Homework 2

Math and Statistics Foundations for Data Science

2022-24790 Sungwoo PARK
Graduate School of Data Science

April 7, 2022

1 Mathematics

Exercise 4.3 (b)

Answer: $E_{\lambda_1} = \text{span}\left\{\frac{1}{\sqrt{5}}\begin{pmatrix}1\\2\end{pmatrix}\right\}$, $E_{\lambda_2} = \text{span}\left\{\frac{1}{\sqrt{5}}\begin{pmatrix}-2\\1\end{pmatrix}\right\}$

By the definition of eigenvalue and eigenvector, there exists a vector $\mathbf{v} \neq \mathbf{0}$ such that satiesfies

$$B\mathbf{v} = \lambda\mathbf{v} \text{ for } \lambda \in \mathbb{R} \tag{1}$$

Then, by the distribution rule of the matrix multiplication, it is equivalent to

$$(B - \lambda I_2) = \mathbf{0} \text{ for } B \in \mathbb{R}^{2\times 2} \tag{2}$$

The fact that $\mathbf{v} \neq \mathbf{0}$ implies that the equation $(B - \lambda I_2) = \mathbf{0}$ does not have the unique solution, and that $\det(B - \lambda I_2) = 0$. Using this, we can obtain the eigenvalues by

$$\det(B - \lambda I_2) = \det\begin{pmatrix}-2-\lambda & 2\\ 2 & 1-\lambda\end{pmatrix} = 0 \tag{3}$$

The equation 3 is equivalent to $\lambda^2 + \lambda - 6 = (\lambda - 2)(\lambda + 3) = 0$, thus $\lambda_1 = 2$ and $\lambda_2 = -3$. By plugging these into the equation 2 again, we can obtain an eigenvector associated with each eigenvalue: $\mathbf{v}_1 = \frac{1}{\sqrt{5}}\begin{pmatrix}1\\2\end{pmatrix}$ and $\mathbf{v}_2 = \frac{1}{\sqrt{5}}\begin{pmatrix}-2\\1\end{pmatrix}$. Thus, the eigenspace associated with each eigenvalue is a vector space spanned by the corresponding eigenvector.

Exercise 4.8

Answer: $\sigma_1 = 5, \sigma_2 = 3, \sigma_3 = 0$

By the singular value theorem, let $A = U\Sigma\mathbf{v}^T$, where $U \in \mathbb{R}^{2\times2}$, $\Sigma \in \mathbb{R}^{2\times3}$, and an orthogonal matrix $\mathbf{v}^T \in \mathbb{R}^{3\times3}$. Then since $A^TA = \left(U\Sigma\mathbf{v}^T\right)^T\left(U\Sigma\mathbf{v}^T\right) = \mathbf{v}\Sigma^T\Sigma\mathbf{v}^T$ is symmetric, thus diagonalizable, and $\Sigma^T\Sigma \in \mathbb{R}^{3\times3}$ is the diagonal matrix whose elements, σ_i^2 for $i \in \{1,2,3\}$, are the eigenvalues of A^TA.

The eigenvalues of $A^TA = \begin{pmatrix} 13 & 12 & 2 \\ 12 & 13 & -2 \\ 2 & -2 & 8 \end{pmatrix}$ is obtained by solving the equation

$$\det\left(A^TA - \lambda I_3\right) = \det\begin{pmatrix} 13-\lambda & 12 & 2 \\ 12 & 13-\lambda & -2 \\ 2 & -2 & 8-\lambda \end{pmatrix} = 0 \tag{4}$$

By expanding, this equation is equivalent to $-\lambda^3 + 34\lambda^2 - 225\lambda = -\lambda\left(\lambda - 9\right)\left(\lambda - 25\right) = 0$. Thus, $\lambda_1 = 25, \lambda_2 = 9$, and $\lambda_3 = 0$. Note that since A^TA is positive semi-definite, the singular values cannot be negative. Therefore, $\sigma_1 = \sqrt{\lambda_1}, \sigma_2 = \sqrt{\lambda_2}$, and $\sigma_3 = \sqrt{\lambda_3}$.

Exercise 5.2

Let $y = \exp\left(-x\right)$, then $f\left(y\right) = \frac{1}{1+y} = \left(1+y\right)^{-1}$. Therefore, by applying the chain rule,

$$\frac{df}{dx} = \frac{\partial f}{\partial y}\cdot\frac{\partial y}{\partial x} = -\left(1+y\right)^{-2}\cdot-\exp\left(-x\right) \tag{5}$$

By plugging $y = \exp\left(-x\right)$, the derivative is finally obtained as

$$\frac{\exp\left(-x\right)}{\left(1+\exp\left(-x\right)\right)^2} \tag{6}$$

Exercise 5.3

Let $y = -\frac{1}{2\sigma^2}\left(x-\mu\right)^2$, then $f\left(y\right) = \exp\left(y\right)$. Therefore, by applying the chain rule,

$$\frac{df}{dx} = \frac{\partial f}{\partial y}\cdot\frac{\partial y}{\partial x} = -\exp\left(y\right)\cdot\left[-\frac{1}{\sigma^2}\left(x-\mu\right)\right] \tag{7}$$

