Review of probability and statistics

Machine Learning I (2023-2024)
UMONS

Exercise 1

An economics consulting firm has created a model to predict recessions. The model predicts a recession with probability 80% when a recession is indeed coming and with probability 10% when no recession is coming. The (unconditional) probability of falling into a recession is 20%. If the model predicts a recession, what is the probability that a recession will indeed come?

Solution:

Let $R \in \{0,1\}$ be a Bernouilli random variable indicating whether we fell into a recession ($R = 1$ means we fell into a recession, $R = 0$ means we did not). Let $M \in \{0,1\}$ be another Bernouilli random variable indicating the outcome of the prediction model ($M = 1$ means that the model predicted that a recession was coming, $M = 0$ means that it did not).

We know that $\mathbb{P}(R=1) = 0.2$, $\mathbb{P}(M=1|R=1) = 0.8$ and $\mathbb{P}(M=1|R=0) = 0.1$. We are interested in finding the probability that a recession will come, conditional on the fact that the model predicted it, i.e. we are looking for $\mathbb{P}(R=1|M=1)$. We have:

$$\begin{aligned}
\mathbb{P}(R=1|M=1) &= \frac{\mathbb{P}(R=1,M=1)}{\mathbb{P}(M=1)} && \text{(Conditional probability)} \\
&= \frac{\mathbb{P}(M=1|R=1)\mathbb{P}(R=1)}{\mathbb{P}(M=1)} && \text{(Conditional probability)} \\
&= \frac{\mathbb{P}(M=1|R=1)\mathbb{P}(R=1)}{\mathbb{P}(M=1|R=1)\mathbb{P}(R=1)+\mathbb{P}(M=1|R=0)\mathbb{P}(R=0)} && \text{(Law of total probability)} \\
&= \frac{\mathbb{P}(M=1|R=1)\mathbb{P}(R=1)}{\mathbb{P}(M=1|R=1)\mathbb{P}(R=1)+(1-\mathbb{P}(M=0|R=0))(1-\mathbb{P}(R=1))} \\
&= \frac{0.8\times 0.2}{0.8\times 0.2+0.1\times 0.8} \\
&= \frac{2}{3}
\end{aligned}$$

Exercise 2

Answer the questions for the following joint distributions between random variables X and Y.

2.1

Given the following joint PMF:

	$X=0$	$X=1$
$Y=0$	0.14	0.26
$Y=1$	0.21	0.39

a) Compute the marginal PMF of X and the marginal PMF of Y.

b) Compute the conditional PMF of Y given $X=0$.

c) Given $s_1(X,Y) = X^2 + 3Y + 1$, compute the joint expectation $\mathbb{E}_{XY}[s_1(X,Y)]$ and the conditional expectation $\mathbb{E}_{Y|X}[s_1(X,Y)|X=0]$.

d) Given $s_2(X,Y) = XY^3 - 4X + 2Y$, compute the joint expectation $\mathbb{E}_{XY}[s_2(X,Y)]$ and the conditional expectation $\mathbb{E}_{XY}[s_2(X,Y)|Y=1]$.

e) Are X and Y independent?

Solution:

Marginal PMFs

$p_X(0) = p_{XY}(0,0) + p_{XY}(0,1) = 0.14 + 0.21 = 0.35$
$p_X(1) = p_{XY}(1,0) + p_{XY}(1,1) = 0.26 + 0.39 = 0.65$
$p_Y(0) = p_{XY}(0,0) + p_{XY}(1,0) = 0.14 + 0.26 = 0.4$
$p_Y(1) = p_{XY}(0,1) + p_{XY}(1,1) = 0.21 + 0.39 = 0.6$

Conditional PMFs

$p_{Y|X}(1|0) = \frac{p_{XY}(0,1)}{p_X(0)} = \frac{0.21}{0.35} = 0.6$
$p_{Y|X}(0|0) = \frac{p_{XY}(0,0)}{p_X(0)} = \frac{0.14}{0.35} = 0.4$

