ST451 Bayesian Machine Learning
Week 7

Exercises

1. Consider an observation from a trinomial random variable $x = (x_1, x_2, x_3)$ and parameters $\theta = (\theta_1, \theta_2, \theta_3)$. The likelihood for this observation is proportional to

$$f(x|\theta) = f(x_1, x_2, x_3|\theta_1, \theta_2, \theta_3) = \frac{\Gamma(n+1)}{\Gamma(x_1+1)\Gamma(x_2+1)\Gamma(x_3+1)}\theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3},$$

for $0 < \theta_i < 1$, $x_i \in \{0, 1, \ldots, n\}$, $i \in \{1, 2, 3\}$, $x_1 + x_2 + x_3 = n$ and $\theta_1 + \theta_2 + \theta_3 = 1$. It is also known that $E(x_i) = n\theta_i$.

   (a) Assign the Dirichlet$(\alpha_1, \alpha_2, \alpha_3)$ distribution to θ with

$$\pi(\theta) = \pi(\theta_1, \theta_2, \theta_3) = \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\theta_3^{\alpha_3-1},$$

   for $0 < \theta_i < 1$, $\alpha_i > 0$, $i \in \{1, 2, 3\}$, $\theta_1 + \theta_2 + \theta_3 = 1$. Derive the posterior of θ.

   (b) Find the Jeffreys' prior for θ and use it to obtain the corresponding posterior distribution.

   (c) Let $y = (y_1, y_2, y_3)$ represent a future observation from the same model. Write down the posterior predictive distribution of y based on a prior of your choice. Describe a procedure to simulate from this posterior predictive distribution based on random samples from the posterior distribution of θ.

Answer:

   (a) The likelihood can be written as

$$f(x|\theta) = f(x_1, x_2, x_3|\theta_1, \theta_2, \theta_3) \propto \theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3},$$

   for $0 < \theta_i < 1$, $x_i \in \{0, 1, \ldots, n\}$, $i \in \{1, 2, 3\}$, $x_1 + x_2 + x_3 = n$ and $\theta_1 + \theta_2 + \theta_3 = 1$.
   The prior is written as

$$\pi(\theta) = \pi(\theta_1, \theta_2, \theta_3) \propto \theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\theta_3^{\alpha_3-1},$$

   for $0 < \theta_i < 1$, $\alpha_i > 0$, $i \in \{1, 2, 3\}$, $\theta_1 + \theta_2 + \theta_3 = 1$.
   Hence the posterior is proportional to

$$\pi(\theta_1, \theta_2, \theta_3|x) \propto \theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3}\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\theta_3^{\alpha_3-1} = \theta_1^{x_1+\alpha_1-1}\theta_2^{x_2+\alpha_2-1}\theta_3^{x_3+\alpha_3-1}$$

   which is proportional to pdf of a Dirichlet$(x_1 + \alpha_1, x_2 + \alpha_2, x_3 + \alpha_3)$ for $\theta = (\theta_1, \theta_2, \theta_3)$, with $0 < \theta_i < 1$, $a_i > 0$, $i \in \{1, 2, 3\}$, $\theta_1 + \theta_2 + \theta_3 = 1$.

   (b) In order to find the Jeffreys' prior we need to find Fisher's information matrix. Consider the log-likelihood for $\theta = (\theta_1, \theta_2, \theta_3)$, with $0 < \theta_i < 1$, $a_i > 0$, $i \in \{1, 2, 3\}$, $\theta_1 + \theta_2 + \theta_3 = 1$, which is

$$l(x|\theta_1, \theta_2, \theta_3) = x_1 \log(\theta_1) + x_2 \log(\theta_2) + x_3 \log(\theta_3)$$

   Note also that for

$$\frac{\partial}{\theta_i} l(x|\theta) = \frac{x_i}{\theta_i},\ i = 1, 2, 3$$

$$\frac{\partial^2}{\theta_i^2} l(x|\theta) = -\frac{x_i}{\theta_i^2}\ i = 1, 2, 3$$

and

$$\frac{\partial^2}{\theta_i \theta_j} l(x|\theta) = 0,\ i, j = 1, 2, 3\ i \neq j.$$

Hence the Fisher's information matrix is diagonal with entries

$$[\mathcal{I}(\theta)]_{ii} = -E\left(-\frac{x_i}{\theta_i^2}\right) = \frac{n}{\theta_i},\ i = 1, 2, 3.$$

