AI 701 Bayesian machine learning, Fall 2020

Homework assignment 1

1. (20 points) Let X be a $\mathbb{R}$-valued continuous random variable with Probability Density Function (PDF)

$$f(x) = \begin{cases} \exp(\theta - x), & x > \theta, \\ 0, & x \leq \theta. \end{cases} \tag{1}$$

   (a) (3 points) Compute $\mathbb{E}[X]$ and $\mathrm{Var}(X)$.

Solution:

$$\begin{aligned} \mathbb{E}[X] &= e^{\theta} \int_{\theta}^{\infty} x e^{-x} \mathrm{d}x \\ &= e^{\theta} \left(\left[-x e^{-x} \right]_{\theta}^{\infty} + \int_{\theta}^{\infty} e^{-x} \mathrm{d}x \right) \\ &= \theta + 1. \end{aligned} \tag{S.1}$$

$$\begin{aligned} \mathrm{Var}(X) &= e^{\theta} \int_{\theta}^{\infty} x^2 e^{-x} \mathrm{d}x - (\theta+1)^2 \\ &= e^{\theta} \left(\left[-x^2 e^{-x} \right]_{\theta}^{\infty} + 2 \int_{\theta}^{\infty} x e^{-x} \mathrm{d}x \right) - (\theta+1)^2 \\ &= 1. \end{aligned} \tag{S.2}$$

   (b) (3 points) Let $X_1, \ldots, X_n$ be i.i.d. copies of X. Show that

$$\hat{\theta}_n := \frac{1}{n} \sum_{i=1}^{n} (X_i - 1), \tag{2}$$

is an unbiased estimator of θ.

Solution:

$$\mathbb{E}[\hat{\theta}_n] = \frac{1}{n} \sum_{i=1}^{n} (\mathbb{E}[X_i] - 1) = \theta. \tag{S.3}$$

   (c) (7 points) Derive the $100(1-\alpha)\%$ confidence interval of θ for $\alpha \in (0, 1)$.

Solution: Since $\mathbb{E}[\hat{\theta}_n] = \theta$ and $\mathrm{Var}(\hat{\theta}_n) = 1/n$, by the central limit theorem,

$$\frac{\hat{\theta}_n - \theta}{1/\sqrt{n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1). \tag{S.4}$$

Let $Z(x)$ be an inverse CDF of $\mathcal{N}(0,1)$. We have

$$\begin{aligned} &\mathbb{P}\left(-Z\left(1-\frac{\alpha}{2}\right) < \frac{\hat{\theta}_n - \theta}{1/\sqrt{n}} < Z\left(1-\frac{\alpha}{2}\right)\right) \to 1-\alpha \\ \Rightarrow\ &\mathbb{P}\left(\hat{\theta}_n - \frac{1}{\sqrt{n}}Z\left(1-\frac{\alpha}{2}\right) < \theta < \hat{\theta}_n + \frac{1}{\sqrt{n}}Z\left(1-\frac{\alpha}{2}\right)\right) \to 1-\alpha. \end{aligned} \tag{S.5}$$

Hence, the confidence interval is

$$\mathrm{CI}_\alpha(\hat{\theta}_n) = \left(\hat{\theta}_n - \frac{1}{\sqrt{n}}Z\left(1-\frac{\alpha}{2}\right),\ \hat{\theta}_n + \frac{1}{\sqrt{n}}Z\left(1-\frac{\alpha}{2}\right)\right). \tag{S.6}$$

(d) (7 points) Assume we have a set of observations $\mathcal{D} = \{10.0, 12.0, 15.0\}$. Using the fact that

$$\int_{-\infty}^{1.96} \mathcal{N}(x|0,1)\mathrm{d}x \approx 0.975, \tag{3}$$

compute the 95% confidence interval of θ. Do you find anything counter-intuitive? If so, what is weird about the confidence interval?

