Probabilistic models, Spring 2013
Exercise 5: Solutions

---

1. Let $\#(cond)$ denote the number of samples in training data for which the condition *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:
$P(Y=1) = 0.6000$
$P(Y=2) = 0.3000$
$P(Y=3) = 0.1000$
$P(X_1 = 1 \mid Y=1) = 0.6667$
$P(X_1 = 1 \mid Y=2) = 0.3333$
$P(X_1 = 1 \mid Y=3) = 1.0000$
$P(X_2 = 1 \mid Y=1) = 0.3333$
$P(X_2 = 1 \mid Y=2) = 1.0000$
$P(X_2 = 1 \mid Y=3) = 1.0000$
$P(X_3 = 1 \mid Y=1) = 0.1667$
$P(X_3 = 1 \mid Y=2) = 0.6667$
$P(X_3 = 1 \mid Y=3) = 0.0000$
$P(X_4 = 1 \mid Y=1) = 0.8333$
$P(X_4 = 1 \mid Y=2) = 0.0000$
$P(X_4 = 1 \mid Y=3) = 0.0000$

b) For the BDeu prior with equivalent sample size 1 we have $\alpha_{ijk} = 1/6$ and $\alpha_{ij} = 1/3$ for X_i:s, and $\alpha_{yjk} = 1/3$ and $\alpha_{yj} = 1$ for Y. Thus the expected parameters are given by

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

and

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

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:

				ML parameters			Expected BDeu parametes		
X_1	X_2	X_3	X_4	$P(Y=1\|X)$	$P(Y=2\|X)$	$P(Y=3\|X)$	$P(Y=1\|X)$	$P(Y=2\|X)$	$P(Y=3\|X)$
0	0	0	0	1.0000	0.0000	0.0000	0.8048	0.1353	0.0599
0	0	0	1	1.0000	0.0000	0.0000	0.9956	0.0020	0.0024
0	0	1	0	1.0000	0.0000	0.0000	0.4115	0.5691	0.0194
0	0	1	1	1.0000	0.0000	0.0000	0.9824	0.0161	0.0015
0	1	0	0	0.1220	0.8780	0.0000	0.1228	0.7542	0.1230
0	1	0	1	1.0000	0.0000	0.0000	0.9047	0.0661	0.0293
0	1	1	0	0.0137	0.9863	0.0000	0.0192	0.9687	0.0122
0	1	1	1	1.0000	0.0000	0.0000	0.6169	0.3705	0.0126
1	0	0	0	1.0000	0.0000	0.0000	0.7587	0.0357	0.2056
1	0	0	1	1.0000	0.0000	0.0000	0.9908	0.0006	0.0087
1	0	1	0	1.0000	0.0000	0.0000	0.6416	0.2484	0.1100
1	0	1	1	1.0000	0.0000	0.0000	0.9900	0.0046	0.0055
1	1	0	0	0.1220	0.2195	0.6585	0.1570	0.2701	0.5729
1	1	0	1	1.0000	0.0000	0.0000	0.8786	0.0180	0.1034
1	1	1	0	0.0526	0.9474	0.0000	0.0573	0.8105	0.1322
1	1	1	1	1.0000	0.0000	0.0000	0.8048	0.1353	0.0599

b) – c) We get the following prediction probabilities $p_i = P(Y = i \mid X)$ and predicted classes $\hat{Y} = \arg\max_i p_i$ as well as the losses $L_{0/1}$ and $L_{\log}$:

Test data					ML parameters						Expected BDeu parametes					
X_1	X_2	X_3	X_4	Y	p_1	p_2	p_3	$\hat{Y}$	$L_{0/1}$	$L_{\log}$	p_1	p_2	p_3	$\hat{Y}$	$L_{0/1}$	$L_{\log}$
0	0	0	0	1	1.000	0.000	0.000	1	0	0	0.805	0.135	0.060	1	0	0.217
0	0	1	0	2	1.000	0.000	0.000	1	1	Inf	0.412	0.569	0.019	2	0	0.564
0	0	1	1	1	1.000	0.000	0.000	1	0	0	0.982	0.016	0.001	1	0	0.018
0	1	1	1	2	1.000	0.000	0.000	1	1	Inf	0.617	0.370	0.013	1	1	0.993
1	0	1	0	3	1.000	0.000	0.000	1	1	Inf	0.642	0.248	0.110	1	1	2.207
1	1	1	1	2	1.000	0.000	0.000	1	1	Inf	0.805	0.135	0.060	1	1	2.000
								Sum:	4	Inf				Sum:	3	5.999

