CHAPTER

MULTIDIMENSIONAL PROBABILITY DISTRIBUTIONS 5

CHAPTER CONTENTS

So far, properties of one-dimensional random variable x are discussed. When multiple random variables are available, we may be interested in knowing the dependency of one variable on another, which will give us more information. In this chapter, the relation between two random variables x and y is discussed.

5.1 JOINT PROBABILITY DISTRIBUTION

The probability that discrete random variables x and y take a value in a countable set is denoted by $\Pr(x, y)$. The function that describes the mapping from any realized value of the random variables to probability is called the *joint probability distribution*, and its probability mass function $f(x, y)$ is called the *joint probability mass function*:

$$\Pr(x, y) = f(x, y).$$

Similarly to the one-dimensional case, $f(x, y)$ should satisfy

$$f(x, y) \geq 0 \quad \text{and} \quad \sum_{x,y} f(x, y) = 1.$$

The probability mass functions of x or y alone, $g(x)$ and $h(y)$, can be expressed by using $f(x, y)$ as

$$g(x) = \sum_{y} f(x, y) \quad \text{and} \quad h(y) = \sum_{x} f(x, y).$$

These are called the *marginal probability mass functions* and the corresponding probability distributions are called the *marginal probability distributions*. The process of

An Introduction to Statistical Machine Learning. DOI: 10.1016/B978-0-12-802121-7.00016-9

computing a marginal probability distribution from the joint probability distribution is called *marginalization*.

When x and y are continuous random variables, the *joint probability density function* $f(x,y)$ is defined as

$$\Pr(a \leq x \leq b,\ c \leq y \leq d) = \int_c^d \int_a^b f(x,y)\mathrm{d}x\mathrm{d}y.$$

Similarly to the one-dimensional case, $f(x,y)$ should satisfy

$$f(x,y) \geq 0 \quad \text{and} \quad \iint f(x,y)\mathrm{d}x\mathrm{d}y = 1.$$

The probability density functions of x or y alone, $g(x)$ and $h(y)$, can be expressed by using $f(x,y)$ as

$$\Pr(a \leq x \leq b) = \int_a^b \int f(x,y)\mathrm{d}y\mathrm{d}x = \int_a^b g(x)\mathrm{d}x,$$
$$\Pr(c \leq y \leq d) = \int_c^d \int f(x,y)\mathrm{d}x\mathrm{d}y = \int_c^d h(y)\mathrm{d}y,$$

where

$$g(x) = \int f(x,y)\mathrm{d}y \quad \text{and} \quad h(y) = \int f(x,y)\mathrm{d}x$$

are the *marginal probability density functions*.

5.2 CONDITIONAL PROBABILITY DISTRIBUTION

For discrete random variables x and y, the probability of x given y is denoted by $\Pr(x|y)$ and called the *conditional probability distribution*. Since $\Pr(x|y)$ is the probability that x occurs after y occurs, it is given by

$$\Pr(x|y) = \frac{\Pr(x,y)}{\Pr(y)}. \tag{5.1}$$

Based on this, the *conditional probability mass function* is given by

$$g(x|y) = \frac{f(x,y)}{h(y)}. \tag{5.2}$$

Since the conditional probability distribution is a probability distribution, its expectation and variance can also be defined, which are called the *conditional expectation* and *conditional variance*, respectively,

$$E[x|y] = \sum_x x\, g(x|y) \quad \text{and} \quad V[x|y] = E\Big[(x - E[x|y])^2|y\Big].$$

Table 5.1 Example of Contingency Table

$x \setminus y$	Sleepy during the Lecture	Not Sleepy during the Lecture	Total
Like statistics and probability	20	40	60
Dislike statistics and probability	20	20	40
Total	40	60	100

When x and y are continuous random variables, $\Pr(y) = 0$ and thus the conditional probability cannot be defined by Eq. (5.1). However, the *conditional probability density function* can be defined in the same way as Eq. (5.2), and the conditional expectation is given by

$$E[x|y] = \int x\, g(x|y)\mathrm{d}x.$$

5.3 CONTINGENCY TABLE

A *contingency table* summarizes information of multiple discrete random variables. An example of the contingency table is given in Table 5.1: x is the random variable representing students' likes and dislikes of probability and statistics, while y is the random variable representing their drowsiness during the lecture.

A contingency table corresponds to a probability mass function. The right-most column is called the *row marginal total*, the bottom row is called the *column marginal total*, and the right-bottom cell is called the *grand total*. The row marginal total and the column marginal total correspond to marginal probability mass functions, while each row and each column correspond to conditional probability mass functions. Hypothesis testing using the contingency table will be explained in Section 10.4.