Solution:

$$\mathrm{CI}_{0.05}(\hat{\theta}_n) = \left(\frac{10+12+15-3}{3} - \frac{1.96}{\sqrt{3}},\ \frac{10+12+15-3}{3} + \frac{1.96}{\sqrt{3}}\right) = (10.20, 12.46). \tag{S.7}$$

This is obviously weird. By the definition of PDF, θ should be smaller than any of the observed data $\mathcal{D}$, meaning $\theta < 10$. However, the confidence interval says that $\theta \in (10.20, 12.46)$.

2. (10 points) Let $(E, \mathcal{E}, \mu)$ be a measure space. Using the definition of measures, show that

   1. For any $A \subset B$, $\mu(A) \leq \mu(B)$.
   2. For any $A_1, \ldots, A_n$, $\mu(\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^n, \mu(A_i)$.

Solution:

1. $B = A \cup (B \setminus A)$ and $A \cap (B \setminus A)$ = 0. Hence, $\mu(B) = \mu(A) + \mu(B \setminus A) \geq \mu(A)$.
2. Define

$$B_1 = A_1, \quad B_i = A_i \setminus \bigcup_{j=1}^{i-1} A_j. \tag{S.8}$$

Then $\bigcup_{i=1}^n A_i = \bigcup_{i=1}^n B_i$. Since $(B_i)_{i=1}^n$ are disjoint and $B_i \leq A_i$ for all $i = 1, \ldots, n$, by the 1 we have

$$\mu\left(\bigcup_{i=1}^n B_i\right) = \sum_{i=1}^n \mu(B_i) \leq \sum_{i=1}^n \mu(A_i). \tag{S.9}$$

3. (5 points) Let $(X_i)_{i\geq 1}$ be an i.i.d. sequence of random variables with $\mathbb{E}[X_1] < \infty$ and $\mathrm{Var}(X_1) < \infty$. Show that

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{\mathrm{p}} \mathbb{E}[X_1]. \tag{4}$$

Hint: use Chebyshev's inequality, for a random variable X with $\mathbb{E}[X] = \mu < \infty$ and $\mathrm{Var}(X) = \sigma^2 < \infty$,

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \tag{5}$$

Solution: Let $\bar{X}_n := \frac{X_1+\cdots+X_n}{n}$, $\mu := \mathbb{E}[X_1]$ and $\sigma^2 := \mathrm{Var}(X_1)$. Then $\mathbb{E}[\bar{X}_n] = \mu$ and $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$. By Chebyshev's inequality, for any $\varepsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon} \to 0, \tag{S.10}$$

as $n \to \infty$. Hence $\bar{X}_n \xrightarrow{\mathrm{p}} \mu$.

4. (15 points) Let $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ be a sequence of random variables defined on a common probability space. Show that if $X_n \xrightarrow{\mathrm{p}} X$ and $Y_n \xrightarrow{\mathrm{p}} Y$ for some random variables X and Y, then $X_n + Y_n \xrightarrow{\mathrm{p}} X + Y$.

Solution:

$$\begin{aligned}\mathbb{P}(|X_n + Y_n - X - Y| > \varepsilon) &\leq \mathbb{P}(|X_n - X| + |Y_n - Y| > \varepsilon) \\ &\leq \mathbb{P}\bigg(|X_n - X| > \frac{\varepsilon}{2} \cup \mathbb{P}(|Y_n - Y| > \frac{\varepsilon}{2}\bigg) \\ &\leq \mathbb{P}\bigg(|X_n - X| > \frac{\varepsilon}{2}\bigg) + \mathbb{P}\bigg(|Y_n - Y| > \frac{\varepsilon}{2}\bigg) \to 0 \end{aligned} \tag{S.11}$$

since $X_n \xrightarrow{\mathrm{p}} X$ and $Y_n \xrightarrow{\mathrm{p}} Y$.

