Contents

CHAPTER 1 Data Mining and Analysis

1.7 EXERCISES

Q1. Show that the mean of the centered data matrix $\mathbf{Z}$ in Eq. (1.5) is $\mathbf{0}$.

Answer: Each centered point is given as: $\mathbf{z}_i = \mathbf{x}_i - \boldsymbol{\mu}$. Their mean is therefore:

$$
\begin{aligned}
\frac{1}{n}\sum_{i=0}^{n} \mathbf{z}_i &= \frac{1}{n}\sum_{i=0}^{n} (\mathbf{x}_i - \boldsymbol{\mu}) \\
&= \frac{1}{n}\sum_{i=0}^{n} \mathbf{x}_i - \frac{1}{n} \cdot n \cdot \boldsymbol{\mu} \\
&= \boldsymbol{\mu} - \boldsymbol{\mu} = \mathbf{0}
\end{aligned}
$$

Q2. Prove that for the L_p-distance in Eq. (1.2), we have

$$
\delta_\infty(\mathbf{x}, \mathbf{y}) = \lim_{p \to \infty} \delta_p(\mathbf{x}, \mathbf{y}) = \max_{i=1}^{d} \big\{ |x_i - y_i| \big\}
$$

for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Answer: We have to show that

$$
\lim_{p \to \infty} \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}} = \max_{i=1}^{d} \big\{ |x_i - y_i| \big\}
$$

Assume that dimension a is the max, and let $m = |x_a - y_a|$. For simplicity, we assume that $|x_i - y_i| < m$ for all $i \neq a$.

If we divide and multiply the left hand side with m^p we get:

$$
\left(m^p \sum_{i=1}^{d} \left(\frac{|x_i - y_i|}{m} \right)^p \right)^{\frac{1}{p}} = m \left(1 + \sum_{i \neq a} \left(\frac{|x_i - y_i|}{m} \right)^p \right)^{\frac{1}{p}}
$$

As $p \to \infty$, each term $\left(\frac{|x_i - y_i|}{m}\right)^p \to 0$, since $m > |x_i - y_i|$ for all $i \neq a$. The finite summation $\sum_{i \neq a} \left(\frac{|x_i - y_i|}{m}\right)^p$ converges to 0 as $p \to \infty$, as does $1/p$.

Thus $\delta_\infty(\mathbf{x}, \mathbf{y}) = m \times 1^0 = m = |x_a - y_a| = \max_{i=1}^d \{|x_i - y_i|\}$

Note that the same result is obtained even if we assume that dimensions other than a achieve the maximum value m. In the worst case, we have $m = |x_i - y_i|$ for all d dimensions. In this case, the expression on LHS becomes

$$\lim_{p \to \infty} m \left(\sum_{i=1}^d 1^p \right)^{1/p} = \lim_{p \to \infty} m d^{1/p} = \lim_{p \to \infty} m d^0 = m$$

PART ONE DATA ANALYSIS FOUNDATIONS

CHAPTER 2 Numeric Attributes

2.7 EXERCISES

Q1. True or False:

(a) Mean is robust against outliers.

Answer: False

(b) Median is robust against outliers.

Answer: True

(c) Standard deviation is robust against outliers.

Answer: False

Q2. Let X and Y be two random variables, denoting age and weight, respectively. Consider a random sample of size $n = 20$ from these two variables

$$X = (69, 74, 68, 70, 72, 67, 66, 70, 76, 68, 72, 79, 74, 67, 66, 71, 74, 75, 75, 76)$$

$$Y = (153, 175, 155, 135, 172, 150, 115, 137, 200, 130, 140, 265, 185, 112, 140, 150, 165, 185, 210, 220)$$

(a) Find the mean, median, and mode for X.

Answer: The mean, median, and mode are:

$$\mu = \frac{1}{20}\sum_{i=1}^{2} 0x_i = 1429/20 = 71.45$$

$$\text{median} = (71 + 72)/2 = 71.5$$

$$\text{mode} = 74$$

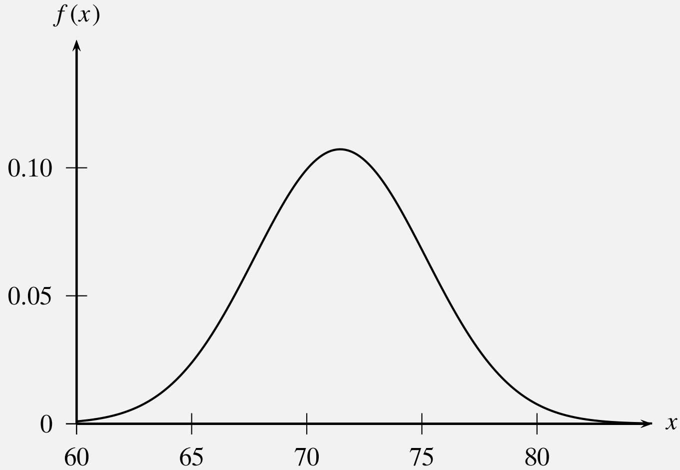
(b) What is the variance for Y?

Answer: The mean of Y is $\mu_Y = 3294/20 = 164.7$. The variance is:

$$\sigma_Y^2 = \frac{1}{20}\sum_{i=1}^{2} 0 y_i - \mu_Y = 27384.2/20 = 1369.21$$

(c) Plot the normal distribution for X.

Answer: The mean for X is $\mu_X = 71.45$, and the variance is $\sigma_X^2 = 13.8475$, with a standard deviation of $\sigma_X = 3.72$.



(d) What is the probability of observing an age of 80 or higher?

Answer: If we leverage the empirical probability mass function, we get:

$$P(X \geq 80) = 0/20 = 0$$

since we do not have anyone with age 80 or more in our sample.
We can use the normal distribution modeling, with parameters $\mu_X = 71.45$ and $\sigma_X^2 = 3.72$ to get:

$$P(X \geq 80) = \int_{80}^{\infty} N(x|\mu_X, \sigma_X) = 0.010769$$

(e) Find the 2-dimensional mean $\hat{\boldsymbol{\mu}}$ and the covariance matrix $\widehat{\boldsymbol{\Sigma}}$ for these two variables.

Answer: The mean and covariance matrices are:

$$\boldsymbol{\mu} = (\mu_X, \mu_Y)^T = (71.45, 164.7)^T$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} 13.8475 & 122.435 \\ 122.435 & 1369.21 \end{pmatrix}$$

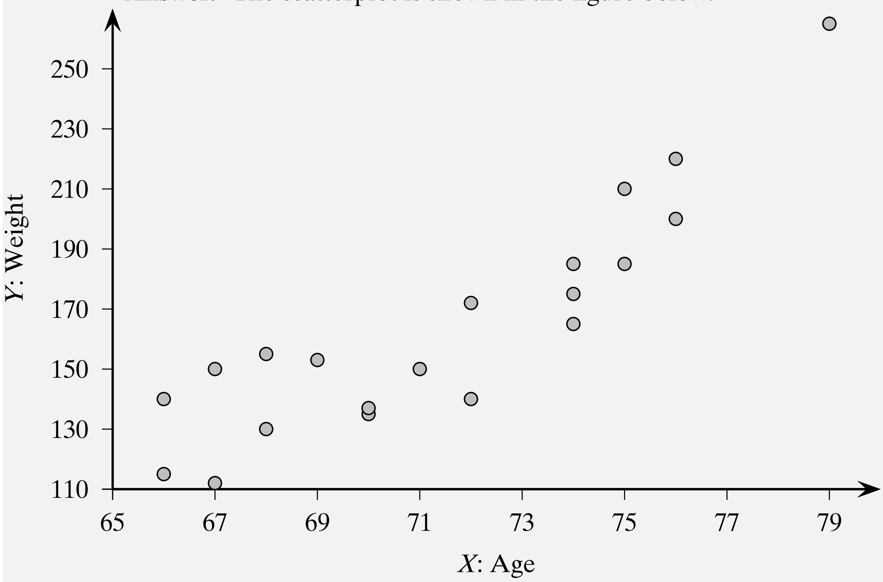
(f) What is the correlation between age and weight?

Answer:

$$\rho_{XY} = \sigma_{XY}/(\sigma_X \sigma_Y) = \frac{122.435}{\sqrt{13.845 \cdot 1369.21}} = 0.889$$

(g) Draw a scatterplot to show the relationship between age and weight.

Answer: The scatterplot is shown in the figure below.



