Chapter 3. Generative models for discrete data

3.1.
$\log p(D|\theta) = N_1 \log\theta + N_0 \log(1-\theta)$.

$$\frac{d}{d\theta}\log p(D|\theta) = 0 \Leftrightarrow \frac{N_1}{\theta} = \frac{N_0}{1-\theta} \Leftrightarrow \theta = \frac{N_1}{N_0+N_1}.$$

3.2.

$$\begin{aligned} p(D) &= \frac{(\alpha_1)\cdots(\alpha_1+N_1-1)(\alpha_0)\cdots(\alpha_0+N_0-1)}{(\alpha)\cdots(\alpha+N-1)} \\ &= \frac{\Gamma(\alpha_1+N_1)}{\Gamma(\alpha_1)}\frac{\Gamma(\alpha_0+N_0)}{\Gamma(\alpha_0)}\frac{\Gamma(\alpha_1+\alpha_0)}{\Gamma(\alpha_1+\alpha_0+N_0+N_1)}. \end{aligned}$$

3.3.

$$p(\tilde{x}=1|n=1,D) = \frac{\mathrm{B}(\alpha_1'+1,\alpha_0')}{\mathrm{B}(\alpha_1',\alpha_0')} = \frac{\Gamma(\alpha_1'+1)\Gamma(\alpha_0')}{\Gamma(\alpha_1'+\alpha_0'+1)}\frac{\Gamma(\alpha_0'+\alpha_1')}{\Gamma(\alpha_0')\Gamma(\alpha_1')} = \frac{\alpha_1'}{\alpha_0'+\alpha_1'}.$$

3.4.

$$p(\theta|X<3) = \frac{p(X<3|\theta)p(\theta)}{\int_\theta p(X<3|\theta)p(\theta)d\theta} = \frac{(6\theta^2+3\theta+1)(1-\theta)^3}{\int_0^1(6\theta^2+3\theta+1)(1-\theta)^3d\theta} = (1-\theta)^3(12\theta^2+6\theta+2).$$

This satisfies $p(\theta|X<3) \propto p(\theta, X<3)$.

3.5.

$$p_\theta(\theta) = |\frac{d\phi}{d\theta}|p_\phi(\phi) \propto |\frac{d\phi}{d\theta}| = \mathrm{Beta}(0,0) = \frac{1}{\theta(1-\theta)}.$$

3.6.
For observed samples $x_1, \cdots, x_n$, the likelihood is

$$\prod_i \mathrm{Poi}(x_i|\lambda) = e^{-\lambda n}\frac{\lambda^{\sum_i x_i}}{\prod_i x_i!}.$$

The log-likelihood is $l = -\lambda n + \sum_i x_i \log \lambda - \sum_i \log(x_i!\,)$.

$$\frac{dl}{d\lambda} = 0 \Leftrightarrow \lambda = \frac{\sum_i x_i}{n}.$$

3.7.
a. Let $D = \{x_1, \cdots, x_n\}$ be i.i.d samples from $\mathrm{Poi}(x|\lambda)$.
Likelihood:

$$p(D|\lambda) = e^{-n\lambda} \frac{\lambda^{\sum_i x_i}}{\prod_i x_i!}$$

Prior:

$$p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-\lambda b}$$

Posterior:

$$p(\lambda|D) \propto p(D|\lambda)p(\lambda) \propto e^{-\lambda(b+n)} \lambda^{(a+\sum_i x_i)-1} \propto \mathrm{Ga}(\lambda|a + \sum_i x_i, b+n)$$

Therefore, $p(\lambda|D) = \mathrm{Ga}(\lambda|a + \sum_i x_i, b+n)$.

b.

$$\hat{\lambda} = \frac{a + \sum_i x_i}{b+n}.$$

As $a, b \to 0$, this converges to $\frac{a}{b}$, which is equal to MLE.