5. (15 points) Let $X_i \sim \mathrm{Gamma}(\alpha_i, 1)$ for $i = 1, \ldots, n$ with PDF

$$f_X(x) = \frac{x^{\alpha_i - 1} e^{-x}}{\Gamma(\alpha_i)}, \tag{6}$$

where

$$\Gamma(z) := \int_0^\infty t^{z-1} e^{-t} \mathrm{d}t, \tag{7}$$

denotes the gamma function. Show that

$$Y := \left(\frac{X_1}{\sum_{i=1}^n X_i}, \ldots, \frac{X_n}{\sum_{i=1}^n X_i}\right), \tag{8}$$

is a Dirichlet random variable with PDF

$$f(y) = \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^n y_i^{\alpha_i - 1}. \tag{9}$$

Hint: use the change of variable

$$(X_1, \dots, X_n) \to (Y_1, \dots, Y_{n-1}, Z), \tag{10}$$

where

$$Y_i = \frac{X_i}{Z} \text{ for } i = 1, \dots, n-1, \quad Z = \sum_{i=1}^n X_i. \tag{11}$$

Solution: The density of $(Y_1, \dots, Y_{n-1}, Z)$ is

$$f_{Y,Z}(y, z) = f_X(x)|J(y, z)|, \tag{S.12}$$

where

$$J(y, z) = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_{n-1}} & \frac{\partial x_1}{\partial z} \\ \vdots & \ddots & \vdots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_{n-1}} & \frac{\partial x_n}{\partial z}. \end{bmatrix} \tag{S.13}$$

Since $x_n = z(1 - \sum_{i=1}^{n-1} y_j)$, we have

$$|J(y, z)| = \left| \begin{bmatrix} z & 0 & \dots & 0 & y_1 \\ 0 & z & \dots & 0 & y_2 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & z & y_{n-1} \\ -z & -z & \dots & -z & 1 - \sum_{i=1}^{n-1} y_i \end{bmatrix} \right| = z^{n-1}. \tag{S.14}$$

Let $y_n := 1 - \sum_{i=1}^{n-1} y_i$. We have

$$f_{Y,Z}(y, z) = \prod_{i=1}^n \frac{(zy_i)^{\alpha_i - 1} e^{-zy_i}}{\Gamma(\alpha_i)} \times z^{n-1} = \prod_{i=1}^n \frac{y_i^{\alpha_i - 1}}{\Gamma(\alpha_i)} \times z^{\sum_{i=1}^n \alpha_i - 1} e^{-z}. \tag{S.15}$$

Marginalizing out z gives

$$f_Y(z) = \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^n y_i^{\alpha_i - 1}. \tag{S.16}$$

6. (20 points) Consider the following Gaussian-Gaussian-inverse-Wishart model.

$$\theta := (\mu, \Sigma), \tag{12}$$

$$p(x|\theta) = \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\bigg(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\bigg) \tag{13}$$

$$\begin{aligned} p(\theta) &= \mathcal{N}(\mu|m, \Sigma/r)\mathcal{W}^{-1}(\Sigma|\Psi, \nu) \\ &= \frac{r^{d/2}}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\bigg(-\frac{r}{2}(\mu-m)^\top \Sigma^{-1}(\mu-m)\bigg) \\ &\times \frac{|\Psi|^{\nu/2}}{2^{\nu d/2}\Gamma_d(\nu/2)}|\Sigma|^{-\frac{\nu+d+1}{2}} \exp\bigg(-\frac{1}{2}\mathrm{tr}(\Psi\Sigma^{-1})\bigg), \end{aligned} \tag{14}$$

where

$$\Gamma_d(z) := \pi^{\frac{d(d-1)}{4}} \prod_{j=1}^d \Gamma(z + (1-j)/2). \tag{15}$$

Let $X := \{x_1, \dots, x_n\}$ be a set of observations.

(a) (7 points) Compute the posterior $p(\theta|X)$. (Hint: use the fact $x^\top Ay = \mathrm{tr}(Ayx^\top)$).

