# AI 701 Bayesian machine learning, Fall 2021

# Homework assignment 1

1. (15 points) Suppose we toss a coin $n$ times. Let $X$ be a random variable denoting the number of heads among $n$ coin tosses. We simply assume that $X$ follows the binomial distribution with parameter $\theta$,

$$\mathbb{P}(X = x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}. \tag{1}$$

   (a) (5 points) We assume that $\theta$ itself is a random variable following the beta distribution with Probability Density Function (PDF) with parameters $a, b > 0$,

$$f(\theta; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\mathbb{1}_{\{0\le\theta\le1\}}, \tag{2}$$

   where $\Gamma(t) = \int_0^\infty z^{t-1}e^{-z}\mathrm{d}z$ is the gamma function. Compute the marginal likelihood of $x$ under this prior setting. That is, compute

$$\mathbb{P}(X = x; a, b) = \int_0^1 \mathbb{P}(X = x|\theta)f(\theta; a, b)\mathrm{d}\theta. \tag{3}$$

**Solution:**

$$\begin{aligned}
\mathbb{P}(X = x; a, b) &= \int_0^1 \mathbb{P}(X = x|\theta)f(\theta; a, b)\mathrm{d}\theta \\
&= \int_0^1 \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta \\
&= \binom{n}{x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \theta^{a+x-1}(1-\theta)^{b+n-x-1}\mathrm{d}\theta \\
&= \binom{n}{x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)}.
\end{aligned} \tag{S.1}$$

   (b) (10 points) Assume we toss the coin $n = 10$ times and see $x = 9$ heads. To see if the coin is fair, we compare two prior distributions for $\theta$:

$$\mathcal{M}_1 : \mathbb{P}(\theta) = \delta_{1/2}(\theta), \quad \mathcal{M}_2 : f(\theta; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\mathbb{1}_{\{0\le\theta\le1\}}. \tag{4}$$

   Set $a = b = 1$ and compute the Bayes factor

$$\frac{\mathbb{P}(X = x|\mathcal{M}_1)}{\mathbb{P}(X = x|\mathcal{M}_2)}. \tag{5}$$

   (Hint) Use the fact that $\Gamma(n) = (n-1)!$ for all $n \in \mathbb{N}$.

**Solution:**

$$\begin{aligned}\mathbb{P}(X=9|\mathcal{M}_1) &= \int_0^1 \mathbb{P}(X=9|\theta)\delta_{1/2}(\theta)\mathrm{d}\theta \\ &= \binom{10}{9}(1/2)^9(1/2)^1 \\ &= \frac{10}{1024} = \frac{5}{512}, \end{aligned} \tag{S.2}$$

$$\begin{aligned}\mathbb{P}(X=9|\mathcal{M}_2) &= \int_0^1 \mathbb{P}(X=9|\theta)f(\theta;a,b)\mathrm{d}\theta \\ &= \binom{10}{9}\frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}\frac{\Gamma(10)\Gamma(2)}{\Gamma(12)} \\ &= \binom{10}{9}\frac{1!}{0!\times 0!}\frac{9!\times 1!}{11!} \\ &= \frac{10}{110} = \frac{1}{11}. \end{aligned} \tag{S.3}$$

Hence, the Bayes factor is

$$\frac{\mathbb{P}(X=9|\mathcal{M}_1)}{\mathbb{P}(X=9|\mathcal{M}_2)} = \frac{10/1024}{10/110} = \frac{110}{1024} = \frac{55}{512} \quad (\approx 0.1074). \tag{S.4}$$

2. (15 points) Let $E$ be a set and $\mathcal{A}$ be a collection of subsets of $E$. The $\sigma$-algebra generated by $\mathcal{A}$, denoted by $\sigma(\mathcal{A})$, is the intersection of all $\sigma$-algebras containing $\mathcal{A}$, or equivalently, the smallest $\sigma$-algebra containing $\mathcal{A}$.
   (a) (5 points) Let $E = \{1,2,3,4\}$ and $\mathcal{A} = \{\{1\},\{2\}\}$. Find $\sigma(\mathcal{A})$.

**Solution:**

$$\sigma(A) = \{\{1\},\{2\},\{1,2\},\{3,4\},\{1,3,4\},\{2,3,4\},\{1,2,3,4\},\varnothing\}. \tag{S.5}$$

   (b) (10 points) Let $E$ be a set and $\mathcal{A},\mathcal{C}$ be collections of subsets of $E$. Prove the followings.
      1. $\mathcal{A} \subset \mathcal{C} \implies \sigma(\mathcal{A}) \subset \sigma(\mathcal{C})$.
      2. $\mathcal{A} \subset \sigma(\mathcal{C}),\ \mathcal{C} \subset \sigma(\mathcal{A}) \implies \sigma(\mathcal{A}) = \sigma(\mathcal{C})$.
      3. $\mathcal{A} \subset \mathcal{C} \subset \sigma(\mathcal{A}) \implies \sigma(\mathcal{A}) = \sigma(\mathcal{C})$.