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 α_{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:

$P(D_1, G_{NB}, \alpha_{NB}) = 0.0208333$
$P(D_2 \mid D_{1:1}, G_{NB}, \alpha_{NB}) = 0.0104167$
$P(D_3 \mid D_{1:2}, G_{NB}, \alpha_{NB}) = 0.0372179$
$P(D_4 \mid D_{1:3}, G_{NB}, \alpha_{NB}) = 0.0279134$
$P(D_5 \mid D_{1:4}, G_{NB}, \alpha_{NB}) = 0.00110544$
$P(D_6 \mid D_{1:5}, G_{NB}, \alpha_{NB}) = 0.0780451$
$P(D_7 \mid D_{1:6}, G_{NB}, \alpha_{NB}) = 0.00635742$
$P(D_8 \mid D_{1:7}, G_{NB}, \alpha_{NB}) = 0.00260417$
$P(D_9 \mid D_{1:8}, G_{NB}, \alpha_{NB}) = 0.0469773$
$P(D_{10} \mid D_{1:9}, G_{NB}, \alpha_{NB}) = 0.00718537$

And the product of the above probabilities is $1.08698 \cdot 10^{19}$.

Or in reverse order:

$P(D_{10}, G_{NB}, \alpha_{NB}) = 0.0208333$
$P(D_9 \mid D_{10:10}, G_{NB}, \alpha_{NB}) = 0.0104167$
$P(D_8 \mid D_{9:10}, G_{NB}, \alpha_{NB}) = 0.00694444$
$P(D_7 \mid D_{8:10}, G_{NB}, \alpha_{NB}) = 0.00398763$
$P(D_6 \mid D_{7:10}, G_{NB}, \alpha_{NB}) = 0.100595$
$P(D_5 \mid D_{6:10}, G_{NB}, \alpha_{NB}) = 0.0111493$
$P(D_4 \mid D_{5:10}, G_{NB}, \alpha_{NB}) = 0.0159505$
$P(D_3 \mid D_{4:10}, G_{NB}, \alpha_{NB}) = 0.00299533$
$P(D_2 \mid D_{3:10}, G_{NB}, \alpha_{NB}) = 0.00798375$
$P(D_1 \mid D_{2:10}, G_{NB}, \alpha_{NB}) = 0.0422796$

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

b) In this case we want to calculate $P(D \mid M_\emptyset) = P(D \mid G_\emptyset, \alpha_\emptyset)$ where $G_\emptyset$ is the empty structure and $\alpha_\emptyset$ are the BDeu prior parameters with equivalent sample size 1.
Using the gamma formula we get $P(D \mid M_\emptyset) = 8.608 \cdot 10^{-20}$.

c)

$$\begin{aligned}
P(M_{NB} \mid D) &= P(M_\emptyset \mid D) && \Leftrightarrow \\
P(D \mid M_{NB})P(M_{NB}) &= P(D \mid M_\emptyset)P(M_\emptyset) && \Leftrightarrow \\
\frac{P(M_{NB})}{P(M_\emptyset)} &= \frac{P(D \mid M_\emptyset)}{P(D \mid M_{NB})} \approx 0.791919 \left(\approx \frac{1}{1.26276}\right)
\end{aligned}$$

---

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 $1 \to X$, $2 \to Y$. Since the variables are binary, we have $r_1 = r_2 = 2$. Let α be the equivalent sample size.

