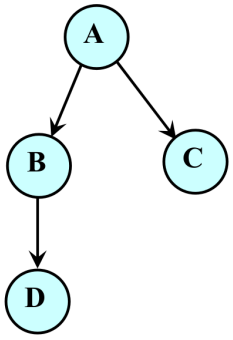
Tutorial 4: Model Accuracy

A data set over four binary variables, A,B, C and D, consists of the following eight data points:

$a_1 \quad b_1 \quad c_1 \quad d_1$
$a_2 \quad b_1 \quad c_1 \quad d_2$
$a_2 \quad b_1 \quad c_1 \quad d_2$
$a_2 \quad b_1 \quad c_2 \quad d_1$
$a_2 \quad b_1 \quad c_1 \quad d_1$
$a_1 \quad b_2 \quad c_1 \quad d_1$
$a_2 \quad b_2 \quad c_1 \quad d_2$
$a_2 \quad b_1 \quad c_2 \quad d_1$

A Bayesian network is proposed to represent the data set.



$$P(A) = [1/4, 3/4]$$

$$P(B|A) = \begin{bmatrix} 1/2 & 5/6 \\ 1/2 & 1/6 \end{bmatrix}$$

$$P(C|A) = \begin{bmatrix} 1 & 2/3 \\ 0 & 1/3 \end{bmatrix}$$

$$P(D|B) = \begin{bmatrix} 2/3 & 1/2 \\ 1/3 & 1/2 \end{bmatrix}$$

The prior probabilities and the link matrices are calculated from the data itself.

1. Compute the MDL score for the above network and data set.
2. If the data set were duplicated exactly (so it contains 16 points) what would be the new value of the MDL score. (PS Dont try to calculate it directly).
3. A different network is proposed with the arc between B and D deleted from the above network. Find the MDL score for this network and determine which of the two is the better network.
4. For a four binary variable problem like this there are only sixteen possible different data points. Calculate the probability distribution over these sixteen points given:
   (a) The data set
   (b) The Bayesian network defined above.
5. Discuss with your friends or the tutors the reasons why these two are different, and why MDL is used as a measure of accuracy rather than calculating a distance (say the Euclidian distance) between the data probability distribution and the network probability distribution