Expectations

1) s_1

$$\begin{aligned}
\mathbb{E}_{XY}[s_1(X,Y)] &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} s_1(x,y)p_{XY}(x,y) \\
&= s_1(0,0)p_{XY}(0,0) + s_1(0,1)p_{XY}(0,1) + s_1(1,0)p_{XY}(1,0) + s_1(1,1)p_{XY}(1,1) \\
&= 0.14 + 4\times 0.21 + 2\times 0.26 + 5\times 0.39 = 3.45
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{Y|X}[s_1(X,Y)|X=0] &= \sum_{y\in\mathcal{Y}} s_1(0,y)p_{Y|X}(y|x=0) \\
&= s_1(0,0)p_{Y|X}(0|0) + s_1(0,1)p_{Y|X}(1|0) \\
&= 0.4 + 4\times 0.6 = 2.8
\end{aligned}$$

2) s_2

$$\begin{aligned}
\mathbb{E}_{XY}[s_2(X,Y)] &= s_2(0,0)p_{XY}(0,0) + s_2(0,1)p_{XY}(0,1) + s_2(1,0)p_{XY}(1,0) + s_2(1,1)p_{XY}(1,1) \\
&= 0 + 2\times 0.21 - 4\times 0.26 - 0.39 = -1.01
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}_{X|Y}[s_2(X,Y)|Y=1] &= \sum_{x\in\mathcal{X}} s_2(x,1)p_{X|Y}(x|y=1) \\
&= s_2(0,1)p_{X|Y}(0|1) + s_2(1,1)p_{X|Y}(1|1) \\
&= s_2(0,1)\frac{p_{XY(0,1)}}{p_Y(1)} + s_2(1,1)\frac{p_{XY(1,1)}}{p_Y(1)} \\
&= 2\times\frac{0.21}{0.6} - \frac{0.39}{0.6} = 0.05
\end{aligned}
$$

Independence

Two discrete random variables X and Y are independent iff

$$p_{XY}(x,y) = p_X(x)p_Y(y) \quad \forall x\in\mathcal{X}, y\in\mathcal{Y}.$$

We must check that the equality holds for all realizations of the random variables X and Y:

$$
\begin{aligned}
&p_{XY}(0,0) = 0.14 = p_X(0)p_Y(0) \\
&p_{XY}(0,1) = 0.21 = p_X(0)p_Y(1) \\
&p_{XY}(1,0) = 0.26 = p_X(1)p_Y(0) \\
&p_{XY}(1,1) = 0.39 = p_X(1)p_Y(1)
\end{aligned}
$$

The random variables X and Y are independent.

2.2

Given the following joint PMF:

	$X=0$	$X=1$	$X=2$
$Y=1$	0.1	0.2	0.3
$Y=2$	0.05	0.15	0.2

a) Compute the marginal PMF of X and the marginal PMF of Y.

b) Compute the conditional PMF of Y given $X=1$.

c) Are X and Y independent?

Solution:

Marginal PMFs

$p_X(0)=0.15$
$p_X(1)=0.35$
$p_X(2)=0.5$
$p_Y(1)=0.6$
$p_Y(2)=0.4$

Conditional PMFs

$p_{Y|X}(1|1)=0.57$
$p_{Y|X}(2|1)=0.43$

Independence

$p_{XY}(0,1)=0.1\neq 0.09=p_X(0)p_Y(1)\rightarrow$ Not independent.

Exercise 3

Alex and Bob each flip a different fair coin twice. Denote "1" as head, and "0" as tail. Let X be the maximum of the two numbers Alex gets, and let Y be the minimum of the two numbers Bob gets.

a) Find the marginal PMF $p_X(x)$ and $p_Y(y)$.

b) Find the joint PMF $p_{X,Y}(x,y)$.

c) Find the conditional PMF $p_{X|Y}(x|y)$. Does $p_{X|Y}(x|y) = p_X(x)$? Why?