3.8
a. Let $x_1 \leq x_2 \leq \cdots \leq x_n$ be order statistics of samples.
Likelihood function is

$$p(D|a) = (2a)^{-n} \mathbf{1}_{\{a : a \geq \max(-x_1, x_n)\}}.$$

Since $\frac{dp(D|a)}{da} < 0$ on $\{a : a \geq \max(-x_1, x_n)\}$, the likelihood is maximized at $\hat{a} = \max(-x_1, x_n)$.

b. The probability is 1 if $x_{n+1} \in [-\hat{a}, \hat{a}]$ and 0 otherwise.

c. This is unsuitable for predicting future data, because it puts zero probability outside the training data.

3.9.
Likelihood:

$$p(\mathcal{D}|\theta) = \frac{1}{\theta^n}\mathbf{1}_{\{\theta:\theta\geq m\}}$$

Prior:

$$p(\theta) = \text{Pareto}(\theta|K, b) = \frac{Kb^k}{\theta^{k+1}}\mathbf{1}_{\{\theta:\theta\geq b\}}$$

Posterior:

$$\begin{aligned} p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{\frac{Kb^k}{\theta^{n+k+1}}\mathbf{1}_{\{\theta:\theta\geq\max(b,m)\}}}{\int_{\max(b,m)}^{\infty}\frac{Kb^k}{\theta^{n+k+1}}\mathcal{D}\theta} = \frac{\frac{Kb^k}{\theta^{n+k+1}}\mathbf{1}_{\{\theta:\theta\geq\max(b,m)\}}}{\frac{Kb^k}{(n+k)\max(b.m)^{n+k}}} \\ = \frac{(n+k)\max(b,m)^{n+k}}{\theta^{n+k+1}}\mathbf{1}_{\{\theta:\theta\geq\max(b,m)\}} = \text{Pareto}(\theta|k+n, \max(b,m)). \quad (1) \end{aligned}$$

3.10.
a.

$$p(\theta|\mathcal{D}) = \text{Pareto}(\theta|1, 100) = \frac{100}{\theta^2}\mathbf{1}_{\{\theta\geq 100\}}.$$

b. Mean: Does not exist.
Mode: 100.
Median: 200.

c.

$$\begin{aligned} p(x|\mathcal{D}, \alpha) = \int p(x|\theta)p(\theta|\mathcal{D}, \alpha)d\theta = \int \frac{1}{\theta}\mathbf{1}_{\{\theta:0\leq x\leq\theta\}}\text{Pareto}(\theta|1, m)d\theta \\ = \int \frac{m}{\theta^3}\mathbf{1}_{\{\theta:\theta\geq\max(x,m)\}}d\theta = \int_{\max(x,m)}^{\infty}\frac{m}{\theta^3}d\theta = \frac{m}{2\max(x,m)^2}\mathbf{1}_{\{x:x\geq 0\}}. \quad (2) \end{aligned}$$

d.
$p(x = 100|\mathcal{D}, \alpha) = \frac{1}{200}$.
$p(x = 50|\mathcal{D}, \alpha) = \frac{1}{200}$.
$p(x = 150|\mathcal{D}, \alpha) = \frac{1}{225}$.

3.11.
a.

Log-likelihood:

$$\log p(\mathcal{D}|\theta) = \log(\theta^n e^{-\theta \sum_i x_i}) = n \log \theta - \theta \sum_i x_i.$$

$$\frac{d \log p(\mathcal{D}|\theta)}{d\theta} = \frac{n}{\theta} - \sum_i x_i \Rightarrow \hat{\theta} = \frac{n}{\sum_i x_i} = \frac{1}{\bar{x}}.$$

b. $\hat{\theta} = \frac{1}{5}$.

c. $\text{Expon}(\theta|\lambda) = \text{Ga}(\theta|1, \lambda)$. Since we seek $\hat{\lambda}$ such that $\frac{1}{\hat{\lambda}} = \frac{1}{3}$, $\hat{\lambda} = 3$.

d.

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta) = \theta^n e^{-\theta \sum_i x_i} \cdot 3e^{-3\theta} \propto \text{Ga}(\theta|n+1, \sum_i x_i + 3).$$

Therefore, $p(\theta|\mathcal{D}, \hat{\lambda}) = \text{Ga}(\theta|n+1, \sum_i x_i + 3) = \text{Ga}(\theta|4, 18)$.

e. Yes.

f.

$$\frac{n+1}{\sum_i x_i + 3} = \frac{4}{18} = \frac{2}{9}.$$

g. Because our prior changed the prediction. Using MLE seems more reasonable than using Bayesian approach, because we don't know much informations other than given data.