Now for the structure $X \to Y$ we have $q_1 = 1$, $q_2 = 2$, $\alpha_{111} = \alpha/2$, $\alpha_{112} = \alpha/2$, $\alpha_{11} = \alpha$, $\alpha_{211} = \alpha/4$, $\alpha_{212} = \alpha/4$, $\alpha_{21} = \alpha/2$, $\alpha_{221} = \alpha/4$, $\alpha_{222} = \alpha/4$, $\alpha_{22} = \alpha/2$. Thus we get

$$\begin{aligned}
P(D \mid G_{X\to Y}) &= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij} + \alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})} \\
&= \left(\frac{\Gamma(\alpha)}{\Gamma(M+\alpha)} \cdot \left(\frac{\Gamma(M_{11} + M_{12} + \alpha/2)}{\Gamma(\alpha/2)} \cdot \frac{\Gamma(M_{21} + M_{22} + \alpha/2)}{\Gamma(\alpha/2)} \right) \right) \cdot \\
&\quad \left[\left(\frac{\Gamma(\alpha/2)}{\Gamma(M_{11} + M12 + \alpha/2)} \cdot \left(\frac{\Gamma(M_{11} + \alpha/4)}{\Gamma(\alpha/4)} \cdot \frac{\Gamma(M_{12} + \alpha/4)}{\Gamma(\alpha/4)} \right) \right) \cdot \right. \\
&\quad \left. \left(\frac{\Gamma(\alpha/2)}{\Gamma(M_{21} + M22 + \alpha/2)} \cdot \left(\frac{\Gamma(M_{21} + \alpha/4)}{\Gamma(\alpha/4)} \cdot \frac{\Gamma(M_{22} + \alpha/4)}{\Gamma(\alpha/4)} \right) \right) \right] \\
&= \frac{\Gamma(\alpha)}{\Gamma(M+\alpha)} \cdot \frac{\Gamma(M_{11} + \alpha/4)}{\Gamma(\alpha/4)} \cdot \frac{\Gamma(M_{12} + \alpha/4)}{\Gamma(\alpha/4)} \cdot \frac{\Gamma(M_{21} + \alpha/4)}{\Gamma(\alpha/4)} \cdot \frac{\Gamma(M_{22} + \alpha/4)}{\Gamma(\alpha/4)}.
\end{aligned}$$

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

$$P(D \mid G_{Y\to X}) = \frac{\Gamma(\alpha)}{\Gamma(M+\alpha)} \cdot \frac{\Gamma(M_{11} + \alpha/4)}{\Gamma(\alpha/4)} \cdot \frac{\Gamma(M_{12} + \alpha/4)}{\Gamma(\alpha/4)} \cdot \frac{\Gamma(M_{21} + \alpha/4)}{\Gamma(\alpha/4)} \cdot \frac{\Gamma(M_{22} + \alpha/4)}{\Gamma(\alpha/4)}.$$

Thus $P(D \mid G_{X\to Y}) = P(D \mid G_{Y\to X})$ 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:

Class	Score
1	-184.20
2	-183.29
3	-186.66
4	-183.52
5	-182.62
6	-185.98
7	-185.76
8	**-177.45**
9	-180.82
10	-180.59
11	-179.91

The highes 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

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

For this particular case the joint distribution is

$$P(\theta \mid D, G, \alpha) = P(\theta_X \mid N_X, \alpha_X) \cdot \prod_{\substack{x \in \{0,1\} \\ z \in \{0,1\}}} P(\theta_{Y|X=x,Z=z} \mid N_{Y,X=x,Z=z}, \alpha_{Y|X=x,Z=z}) \cdot P(\theta_Z \mid N_Z, \alpha_Z),$$

where $\alpha_X = 1/2$ and $\alpha_{Y|X=x,Z=z} = 1/8$.
The distributions for individual parameters are now calculated as follows, for example,

$$\begin{aligned} P(\theta_X \mid D, \alpha) &= P(\theta_X \mid N_X, \alpha_X) \\ &= \text{Dir}(N_{X=0} + \alpha_{X=0}, N_{X=1} + \alpha_{X=1}) \\ &= \text{Dir}(68 + 1/2, 32 + 1/2) = \text{Beta}(68.5, 32.5) \end{aligned}$$

and

$$\begin{aligned} P(\theta_{Y|X=0,Z=0} \mid D, \alpha) &= P(\theta_{Y|X=0,Z=0} \mid N_{Y,X=0,Z=0}, \alpha_{Y,X=0,Z=0}) \\ &= \text{Dir}(N_{Y=0,X=0,Z=0} + \alpha_{Y=0|X=0,Z=0}, N_{Y=1,X=0,Z=0} + \alpha_{Y=1|X=0,Z=0}) \\ &= \text{Dir}(10 + 1/8, 1 + 1/8) = \text{Beta}(10.125, 1.125). \end{aligned}$$