5.4 BAYES' THEOREM

When probability $\Pr(y|x)$ of effect y given cause x is known, *Bayes' theorem* allows us to compute the probability $\Pr(x|y)$ of cause x given effect y as

$$\Pr(x|y) = \frac{\Pr(y|x)\Pr(x)}{\Pr(y)}.$$

Since $\Pr(x)$ is the probability of cause x before effect y is known, it is called the *prior probability* of x. On the other hand, $\Pr(x|y)$ is the probability of cause x after effect y is known, and it is called the *posterior probability* of x. Bayes' theorem can be immediately proved by the definition of joint probability distributions:

$$\Pr(x|y)\Pr(y) = \Pr(x,y) = \Pr(y|x)\Pr(x).$$

Bayes' theorem also holds for continuous random variables x and y, using probability density functions as

$$g(x|y) = \frac{h(y|x)g(x)}{h(y)}.$$

Let us illustrate the usefulness of Bayes' theorem through an example of a *polygraph*. Let x be a random variable representing whether a subject's word is true or false, and let y be its prediction by a polygraph. The polygraph has excellent performance, such that

$$\begin{aligned}\Pr(y = \text{false}|\ x = \text{false}) &= 0.99,\\ \Pr(y = \text{true}|\ x = \text{true}) &= 0.95.\end{aligned}$$

Suppose that the prior probability is

$$\Pr(x = \text{false}) = 0.001.$$

If the polygraph says that the subject's word is false, can we believe its decision? The reliability of the polygraph can be evaluated by comparing

$$\Pr(x = \text{false}|\ y = \text{false}) \quad \text{and} \quad \Pr(x = \text{true}|\ y = \text{false}).$$

Since marginalization of x yields

$$\begin{aligned}\Pr(y = \text{false}) &= \Pr(y = \text{false}|\ x = \text{false})\Pr(x = \text{false})\\ &\quad + \Pr(y = \text{false}|\ x = \text{true})\Pr(x = \text{true})\\ &= \Pr(y = \text{false}|\ x = \text{false})\Pr(x = \text{false})\\ &\quad + \big(1 - \Pr(y = \text{true}|\ x = \text{true})\big)\big(1 - \Pr(x = \text{false})\big)\\ &= 0.99 \times 0.001 + (1 - 0.95) \times (1 - 0.001)\\ &\approx 0.051,\end{aligned}$$

Bayes' theorem results in

$$\begin{aligned}\Pr(x = \text{false}|\ y = \text{false}) &= \frac{\Pr(y = \text{false}|\ x = \text{false})\Pr(x = \text{false})}{\Pr(y = \text{false})}\\ &\approx \frac{0.99 \times 0.001}{0.051} \approx 0.019.\end{aligned}$$

Thus,

$$\begin{aligned}\Pr(x = \text{true}|\ y = \text{false}) &= 1 - \Pr(x = \text{false}|\ y = \text{false})\\ &\approx 0.981\end{aligned}$$

holds and therefore

$$\Pr(x = \text{false}|\ y = \text{false}) \ll \Pr(x = \text{true}|\ y = \text{false}).$$

Consequently, we conclude that the output of the polygraph, $y = \text{false}$, is not reliable.

The above analysis shows that, if the prior probability that the subject tells a lie, $\Pr(x = \text{false})$, is small, the output of the polygraph, $y = \text{false}$, is not reliable. If

$$\Pr(x = \text{false}) > 0.048,$$

it follows that

$$\Pr(x = \text{false}|\, y = \text{false}) > \Pr(x = \text{true}|\, y = \text{false}),$$

showing that the output of the polygraph, $y = \text{false}$, becomes reliable.

5.5 COVARIANCE AND CORRELATION

If random variables x and y are related to each other, change in one variable may affect the other one.

The variance of random variables x and y, $V[x+y]$, does not generally agree with the sum of each variance, $V[x] + V[y]$. Indeed, $V[x + y]$ and $V[x] + V[y]$ are related as

$$V[x + y] = V[x] + V[y] + 2\mathrm{Cov}[x, y],$$

where $\mathrm{Cov}[x, y]$ is the *covariance* of x and y defined by

$$\mathrm{Cov}[x, y] = E\Big[\big(x - E[x]\big)\big(y - E[y]\big)\Big].$$

Increasing x tends to increase y if $\mathrm{Cov}[x, y] > 0$, while increasing x tends to decrease y if $\mathrm{Cov}[x, y] < 0$. If $\mathrm{Cov}[x, y] \approx 0$, x and y are unrelated to each other.