Q3. Show that the identity in Eq. (2.15) holds, that is,

$$\sum_{i=1}^{n}(x_i - \mu)^2 = n(\hat{\mu} - \mu)^2 + \sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Answer: Consider the RHS

$$\begin{aligned}
n(\hat{\mu} - \mu)^2 + \sum_{i=1}^{n}(x_i - \hat{\mu})^2 &= n(\hat{\mu}^2 - 2\hat{\mu}\mu + \mu^2) + \sum_{i=1}^{n}(x_i^2 - 2x_i\hat{\mu} + \hat{\mu}^2) \\
&= n\hat{\mu}^2 - 2n\hat{\mu}\mu + n\mu^2 + \left(\sum_{i=1}^{n} x_i^2\right) - 2n\hat{\mu}^2 + n\hat{\mu}^2 \\
&= \left(\sum_{i=1}^{n} x_i^2\right) - 2n\hat{\mu}\mu + n\mu^2 \\
&= \left(\sum_{i=1}^{n} x_i^2\right) - 2n\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\mu + \sum_{i=1}^{n}\mu^2 \\
&= \sum_{i=1}^{n}(x_i - \mu)^2
\end{aligned}$$

Q4. Prove that if x_i are independent random variables, then

$$var\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} var(x_i)$$

This fact was used in Eq. (2.12).

Answer: We assume for simplicity that all the variables are discrete. A similar approach can be used for continuous variables.

Consider the random variable $x_1 + x_2$. Its mean is given as

$$\mu_{x_1+x_2} = \sum_{x_1=a}\sum_{x_2=b} (a+b) f(a,b)$$

Since x_1 and x_2 are independent, their joint probability mass function is given as:

$$f(x_1, x_2) = f(x_1) \cdot f(x_2)$$

Thus, the mean is given as

$$\begin{aligned}
\mu_{x_1+x_2} &= \sum_{x_1=a}\sum_{x_2=b} (a+b) f(a,b) \\
&= \sum_{x_1=a}\sum_{x_2=b} (a+b) f(a) f(b) \\
&= \sum_{x_1=a} f(a) \sum_{x_2=b} (a+b) f(b) \\
&= \sum_{x_1=a} f(a) \left(\sum_{x_2=b} a f(b) + \sum_{x_2=b} b f(b)\right) \\
&= \sum_{x_1=a} f(a) \left(a + \mu_{x_2}\right) \\
&= \mu_{x_1} + \mu_{x_2}
\end{aligned}$$

In general, we can show that the expected value of the sum of the variables x_i is the sum of their expected values, i.e.,

$$E\left[\sum_{i=1}^{n} x_i\right] = \sum_{i=1}^{n} E[x_i]$$

Now, let us consider the variance of the sum of the random variables:

$$\begin{aligned}
var\left(\sum_{i=1}^{n} x_i\right) &= E\left[\left(\sum_{i=1}^{n} x_i - E\left[\sum_{i=1}^{n} x_i\right]\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} E[x_i]\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n} (x_i - E[x_i])\right)^2\right]
\end{aligned}$$

$$= E\left[\sum_{i=1}^{n}(x_i - E[x_i])^2 + 2\sum_{i=1}^{n}\sum_{j>i}(x_i - E[x_i])(x_j - E[x_j])\right]$$

$$= \sum_{i=1}^{n} E\left[(x_i - E[x_i])^2\right] + 2\sum_{i=1}^{n}\sum_{j>i} cov(x_i, x_j)$$

$$= \sum_{i=1}^{n} var(x_i)$$

The last step follows from the fact that $cov(x_i, x_j) = 0$ since they are independent.

Q5. Define a measure of deviation called *mean absolute deviation* for a random variable X as follows:

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \mu|$$

Is this measure robust? Why or why not?

Answer: No, it is not robust, since a single outlier can skew the mean absolute deviation.

Q6. Prove that the expected value of a vector random variable $\mathbf{X} = (X_1, X_2)^T$ is simply the vector of the expected value of the individual random variables X_1 and X_2 as given in Eq. (2.18).

Answer: This follows directly from the definition of expectation of a vector random variable. When both X_1 and X_2 are discrete we have

$$\boldsymbol{\mu} = E[\mathbf{X}] = \sum_{\mathbf{x}} \mathbf{x} f(\mathbf{x}) = \sum_{x_1}\sum_{x_2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} f(x_1, x_2) = \begin{pmatrix} \mu_{X_1} \\ \mu_{X_2} \end{pmatrix}$$