By plugging $y = -\frac{1}{2\sigma^2}\left(x-\mu\right)^2$, the derivative is finally obtained as

$$\exp\left[-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right]\cdot\left[-\frac{1}{\sigma^2}\left(x-\mu\right)\right] \tag{8}$$

Exercise 5.7

(a)

First, by the chain rule, $\frac{df}{d\boldsymbol{x}} = \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial \boldsymbol{x}}$. Here, since z is defined as the dot product of $\boldsymbol{x}$ itself, z is a scalar, thus the dimension of the former part, $\frac{\partial f}{\partial z}$, equals to 1. Next, since $\boldsymbol{x} = \begin{pmatrix} x_1 & x_2 & \dots & x_D \end{pmatrix}^T$, $z = \boldsymbol{x}^T\boldsymbol{x}$ can be represented as $z = x_1^2 + x_2^2 + \cdots + x_D^2$.

Therefore, z is differentiated by every element in $\mathbf{x}$, thus the dimension of the latter part, $\frac{\partial z}{\partial \boldsymbol{x}}$, equals to D. More precisely, $\frac{\partial z}{\partial \boldsymbol{x}} \in \mathbb{R}^{1\times D}$ to satisfy the matrix multiplication.

Now, since $f(z) = \log(1+z)$, $\frac{\partial f}{\partial z} = \frac{1}{1+z}$ and $\frac{\partial z}{\partial \boldsymbol{x}} = \begin{pmatrix} 2x_1 & 2x_2 & \dots & 2x_D \end{pmatrix} = 2\boldsymbol{x}^T$. By plugging in $z = \boldsymbol{x}^T\boldsymbol{x}$, the derivative is:

$$\frac{1}{1+\boldsymbol{x}^T\boldsymbol{x}} \cdot 2\boldsymbol{x}^T = \frac{2\boldsymbol{x}^T}{1+\boldsymbol{x}^T\boldsymbol{x}} \tag{9}$$

(b)

Here $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \in \mathbb{R}^E$ is a vector, thus $f(\boldsymbol{z})$ is a vector as well, and its dimension is E. Thus $\frac{\partial f}{\partial \boldsymbol{z}} \in \mathbb{R}^{E\times E}$, and this implies that $\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} \in \mathbb{R}^{E\times D}$ to satisfy the matrix multiplication. Therefore, the dimension of the derivative $\frac{df}{d\boldsymbol{x}}$ equals to $E \times D$.

Now, let's compute the actual derivative. Since

$$f(\boldsymbol{z}) = \begin{pmatrix} \sin(z_1) \\ \vdots \\ \sin(z_i) \\ \vdots \\ \sin(z_E) \end{pmatrix} \tag{10}$$

thus the partial derivative with respect to z_i, equals to $\begin{pmatrix} 0 & \dots & \cos(z_i) & \dots & 0 \end{pmatrix}^T$. This means that the partial derivative $\frac{\partial f}{\partial \boldsymbol{z}}$, which yields a matrix, is a diagonal matrix where each diagonal element is $\cos(z_i)$ for $i \in \{1, \dots, E\}$.

Next, an ith element in $\boldsymbol{z}$, z_i, is calculated as

$$z_i = A_{i1}x_1 + \cdots + A_{iD}x_D + b_i = \sum_{j=1}^{D} A_{ij}x_j + b_i \tag{11}$$

thus $\frac{\partial z_i}{\partial x_j} = A_{ij}$. Expanding this to every column, $\frac{\partial z_i}{\partial \boldsymbol{x}} = \begin{pmatrix} A_{i1} & \dots & A_{iD} \end{pmatrix}$. Therefore, we can

infer that $\frac{\partial z_i}{\partial x_j}$ equals to expanding $\frac{\partial z_i}{\partial \boldsymbol{x}}$ to every $i \in \{1, \dots, E\}$, which yields

$$\begin{pmatrix} A_{11} & \dots & A_{1D} \\ \vdots & \ddots & \vdots \\ A_{E1} & \dots & A_{ED} \end{pmatrix} = \boldsymbol{A} \tag{12}$$

Finally, by the chain rule, $\frac{df}{d\boldsymbol{x}}$ is calculated as

$$\operatorname{diag}(\cos(\boldsymbol{z})) \cdot \boldsymbol{A} = \operatorname{diag}(\cos(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})) \cdot \boldsymbol{A} \tag{13}$$

2 Probability

Exercise 1.11

Since A and B are independent, $\Pr(A \cap B) = \Pr(A)\Pr(B)$. Then,

$$\Pr(A^c \cap B^c) = \Pr((A \cup B)^c) = 1 - \Pr(A \cup B) = 1 - [\Pr(A) + \Pr(B) - \Pr(A \cap B)] \tag{14}$$

By the independence,

$$1 - [\Pr(A) + \Pr(B) - \Pr(A \cap B)] = 1 - [\Pr(A) + \Pr(B) - \Pr(A)\Pr(B)] \tag{15}$$

and this can be factorized as $(1 - \Pr(A))(1 - \Pr(B)) = \Pr(A^c)\Pr(B^c)$.