The covariance is useful to create a *portfolio* of stocks. Let x and y be the stock prices of companies A and B, respectively. If $\mathrm{Cov}[x, y] > 0$, buying the stocks of both companies increases the variance. Therefore, the property tends to fluctuate and there is chance to gain a big profit (and at the same time, there is the possibility to lose a lot). On the other hand, if $\mathrm{Cov}[x, y] < 0$, buying the stocks of both companies decreases the variance. Therefore, the risk is hedged and the property is more stabilized (and at the same time, there is less opportunities to gain a big profit).

The matrix that summarizes the variance and covariance of x and y is called the *variance–covariance matrix*:

$$\begin{aligned}\boldsymbol{\Sigma} &= E\left[\left\{\begin{pmatrix} x \\ y \end{pmatrix} - E\begin{pmatrix} x \\ y \end{pmatrix}\right\}\left\{\begin{pmatrix} x \\ y \end{pmatrix} - E\begin{pmatrix} x \\ y \end{pmatrix}\right\}^{\top}\right] \\ &= \begin{pmatrix} V[x] & \mathrm{Cov}[x, y] \\ \mathrm{Cov}[y, x] & V[y] \end{pmatrix},\end{aligned}$$

where $^{\top}$ denotes the transposes. Since $\mathrm{Cov}[y, x] = \mathrm{Cov}[x, y]$, the variance–covariance matrix is symmetric.

The *correlation coefficient* between x and y, denoted by $\rho_{x,y}$, is defined as $\mathrm{Cov}[x, y]$ normalized by the product of standard deviations $\sqrt{V[x]}$ and $\sqrt{V[y]}$,

$$\rho_{x,y} = \frac{\mathrm{Cov}[x,y]}{\sqrt{V[x]}\sqrt{V[y]}}.$$

Since the correlation coefficient is a normalized variant of the covariance, it essentially plays the same role as the covariance. However, the correlation coefficient is bounded as

$$-1 \leq \rho_{x,y} \leq 1, \tag{5.3}$$

and therefore the absolute strength of the relation between x and y can be known. Eq. (5.3) can be proved by the generic inequality

$$|E[x]| \leq E[|x|]$$

and Schwarz's inequality explained in Section 8.3.2,

$$|\mathrm{Cov}[x,y]| \leq E\big[|(x-E[x])(y-E[y])|\big] \leq \sqrt{V[x]}\sqrt{V[y]}.$$

Examples of the correlation coefficient are illustrated in Fig. 5.1. If $\rho_{x,y} > 0$, x and y have the same tendency and x and y are said to be *positively correlated*. On the other hand, if $\rho_{x,y} < 0$, x and y have the opposite tendency and x and y are said to be *negatively correlated*. x and y are deterministically proportional to each other if $\rho_{x,y} = \pm 1$, and x and y are unrelated if $\rho_{x,y} \approx 0$. If $\rho_{x,y} = 0$, x and y are said to be *uncorrelated* to each other. Thus, the correlation coefficient allows us to capture the relatedness between random variables. However, if x and y have nonlinear relation, the correlation coefficient can be close to zero, as illustrated in Fig. 5.2.

5.6 INDEPENDENCE

x and y are said to be statistically *independent* if

$$f(x,y) = g(x)h(y) \quad \text{for all } x \text{ and } y.$$

Conversely, if x and y are not independent, they are said to be *dependent*. If x and y are independent, the following properties hold:

- Conditional probability is independent of the condition,

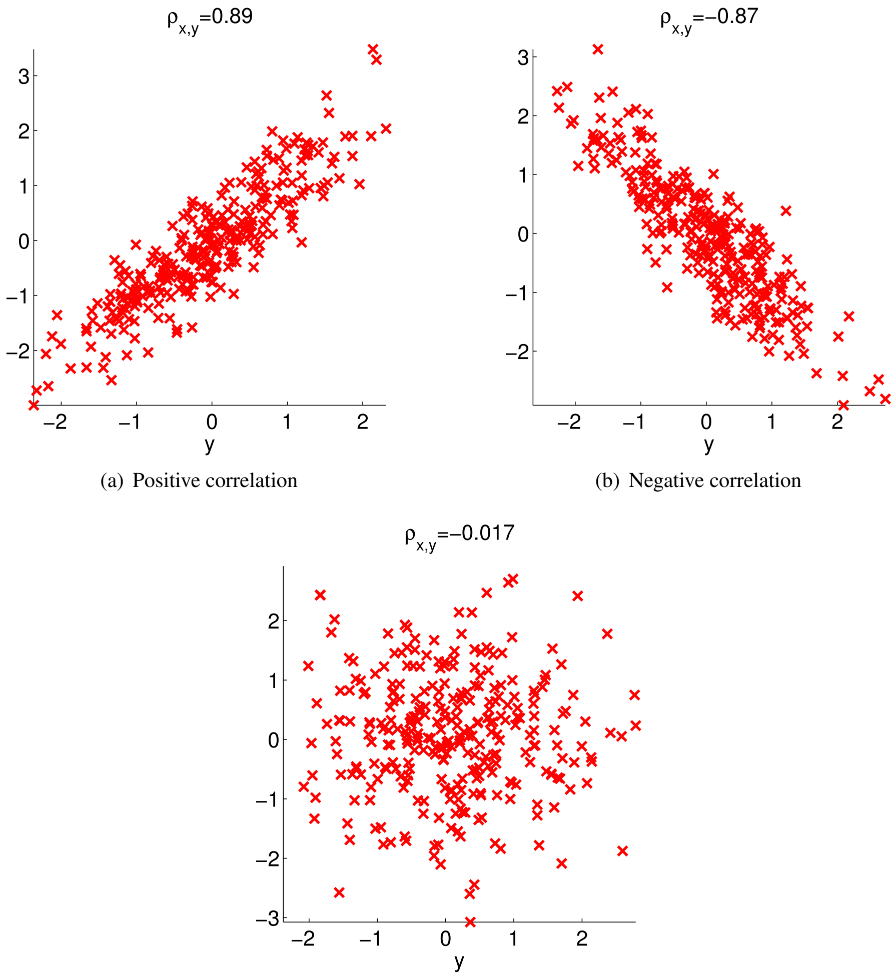
$$g(x|y) = g(x) \quad \text{and} \quad h(y|x) = h(y).$$

- The expectation of the product agrees with the product of the expectations,

$$E[xy] = E[x]E[y].$$

- The moment-generating function of the sum agrees with the product of the moment-generating functions,

$$M_{x+y}(t) = M_x(t)M_y(t).$$



(a) Positive correlation

(b) Negative correlation

(c) No correlation

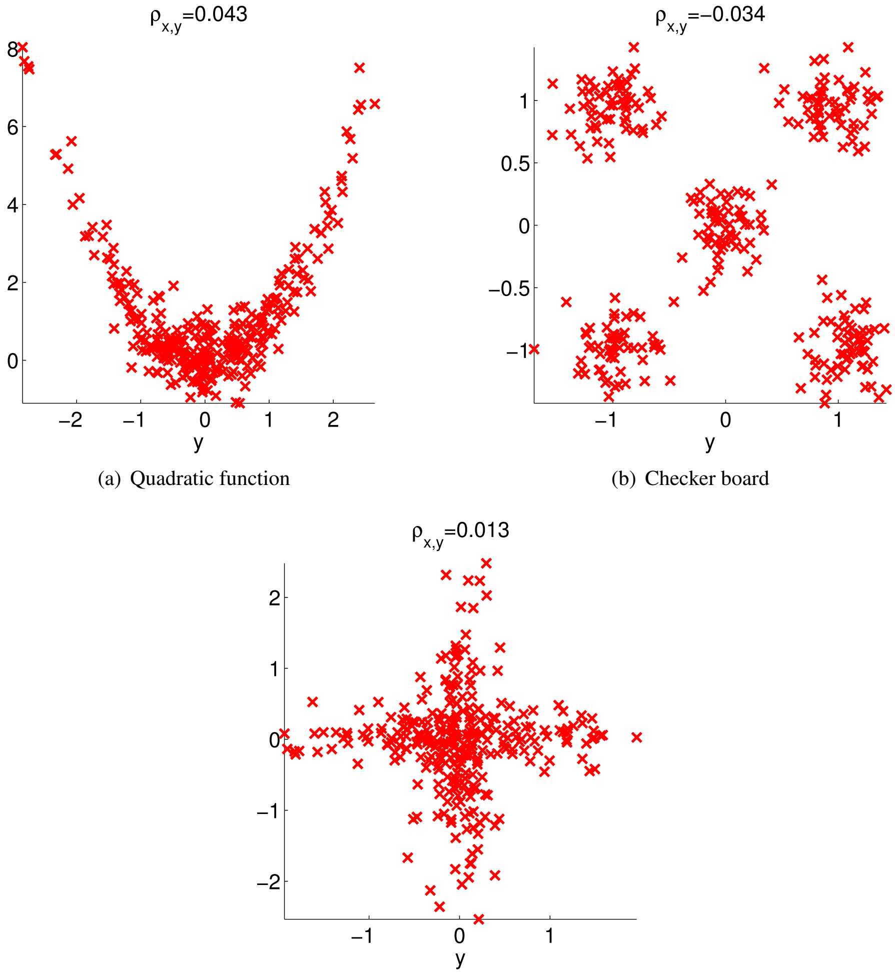
FIGURE 5.1

Correlation coefficient $\rho_{x,y}$. Linear relation between x and y can be captured.