**Solution:**

1. $\mathcal{A} \subset \mathcal{C} \subset \sigma(\mathcal{C})$, meaning that $\sigma(\mathcal{C})$ is a $\sigma$-algebra containing $\mathcal{A}$. Since $\sigma(\mathcal{A})$ is the intersection of all $\sigma$-algebras containing $\mathcal{A}$, $\sigma(\mathcal{A}) \subset \sigma(\mathcal{C})$.
2. Since $\mathcal{A} \subset \sigma(\mathcal{C})$ and $\sigma(\mathcal{A})$ is the smallest $\sigma$-algebra containing $\mathcal{A}$, $\sigma(\mathcal{A}) \subset \sigma(\mathcal{C})$. Similarly, we have $\sigma(C) \subset \sigma(\mathcal{A})$. Hence $\sigma(\mathcal{A}) = \sigma(\mathcal{C})$.
3. $\mathcal{A} \subset \mathcal{C}$ implies $\mathcal{A} \subset \sigma(\mathcal{C})$. Hence, by Item 2, we have $\sigma(\mathcal{A}) = \sigma(\mathcal{C})$.

3. (25 points) Let $X$ be a $\mathbb{R}$-valued random variable. The Moment Generating Function (MGF) of $X$ is defined as

$$M_X(t) = \mathbb{E}[e^{tX}]. \tag{6}$$

An important property of MGF is that it uniquely determines a distribution of a random variable. That is, given two $\mathbb{R}$-valued random variables $X$ and $Y$ (assume that there exist MGF for both $X$ and $Y$),

$$\forall t, M_X(t) = M_Y(t) \implies \forall x, F_X(x) = F_Y(x). \tag{7}$$

One can also show that, if $(X_n)_{n\geq 1}$ is a sequence of random variables whose MGF $M_{X_n}(t)$ converges to the MGF of a random variable $X$, then $(X_n)_{n\geq 1}$ converges in distribution to $X$.

(a) (5 points) Compute the MGF of a standard normal random variable $X$ with PDF,

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{8}$$

**Solution:**

$$\begin{aligned}
M_X(t) &= \int_{-\infty}^{\infty} \exp(tx) f_X(x) \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} + tx\right) \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-t)^2}{2} + \frac{t^2}{2}\right) \mathrm{d}x \\
&= \exp\left(\frac{t^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-t)^2}{2}\right) \mathrm{d}x \\
&= e^{t^2/2}.
\end{aligned} \tag{S.6}$$

(b) (5 points) For $n \in \mathbb{N}$, let $X_1, \ldots, X_n$ be i.i.d. $\mathbb{R}$-valued random variables, and let $M_{X_1}(t)$ be the MGF of $X_1$. Derive the MGF of $Y_n := c(X_1 + \cdots + X_n)$.

**Solution:**

$$M_{Y_n}(t) = \mathbb{E}\left[\exp\left(ct \sum_{i=1}^{n} X_i\right)\right] = \prod_{i=1}^{n} \mathbb{E}[e^{ctX_i}] = \{M_{X_1}(ct)\}^n. \tag{S.7}$$

(c) (15 points) Let $X$ be a $\mathbb{R}$-valued random variable with mean $\mu$ and variance $\sigma^2$. For $n \in \mathbb{N}$, define

$$Y_i := \frac{X_i - \mu}{\sigma}, \quad Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i. \tag{9}$$

Show that $Z_n$ converges in distribution to a standard normal distribution as $n \to \infty$.

(Hint) Show that the MGF of $Z_n$ converges to the MGF of a standard normal distribution as $n \to \infty$. Use the Taylor's theorem to analyze the limit.

**Solution:** Using the result above, we have

$$M_{Z_n}(t) = \{M_{Y_1}(t/\sqrt{n})\}^n. \tag{S.8}$$

By Taylor's theorem (expanding around $t = 0$),

$$M_{Y_1}(t/\sqrt{n}) = M_{Y_1}(0) + \frac{M'_{Y_1}(0)}{\sqrt{n}}t + \frac{M''_{Y_1}(0)}{2n}t^2 + R_n, \tag{S.9}$$

with $R_n/(t^2/n) \to 0$ as $n \to \infty$. Since

$$M_{Y_1}(0) = 1, \quad M'_{Y_1}(0) = \mathbb{E}[Y_1] = 0, \quad M''_{Y_1}(0) = \mathbb{E}[Y_1^2] = 1, \tag{S.10}$$

we have

$$M_{Y_1}(t/\sqrt{n}) = 1 + \frac{t^2}{2n} + R_n = 1 + \frac{t^2}{2n}\left(1 + 2R_n/(t^2/n)\right). \tag{S.11}$$

