1. Let \( \#(\text{cond}) \) denote the number of samples in training data for which the condition \( \text{cond} \) is true.

a) The maximum likelihood parameter are now given by

\[
P(Y = k) = \hat{\theta}_{Y=k} = \frac{\#(Y = k)}{\#(Y = 1) + \#(Y = 2) + \#(Y = 3)}
\]

and

\[
P(X_i = k \mid Y = j) = \hat{\theta}_{X_i=k,Y=j} = \frac{\#(X_i = k, Y = j)}{\#(X_i = 0, Y = j) + \#(X_i = 1, Y = j)}.
\]

Using these we get the following values:

- \( P(Y = 1) = 0.6000 \)
- \( P(Y = 2) = 0.3000 \)
- \( P(Y = 3) = 0.1000 \)
- \( P(X_1 = 1 \mid Y = 1) = 0.6667 \)
- \( P(X_1 = 1 \mid Y = 2) = 0.3333 \)
- \( P(X_1 = 1 \mid Y = 3) = 1.0000 \)
- \( P(X_2 = 1 \mid Y = 1) = 0.3333 \)
- \( P(X_2 = 1 \mid Y = 2) = 1.0000 \)
- \( P(X_2 = 1 \mid Y = 3) = 0.1667 \)
- \( P(X_3 = 1 \mid Y = 1) = 0.6667 \)
- \( P(X_3 = 1 \mid Y = 2) = 0.0000 \)
- \( P(X_4 = 1 \mid Y = 1) = 0.8333 \)
- \( P(X_4 = 1 \mid Y = 2) = 0.0000 \)
- \( P(X_4 = 1 \mid Y = 3) = 0.0000 \)

b) For the BDeu prior with equivalent sample size 1 we have \( \alpha_{ijk} = 1/6 \) and \( \alpha_k = 1/3 \) for \( X_i \)'s, and \( \alpha_{ijk} = 1/3 \) and \( \alpha_{jk} = 1 \) for \( Y \). Thus the expected parameters are given by

\[
P(Y = k) = E[\theta_{Y=k} \mid D] = \frac{\#(Y = k) + 1/3}{\#(Y = 1) + \#(Y = 2) + \#(Y = 3) + 1}
\]

and

\[
P(X_i = k \mid Y = j) = E[\theta_{X_i=k,Y=j} \mid D] = \frac{\#(X_i = k, Y = j) + 1/6}{\#(X_i = 0, Y = j) + \#(X_i = 1, Y = j) + 1/3}.
\]

Using these we get the following values:

- \( P(Y = 1) = 0.5758 \)
- \( P(Y = 2) = 0.3030 \)
- \( P(Y = 3) = 0.1212 \)
- \( P(X_1 = 1 \mid Y = 1) = 0.6579 \)
- \( P(X_1 = 1 \mid Y = 2) = 0.3500 \)
- \( P(X_1 = 1 \mid Y = 3) = 0.8750 \)
- \( P(X_2 = 1 \mid Y = 1) = 0.3421 \)
- \( P(X_2 = 1 \mid Y = 2) = 0.9500 \)
- \( P(X_2 = 1 \mid Y = 3) = 0.8750 \)
- \( P(X_3 = 1 \mid Y = 1) = 0.1842 \)
- \( P(X_3 = 1 \mid Y = 2) = 0.6500 \)
- \( P(X_3 = 1 \mid Y = 3) = 0.1250 \)
- \( P(X_4 = 1 \mid Y = 1) = 0.8158 \)
- \( P(X_4 = 1 \mid Y = 2) = 0.0500 \)
- \( P(X_4 = 1 \mid Y = 3) = 0.1250 \)
2. a) By using 
\[ P(Y \mid x_1, x_2, x_3, x_4) \propto P(Y)P(x_1 \mid Y)P(x_2 \mid Y)P(x_3 \mid Y)P(x_4 \mid Y) \]
we get the following classification distributions:

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>ML parameters</th>
<th>Expected BDeu parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0.8048</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1.0000</td>
<td>0.9956</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1.0000</td>
<td>0.4115</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1.0000</td>
<td>0.9824</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.1220</td>
<td>0.1228</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1.0000</td>
<td>0.9047</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.0137</td>
<td>0.0192</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.0000</td>
<td>0.6169</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0.7587</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1.0000</td>
<td>0.9908</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1.0000</td>
<td>0.6416</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1.0000</td>
<td>0.9900</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.1220</td>
<td>0.1370</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1.0000</td>
<td>0.8786</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.0526</td>
<td>0.0573</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.0000</td>
<td>0.8048</td>
</tr>
</tbody>
</table>

b) – c) We get the following prediction probabilities $p_i = P(Y = i \mid X)$ and predicted classes $\hat{Y} = \arg \max_i p_i$