3.12.
a.

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta} = \frac{\theta^{N_1}(1-\theta)^{N_0}}{0.5^{N_1+N_0} + 0.4^{N_1}0.6^{N_0}}$$

for $\theta = 0.5$ or 0.4, and $p(\theta|\mathcal{D}) = 0$ otherwise.
To calculate MAP estimate for θ,

$$p(\theta = 0.5|\mathcal{D}) > p(\theta = 0.4|\mathcal{D}) \Leftrightarrow 0.5^{N_1+N_0} > 0.4^{N_1}0.6^{N_0} \Leftrightarrow \frac{N_1}{N_0} > \frac{\log 1.2}{\log 1.25} \approx 0.4497.$$

Therefore, $\hat{\theta} = 0.5$ if $\frac{N_1}{N} > \frac{\log 1.2}{\log 1.5}$ and $\hat{\theta} = 0.4$ if $\frac{N_1}{N} < \frac{\log 1.2}{\log 1.5}$. (There are no cases in which $\frac{N_1}{N} = \frac{\log 1.2}{\log 1.5}$, since both N, N_1 are integers.)

b. When N is small, the two-point prior would probably give better estimation, because the MAP estimate based on conjugate Beta prior would heavily rely on prior parameter values. When N is large, the Beta prior would probably give better estimation, since by the law of large numbers, the MAP estimate would converge to the true parameter within extremely narrow error.

3.13
The posterior predictive is

$$p(\tilde{\mathcal{D}}|\mathcal{D},\boldsymbol{\alpha}) = \int_{S_K} p(\tilde{\mathcal{D}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} = \int_{S_K} (\prod_k \theta_k^{N_k^{\text{new}}})\frac{1}{\mathrm{B}(\boldsymbol{\alpha}+\boldsymbol{N}^{\textbf{old}})}(\prod_k \theta_k^{\alpha_k+N_k^{\text{old}}-1})d\boldsymbol{\theta}$$

$$= \frac{1}{\mathrm{B}(\boldsymbol{\alpha}+\boldsymbol{N}^{\textbf{old}})}\int_{S_K}\prod_k \theta_k^{N_k^{\text{new}}+N_k^{\text{old}}+\alpha_k-1}d\boldsymbol{\theta}.$$

Now for simplicity, denote $\gamma_k = N_k^{\text{new}} + N_k^{\text{old}} + \alpha_k$.
To compute the integral, first observe the following property: for any $a \neq 0$ and $m, n > 0$, one has

$$\frac{1}{a^{m+n-1}}\int_0^a y^{m-1}(a-y)^{n-1}dy = \frac{1}{a}\int_0^a (\frac{y}{a})^{m-1}(1-\frac{y}{a})^{n-1}dy$$
$$= \int_0^1 x^{m-1}(1-x)^{n-1}dx = \mathrm{B}(m,n). \quad (3)$$

Therefore, the following holds:

$$\int_0^a y^{m-1}(a-y)^{n-1}dy = a^{m+n-1}\mathrm{B}(m,n).$$

Now go back to compute our main integral:

$$\begin{aligned}
\mathrm{B}(\boldsymbol{\alpha}+\boldsymbol{N}^{\mathbf{old}})p(\tilde{\mathcal{D}}|\mathcal{D},\boldsymbol{\alpha}) &= \int_{S_K}\prod_k \theta_k^{\gamma_k-1}d\boldsymbol{\theta} \\
&= \int_0^1 \theta_1^{\gamma_1-1}\int_0^{1-\theta_1}\theta_2^{\gamma_2-1}\cdots\int_0^{1-\theta_1-\cdots-\theta_{K-1}}\theta_K^{\gamma_K-1}d\theta_K\cdots d\theta_2 d\theta_1 \\
&= \frac{1}{\gamma_K}\int_0^1 \theta_1^{\gamma_1-1}\int_0^{1-\theta_1}\theta_2^{\gamma_2-1}\cdots\int_0^{1-\theta_1-\cdots-\theta_{K-2}}\theta_{K-1}^{\gamma_{K-1}-1}(1-\theta_1-\cdots-\theta_{K-1})^{\gamma_K}d\theta_{K-1}\cdots d\theta_2 d\theta_1 \\
&= \frac{\mathrm{B}(\gamma_{K-1},\gamma_K+1)}{\gamma_K}\int_0^1 \theta_1^{\gamma_1-1}\int_0^{1-\theta_1}\theta_2^{\gamma_2-1}\cdots\int_0^{1-\theta_1-\cdots-\theta_{K-3}}\theta_{K-2}^{\gamma_{K-2}-1} \\
&\qquad (1-\theta_1-\cdots-\theta_{K-2})^{\gamma_{K-1}+\gamma_K}d\theta_{K-2}\cdots d\theta_2 d\theta_1 = \cdots \\
&= \frac{\mathrm{B}(\gamma_{K-1},\gamma_K+1)\mathrm{B}(\gamma_{K-2},\gamma_K+\gamma_{K-1}+1)\cdots\mathrm{B}(\gamma_1,\gamma_K+\cdots+\gamma_2+1)}{\gamma_K} \\
&= \frac{\Gamma(\gamma_1)\Gamma(\gamma_2)\cdots\Gamma(\gamma_K)}{\Gamma(\sum_i\gamma_i)\sum_i\gamma_i} = \frac{\mathrm{B}(\boldsymbol{\gamma})}{\sum_i\gamma_i}
\end{aligned} \tag{4}$$

Calculating this gives

$$p(\tilde{\mathcal{D}}|\mathcal{D},\boldsymbol{\alpha}) = \frac{\mathrm{B}(\boldsymbol{\gamma})}{\mathrm{B}(\boldsymbol{\alpha}+\boldsymbol{N}^{\mathbf{old}})\sum_i\gamma_i} = \frac{\mathrm{B}(\boldsymbol{\alpha}+\boldsymbol{N}^{\mathbf{old}}+\boldsymbol{N}^{\mathbf{new}})}{\mathrm{B}(\boldsymbol{\alpha}+\boldsymbol{N}^{\mathbf{old}})\sum_i(N_i^{\mathrm{new}}+N_i^{\mathrm{old}}+\alpha_i)} \tag{5}$$

3.14.
a. By the equation 3.51,

$$p(x_{2001}=e|\mathcal{D}) = \frac{270}{2270} \approx 0.1189.$$

b.

$$p(x_{2001}=p,x_{2002}=a|\mathcal{D}) = p(x_{2001}=p|\mathcal{D})p(x_{2002}=a|\mathcal{D}) = \frac{110}{2270}\cdot\frac{97}{2270} \approx 0.00207.$$

3.15.

$$\frac{\alpha_1}{\alpha_1+\alpha_2} = m, \frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)} = v.$$

Solving this for α_1, α_2 gives

$$\alpha_1 = \frac{m^2(1-m)}{v} - m, \alpha_2 = \frac{m(1-m)^2}{v} - (1-m).$$

If $m = 0.7$ and $v = 0.2^2$, $\alpha_1 = \frac{119}{40}, \alpha_2 = \frac{51}{40}$.

3.16.
Since $\alpha_2 = \alpha_1 \frac{1-m}{m}$, we have

$$\int_l^u \frac{x^{\alpha_1 - 1}(1-x)^{\alpha_1 \frac{1-m}{m} - 1}}{\mathrm{B}(\alpha_1, \frac{1-m}{m}\alpha_1)} dx = 0.95.$$

Since m, l, u are known and only α_1 is unknown, we can compute α_1. For example, if $m = 0.15, l = 0.05, u = 0.3$, our equation becomes

$$\int_{0.05}^{0.3} \frac{x^{\alpha_1 - 1}(1-x)^{\alpha_1 \frac{0.85}{0.15} - 1}}{\mathrm{B}(\alpha_1, \frac{0.85}{0.15}\alpha_1)} dx = 0.95.$$

Numerical computation gives $\alpha_1 \approx 4.5$ (and $\alpha_2 \approx 25.6$).
This is approximately equivalent to the Beta prior corresponding to sample size $N_1 = 4, N = 29$.