Using this, we see that

$$\begin{aligned} \lim_{n\to\infty} M_{Z_n}(t) &= \lim_{n\to\infty} \left\{M_{Y_1}(t/\sqrt{n})\right\}^n \\ &= \lim_{n\to\infty} \left\{1 + \frac{t^2}{2n}\left(1 + 2R_n/(t^2/n)\right)\right\}^n \\ &= e^{\frac{t^2}{2}}. \end{aligned} \tag{S.12}$$

That is, $M_{Z_n}(t)$ converges to the MGF of a standard normal distribution as $n \to \infty$. Hence we see that $Z_n$ converges in distribution to a standard normal distribution.

4. (25 points) Let $X$ be a random variable with finite mean $\mu$ and variance $\sigma^2$. Let $X_1, \ldots, X_n$ be i.i.d. copies of $X$, and consider the following estimator of $\sigma$:

$$\hat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2, \quad \bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i. \tag{10}$$

(a) (10 points) Show that $\hat{\sigma}_n^2$ is consistent; that is, show that $\hat{\sigma}_n^2 \xrightarrow{\mathrm{p}} \sigma^2$ as $n \to \infty$.

**Solution:**

$$\begin{aligned} \hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^n (X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2) \\ &= \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2. \end{aligned} \tag{S.13}$$

By the weak law of large numbers, $\frac{1}{n}X_i^2 \xrightarrow{\mathrm{p}} \mathbb{E}[X^2]$ and $\bar{X}_n \xrightarrow{\mathrm{p}} \mu$. By the basic property of convergence in probability, we have

$$\hat{\sigma}_n^2 \xrightarrow{\mathrm{p}} \mathbb{E}[X^2] - \mu^2 = \sigma^2. \tag{S.14}$$

(b) (10 points) Show that $\hat{\sigma}_n^2$ is biased; that is, show that $\mathbb{E}[\hat{\sigma}_n^2] \neq \sigma^2$.

**Solution:**

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}_n^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i^2\right] - \mathbb{E}[\bar{X}_n^2] \\
&= \sigma^2 + \mu^2 - \frac{1}{n^2}\mathbb{E}\left[\sum_{i,j} X_i X_j\right] \\
&= \sigma^2 + \mu^2 - \frac{1}{n^2}(n(\sigma^2+\mu^2) + (n^2-n)\mu^2) \\
&= \frac{n-1}{n}\sigma^2 \neq \sigma^2. \qquad \text{(S.15)}
\end{aligned}$$

(c) (5 points) Propose a simple fix to make $\hat{\sigma}_n^2$ unbiased.

**Solution:** We can set

$$\tilde{\sigma}_n^2 := \frac{n}{n-1}\hat{\sigma}_n^2. \qquad \text{(S.16)}$$

Then it is easy to see that $\mathbb{E}[\tilde{\sigma}_n^2] = \sigma^2$.

5. (20 points) Let $X$ be a $\mathbb{R}$-valued random variable. Show that $X_n \xrightarrow{\text{d}} c$ where $c$ is a constant implies $X_n \xrightarrow{\text{p}} c$.

**Solution:** We need to show that $X_n \xrightarrow{\text{p}} c$, i.e., for any $\varepsilon > 0$,

$$\mathbb{P}(|X_n - c| > \varepsilon) = \mathbb{P}(X_n > c+\varepsilon) + \mathbb{P}(X_n < c-\varepsilon) \to 0, \qquad \text{(S.17)}$$

as $n \to \infty$.

1. Note that

$$\mathbb{P}(X_n > c+\varepsilon) = 1 - \mathbb{P}(X_n \le c+\varepsilon) = 1 - F_{X_n}(c+\varepsilon). \qquad \text{(S.18)}$$

Since $X_n \xrightarrow{\text{d}} c$, we have $F_{X_n}(c+\varepsilon) \to 1$, so $\mathbb{P}(X_n > c+\varepsilon) \to 0$.

2. Also,

$$\mathbb{P}(X_n < c-\varepsilon) \le \mathbb{P}(X_n \le c-\varepsilon/2) = F_{X_n}(c-\varepsilon/2). \qquad \text{(S.19)}$$

Again, since $X_n \xrightarrow{\text{d}} c$, we have $F_{X_n}(c-\varepsilon/2) \to 0$, and thus $\mathbb{P}(X_n < c-\varepsilon) \to 0$.

Combining these two, we get $\mathbb{P}(|X_n - c| > \varepsilon) \to 0$ as desired.