<table>
<thead>
<tr>
<th>Test data</th>
<th>ML parameters</th>
<th>Expected BDeu parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$X_2$</td>
<td>$X_3$</td>
</tr>
<tr>
<td>0 0 0 0 1</td>
<td>1.000 0.000 0.000</td>
<td>0.805 0.135 0.060</td>
</tr>
<tr>
<td>0 0 1 0 2</td>
<td>1.000 0.000 0.000</td>
<td>0.412 0.569 0.019</td>
</tr>
<tr>
<td>0 0 1 1 1</td>
<td>1.000 0.000 0.000</td>
<td>0.982 0.016 0.001</td>
</tr>
<tr>
<td>0 1 1 0 3</td>
<td>1.000 0.000 0.000</td>
<td>0.642 0.248 0.110</td>
</tr>
<tr>
<td>1 1 1 2 1</td>
<td>1.000 0.000 0.000</td>
<td>0.805 0.135 0.060</td>
</tr>
</tbody>
</table>

The total losses are therefore:
For ML parameters: $L_{0/1} = 4, L_{log} = \infty$
For expected BDeu parameters: $L_{0/1} = 3, L_{log} = 6.00$

3. a) We want to calculate $P(D \mid M_{NB}) = P(D \mid G_{NB}, \alpha_{NB})$ where $G_{NB}$ is the Naive Bayes structure and $\alpha_{NB}$ are the BDeu prior parameters for equivalent sample size 1.
Using the gamma formula we get $P(D \mid M_{NB}) = 1.08698 \cdot 10^{-19}$.

The same result can also be obtained by taking the product of the predictive probabilities for the data samples given the previous samples. For example:

$P(D_1 \mid G_{NB}, \alpha_{NB}) = 0.020833$
$P(D_2 \mid D_1, G_{NB}, \alpha_{NB}) = 0.0104167$
$P(D_3 \mid D_1, G_{NB}, \alpha_{NB}) = 0.0372179$
$P(D_4 \mid D_1, G_{NB}, \alpha_{NB}) = 0.0279134$
$P(D_5 \mid D_1, G_{NB}, \alpha_{NB}) = 0.00110544$
$P(D_6 \mid D_1, G_{NB}, \alpha_{NB}) = 0.0780451$
$P(D_7 \mid D_1, G_{NB}, \alpha_{NB}) = 0.00635742$
$P(D_8 \mid D_1, G_{NB}, \alpha_{NB}) = 0.00260417$
$P(D_9 \mid D_1, G_{NB}, \alpha_{NB}) = 0.0469773$
$P(D_{10} \mid D_1, G_{NB}, \alpha_{NB}) = 0.00718537$
And the product of the above probabilities is $1.08698 \cdot 10^{19}$.

Or in reverse order:

$$P(D_{10}, G_{NB}, \alpha_{NB}) = 0.0208333$$

$$P(D_0 \mid D_{10}, G_{NB}, \alpha_{NB}) = 0.0104167$$

$$P(D_8 \mid D_{10}, G_{NB}, \alpha_{NB}) = 0.00694444$$

$$P(D_7 \mid D_{10}, G_{NB}, \alpha_{NB}) = 0.00398763$$

$$P(D_6 \mid D_{10}, G_{NB}, \alpha_{NB}) = 0.100595$$

$$P(D_5 \mid D_{10}, G_{NB}, \alpha_{NB}) = 0.0111493$$

$$P(D_4 \mid D_{10}, G_{NB}, \alpha_{NB}) = 0.0159505$$

$$P(D_3 \mid D_{10}, G_{NB}, \alpha_{NB}) = 0.00299533$$

$$P(D_2 \mid D_{10}, G_{NB}, \alpha_{NB}) = 0.00798375$$

$$P(D_1 \mid D_{210}, G_{NB}, \alpha_{NB}) = 0.0422796$$

Again the product is $1.08698 \cdot 10^{19}$.

b) In this case we want to calculate $P(D \mid M_0) = P(D \mid G_0, \alpha_0)$ where $G_0$ is the empty structure and $\alpha_0$ are the BDeu prior parameters with equivalent sample size 1.

Using the gamma formula we get $P(D \mid M_0) = 8.608 \cdot 10^{-20}$.

c)  

$$P(M_{NB} \mid D) = P(M_0 \mid D)$$

$$P(D \mid M_{NB})P(M_{NB}) = P(D \mid M_0)P(M_0)$$

$$\frac{P(M_{NB})}{P(M_0)} \approx 0.791919 \left( \approx \frac{1}{1.26276} \right)$$

4. Let $M_{ij}$ be the number of data samples where $X = i$ and $Y = j$. Thus the total number of samples is $M = M_{00} + M_{01} + M_{10} + M_{11}$. Let the variables be numbered as $1 \rightarrow X$, $2 \rightarrow Y$. Since the variables are binary, we have $r_1 = r_2 = 2$. Let $\alpha$ be the equivalent sample size.