Solution:

For both Alex and Bob, the sample space for flipping a fair coin twice is $\Omega = \{00, 01, 10, 11\}$. if X and Y are the random variables respectively denoting the maximum of the two numbers Alex gets and the minimum of the two numbers Bob gets, then $\mathcal{X} = \{0,1\}$ and $\mathcal{Y} = \{0,1\}$.

a) From the sample space, we find:

$p_X(0) = \frac{1}{4}$
$p_X(1) = \frac{3}{4}$
$p_Y(0) = \frac{3}{4}$
$p_Y(1) = \frac{1}{4}$

b) By definition of the joint PMF and as the random variables X and Y are independent, we have:

$p_{XY}(0,0) = p_X(0)p_Y(0) = \frac{1}{4} \times \frac{3}{4} = \frac{3}{16}$

$p_{XY}(0,1) = p_X(0)p_Y(1) = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$

$p_{XY}(1,0) = p_X(1)p_Y(0) = \frac{3}{4} \times \frac{3}{4} = \frac{9}{16}$

$p_{XY}(1,1) = p_X(0)p_Y(0) = \frac{3}{4} \times \frac{1}{4} = \frac{3}{16}$

We can check that $\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p_{XY}(x,y) = \frac{3}{16} + \frac{1}{16} + \frac{9}{16} + \frac{3}{16} = 1$.

c) As the variables X and Y are independent, we have that $p_{X|Y}(x|y) = p_X(x)$ and $p_{Y|X}(y|x) = p_Y(y)$. Therefore,

$p_{X|Y}(0|0) = p_{X|Y}(0|1) = p_X(0) = \frac{1}{4}$.

Exercise 4

We have a population of people, 47% of whom were men and the remaining 53% were women. Suppose that the average height of the men was 70 inches, and the women was 71 inches. What is the average height of the entire population? [Hint: Use the law of total expectation]

Solution:

Let M be a Bernouilli random variable with support $\mathcal{M} \in \{0, 1\}$ indicating whether an individual is either male or female ($M = 1$ means that the individual is male, $M = 0$ means that the individual is female). Let H be a continuous random variable with support $\mathcal{H} \in \mathbb{R}^+$ indicating the height of an individual of the population.

We know that $p_M(1) = 0.47$ and that $p_M(0) = 0.53$. Moreover, $\mathbb{E}[H|M = 1] = 70$ and $\mathbb{E}[H|M = 0] = 71$. We are interested in finding the average height of the entire population, i.e $\mathbb{E}[H]$.

From the law of total expectation, we have:

$$\begin{aligned}
\mathbb{E}[H] &= \mathbb{E}[H|M = 1]p_M(1) + \mathbb{E}[H|M = 0]p_M(0) \\
&= 70 \times 0.47 + 71 \times 0.53 \\
&= 70.53 \text{ inches}
\end{aligned}$$

Exercise 5

Let $X_1, X_2, \dots, X_n \in \mathbb{R}$ be a collection of n random variables, and $a_1, a_2, \dots, a_n$, a set of constants, we have

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \,\mathrm{Cov}(X_i, X_j).$$

Prove the above fact. You can use the fact that, for a set of numbers $e_1, e_2, \dots, e_n$,

$$\left(\sum_{i=1}^{n} e_i\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} e_i e_j.$$

Solution:

By expanding the expression of the variance, and by successively applying the properties of the expectation, we get :

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) &= \mathbb{E}\left[\left(\sum_{i=1}^{n} a_i X_i - \mathbb{E}\left[\left(\sum_{i=1}^{n} a_i X_i\right)\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} a_i X_i\right)^2 - 2\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right]\left(\sum_{i=1}^{n} a_i X_i\right) + \left(\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right]\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j X_i X_j - 2\left(\sum_{i=1}^{n} a_i \mathbb{E}\left[X_i\right]\right)\left(\sum_{i=1}^{n} a_i X_i\right) + \left(\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right]\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j X_i X_j\right] - 2\left(\sum_{i=1}^{n} a_i \mathbb{E}\left[X_i\right]\right)\left(\sum_{i=1}^{n} a_i \mathbb{E}[X_i]\right) + \left(\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right]\right)^2 \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j X_i X_j\right] - 2\left(\sum_{i=1}^{n} a_i \mathbb{E}[X_i]\right)^2 + \left(\sum_{i=1}^{n} a_i \mathbb{E}[X_i]\right)^2 \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j X_i X_j\right] - \left(\sum_{i=1}^{n} a_i \mathbb{E}[X_i]\right)^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathbb{E}[X_i X_j] - \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathbb{E}[X_i]\mathbb{E}[X_j] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \Big(\mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]\Big) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \,\mathrm{Cov}(X_i, X_j)
\end{aligned}$$