Likewise, when both X_1 and X_2 are continuous we have

$$\boldsymbol{\mu} = E[\mathbf{X}] = \int\int_{\mathbf{x}} \mathbf{x} f(\mathbf{x}) d\mathbf{x} = \int_{x_1}\int_{x_2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} f(x_1, x_2) dx_1 dx_2 = \begin{pmatrix} \mu_{X_1} \\ \mu_{X_2} \end{pmatrix}$$

In more detail, assume that both X_1 and X_2 are discrete, we have

$$\boldsymbol{\mu} = E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \sum_{x_1, x_2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} f(x_1, x_2) = \begin{pmatrix} \sum_{x_1, x_2} x_1\ f(x_1, x_2) \\ \sum_{x_1, x_2} x_2\ f(x_1, x_2) \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{x_1} x_1 \sum_{x_2} f(x_1, x_2) \\ \sum_{x_2} x_2 \sum_{x_1} f(x_1, x_2) \end{pmatrix} = \begin{pmatrix} \sum_{x_1} x_1\ f(x_1) \\ \sum_{x_2} x_2\ f(x_2) \end{pmatrix} = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_{X_1} \\ \mu_{X_2} \end{pmatrix}$$

where $f(x_1, x_2) = p(X_1 = x_1, X_2 = x_2)$ is the joint probability mass function of X_1 and X_2, and $f(x_1) = \sum_{x_2} f(x_1, x_2)$ and $f(x_2) = \sum_{x_1} f(x_1, x_2)$ are the *marginal probability distributions* of X_1 and X_2, respectively. Note that X_1 and X_2 do not have to be independent for the above to hold.

Q7. Show that the correlation [Eq. (2.23)] between any two random variables X_1 and X_2 lies in the range $[-1, 1]$.

Answer: The Cauchy-Schwartz inequality states that for any two vectors $\mathbf{x}$ and $\mathbf{y}$ in an inner product space, they satisfy:

$$|\langle \mathbf{x}, \mathbf{y} \rangle|^2 \leq \langle \mathbf{x}, \mathbf{x} \rangle \cdot \langle \mathbf{y}, \mathbf{y} \rangle$$

Define the inner product between two random variables X_1 and X_2 as follows:

$$\langle X_1, X_2 \rangle = E[X_1 X_2]$$

Expectation is a valid inner product since it satisfies the three conditions: i) symmetric: $E[X_1X_2] = E[X_2X_2]$, ii) positive-semidefinite: $E[X_1X_2] = E[X_1^2] \geq 0$, and iii) linear: $E[(aX_1)X_2] = aE[X_1X_2]$ and $E[(X_1 + Z)X_2] = E[X_1X_2] + E[ZX_2]$.

Then, we have

$$\begin{aligned}
|\sigma_{12}| &= \left|cov(X_1, X_2)\right|^2 \\
&= \left|E[(X_1 - \mu_1)(X_2 - \mu_2)]\right|^2 \\
&= \left|\langle (X_1 - \mu_1)(X_2 - \mu_2) \rangle\right|^2 \\
&\leq \langle X_1 - \mu_1, X_1 - \mu_1 \rangle \cdot \langle X_2 - \mu_2, X_2 - \mu_2 \rangle \\
&= E[X_1 - \mu_1] \cdot E[X_2 - \mu_2] \\
&= \sigma_1 \cdot \sigma_2
\end{aligned}$$

Since $|\sigma_{12}| \leq \sigma_1 \cdot \sigma_2$, it follows that the correlation $\rho_{12} = \sigma_{12}/\sigma_1\sigma_2$ lies in the range $[-1, 1]$.

Q8. Given the dataset in Table 2.1, compute the covariance matrix and the generalized variance.

Table 2.1. Dataset for Q8

	X_1	X_2	X_3
$\mathbf{x}_1$	17	17	12
$\mathbf{x}_2$	11	9	13
$\mathbf{x}_3$	11	8	19

Answer: The covariance matrix is:

$$\mathbf{\Sigma} = \begin{pmatrix} 8.0 & 11.33 & -5.33 \\ 11.33 & 16.22 & -8.56 \\ -5.33 & -8.56 & 9.56 \end{pmatrix}$$

The generalized variance is:

$$\det(\mathbf{\Sigma}) = -1.38 \times 10^{-13}$$

Q9. Show that the outer-product in Eq. (2.31) for the sample covariance matrix is equivalent to Eq. (2.29).

