Assignment 1: Probability (Partial Solution)

AIGS/CSED515 Machine Learning

Instructor: Jungseul Ok

jungseul@postech.ac.kr

Due: 20:00pm Sep 30, 2020

Remarks

- Group study and open discussion via LMS board are encouraged, however, assignment that your hand-in must be **of your own work**, and **hand-written**.
- Submit a scanned copy of your answer on LMS online in **a single PDF file**.
- Delayed submission may get some penalty in score: 5% off for delay of 0 ∼ 4 hours; 20% off for delay of 4 ∼ 24 hours; and delay longer than 24 hours will not be accepted.

1. [20pt; marginalization] Consider the following bivariate distribution $p(x, y)$ of two discrete random variables X and Y.

Y X	1	2	3
1	0.01	0.05	0.1
2	0.02	0.1	0.05
3	0.03	0.05	0.03
4	0.1	0.07	0.05
5	0.1	0.2	0.04

Compute:

(a) The marginal distribution $p(x)$ and $p(y)$.

(b) The expectation $\mathbb{E}[X]$ and $\mathbb{E}[Y]$.

(c) The conditional distributions $p(x|Y = 1)$ and $P(y \mid X = 3)$

(d) The conditional expectation $\mathbb{E}[X \mid Y = 1]$ and $\mathbb{E}[Y \mid X = 3]$

2. [10pt; Bayes' theorem; from Murphy's book] After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease? (Show your calculations as well as giving the final result.)

3. [20pt] Consider two random variables X, Y with joint distribution $p(x, y)$ and finite supports $\mathcal{X}$ and $\mathcal{Y}$. Prove:

(a) The expected value of the conditional expected value of X given Y is the same as the expected value of X, i.e.,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] \, .$$

sol) From the definition of $\mathbb{E}_Y[\mathbb{E}_X[x|y]]$, we have

$$\begin{aligned}
\mathbb{E}_Y[\mathbb{E}_X[x|y]] &= \sum_y \mathbb{E}_X[X|Y = y]\mathbb{P}[Y = y] \\
&= \sum_y \sum_x x \cdot \mathbb{P}[X = x|Y = y]\mathbb{P}[Y = y] \\
&= \sum_y \sum_x x \cdot \mathbb{P}[X = x, Y = y] \\
&= \sum_y \sum_x x \cdot p(x, y) \\
&= \sum_x \left(x \left(\sum_y p(x, y) \right) \right) \\
&= \sum_x x\mathbb{P}[X = x] = \mathbb{E}_X(x)
\end{aligned}$$

which concludes the proof.

(b) The covariance can be computed as follows:

$$\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \, .$$

sol) From the definition of $\text{cov}(X, Y)$, it follows that

$$\begin{aligned}
\text{cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\
&= \mathbb{E}[XY - X\,\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\,\mathbb{E}[Y]] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] + \mathbb{E}[X]\,\mathbb{E}[Y] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]
\end{aligned}$$

where the third equality is from the fact that $\mathbb{E}[X]$, $\mathbb{E}[Y]$ and $\mathbb{E}[X]\,\mathbb{E}[Y]$ are constant. This concludes the proof.

4. [40pt; Bayes' theorem; from Murphy's book] Consider a data set $\mathcal{D} = \{x_n\}_{n=1,\ldots,N}$, where each $x_n \in \{0,1\}$ is independently drawn from Bernoulli distribution with mean θ, i.e., $p(x_i = 1 \mid \theta) = \theta$. Denote by $\bar{x}$ the sample mean, i.e., $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$, and answer the following questions:

(a) Obtain the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$ of θ for given $\mathcal{D}$.

sol)

Let $\mathcal{L}(\theta)$ be the log-likelihood:

$$\begin{aligned}\mathcal{L}(\theta) &= \log\left(\prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n}\right) \\ &= \left(\sum_{n=1}^{N} x_n\right)\log\theta + \left(N - \sum_{n=1}^{N} x_n\right)\log(1-\theta)\ .\end{aligned}$$