Exercise 6

We observe a sample of real values $y_1, y_2, \dots, y_n$ where $y_i \geq 0$ for $i = 1, 2, \dots, n$. Let us assume they are all i.i.d. observations of a random variable Y with an exponential distribution:

$$p(y;\alpha) = \alpha e^{-\alpha y}$$

where $\alpha > 0$ is called the rate.

a) Write down the formula of the likelihood function as a function of the observed data and the unknown parameter α.

b) Write down the formula of the log-likelihood

c) Compute the maximum likelihood estimate (MLE) of α.

Solution:

a)

$$\begin{aligned}
\mathcal{L}(\alpha) &= \mathcal{L}(\alpha; y_1, ..., y_n) \\
&= p(y_1, y_2, ..., y_n; \alpha) \\
&= p(y_1;\alpha)p(y_2;\alpha)...p(y_n;\alpha) \qquad (y_i,\ i = 1, ..., n \text{ are i.i.d. random variables.}) \\
&= \prod_{i=1}^{n} p(y_i;\alpha) \\
&= \prod_{i=1}^{n} \alpha e^{-\alpha y_i}.
\end{aligned}$$

b)

$$\begin{aligned}
\log\, \mathcal{L}(\alpha) = \log\, &\left(\prod_{i=1}^{n} \alpha e^{-\alpha y_i}\right) \\
&= \sum_{i=1}^{n} (\log\, \alpha - \alpha y_i) \\
&= n(\log \alpha - \alpha \bar{y}),
\end{aligned}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

c)

$$\begin{aligned}
\hat{\alpha} &= \underset{\alpha \in \mathbb{R}}{\text{argmax}}\ \mathcal{L}(\alpha) \\
&= \underset{\alpha \in \mathbb{R}}{\text{argmax}}\ \log \mathcal{L}(\alpha) \\
&= \underset{\alpha \in \mathbb{R}}{\text{argmax}}\ \big(n\,(\log\, \alpha - \alpha \bar{y})\big)
\end{aligned}$$

Taking the derivative with respect to α and equaling to zero:

$$\begin{aligned}
&\tfrac{\partial \log \mathcal{L}(\alpha)}{\partial \alpha} = 0 \\
\iff &n\left(\frac{1}{\alpha} - \bar{y}\right) = 0 \\
\iff &\alpha = \frac{1}{\bar{y}}
\end{aligned}$$

Thus we have that the MLE $\hat{\alpha} = \frac{1}{\bar{y}}$. To check that $\hat{\alpha}$ is indeed a maximum, we can verify that the second derivative of the log-likelihood is always negative :

$$\begin{aligned}
&\tfrac{\partial^2 \log \mathcal{L}(\alpha)}{\partial \alpha^2} < 0 \\
\iff &-\frac{n}{\alpha^2} < 0 \qquad \forall \alpha \in \mathbb{R}
\end{aligned}$$

Exercise 7

We observe a sample of i.i.d. pairs $(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)$ where $(x_i,y_i) \in \mathcal{X}\times\mathbb{R}$ for $i = 1,2,\dots,n$. We assume that the conditional PDF $p(y;x)$ is normally distributed with a variance fixed at σ^2. Given an input x, the mean $\mu_\theta(x)$ is determined by a model μ_θ with parameters $\theta \in \Theta$. More specifically, the conditional PDF is given by:

$$\begin{aligned} p(y;x,\theta) &= \mathcal{N}(y;\mu_\theta(x),\sigma^2) \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu_\theta(x)}{\sigma}\right)^2}. \end{aligned}$$

Note that the specific sets $\mathcal{X}$ and Θ are not relevant to our problem. For example, we could have $\mathcal{X} = \mathbb{R}$ and $\Theta = \mathbb{R}^2$ for a uni-dimensional linear regression task with one coefficient and one bias.

a) Write down the formula of the likelihood function as a function of the observed data and parameters θ.

b) Write down the formula of the log-likelihood.

c) Can you prove that maximizing the likelihood is equivalent to minimizing the mean squared error $\frac{1}{n}\sum_{i=1}^n(\mu_\theta(x_i)-y_i)^2$ (with respect to θ)?

