Student Number: S5132483 Last Name: Perry First Name: Sebastian

1)

a. A set of items, sub sequences, or structures which appear frequently in a given data set.

b. Correlation can be used to measure the interestingness of a rule

c. The difference between supervised and un-supervised learning is whether or not the given training data has labels. In supervised learning the machine is given labelled training data which it is meant to be able to use in classifying all future instances. In un-supervised learning the training data has no labels, with the intention that the computer look at the data with the aim of establishing classes and clusters

d. Overfitting is when a model fits its training data too well, such that it might cause it to miss real world instances. Underfitting is when the fit is not close enough or too simple, such that instances that shouldn't fit may do so.

e. Challenges with outlier detection can include:
   - The difference between outlier and normal can often be blurred
   - Outlier definition changes wildly based on the context of what the attribute is about
   - Noise may blur the distinction between outlier and normal
   - It can be hard to justify why a data point is an outlier

f. Time Series data mining is mining on datasets that consist of sequences of values changing with time. Sequence Data Mining is mining on datasets that consist or a sequence of events with or without a notion of time attached

2)

Transaction	Products
1	laptop, camera, hard-drive
2	laptop, DVD
3	DVD, speakers
4	laptop, camera, hard-drive
5	CD, hard-drive
6	DVD, hard-drive
7	CD, DVD
8	laptop, camera, TV
9	TV, speakers
10	laptop, camera

a.

CL_1	sup_count
laptop	5
camera	4
hard-drive	4
DVD	4
speakers	2
CD	2
TV	2

FL_1	sup_count
laptop	5
camera	4
hard-drive	4
DVD	4

CL_2	sup_count
laptop, camera	4
laptop, hard-drive	2
laptop, DVD	1
camera, hard-drive	2
camera, DVD	0
hard-drive, DVD	1

FL_2	sup_count
laptop, camera	4

CL_3	sup_count
laptop, camera, hard-drive	2
laptop, camera, TV	1

b.

Rules	Confidence
{laptop}->{camera}	$= \frac{SC(laptop, camera)}{SC(laptop)} = \frac{4}{5} = 0.8$
{camera}->{laptop}	$= \frac{SC(camera, laptop)}{SC(camera)} = \frac{4}{4} = 1$

3)

a.

Rule	rule	Probability
P(A|+)	$= \frac{P(+,A)}{P(+)} = \frac{3}{5} =$	0.6
P(B|+)	$= \frac{P(+,B)}{P(+)} = \frac{1}{5} =$	0.2
P(C|+)	$= \frac{P(+,C)}{P(+)} = \frac{4}{5} =$	0.8
P(A|-)	$= \frac{P(-,A)}{P(-)} = \frac{2}{5} =$	0.4
P(B|-)	$= \frac{P(-,B)}{P(-)} = \frac{2}{5} =$	0.4
P(C|-)	$= \frac{P(-,C)}{P(-)} = \frac{5}{5} =$	1

b.

$$Let\ test\ X = (A = 0, B = 1, C = 0)$$

$$classification\ of\ X = the\ larger\ of\ P(X|+)P(+) and\ P(X|-)P(-)$$

$$P(X|+)P(+) = \left(\frac{2}{5} * \frac{1}{5} * \frac{1}{5}\right) * \left(\frac{5}{10}\right) \qquad P(X|-)P(-) = \left(\frac{3}{5} * \frac{2}{5} * \frac{0}{5}\right) * \left(\frac{5}{10}\right)$$

$$= (0.016) * (0.5) \qquad = (0) * (0.5)$$

$$0.008 > 0$$

$$therefore\ in\ accordance\ to\ Bayes\ approach\ X\ would\ be\ classified\ as\ " + "$$

4)

a.

Original	<{1,3}{3}{2,3}{4,5}>
Count	**4-subsequences**
1	{1}{3}{2}{4}
2	{1}{3}{2}{5}
3	{1}{3}{3}{4}
4	{1}{3}{3}{5}
5	{3}{3}{2}{4}
6	{3}{3}{2}{5}
7	{3}{3}{3}{4}
8	{3}{3}{3}{5}
9	{1,3}{3}{2}
10	{1,3}{3}{3}

b.

Original	<{1,2,3}{2,3}{2,3,5}{4}>
Count	**3-element sub-sequences**
1	{1}{2}{2}
2	{1}{2}{3}
3	{1}{2}{5}
4	{1}{3}{2}
5	{1}{3}{3}
6	{1}{3}{5}
7	{2}{2}{2}
8	{2}{2}{3}
9	{2}{2}{5}
10	{2}{3}{2}