- Uncorrelated,

$$\mathrm{Cov}[x, y] = 0.$$

Although independence and no correlation both mean that x and y are "unrelated," independence is stronger than no correlation. Indeed, independence implies no correlation but not *vice versa*. For example, random variables x and y whose joint



(a) Quadratic function

(b) Checker board

(c) Cross

FIGURE 5.2

Correlation coefficient for nonlinear relations. Even when there is a nonlinear relation between x and y, the correlation coefficient can be close to zero if the probability distribution is symmetric.

probability density function is given by

$$f(x, y) = \begin{cases} 1 & (|x| + |y| \leq \frac{1}{\sqrt{2}}) \\ 0 & (\text{otherwise}) \end{cases}$$

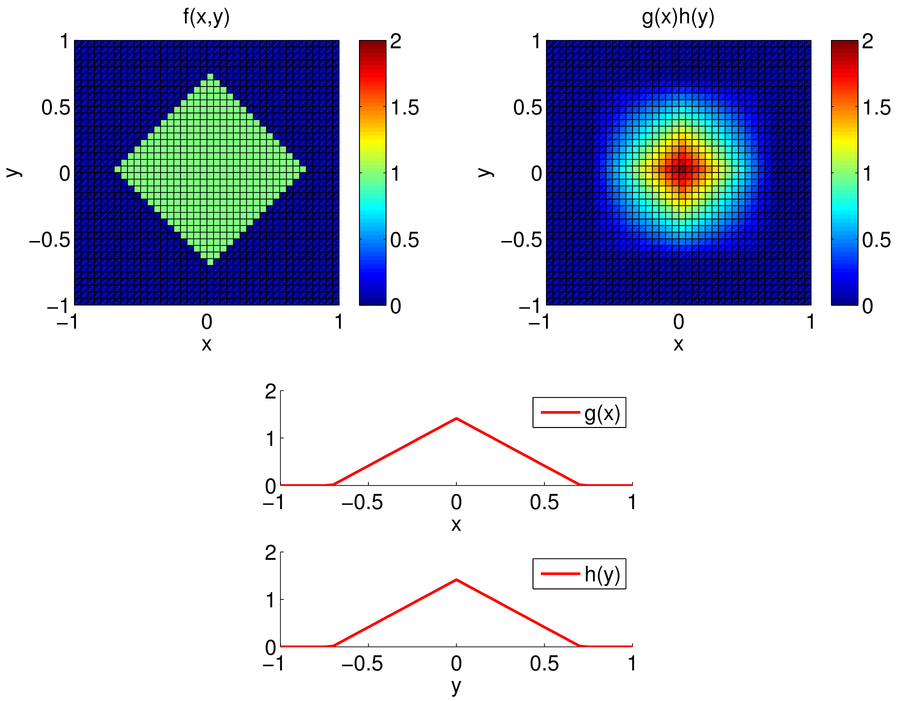


FIGURE 5.3

Example of x and y which are uncorrelated but dependent.

have no correlation but dependent (see Fig. 5.3). More precisely, the uncorrelatedness of x and y can be confirmed by

$$\begin{aligned}\mathrm{Cov}[x,y] &= E\Big[\big(x - E[x]\big)\big(y - E[y]\big)\Big] = E[xy] \\ &= \int_{-\frac{1}{\sqrt{2}}}^{\frac{1}{\sqrt{2}}} x \left(\int_{-\frac{1}{\sqrt{2}}+|x|}^{\frac{1}{\sqrt{2}}-|x|} y\mathrm{d}y \right) \mathrm{d}x = -\int_{-\frac{1}{\sqrt{2}}}^{\frac{1}{\sqrt{2}}} \sqrt{2}x|x|\mathrm{d}x = 0,\end{aligned}$$

while dependence of x and y, $f(x,y) \neq g(x)h(y)$, can be confirmed by

$$\begin{aligned} g(x) &= \max\big(0, \sqrt{2} - 2|x|\big), \\ h(y) &= \max\big(0, \sqrt{2} - 2|y|\big). \end{aligned}$$