + \sigma(\{m,vg\})}{\sigma(\{\bar{m}\}) + \sigma(\{m\})} \leq 1$$

$$\Rightarrow \sigma(\{m,vg\}) \cdot \sigma(\{\bar{m}\}) + \cancel{\sigma(\{m,vg\}) \cdot \sigma(\{m\})} \quad ④$$
$$< \sigma(\{m\}) \cdot \sigma(\{\bar{m}, vg\}) + \cancel{\sigma(\{m\}) \cdot \sigma(\{m, vg\})}$$

$$\Rightarrow \frac{\sigma(\{m,vg\})}{\sigma(\{m\})} < \frac{\sigma(\{\bar{m}, vg\})}{\sigma(\{\bar{m}\})}$$

so: $$c(\{m\} \rightarrow \{vg\}) < c(\{\bar{m}\} \rightarrow \{vg\})$$

from ④ $$\Rightarrow \sigma(\{m,vg\}) \cdot \sigma(\{\bar{m}\}) + \sigma(\{\bar{m},vg\}) \cdot \sigma(\{\bar{m}\})$$
$\cancel{s(\{vg\}) =}$
$$< \sigma(\{m\}) \cdot \sigma(\{\bar{m},vg\}) + \sigma(\{\bar{m}\}) \cdot \sigma(\{\bar{m},vg\})$$

$$\Rightarrow \frac{\sigma(\{\bar{m},vg\}) + \sigma(\{m,vg\})}{\sigma(\{\bar{m}\}) + \sigma(\{m\})} < \frac{\sigma(\{\bar{m}, vg\})}{\sigma(\{\bar{m}\})}$$

$$\Rightarrow s(\{vg\}) < c(\{\bar{m}\} \rightarrow \{vg\}) \quad \square$$

problem 3. $\{6, 12, 18, 24, 30, 42, 48\}$

a). i. $\{18, 45\}$.

$$|30-18| < |30-45|.$$

so: first cluster: $\{6, 12, 18, 24, 30\}$.

squared error: $\sum_{i \in C_1} (i-18)^2 = 360$

second cluster: $\{42, 48\}$.

squared error: $\sum_{i \in C_2} (i-45)^2 = 18$.

total error: 378.

ii). $\{15, 40\}$.

cluster 1. $\{6, 12, 18, 24\}$, error = 180

custer 2: $\{30, 42, 48\}$, error = 168

total error: 348

b) yes. both are stable,
since the average for each cluster in each set equals the initial centroid value set

e.g. $\frac{6+12+18+24+30}{5} = 18$; $\frac{42+48}{2} = 45$

c). the two clusters are $\{6, 12, 18, 24, 30\}$ & $\{42, 48\}$.

d) single link produces more natural clustering. since clusters generated by single link has uniform distance between nodes.

e) Contiguous and also density.

single link: $\{6, 12, 18, 24, 30\}$ & $\{42, 48\}$

k-means: $\{6, 12, 18, 24\}$ & $\{30, 42, 48\}$

f). the well-known drawback of k-means is its bad performance at dividing the clusters with big-different-sizes.

the k-mean algorithm tends to break the larger cluster as we can see from this example.