Homework 1: Interpretable Machine Learning Solution

Nir Goren 313452781 Nir Endy 205686397

July 2, 2024

1 Problem 1

1.1 Part 1

$$\begin{aligned}
\hat{f}_{1,PDP}(x_1) &= \int_{-\infty}^{\infty} \hat{f}(x_1, x_2) dP(x_2) \\
&= \int_{-\infty}^{\infty} (x_1^2 + ax_1x_2 + cx_2^2) \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x_2^2}{2}} dx_2 \\
&= x_1^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_2^2}{2}} dx_2 + ax_1 \int_{-\infty}^{\infty} x_2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x_2^2}{2}} dx_2 + c \int_{-\infty}^{\infty} x_2^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x_2^2}{2}} dx_2 \\
&= x_1^2 \cdot 1 + ax_1 \cdot \mathbb{E}[x_2] + c \cdot Var[x_2] \\
&= x_1^2 + ax_1 \cdot 0 + c \cdot 1 \\
&= x_1^2 + c
\end{aligned}$$

$$\begin{aligned}
\hat{f}_{2,PDP}(x_2) &= \int_{-\infty}^{\infty} \hat{f}(x_1, x_2) dP(x_1) \\
&= \int_{-\infty}^{\infty} (x_1^2 + ax_1x_2 + cx_2^2) \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} dx_1 \\
&= \int_{-\infty}^{\infty} x_1^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} dx_1 + ax_2 \int_{-\infty}^{\infty} x_1 \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} dx_1 + cx_2^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} dx_1 \\
&= Var[x_1] + ax_2 \cdot \mathbb{E}[x_1] + cx_2^2 \cdot 1 \\
&= 1 + ax_2 \cdot 0 + cx_2^2 \\
&= 1 + cx_2^2
\end{aligned}$$

1.2 Part 2

$$\begin{aligned}
\hat{f}_{1,CE}(x_1) &= \int_{-\infty}^{\infty} \hat{f}(x_1, x_2) dP(x_2 | X_1 = x_1) \\
&= \int_{-\infty}^{\infty} (x_1^2 + ax_1x_2 + cx_2^2) \cdot \frac{1}{\sqrt{2\pi(1-r^2)}} e^{-\frac{(x_2 - rx_1)^2}{2(1-r^2)}} dx_2 \\
&= x_1^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-r^2)}} e^{-\frac{(x_2 - rx_1)^2}{2(1-r^2)}} dx_2 + ax_1 \int_{-\infty}^{\infty} x_2 \frac{1}{\sqrt{2\pi(1-r^2)}} e^{-\frac{(x_2 - rx_1)^2}{2(1-r^2)}} dx_2 \\
&+ c \int_{-\infty}^{\infty} x_2^2 \frac{1}{\sqrt{2\pi(1-r^2)}} e^{-\frac{(x_2 - rx_1)^2}{2(1-r^2)}} dx_2 \\
&= x_1^2 \cdot 1 + ax_1 \mathbb{E}[x_2 | X_1 = x_1] + c \cdot (Var[x_2 | X_1 = x_1] - (rx_1)^2) \\
&= x_1^2 + ax_1 \cdot rx_1 + c \cdot (1 - r^2 + r^2 x_1^2)
\end{aligned}$$

$$\begin{aligned}
\hat{f}_{2,CE}(x_2) &= \int_{-\infty}^{\infty} \hat{f}(x_1, x_2) dP(x_1 | X_2 = x_2) \\
&= \int_{-\infty}^{\infty} (x_1^2 + ax_1x_2 + cx_2^2) \cdot \frac{1}{\sqrt{2\pi(1-r^2)}} e^{-\frac{(x_1 - rx_2)^2}{2(1-r^2)}} dx_1 \\
&= \int_{-\infty}^{\infty} x_1^2 \frac{1}{\sqrt{2\pi(1-r^2)}} e^{-\frac{(x_1 - rx_2)^2}{2(1-r^2)}} dx_1 + ax_2 \int_{-\infty}^{\infty} x_1 \frac{1}{\sqrt{2\pi(1-r^2)}} e^{-\frac{(x_1 - rx_2)^2}{2(1-r^2)}} dx_1 \\
&+ cx_2^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-r^2)}} e^{-\frac{(x_1 - rx_2)^2}{2(1-r^2)}} dx_1 \\
&= Var[x_1 | X_2 = x_2] + r^2 x_2^2 + ax_2 \cdot \mathbb{E}[x_1 | X_2 = x_2] + cx_2^2 \\
&= 1 - r^2 + r^2 x_2^2 + ax_2 \cdot rx_2 + cx_2^2
\end{aligned}$$