Solution: The log-joint likelihood has the form

$$\begin{aligned} &\log p(X, \theta) = \sum_{i=1}^n \log p(x_i|\theta) + \log p(\theta) \\ &= -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^n (x_i^\top \Sigma x_i - 2\mu^\top\Sigma^{-1}x_i + \mu^\top\Sigma^{-1}\mu) \\ &- \frac{1}{2}\log|\Sigma| - \frac{1}{2}(r\mu^\top\Sigma^{-1}\mu - 2r\mu^\top\Sigma^{-1}m + rm^\top\Sigma^{-1}m) \\ &- \frac{\nu+d+1}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}(\Psi\Sigma^{-1}) + \mathrm{const} \\ &= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\bigg((r+n)\mu^\top\Sigma^{-1}\mu - 2(r+n)\mu^\top\Sigma^{-1}\bigg(\frac{rm + \sum_{i=1}^n x_i}{r+n}\bigg)\bigg) \\ &- \frac{\nu+n+d+1}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\bigg(\Sigma^{-1}\bigg(\Psi + \sum_{i=1}^n x_i x_i^\top + rmm^\top\bigg)\bigg) + \mathrm{const}. \end{aligned} \tag{S.17}$$

Now define

$$r_n := r + n, \qquad \nu_n := \nu + n \tag{S.18}$$

$$m_n := \frac{rm + \sum_{i=1}^n x_i}{r+n}, \qquad \Psi_n := \Psi + \sum_{i=1}^n x_i x_i^\top + rmm^\top - r_n m_n m_n^\top. \tag{S.19}$$

Then

$$\begin{aligned} &\log p(X, \theta) \\ &= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(r_n\mu^\top\Sigma\mu - 2r_n\mu^\top\Sigma^{-1}m_n + r_n m_n^\top\Sigma^{-1}m_n) \\ &- \frac{\nu_n+d+1}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}\Psi_n) + \mathrm{const}. \end{aligned} \tag{S.20}$$

Hence, we can see that

$$p(\theta|X) = \mathcal{N}(\mu|m_n, \Sigma/r_n)\mathcal{W}^{-1}(\Sigma|\Psi_n, \nu_n). \tag{S.21}$$

(b) (6 points) Compute the marginal likelihood $p(X)$.

Solution:

$$p(X) = \frac{p(X,\theta)}{p(\theta|X)} = \pi^{-nd/2}\frac{\tau^{d/2}}{\tau_n^{d/2}}\frac{\Gamma_d(\nu_n/2)}{\Gamma_d(\nu/2)}\frac{|\Psi|^{\nu/2}}{|\Psi_n|^{\nu_n/2}} \tag{S.22}$$

(c) (7 points) Show that $p(x|\theta)$ and $p(\theta)$ belong to conjugate exponential families. That is, show that they can be represented as

$$p(x|\theta) = \exp(T(x)^\top \eta - \mathbb{1}^\top A(\eta) - B(x)) \tag{16}$$
$$p(\theta) = \exp(\eta^\top \chi - \zeta^\top A(\eta) - \mathbb{1}^\top C(\chi, \zeta)). \tag{17}$$

(Hint: use the fact $\mathrm{tr}(A^\top B) = \mathrm{vec}(A)^\top \mathrm{vec}(B)$.

Solution: The log-likelihood can be written as (all the matrices inside inner products are vectorized)

$$\log p(x|\theta) = \exp\left(\begin{bmatrix} x \\ xx^\top \end{bmatrix}^\top \begin{bmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\Sigma^{-1} \end{bmatrix} - \mathbb{1}^\top \begin{bmatrix} \frac{1}{2}\mu^\top\Sigma^{-1}\mu \\ \frac{1}{2}\log|\Sigma| \end{bmatrix} - \frac{d}{2}\log 2\pi\right), \tag{S.23}$$

and hence

$$T(x) = [x, xx^\top]^\top, \qquad \eta = \left[\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}\right]^\top, \tag{S.24}$$

$$A(\eta) = \left[\frac{1}{2}\mu^\top\Sigma^{-1}\mu, \frac{1}{2}\log|\Sigma|\right]^\top, \qquad B(x) = \frac{d}{2}\log 2\pi. \tag{S.25}$$