By setting the derivative to zero, i.e., $\frac{\partial \mathcal{L}(\theta)}{\partial\theta} = \frac{1}{\theta}\left(\sum_{n=1}^{N} x_n\right) - \frac{1}{1-\theta}\left(N - \sum_{n=1}^{N} x_n\right) = 0$, we obtain

$$\hat{\theta}_{\text{MLE}} = \frac{1}{N}\sum_{n=1}^{N} x_n = \bar{x}\ .$$

(b) Compute (i) the bias $\mathbb{E}[\hat{\theta}_{\text{MLE}} - \theta]$, (ii) variance $\mathbb{E}[(\hat{\theta}_{\text{MLE}} - \mathbb{E}[\hat{\theta}_{\text{MLE}}])^2]$, and (iii) mean squared error $\mathbb{E}[(\hat{\theta}_{\text{MLE}} - \theta)^2]$ of the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$ for given θ.

sol)

(i) The bias of $\hat{\theta}_{\text{MLE}}$ can be computed as follows:

$$\begin{aligned}\mathbb{E}[\hat{\theta}_{\text{MLE}} - \theta] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} x_n\right] - \theta \\ &= \theta - \theta = 0\ .\end{aligned}$$

(ii) To obtain the variance of $\hat{\theta}_{\text{MLE}}$, we first compute the individual variance:

$$\begin{aligned}\mathbb{E}[(x_n - \theta)^2] &= \mathbb{E}[x_n^2 - 2x_n + \theta^2] \\ &= \mathbb{E}[x_n - 2x_n + \theta^2] \\ &= -\theta + \theta^2 = \theta(1-\theta)\ ,\end{aligned}$$

where the second equality is from the fact that x_n is binary, i.e., $x_n^2 = x_n$.

Therefore, the variance of $\hat{\theta}_{\text{MLE}}$ is obtained as follows:

$$\begin{aligned}
\mathbb{E}[(\hat{\theta}_{\text{MLE}} - \mathbb{E}[\hat{\theta}_{\text{MLE}}])^2] &= \mathbb{E}[(\hat{\theta}_{\text{MLE}} - \theta)^2] \\
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N} x_n - \theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N} (x_n - \theta)\right)^2\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\sum_{n=1}^{N}\sum_{m=1}^{N} (x_n - \theta)(x_m - \theta)\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\sum_{n=1}^{N} (x_n - \theta)^2\right] = \frac{\theta(1-\theta)}{N} \ ,
\end{aligned}$$

where the second last equality is from the facts that different x_n's are independent and $\mathbb{E}[x_n - \theta] = 0$, and the last one is from the aforementioned computation of individual variance.

(iii) The mean squared error $\mathbb{E}[(\hat{\theta}_{\text{MLE}} - \theta)^2]$ is directly given from the computation of the variance:

$$\mathbb{E}[(\hat{\theta}_{\text{MLE}} - \theta)^2] = \frac{\theta(1-\theta)}{N} \ .$$

(c) Recall that a random variable $y \in [0,1]$ from $\text{Beta}(\alpha, \beta)$ with $\alpha, \beta > 0$ has the probability density function in the following form:

$$\begin{aligned}
p(y \mid \alpha, \beta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \\
&\propto y^{\alpha-1}(1-y)^{\beta-1} \ ,
\end{aligned}$$

where $\Gamma(\alpha) = \int_0^\infty z^{\alpha-1}e^{-z}dz$, and thus the expected value of y is $\frac{\alpha}{\alpha+\beta}$. Assume that θ is drawn from $\text{Beta}(v, v)$ for a given $v > 0$. Obtain the Bayes estimator $\hat{\theta}_{\text{BE}} = \mathbb{E}[\theta \mid \mathcal{D}, v]$.

sol)

To obtain the Bayes estimator, we write the posterior as follows:

$$\begin{aligned}
p(\theta \mid \mathcal{D}, v) &\propto p(\mathcal{D} \mid \theta) p(\theta \mid v) \\
&\propto \left(\prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n}\right) \theta^{v-1}(1-\theta)^{v-1} \\
&= \theta^{v-1+N\bar{x}}(1-\theta)^{v-1+N-N\bar{x}} \ ,
\end{aligned}$$

which is Beta distribution with $(\alpha, \beta) = (v + N\bar{x}, v + N(1-\bar{x}))$. Recalling the fact that the expected value of random variable from $\text{Beta}(\alpha, \beta)$ is $\frac{\alpha}{\alpha+\beta}$, we have

$$\hat{\theta}_{\text{BE}} = \frac{N\bar{x} + v}{N + 2v} \ .$$

(d) Compute (i) the bias $\mathbb{E}[\hat{\theta}_{\mathrm{BE}} - \theta]$, (ii) variance $\mathbb{E}[(\hat{\theta}_{\mathrm{BE}} - \mathbb{E}[\hat{\theta}_{\mathrm{BE}}])^2]$, and (iii) mean squared error $\mathbb{E}[(\hat{\theta}_{\mathrm{BE}} - \theta)^2]$ of the Bayes estimator $\hat{\theta}_{\mathrm{BE}}$ for given θ.