Now for the structure $X \rightarrow Y$ we have $q_1 = 1$, $q_2 = 2$, $\alpha_{111} = \alpha/2$, $\alpha_{112} = \alpha/2$, $\alpha_{11} = \alpha$, $\alpha_{211} = \alpha/4$, $\alpha_{212} = \alpha/2$, $\alpha_{221} = \alpha/4$, $\alpha_{222} = \alpha/4$, $\alpha_{22} = \alpha/2$. Thus we get

$$P(D \mid G_{X \rightarrow Y}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}) + \alpha} \prod_{k=1}^{r_2} \frac{\Gamma(N_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{jk})}$$

$$= \left( \frac{\Gamma(\alpha)}{\Gamma(M + \alpha)} \cdot \left( \frac{\Gamma(M_1 + M_{12} + \alpha/2)}{\Gamma(M/2)} \cdot \frac{\Gamma(M_2 + M_{22} + \alpha/2)}{\Gamma(M/2)} \right) \right)$$

Similarly, for the structure $Y \rightarrow X$ we get

$$P(D \mid G_{Y \rightarrow X}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}) + \alpha} \prod_{k=1}^{r_2} \frac{\Gamma(N_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{jk})}$$

Thus $P(D \mid G_{X \rightarrow Y}) = P(D \mid G_{Y \rightarrow X})$ regardless of the data and equivalent sample size.
5. a) Since the BDeu score is the same for all network in a same equivalence class, we only need to compute the likelihood for one network from each equivalence class. As listed in Problem 3 of Exercise 3, there are 11 equivalence classes. By selecting the first network from each class and computing the log score, that is \( \log P(D \mid G, \alpha = 1) \), we get the following results:

<table>
<thead>
<tr>
<th>Class</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-184.20</td>
</tr>
<tr>
<td>2</td>
<td>-183.29</td>
</tr>
<tr>
<td>3</td>
<td>-186.66</td>
</tr>
<tr>
<td>4</td>
<td>-183.52</td>
</tr>
<tr>
<td>5</td>
<td>-182.62</td>
</tr>
<tr>
<td>6</td>
<td>-185.98</td>
</tr>
<tr>
<td>7</td>
<td>-185.76</td>
</tr>
<tr>
<td>8</td>
<td><strong>-177.45</strong></td>
</tr>
<tr>
<td>9</td>
<td>-180.82</td>
</tr>
<tr>
<td>10</td>
<td>-180.59</td>
</tr>
<tr>
<td>11</td>
<td>-179.91</td>
</tr>
</tbody>
</table>

The highest scoring class 8 contains only one network \( X \rightarrow Y \leftarrow Z \), which therefore maximizes the marginal likelihood. The likelihood of this network is \( P(D \mid G, \alpha) \approx e^{-177.45} \approx 8.599 \cdot 10^{-78} \).

b) The joint posterior distribution of the parameters is

\[
P(\theta \mid D, G, \alpha) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} P(\theta_{ij} \mid N_{ij}, \alpha_{ij})
\]

where the individual (marginal) probability distributions are

\[
P(\theta_{ij} \mid N_{ij}, \alpha_{ij}) = \text{Dir}(N_{ij1} + \alpha_{ij1}, N_{ij2} + \alpha_{ij2}, \ldots, N_{ijr} + \alpha_{ijr}).
\]

For this particular case the joint distribution is

\[
P(\theta \mid D, G, \alpha) = P(\theta_X \mid N_X, \alpha_X) \cdot \prod_{z \in \{0,1\}} P(\theta_Y \mid \alpha_Y) \cdot P(\theta_Z \mid \alpha_Z),
\]

where \( \alpha_X = 1/2 \) and \( \alpha_Y = 1/8 \).

The distributions for individual parameters are now calculated as follows, for example,

\[
P(\theta_X \mid D, \alpha) = P(\theta_X \mid N_X, \alpha_X)
\]

\[
= \text{Dir}(N_{X=0} + \alpha_X = 0, N_{X=1} + \alpha_X = 1)
\]

\[
= \text{Dir}(68 + 1/2, 32 + 1/2) = \text{Beta}(68.5, 32.5)
\]

and

\[
P(\theta_Y \mid X=0, Z=0 \mid D, \alpha) = P(\theta_Y \mid X=0, Z=0 \mid N_{Y,X=0,Z=0} + \alpha_{Y,X=0,Z=0})
\]

\[
= \text{Dir}(N_{Y,X=0,Z=0} + \alpha_{Y,X=0,Z=0}, N_{Y,X=1,Z=0} + \alpha_{Y,X=1,Z=0})
\]

\[
= \text{Dir}(10 + 1/8 + 1/8) = \text{Beta}(10.125, 1.125).
\]