The Jeffreys' prior is then proportional to

$$\pi^J(\theta_1, \theta_2, \theta_3) \propto \det\big(\mathcal{I}(\theta)\big)^{1/2} \propto \theta_1^{-1/2}\theta_2^{-1/2}\theta_3^{-1/2} = \theta_1^{1/2-1}\theta_2^{1/2-1}\theta_3^{1/2-1},$$

for $\theta = (\theta_1, \theta_2, \theta_3)$, with $0 < \theta_i < 1$, $\alpha_i > 0$, $i \in \{1, 2, 3\}$, $\theta_1 + \theta_2 + \theta_3 = 1$. This can be recognized as a Dirichlet$(1/2, 1/2, 1/2)$ distribution.

Hence, given the result of the previous part, the posterior will be a Dirichlet$(x_1 + 1/2, x_3 + 1/2, x_3 + 1/2)$ distribution.

(c) Note that if $X \sim$ Dirichlet(A, B, C) we can write

$$\int_{\mathcal{X}} \frac{\Gamma(A + B + C)}{\Gamma(A)\Gamma(B)\Gamma(C)} \theta_1^{A-1}\theta_2^{B-1}\theta_3^{C-1} dX = 1$$

where $\mathcal{X}$ denotes the range of X. The above can be re-written as

$$\int_{\mathcal{X}} \theta_1^{A-1}\theta_2^{B-1}\theta_3^{C-1} dX = \frac{\Gamma(A)\Gamma(B)\Gamma(C)}{\Gamma(A + B + C)}$$

The predictive distribution, based on the Jeffreys prior, for $y = (y_1, y_2, y_3)$ such that $y_i \in \{0, 1, \ldots, n\}$, $i \in \{1, 2, 3\}$, $y_1 + y_2 + y_3 = n$ can be written as

$$\begin{aligned} f(y|x) = \int_\Theta f(y|\theta)\pi(\theta|x)d\theta = \int_\Theta \frac{\Gamma(n+1)}{\Gamma(y_1+1)\Gamma(y_2+1)\Gamma(y_3+1)} \theta_1^{y_1}\theta_2^{y_2}\theta_3^{y_3} \\ \frac{\Gamma(x_1 + x_2 + x_3 + 3/2)}{\Gamma(x_1 + 1/2)\Gamma(x_2 + 1/2)\Gamma(x_3 + 1/2)} \theta_1^{x_1+1/2-1}\theta_2^{x_2+1/2-1}\theta_3^{x_3+1/2-1} d\theta = \\ = \frac{\Gamma(n+1)}{\Gamma(y_1+1)\Gamma(y_2+1)\Gamma(y_3+1)} \frac{\Gamma(x_1 + x_2 + x_3 + 3/2)}{\Gamma(x_1 + 1/2)\Gamma(x_2 + 1/2)\Gamma(x_3 + 1/2)} \\ \int_\Theta \theta_1^{y_1+x_1+1/2-1}\theta_2^{y_2+x_2+1/2-1}\theta_3^{y_3+x_3+1/2-1} d\theta = \\ = \frac{\Gamma(n+1)}{\Gamma(y_1+1)\Gamma(y_2+1)\Gamma(y_3+1)} \frac{\Gamma(x_1 + x_2 + x_3 + 3/2)}{\Gamma(x_1 + 1/2)\Gamma(x_2 + 1/2)\Gamma(x_3 + 1/2)} \\ \frac{\Gamma(y_1 + x_1 + 1/2)\Gamma(y_2 + x_2 + 1/2)\Gamma(y_3 + x_3 + 1/2)}{\Gamma\left(3/2 + \sum_{i=1}^3 x_i + \sum_{i=1}^3 y_i\right)} \end{aligned}$$

In order to simulate from the above distribution given posterior samples from the posterior θ, $\{\theta^{(i)} : i = 1, \ldots N\}$ we can do the following. Take each simulated $\theta^{(i)}$ and use it to simulate a triplet $y^{(i)}$ from a Trinomial$(n, \theta^{(i)})$ distribution. The samples $\{y^{(i)} : i = 1, \ldots N\}$ will then be draws from the above posterior predictive distribution.