Solution:

a)

$$\begin{aligned} \mathcal{L}(\theta) &= p(y_1,\dots,y_n;x_1,\dots,x_n,\theta) \\ &= \prod_{i=1}^n p(y_i;x_i,\theta) \\ &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i-\mu_\theta(x_i)}{\sigma}\right)^2} \end{aligned}$$

b)

$$\begin{aligned} \log\mathcal{L}(\theta) &= \sum_{i=1}^n \log p(y_i;x_i,\theta) \\ &= \sum_{i=1}^n -\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{y_i-\mu_\theta(x_i)}{\sigma}\right)^2 \end{aligned}$$

c)

$$\begin{aligned} \arg\max_\theta \mathcal{L}(\theta) &= \arg\max_\theta \log\mathcal{L}(\theta) \\ &= \arg\max_\theta \sum_{i=1}^n -\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{y_i-\mu_\theta(x_i)}{\sigma}\right)^2 \\ &= \arg\max_\theta \sum_{i=1}^n -\frac{1}{2}\left(\frac{y_i-\mu_\theta(x_i)}{\sigma}\right)^2 \quad \text{(The first term is constant with respect to } \theta\text{)} \\ &= \arg\max_\theta -\frac{1}{2\sigma^2}\sum_{i=1}^n \left(y_i-\mu_\theta(x_i)\right)^2 \quad \text{(By linearity of the sum)} \\ &= \arg\max_\theta -\frac{1}{n}\sum_{i=1}^n \left(y_i-\mu_\theta(x_i)\right)^2 \quad \text{(Multiplying by } \frac{2\sigma^2}{n}\text{, which is constant with respect to } \theta \text{ and positive)} \\ &= \arg\min_\theta \frac{1}{n}\sum_{i=1}^n \left(y_i-\mu_\theta(x_i)\right)^2 \quad \text{(Maximizing } x \text{ is equivalent to minimizing } -x\text{)} \end{aligned}$$

Complementary exercise

Find the marginal PDF $f_X(x)$ if the joint PDF $f_{XY}(x,y)$ is defined as:

$$f_{XY}(x,y) = \frac{e^{-|y-x|-x^2/2}}{2\sqrt{2\pi}}$$

Solution:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x,y)dy \\ &= \int_{-\infty}^{\infty} \frac{e^{-|y-x|-x^2/2}}{2\sqrt{2\pi}} dy \\ &= \frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-|y-x|} e^{-x^2/2} dy \\ &= \frac{e^{-x^2/2}}{2\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-|y-x|} dy \end{aligned}$$

We have that : $|y-x| = \begin{cases} y-x, & \text{if } y \geq x \\ x-y, & \text{if } y \leq x \end{cases}$, and thus :

$$\begin{aligned} f_X(x) &= \frac{e^{-x^2/2}}{2\sqrt{2\pi}} \int_{-\infty}^{x} e^{y-x} dy + \int_{x}^{\infty} e^{x-y} dy \\ &= \frac{e^{-x^2/2}}{2\sqrt{2\pi}} \left(e^{-x} \left[e^{y} \right]_{-\infty}^{x} + e^{x} \left[-e^{-y} \right]_{x}^{\infty} \right) \\ &= \frac{e^{-x^2/2}}{2\sqrt{2\pi}} \left(e^{0} - e^{-\infty} - e^{-\infty} + e^{0} \right) \\ &= \frac{e^{-x^2/2}}{\sqrt{2\pi}} \end{aligned}$$

Complementary exercise

Let p_X be a normal distribution $\mathcal{N}(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$, and $\sigma > 0$. Consider the two scenarios where $n = 10$ or $n = 1000$. For each scenario,

1. repeat the following procedure 1000 times:
   (a) Generate n i.i.d. realizations $X_1, X_2, \ldots, X_n$ where $X_i \sim p_X$.
   (b) Compute $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
2. compute the mean and variance of the 1000 values computed in 1(b)
3. plot a histogram of these 1000 values, and add vertical lines at the true mean and the computed mean.

Experiment with different values of μ and σ, and confirm that you obtain $E[\bar{X}_n] = \mu$ and $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$.