This way we can calculate all the distributions:

\[
P(\theta_X \mid D, \alpha) = \text{Beta}(68.5, 32.5)
\]

\[
P(\theta_Z \mid D, \alpha) = \text{Beta}(17.5, 83.5)
\]

\[
P(\theta_Y \mid X=0, Z=0 \mid D, \alpha) = \text{Beta}(10.125, 1.125)
\]

\[
P(\theta_Y \mid X=1, Z=0 \mid D, \alpha) = \text{Beta}(2.125, 4.125)
\]

\[
P(\theta_Y \mid X=0, Z=1 \mid D, \alpha) = \text{Beta}(13.125, 44.125)
\]

\[
P(\theta_Y \mid X=1, Z=1 \mid D, \alpha) = \text{Beta}(18.125, 8.125)
\]
c) The expected parameters are given by $\theta_{ijk} = \frac{N_{ijk} + \alpha_{ijk}}{N_{ij} + \alpha_{ij}} = \frac{N_{ijk} + \alpha_{ijk}}{\sum(N_{ijk} + \alpha_{ijk})}$.

For example

$$E[\theta_{X=0} | D, \alpha] = \frac{68 + \frac{1}{2}}{68 + \frac{1}{2} + 32 + \frac{1}{2}} \approx 0.678.$$ 

and

$$E[\theta_{X=1} | D, \alpha] = \frac{32 + \frac{1}{2}}{68 + \frac{1}{2} + 32 + \frac{1}{2}} \approx 0.322.$$ 

Thus we get

$$E[\theta_{X=0} | D, \alpha] \approx (0.678, 0.322)$$

$$E[\theta_{X=1} | D, \alpha] \approx (0.173, 0.827)$$

where the first probability is for variable value 0 and the second is for variable value 1.

Octave code that computes the scores listed in a):

```octave
#!/usr/bin/octave -q

% computes BDeu score: log P(D | G, alpha)
% n = number of nodes
% pa = parents for each node
% confs = data configurations
% counts = sample count in data for each configuration in confs
% alpha = equivalent sample size
function score = BDeuScore(n, pa, confs, counts, alpha)
    score = 0;
    for i = 1:n
        % for each variable
        r_i = 2;
        for j = 1:q_i
            % get the sample count for parent configuration j
            jrows = true(length(counts), 1);
            for y = 1:length(pa{i})
                jrows = jrows & (confs(:, pa{i}(y)) == bitget(j - 1, y));
                end
            Nij = sum(counts(jrows));
            % update score
            score = score + gammaln(aij) - gammaln(Nij + aij); 
        end
        for k = 1:r_i
            % get the sample count for configuration j and k
            jrows = jrows & (confs(:, i) == (k - 1));
            Nijk = sum(counts(jrows));
            % update score
            score = score + gammaln(Nijk + aijk) - gammaln(aijk); 
        end
    end
end
```
% data
n = 3;
confs = [0 0 0
0 0 1
0 1 0
0 1 1
1 0 0
1 0 1
1 1 0
1 1 1];
counts = [10 13 1 44 2 18 4 8];

% equivalent sample size
alpha = 1;

% compute the score for one network from each equivalence class
fprintf(‘Class %2d: %.2f\n’, 1, BDeuScore(3, {[]}, confs, counts, alpha));
fprintf(‘Class %2d: %.2f\n’, 2, BDeuScore(3, {[]}, confs, counts, alpha));
fprintf(‘Class %2d: %.2f\n’, 3, BDeuScore(3, {[]}, confs, counts, alpha));
fprintf(‘Class %2d: %.2f\n’, 4, BDeuScore(3, {[1]}, confs, counts, alpha));
fprintf(‘Class %2d: %.2f\n’, 5, BDeuScore(3, {[1]}, confs, counts, alpha));
fprintf(‘Class %2d: %.2f\n’, 6, BDeuScore(3, {[2]}, confs, counts, alpha));
fprintf(‘Class %2d: %.2f\n’, 7, BDeuScore(3, {[2]}, confs, counts, alpha));
fprintf(‘Class %2d: %.2f\n’, 8, BDeuScore(3, {[2]}, confs, counts, alpha));
fprintf(‘Class %2d: %.2f\n’, 9, BDeuScore(3, {[1 2]}, confs, counts, alpha));
fprintf(‘Class %2d: %.2f\n’, 10, BDeuScore(3, {[1 2]}, confs, counts, alpha));
fprintf(‘Class %2d: %.2f\n’, 11, BDeuScore(3, {[1 2]}, confs, counts, alpha));