3.17.

$$p(N_1|N) = \int p(N_1|\theta)p(\theta)d\theta = \binom{N}{N_1} \int_0^1 \theta^{N_1}(1-\theta)^{N-N_1} d\theta = \binom{N}{N_1} \mathrm{B}(N_1+1, N-N_1+1)$$
$$= \frac{N!}{N_1!\,(N-N_1)!} \frac{\Gamma(N_1+1)\Gamma(N-N_1+1)}{\Gamma(N+2)} = \frac{1}{N+1}.$$

3.18.

$$BF_{1,0} = \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)}.$$

Since M_0 is the null hypothesis that the coin is fair, the marginal likelihood under M_0 is simply

$$p(\mathcal{D}|M_0) = (\frac{1}{2})^N.$$

The marginal likelihood under M_1 is

$$p(\mathcal{D}|M_1) = \frac{1}{N+1}.$$

Thus the Bayes factor is $BF_{1,0} = \frac{2^N}{N+1} = \frac{2^{10}}{11} \approx 93.1$.
If $N = 100$, the Bayes factor is $BF_{1,0} = \frac{2^{100}}{101} \approx 1.152 \times 10^{29}$.

3.19.
a.

$$\log \frac{p(c=1|\mathbf{x}_i)}{p(c=2|\mathbf{x}_i)} = \log \frac{p(\mathbf{x}_i|c=1)p(c=1)}{p(\mathbf{x}_i|c=2)p(c=2)} = \log \frac{p(\mathbf{x}_i|c=1)}{p(\mathbf{x}_i|c=2)}$$
$$= \log p(\mathbf{x}_i|c=1) - \log p(\mathbf{x}_i|c=2) = \sum_w [x_{iw} \log \frac{\theta_{1w}(1-\theta_{2w})}{\theta_{2w}(1-\theta_{1w})}] - \sum_w \log \frac{1-\theta_{1w}}{1-\theta_{2w}}. \tag{6}$$

b. The posterior odds ratio would be 1, therefore,

$$x_{iw} \log \frac{\theta_{1w}(1-\theta_{2w})}{\theta_{2w}(1-\theta_{1w})} = \log \frac{1-\theta_{1w}}{1-\theta_{2w}}.$$

c. We have

$$\hat{\theta}_{1w} = \frac{1+n_1}{2+n_1}, \hat{\theta}_{2w} = \frac{1+n_2}{2+n_2}.$$

Since $x_{1w} = x_{2w} = 1$, the posterior odds ratio condition becomes

$$\log \frac{\theta_{1w}(1-\theta_{2w})}{\theta_{2w}(1-\theta_{1w})} = \log \frac{1-\theta_{1w}}{1-\theta_{2w}} \Leftrightarrow \frac{\theta_{1w}(1-\theta_{2w})}{\theta_{2w}(1-\theta_{1w})} = \frac{1-\theta_{1w}}{1-\theta_{2w}}.$$

Substituting $\theta_{1w} = \hat{\theta}_{1w}$ and $\theta_{2w} = \hat{\theta}_{2w}$ gives $(n_1+1)(n_1+2) = (n_2+1)(n_2+2)$. Since $f(x) = (x+1)(x+2)$ monotonically increases and $n_1 \neq n_2$, the word would not be ignored.

d. Use multinomial NB model which take accounts the number of occurence of each word in the documents, rather than using Bernoulli NB model.