Therefore, $\Pr(A^c \cap B^c) = \Pr(A^c)\Pr(B^c)$, thus A^c and B^c are independent as well.

Exercise 1.15

(a)

Let X denote a random variable representing the number of children who have blue eyes. Knowing that at least one child has blue eyes, the conditional probability of at least two children having blue eyes is denoted as

$$\Pr(X \geq 2 \mid X \geq 1) = \frac{\Pr(X \geq 2 \cap X \geq 1)}{\Pr(X \geq 1)}$$

For the numerator, it is equal to $\Pr(X \geq 2) = 1 - \Pr(X < 2)$, where $\Pr(X < 2) = \Pr(X = 0) + \Pr(X = 1)$. Since there are three children and independence is assumed among them, $\Pr(X = 1)$ is calculated as

$$\binom{3}{1}\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^2 = \frac{27}{64} \tag{16}$$

Since $\Pr(X=0) = \frac{27}{64}$ as well, $\Pr(X \geq 2) = 1 - \frac{27}{64} - \frac{27}{64} = \frac{10}{64}$. For the denominator, $\Pr(X \geq 1) = 1 - \Pr(X=0) = \frac{37}{64}$. Therefore, the given conditional probability is

$$\Pr(X \geq 2 \mid X \geq 1) = \frac{\Pr(X \geq 2 \cap X \geq 1)}{\Pr(X \geq 1)} = \frac{\frac{10}{64}}{\frac{37}{64}} = \frac{10}{37}$$

(b)

Let C_i denote an event that the child i has blue eyes, where C_1 denotes the event of the youngest one, then

$$\Pr(C_1) = \Pr(C_2) = \Pr(C_3) = \frac{1}{4}$$

Now, knowing that the youngest has blue eyes, at least one child of the rest two should have blue eyes, and that probability equals to $\Pr(C_2 \cup C_3)$. More precisely, the conditional distribution is

$$\Pr(C_2 \cup C_3 \mid C_1) = \frac{\Pr(C_1 \cap (C_2 \cup C_3))}{\Pr(C_1)} = \frac{\Pr((C_1 \cap C_2) \cup (C_1 \cap C_3))}{\Pr(C_1)} \tag{17}$$

The numerator equals to $\Pr(C_1 \cap C_2) + \Pr(C_1 \cap C_3) - \Pr(C_1 \cap C_2 \cap C_3)$. Since independence is assumed among the children, it can be factorized as

$$\Pr(C_1)\left[\Pr(C_2) + \Pr(C_3) - \Pr(C_2 \cap C_3)\right] \tag{18}$$

where $\Pr(C_1)$ is canceled out. Therefore, the conditional proabability is computed as

$$\frac{1}{4} + \frac{1}{4} - \frac{1}{16} = \frac{7}{16}$$

Exercise 1.19

Let $\Pr(M)$ be the probability of using `Macintosh`, $\Pr(W)$ be that of using `Windows`, and $\Pr(L)$ be that of using `Linux`. Here, $\Pr(M) = 0.3$, $\Pr(W) = 0.5$, and $\Pr(L) = 0.2$. Then let V denote an event that a user gets viruses. Now, each conditional probability is:

$$\Pr(V \mid M) = 0.65,\ \Pr(V \mid W) = 0.82,\ \text{and}\ \Pr(V \mid L) = 0.5$$

Here we want to obtain the conditional probability $\Pr(W \mid V)$. It can be obtained by applying the Bayes theorem as below:

$$\Pr(W \mid V) = \frac{\Pr(W \cap V)}{\Pr(V)} = \frac{\Pr(W \cap V)}{\Pr(W \cap V) + \Pr(M \cap V) + \Pr(L \cap V)} \tag{19}$$

This is expanded as

$$\frac{\Pr(W)\Pr(V \mid W)}{\Pr(W)\Pr(V \mid W) + \Pr(M)\Pr(V \mid M) + \Pr(L)\Pr(V \mid L)} \tag{20}$$

Therefore, the actual probability is

$$\frac{0.5 \times 0.82}{0.5 \times 0.82 + 0.3 \times 0.65 + 0.2 \times 0.5} \approx 0.5816 \tag{21}$$

We can see that with the information about getting viruses, the probability of a user to use a `Windows` has gone up from 0.5 to around 0.58.