sol)

(i) The bias of $\hat{\theta}_{\mathrm{BE}}$ can be computed as follows:

$$\begin{aligned}\mathbb{E}[\hat{\theta}_{\mathrm{BE}} - \theta] &= \mathbb{E}\left[\frac{N\bar{x} + v}{N + 2v} - \theta\right] \\ &= \frac{N\theta + v}{N + 2v} - \theta = \frac{v(1 - 2\theta)}{N + 2v} \,.\end{aligned}$$

(ii) Recall the variance of MLE:

$$\mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N} x_n - \theta\right)^2\right] = \frac{\theta(1-\theta)}{N} \,.$$

Then, the variance of $\hat{\theta}_{\mathrm{BE}}$ is obtained as follows:

$$\begin{aligned}\mathbb{E}[(\hat{\theta}_{\mathrm{BE}} - \mathbb{E}[\hat{\theta}_{\mathrm{BE}}])^2] &= \mathbb{E}\left[\left(\frac{N\bar{x} + v}{N + 2v} - \frac{N\theta + v}{N + 2v}\right)^2\right] \\ &= \left(\frac{N}{N + 2v}\right)^2 \mathbb{E}\left[(\bar{x} - \theta)^2\right] \\ &= \left(\frac{N}{N + 2v}\right)^2 \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N}(x_n - \theta)\right)^2\right] = \frac{N\theta(1-\theta)}{(N + 2v)^2} \,.\end{aligned}$$

(iii) The mean squared error $\mathbb{E}[(\hat{\theta}_{\mathrm{BE}} - \theta)^2]$ is given as follows:

$$\begin{aligned}\mathbb{E}[(\hat{\theta}_{\mathrm{BE}} - \theta)^2] &= \mathbb{E}\left[\left(\left(\hat{\theta}_{\mathrm{BE}} - \mathbb{E}[\hat{\theta}_{\mathrm{BE}}]\right) + \left(\mathbb{E}[\hat{\theta}_{\mathrm{BE}}] - \theta\right)\right)^2\right] \\ &= \mathbb{E}\left[\left(\hat{\theta}_{\mathrm{BE}} - \mathbb{E}[\hat{\theta}_{\mathrm{BE}}]\right)^2 + \left(\mathbb{E}[\hat{\theta}_{\mathrm{BE}}] - \theta\right)^2 + 2\left(\hat{\theta}_{\mathrm{BE}} - \mathbb{E}[\hat{\theta}_{\mathrm{BE}}]\right)\left(\mathbb{E}[\hat{\theta}_{\mathrm{BE}}] - \theta\right)\right] \\ &= \underbrace{\mathbb{E}\left[\left(\hat{\theta}_{\mathrm{BE}} - \mathbb{E}[\hat{\theta}_{\mathrm{BE}}]\right)^2\right]}_{\text{Variance}} + \underbrace{\left(\mathbb{E}[\hat{\theta}_{\mathrm{BE}}] - \theta\right)^2}_{\text{Bias}^2} \\ &= \frac{N\theta(1-\theta)}{(N + 2v)^2} + \left(\frac{v(1 - 2\theta)}{N + 2v}\right)^2 = \frac{N\theta(1-\theta) + v^2(1 - 2\theta)^2}{(N + 2v)^2} \,.\end{aligned}$$

(e) (i) Show that the maximum likelihood estimator $\hat{\theta}_{\mathrm{MLE}}$ is a special case of the Bayes estimator $\hat{\theta}_{\mathrm{BE}}$ with a particular choice of v, and (ii) discuss the role of hyper-parameter v in the Bayes estimator.

sol)

(i) As $v \to 0$, the Bayes estimator converges to the maximum likelihood estimator.
(ii) The hyper-parameter v controls the strength of belief on that the true parameter θ is around $1/2$, where as v increases, the prior distribution of θ concentrates around $1/2$.