3.20.
a.

$$p(\mathbf{x}|y = c) = p(x_1|y = c)p(x_2|x_1, y = c) \cdots p(x_D|x_1, \cdots, x_{D-1}, y = c).$$

Here all factors are conditional probability density function which is affected by all of its parent nodes in the Bayesian network, so the number of required parameters are $C + 2C + \cdots + 2^{D-1}C = C(2^D - 1)$.

b. Naive Bayes model would be likely to give lower test error, because it is hard to infer class dependence from small set of data.

c. Full Bayes model would be likely to give lower test error, because N is large enough to discriminate large number of features and the class dependence would become more clear.

d. Naive Bayes: Obviously $O(ND)$.
Full Bayes: The whole training algorithm goes as follows:

```
$p_{i,c} = 0, N_c = 0 \forall i, c$;
for $i = 1 : N$ do
    $c = y_i$;
    $N_c = N_c + 1$;
    $x =$ BinaryToIndex($x_0, \cdots, x_i$);
    $p_{x,c} = p_{x,c} + 1$;
end
$p_{x,c} = \frac{p_{x,c}}{N_c}$;
```

where `BinaryToIndex()` is the one-to-one correspondence which maps a binary sequence $(x_0, \cdots, x_i)$ to a single index $\sum_{k=0}^{i} x_k 2^k$. Since time complexity of this function is at most $O(D)$, the whole time complexity is $O(ND)$.

e. Naive Bayes: $O(CD)$.
Full Bayes: The whole testing algorithm goes as follows:

```
$p_{max} = 0$;
$c_{max} = -1$;
for $c = 1 : C$ do
    if $p_{max} < p_{\texttt{BinaryToIndex}(x_0,\cdots,x_N),c}$ then
        $p_{max} = p_{\texttt{BinaryToIndex}(x_0,\cdots,x_N),c}$;
        $c_{max} = c$;
    end
end
$p_{x,c} = \frac{p_{x,c}}{N_c}$;
```

The algorithm runs within time complexity $O(D) + O(C) = O(\max(C, D))$.

f. Naive Bayes: Since we assume the class features to be conditionally independent, we have $p(x_v) = \sum_y p(y)p(x_v|y)$. The testing algorithm goes as follows:

```
for $c = 1 : C$ do
    $L_c = \log \pi_c$;
    for $j = 1 : D$ do
        if $x_j = 1$ and $x_j$ is visible then
            $L_c = L_c + \log \theta_{jc}$;
            else $L_c = L_c + \log(1 - \theta_{jc})$;
        end
    end
end
$p_c = \exp L_c - \text{logsumexp}(L_:)$;
Return $\text{argmax} x_c p_c$;
```

The time complexity of this algorithm is $O(CD)$.

Full Bayes: We can sort the features so that every visible features are at the left of every hidden features. With these assumptions, the testing algorithm goes as follows:

```
    Sort features;
    $px_v|c = 0, p(x_v) = 0$;
    for $j = 0 : 2^h - 1$ do
        $i =$ BinaryToIndex$(x_0, \cdots, x_v) + j$;
        $p_{x_v|c} = p_{x_v|c} + p_{i,c}$;
    end
    for $c = 1 : C$ do
        $p_{x_v} = p_{x_v} + p(c)p_{x_v|c}$;
    end
    $p_{c|x} = \frac{p(c)p_{x_v|c}}{p(x_v)}$;
    Return argmax$p_{c|x}$;
```

The total runtime is $O(D) + O(2^h v) = O(\max(v + h, 2^h v))$.

3.21.

$$I_j = \sum_{x_j} \sum_{y} p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)}$$
$$\sum_y (p(x_j = 0, y) \log \frac{p(x_j = 0, y)}{p(x_j = 0)p(y)} + p(x_j = 1, y) \log \frac{p(x_j = 1, y)}{p(x_j = 1)p(y)})$$
$$\sum_c \theta_{jc} \pi_c \log \frac{\theta_{jc}}{\theta_j} + (1 - \theta_{jc})\pi_c \log \frac{1 - \theta_{jc}}{1 - \theta_j}. \tag{7}$$

3.22.
$\hat{\theta}_{\text{spam}} = \frac{3}{7}$.
$\hat{\theta}_{\text{secret—spam}} = \frac{2}{3}$.
$\hat{\theta}_{\text{secret—non-spam}} = \frac{1}{4}$.
$\hat{\theta}_{\text{sports—non-spam}} = \frac{1}{2}$.
$\hat{\theta}_{\text{dollar—spam}} = \frac{1}{3}$.