Answer: Let $\mathbf{z}_i = \mathbf{x}_i - \hat{\boldsymbol{\mu}}$ denote a centered data point. The outer product form of covariance matrix is given as:

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^T$$

Let us consider the entry in cell (j,k); we have:

$$\widehat{\boldsymbol{\Sigma}}(j,k) = \frac{1}{n}\sum_{i=1}^{n} z_{ij} z_{ik} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \hat{\mu}_j)(x_{ik} - \hat{\mu}_k) = \hat{\sigma}_{jk}$$

which is exactly the covariance between the j-th and k-th attribute.

Q10. Assume that we are given two univariate normal distributions, N_A and N_B, and let their mean and standard deviation be as follows: $\mu_A = 4, \sigma_A = 1$ and $\mu_B = 8, \sigma_B = 2$.

(a) For each of the following values $x_i \in \{5,6,7\}$ find out which is the more likely normal distribution to have produced it.

Answer: If we plug-in x_i in the equation for the normal distribution, we obtain the following:

$$N_A(5) = 0.242 \qquad N_B(5) = 0.065$$
$$N_A(6) = 0.054 \qquad N_B(6) = 0.121$$
$$N_A(7) = 0.004 \qquad N_B(7) = 0.176$$

Based on these values, we can claim that N_A is more likely to have produced 5, but N_B is more likely to have produced 6 and 7.

We can also solve this problem by finding the z-score for each value. We can then assign a point to the distribution for which it has a lower z-score (in terms of absolute value). For example, for 5, we have $z_A(5) = (5-4)/1 = 1$, and $z_B(5) = (5-8)/2 = -1.5$. Since $|z_B| > |z_A|$ we can claim that 5 comes from N_A.

For 6 and 7 we have:

$$z_A(6) = (6-4)/1 = 2 \qquad z_B(6) = (6-8)/2 = -1$$
$$z_A(6) = (7-4)/1 = 3 \qquad z_B(7) = (7-8)/2 = -0.5$$

Thus, these values are more likely to have been generated from N_B.

(b) Derive an expression for the point for which the probability of having been produced by both the normals is the same.

Answer: Plugging in the parameters of N_A and N_B into the equation for the normal distribution, and after setting up the equality, we obtain:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-4)^2}{2}} = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-8)^2}{8}}$$

$$2e^{\frac{(x-8)^2}{8}} = e^{\frac{(x-4)^2}{2}}$$

taking ln on both sides yields

$$\begin{aligned}
\ln(2) + \frac{(x-8)^2}{8} &= \frac{(x-4)^2}{2} \\
\frac{8\ln(2) + x^2 + 64 - 16x}{8} &= \frac{x^2 + 16 - 8x}{2} \\
2\ln(2) + x^2/4 - 4x &= x^2 - 8x \\
\frac{3}{4}x^2 - 4x - 2\ln(2) &= 0 \\
0.75x^2 - 4x - 1.4 &= 0
\end{aligned}$$

We can solve this equation using the general solution for a quadratic equation: $\frac{-b \pm \sqrt{b^2-4ac}}{2a}$. Plugging in the values from above we get $x = 5.67$.

Q11. Consider Table 2.2. Assume that both the attributes X and Y are numeric, and the table represents the entire population. If we know that the correlation between X and Y is zero, what can you infer about the values of Y?

Table 2.2. Dataset for Q11

X	Y
1	a
0	b
1	c
0	a
0	c

Answer: Since the correlation is zero, we have $cov(X,Y) = 0$, which implies that $E[XY] = E[X]E[Y]$. From the data we have

$$E[XY] = (a+c)/5 \qquad E[X] = 2/5 \qquad E[Y] = (2a+2c+b)/5$$

Equating these we get

$$\begin{aligned}
(a+c)/5 &= 2(2a+2c+b)/25 \\
5a + 5c &= 4a + 4c + 2b \\
a + c &= 2b
\end{aligned}$$

Q12. Under what conditions will the covariance matrix $\mathbf{\Sigma}$ be identical to the correlation matrix, whose (i, j) entry gives the correlation between attributes X_i and X_j? What can you conclude about the two variables?

Answer: If the covariance matrix equals the correlation matrix, this means that for all i and j, we have

$$\rho_{ij} = \sigma_{ij}$$
$$\frac{\sigma_{ij}}{\sigma_i \sigma_j} = \sigma_{ij}$$
$$\sigma_i \sigma_j = 1$$

Thus, for the covariance matrix to equal the correlation matrix, X_i and X_j must be perfectly correlated; either $\sigma_i = \sigma_j = 1$ or $\sigma_i = \sigma_j = -1$.