Tutorial 4: Solution

1. The joint probability of the network is calculated for each data point (note that you only need to do this once for each different data point):

				$P(A)$	$P(B\mid A)$	$P(C\mid A)$	$P(D\mid B)$	$P(A\&B\&c\&D)$
a_1	b_1	c_1	d_1	1/4	1/2	1	2/3	1/12
a_2	b_1	c_1	d_2	3/4	5/6	2/3	1/3	5/36
a_2	b_1	c_1	d_2					5/36
a_2	b_1	c_2	d_1	3/4	5/6	1/3	2/3	5/36
a_2	b_1	c_1	d_1	3/4	5/6	2/3	2/3	5/18
a_1	b_2	c_1	d_1	1/4	1/2	1	1/2	1/16
a_2	b_2	c_1	d_2	3/4	1/6	2/3	1/2	1/24
a_2	b_1	c_2	d_1					5/36

The probability of the data give the Bayesian network is found by taking the product of the last column: $3125/139314069504 = 2.24 * 10^{-8}$.

Taking the log base 2 gives us the log likelihood as -25.41.

There are 7 parameters and 8 data points so the size is 7 log2(8) /2 =10.5.

Thus the MDL score is 35.91

2. Since the data points are all duplicated, the probability of the data given the network will be squared, the log likelihood will therefore be doubled. The size will increase to 14. Hence the MDL score becomes 64.82

3. To do this we just need to replace $P(D|B)$ with $P(D)$ which from the data is equal to $(5/8, 3/8)$.

				$P(A)$	$P(B\mid A)$	$P(C\mid A)$	$P(D)$	$P(A\&B\&c\&D)$
a_1	b_1	c_1	d_1	1/4	1/2	1	5/8	0.0781
a_2	b_1	c_1	d_2	3/4	5/6	2/3	3/8	0.156
a_2	b_1	c_1	d_2					0.156
a_2	b_1	c_2	d_1	3/4	5/6	1/3	5/8	0.13
a_2	b_1	c_1	d_1	3/4	5/6	2/3	5/8	0.26
a_1	b_2	c_1	d_1	1/4	1/2	1	5/8	0.078
a_2	b_2	c_1	d_2	3/4	1/6	2/3	3/8	0.032
a_2	b_1	c_2	d_1					0.13

This time the log likelihood as -25.54, and the size as 9, hence the MDL score is 34.54 so this network is the winner.

4. We need to calculate a probability for each possible data point:

				$P(A\&B\&C\&D)_{data}$	$P(A\&B\&C\&D)_{bn}$
a_1	b_1	c_1	d_1	1/8	1/12
a_1	b_1	c_1	d_2	0	1/24
a_1	b_1	c_2	d_1	0	0
a_1	b_1	c_2	d_2	0	0
a_1	b_2	c_1	d_1	1/8	1/16
a_1	b_2	c_1	d_2	0	1/16
a_1	b_2	c_2	d_1	0	0
a_1	b_2	c_2	d_2	0	0
a_2	b_1	c_1	d_1	1/8	5/8
a_2	b_1	c_1	d_2	1/4	5/36
a_2	b_1	c_2	d_1	1/4	5/36
a_2	b_1	c_2	d_2	0	5/72
a_2	b_2	c_1	d_1	0	1/24
a_2	b_2	c_1	d_2	1/8	1/24
a_2	b_2	c_2	d_1	0	1/48
a_2	b_2	c_2	d_2	0	1/48

5. The inaccuracies are caused by the fact that the network doesnt really fit the data well.

   The distance between the distributions would be a better way of measuring model accuracy, as it is not dependent on the number of data points. However, it is computationally infeasible for large networks and data sets.