1.3 Part 3

$$\begin{aligned}
\hat{f}_{1,ALE}(x_1) &= \int_{z_1=x_1^0}^{x_1} \int_{-\infty}^{\infty} \frac{\partial \hat{f}(z_1, x_2)}{\partial x_1} dP(x_2|X_1 = z_1) \\
&= \int_{z_1=x_1^0}^{x_1} \int_{-\infty}^{\infty} (2z_1 + ax_2) \cdot \frac{1}{\sqrt{2\pi(1-r^2)}} e^{-\frac{(x_2 - rz_1)^2}{2(1-r^2)}} dx_2 \\
&= \int_{z_1=x_1^0}^{x_1} (2z_1 + a \cdot \mathbb{E}[x_2|X_1 = z_1]) \\
&= \int_{z_1=x_1^0}^{x_1} (2z_1 + a \cdot rz_1) \\
&= (2 + ar) \int_{z_1=x_1^0}^{x_1} z_1 dz_1 \\
&= (2 + ar) \cdot (\frac{x_1^2}{2} - \frac{x_1^{0^2}}{2})
\end{aligned}$$

$$\begin{aligned}
\hat{f}_{2,ALE}(x_2) &= \int_{z_2=x_2^0}^{x_2} \int_{-\infty}^{\infty} \frac{\partial \hat{f}(x_1, z_2)}{\partial x_2} dP(x_1|X_2 = z_2) \\
&= \int_{z_2=x_2^0}^{x_2} \int_{-\infty}^{\infty} (ax_1 + 2cz_2) \cdot \frac{1}{\sqrt{2\pi(1-r^2)}} e^{-\frac{(x_1 - rz_2)^2}{2(1-r^2)}} dx_1 \\
&= \int_{z_2=x_2^0}^{x_2} (a\mathbb{E}[x_1|X_2 = z_2] + 2cz_2) \\
&= \int_{z_2=x_2^0}^{x_2} (arz_2 + 2cz_2) \\
&= (ar + 2c) \int_{z_2=x_2^0}^{x_2} z_2 dz_2 \\
&= (ar + 2c) \cdot (\frac{x_2^2}{2} - \frac{x_2^{0^2}}{2})
\end{aligned}$$

1.4 Part 4

For $a = 0$, $c = 0$:

$$\hat{f} = x_1^2$$
$$\hat{f}_{2,PDP}(x_2) = 1$$
$$\hat{f}_{2,CE}(x_2) = 1 - r^2 + r^2 x_2^2$$
$$\hat{f}_{2,ALE}(x_2) = 0$$

We see that $\hat{f}_{2,PDP}(x_2)$ is constant, as changing x_2 does not affect the value of $\hat{f}$. $\hat{f}_{2,CE}(x_2)$ is quadratic in x_2 due to the correlation between x_1 and x_2. $\hat{f}_{2,ALE}(x_2)$ is constant, as the partial derivative of $\hat{f}$ with respect to x_2 is 0.

1.5 Part 5

$$\hat{f}_{1,PDP}(x_1) = x_1^2 + c$$
$$\hat{f}_{1,ALE} x_1 = (2 + ar) \cdot (\frac{x_1^2}{2} - \frac{{x_1^0}^2}{2})$$

Both $\hat{f}_{1,PDP}(x_1)$ and $\hat{f}_{1,ALE} x_1$ are quadratic in x_1. $\hat{f}_{1,ALE} x_1$ depends on r and a while $\hat{f}_{1,PDP}(x_1)$ does not. This is due the expected value of x_2 being 0 in the PDP case (indepenent of x_1), thus only the fist term is affected, while in the ALE case the conditional distribution of x_2 given x_1 is taken into account.