2. Assume that the data $x = (x_1, \ldots, x_n)$ are independent scalar random variables and the distribution of each x_i, is given by a mixture of $K = 2$ Normal distributions with parameters $(\mu, \sigma^2) = (\mu_k, \sigma_k^2)$, for $k = 0, 1$. In other words assume that z_i is a binary random variable being in category 1 with probability π and in category 0 with probability $1 - \pi$, and that

$$f(x_i|\mu, \sigma^2, \pi) = (1 - \pi) f(x_i|\mu_0, \sigma_0^2) + \pi f(x_i|\mu_1, \sigma_1^2),$$

where $f(x_i|\lambda_k) = N(x_i|\mu_k, \sigma_k^2)$. Give the details of the EM algorithm that can be used to find maximum likelihood estimate of the $\theta = (\mu, \sigma^2, \pi)$. Define explicitly the E and the M steps.
Answer: Following the lecture notes, in the E step we define the following quantity as

$$Q(\theta, \theta^{old}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma(z_{ik}) \left(\log \pi_k + \log f(x_i|\theta_k) \right)$$

where

$$\log f(x_i|\theta_k) = -\tfrac{1}{2} \log \sigma_k^2 - \frac{(x_i - \mu_k)^2}{2\sigma_k^2}$$

In the M step the aim is to maximise $Q(\theta, \theta^{old})$ with respect to (μ, σ^2, π).

Differentiating $Q(\theta, \theta^{old})$ with respect to π and setting equal to 0 yields

$$\frac{\sum_{i=1}^n \gamma(z_{i0})}{\hat{\pi}} - \frac{\sum_{i=1}^n \gamma(z_{i1})}{1 - \hat{\pi}} = 0, \quad \text{or else}$$

$$\sum_{i=1}^{n} \gamma(z_{i0}) = \hat{\pi} \sum_{i=1}^{n} \gamma(z_{i0}) + \hat{\pi} \sum_{i=1}^{n} \gamma(z_{i1}),, \quad \text{or else}$$

$$\hat{\pi} = \frac{\sum_{i=1}^n \gamma(z_{i0})}{\sum_{i=1}^n \sum_{k=1}^K \gamma(z_{ik})} = \frac{n_0}{n}$$

Differentiating $Q(\theta, \theta^{old})$ with respect to μ_0 and setting equal to 0 yields

$$-\frac{\sum_{i=1}^n \gamma(z_{i0})(x_i - \hat{\mu}_0)}{\sigma_0^2} = 0, \quad \text{or else}$$

$$\hat{\mu}_0 = \frac{\sum_{i=1}^n \gamma(z_{i0}) x_i}{\sum_{i=1}^n \gamma(z_{i0})} = \frac{\sum_{i=1}^n \gamma(z_{i0}) x_i}{n_0}$$

Similarly we obtain $\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^n \gamma(z_{i1}) x_i$.

Differentiating $Q(\theta, \theta^{old})$ with respect to σ_0^2 gives

$$-\frac{\sum_{i=1}^n \gamma(z_{i0})}{\hat{\sigma_0}^2} + \frac{\sum_{i=1}^n \gamma(z_{i0})(x_i - \hat{\mu}_0)^2}{(\hat{\sigma_0}^2)^2} = 0, \quad \text{or else}$$

$$\hat{\sigma_0}^2 \left(\sum_{i=1}^{n} \gamma(z_{i0})(x_i - \hat{\mu}_0)^2 \right) = n_0 (\hat{\sigma_0}^2)^2, \quad \text{or else}$$

$$\hat{\sigma_0}^2 = \frac{1}{n_0} \sum_{i=1}^{n} \gamma(z_{i0})(x_i - \hat{\mu}_0)^2, \text{ where } \hat{\mu}_0 = \frac{\sum_{i=1}^n \gamma(z_{i0}) x_i}{n_0}$$

Similarly we get

$$\hat{\sigma_1}^2 = \frac{1}{n_1} \sum_{i=1}^{n} \gamma(z_{i1})(x_i - \hat{\mu}_1)^2, \text{ where } \hat{\mu}_1 = \frac{\sum_{i=1}^n \gamma(z_{i1}) x_i}{n_1}$$

3. Consider the Water Treatment Plant Data Set from the UCI repository. You can check computer class of week 4 on how to access data from the UCI repository. Choose 4 continuous variables from the dataset to analyse using Gaussian Mixture models. Select the optimal model from a set of models with up to 7 clusters and 4 covariance matrix types (spherical,tied, diagonal and full) and present its output. Also fit a Bayesian Gaussian mixture model and compare the results with a standard Gaussian mixture case.

   Answer: Code from the class can be used.