Review of probability and statistics

Introduction to Machine Learning and Data Science (2024–2025)
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?

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) Are X and Y 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?

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?

Exercise 4

We have a population of people, 47% of whom are men and 53% of whom are women. Suppose that the average height of the men is 70 inches, and the average height of the women is 71 inches. What is the average height of the entire population?

Exercise 5

Let $X_1, X_2, \ldots, X_n \in \mathbb{R}$ be a collection of n random variables, and $a_1, a_2, \ldots, a_n$ be a set of constants. Prove that

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

You can use the fact that, for a set of numbers $e_1, e_2, \ldots, e_n$,

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

Exercise 6

We observe a sample of real values $y_1, y_2, \ldots, y_n$ where $y_i \geq 0$ for $i = 1, 2, \ldots, n$. 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.

Exercise 7

We observe a sample of i.i.d. pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$ for $i = 1, 2, \ldots, n$. 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:

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

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) 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).

Exercise 8

Let $\mu \in \mathbb{R}$ and $\sigma > 0$ be the mean and standard deviation of a normal distribution $\mathcal{N}(\mu, \sigma^2)$. Consider two scenarios: $n = 10$ and $n = 1000$. 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 $X_i \sim \mathcal{N}(\mu, \sigma^2)$.
   (b) Compute $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.
2. compute the empiric 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 $E[\overline{X}_n] = \mu$ and $\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$.

Exercise 9

Let $\Theta \sim \mathcal{U}[0, 2\pi]$.

(a) Assume $X = \cos(\Theta)$ and $Y = \sin(\Theta)$. Are X and Y correlated?

(b) Assume $X = \cos(\frac{\Theta}{4})$ and $Y = \sin(\frac{\Theta}{4})$. Are X and Y correlated?

Hint: You may want to use the trigonometric identity $\sin(2\omega) = 2\sin(\omega)\cos(\omega)$.

Complementary exercise

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

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