The prior is written as

$$\begin{aligned} p(\theta) = \exp\Bigg(& \begin{bmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\Sigma^{-1} \end{bmatrix}^\top \begin{bmatrix} rm \\ \Psi + rmm^\top \end{bmatrix} - \begin{bmatrix} r \\ \nu + d + 2 \end{bmatrix}^\top \begin{bmatrix} \frac{1}{2}\mu^\top\Sigma^{-1}\mu \\ \frac{1}{2}\log|\Sigma| \end{bmatrix} \\ & -\mathbb{1}\begin{bmatrix} -\frac{d}{2}\log r + \frac{d}{2}\log 2\pi + \frac{\nu d}{2}\log 2 \\ \log\Gamma_d(\nu/2) - \frac{\nu}{2}\log|\Psi| \end{bmatrix}\Bigg). \end{aligned} \tag{S.26}$$

Hence,

$$\chi = [rm, \Psi + rmm^\top]^\top \tag{S.27}$$
$$\zeta = [r, \nu + d + 2]^\top, \tag{S.28}$$
$$C(\chi, \zeta) = \begin{bmatrix} -\frac{d}{2}\log r + \frac{d}{2}\log 2\pi + \frac{\nu d}{2}\log 2 \\ \log\Gamma_d(\nu/2) - \frac{\nu}{2}\log|\Psi| \end{bmatrix}. \tag{S.29}$$

Note that you may have multiple valid answers.

7. (15 points) Define U_1, U_2 as follows:

$$U_1, U_2 \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1), \quad \text{accept } (U_1, U_2) \text{ if } U_1^2 + U_2^2 \le 1. \tag{18}$$

Show that

$$X_1 = U_1 \sqrt{\frac{-2\log(U_1^2 + U_2^2)}{U_1^2 + U_2^2}} \sim \mathcal{N}(0,1), \quad X_2 = U_2 \sqrt{\frac{-2\log(U_1^2 + U_2^2)}{U_1^2 + U_2^2}} \sim \mathcal{N}(0,1). \tag{19}$$

Hint: use the intermediate transform

$$U_1 = R\cos\Theta, \quad U_2 = R\sin\Theta, \quad 0 \le R \le 1. \tag{20}$$

Solution: The joint density of (R, Θ) is

$$f_{R,\Theta}(r,\theta) = f_{U_1,U_2}(u_1,u_2) \left| \frac{\partial(u_1,u_2)}{\partial(r,\theta)} \right| = \frac{r}{\pi}. \tag{S.30}$$

Since

$$X_1 = \cos\Theta\sqrt{-2\log R^2}, \quad X_2 = \sin\Theta\sqrt{-2\log R^2}, \quad X_1^2 + X_2^2 = -2\log R^2, \tag{S.31}$$

we have

$$f_{X_1,X_2}(x_1,x_2) = f_{R,\Theta}(r,\theta) \left| \frac{\partial(r,\theta)}{\partial(x_1,x_2)} \right| \tag{S.32}$$

Since

$$\begin{bmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \theta} \\ \frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \theta} \end{bmatrix} = \begin{bmatrix} -\frac{\cos\theta}{r\sqrt{-2\log r^2}} & -\sin\theta\sqrt{-2\log r^2} \\ -\frac{\sin\theta}{r\sqrt{-2\log r^2}} & \cos\theta\sqrt{-2\log r^2} \end{bmatrix}. \tag{S.33}$$

and

$$\left| \frac{\partial(r,\theta)}{\partial(x_1,x_2)} \right| = \left| \frac{\partial(x_1,x_2)}{\partial(r,\theta)} \right|^{-1} = \frac{r}{2}, \tag{S.34}$$

we have

$$\begin{aligned} f_{X_1,X_2}(x_1,x_2) &= \frac{r^2}{2\pi} = \frac{1}{2\pi}\exp\left(-\frac{x_1^2 + x_2^2}{2}\right) \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x_1^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x_2^2}{2}\right). \end{aligned} \tag{S.35}$$