This way we can calculate all the distributions:
$P(\theta_X \mid D, \alpha) = \text{Beta}(68.5, 32.5)$
$P(\theta_Z \mid D, \alpha) = \text{Beta}(17.5, 83.5)$
$P(\theta_{Y|X=0,Z=0} \mid D, \alpha) = \text{Beta}(10.125, 1.125)$
$P(\theta_{Y|X=1,Z=0} \mid D, \alpha) = \text{Beta}(2.125, 4.125)$
$P(\theta_{Y|X=0,Z=1} \mid D, \alpha) = \text{Beta}(13.125, 44.125)$
$P(\theta_{Y|X=1,Z=1} \mid D, \alpha) = \text{Beta}(18.125, 8.125)$

c) The expected parameters are given by $\theta_{ijk} = \frac{N_{ijk}+\alpha_{ijk}}{N_{ij}+\alpha_{ij}} = \frac{N_{ijk}+\alpha_{ijk}}{\sum_k (N_{ijk}+\alpha_{ijk})}$.
For example

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

and

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

Thus we get
$\mathbf{E}[\theta_X \mid D, \alpha] \approx (0.678,\ 0.322)$
$\mathbf{E}[\theta_Z \mid D, \alpha] \approx (0.173,\ 0.827)$
$\mathbf{E}[\theta_{Y|X=0,Z=0} \mid D, \alpha] \approx (0.900,\ 0.100)$
$\mathbf{E}[\theta_{Y|X=1,Z=0} \mid D, \alpha] \approx (0.340,\ 0.660)$
$\mathbf{E}[\theta_{Y|X=0,Z=1} \mid D, \alpha] \approx (0.229,\ 0.771)$
$\mathbf{E}[\theta_{Y|X=1,Z=1} \mid D, \alpha] \approx (0.690,\ 0.310)$
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):

```
#!/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 each variable
    for i = 1:n
        % for each parent configuration
        qi = 2^length(pa{i});
        aij = alpha / qi;
        for j = 1:qi
            % 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);
            % for each value of variable i
            ri = 2;
            aijk = aij / ri;
            for k = 1:ri
                % get the sample count for configuration j and k
                jirows = jrows & (confs(:,i) == (k-1));
                Nijk = sum(counts(jirows));
                % update score
                score = score + gammaln(Nijk + aijk) - gammaln(aijk);
            end
        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, {[],[1],[]}, confs, counts, alpha));
fprintf('Class %2d: %.2f\n', 3, BDeuScore(3, {[],[],[1]}, confs, counts, alpha));
fprintf('Class %2d: %.2f\n', 4, BDeuScore(3, {[],[],[2]}, confs, counts, alpha));
fprintf('Class %2d: %.2f\n', 5, BDeuScore(3, {[],[1],[2]}, confs, counts, alpha));
fprintf('Class %2d: %.2f\n', 6, BDeuScore(3, {[],[3],[1]}, confs, counts, alpha));
fprintf('Class %2d: %.2f\n', 7, BDeuScore(3, {[2],[],[1]}, confs, counts, alpha));
fprintf('Class %2d: %.2f\n', 8, BDeuScore(3, {[],[1 3],[]}, confs, counts, alpha));
fprintf('Class %2d: %.2f\n', 9, BDeuScore(3, {[],[],[1 2]}, confs, counts, alpha));
fprintf('Class %2d: %.2f\n', 10, BDeuScore(3, {[2 3],[],[]}, confs, counts, alpha));
fprintf('Class %2d: %.2f\n', 11, BDeuScore(3, {[],[1],[1 2]}, confs, counts, alpha));
```