Exercise of Supervised Learning: Feature Selection

Yawei Li

yawei.li@stat.uni-muenchen.de

January 19, 2024

Exercise 1: Filter Problems

Let $f(x_1, x_2|\boldsymbol{\mu})$ be the density function of the bivariate Normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma} = \boldsymbol{I}_2$. You are given the following data generation process (DGP):

- the target $Y \sim \mathrm{Bernoulli}(0.5)$,
- the conditional density $p(x_1, x_2|Y = 1) = 0.5(f(x_1, x_2|(1, -1)^T) + f(x_1, x_2|(-1, 1)^T))$,
- the conditional density $p(x_1, x_2|Y = 0) = 0.5(f(x_1, x_2|(1, 1)^T) + f(x_1, x_2|(-1, -1)^T))$.

(Write the formulas on white board)

(a) Sketch the DGP.

Solution: Show the standard solution.

Exercise 1 (b)

(b) Compute $\mathbb{P}\left(Y = 1 | x_1 = \tilde{x}_1\right)$ and $\mathbb{P}\left(Y = 1 | x_2 = \tilde{x}_2\right)$.

Hint: x_1, x_2 are generated based on Y.

$$\begin{aligned}
\mathbb{P}\left(Y = 1 | x_1 = \tilde{x}_1\right) &= \frac{p(x_1 = \tilde{x}_1 | Y = 1)\mathbb{P}\left(Y = 1\right)}{p(x_1 = \tilde{x}_1 | Y = 1)\mathbb{P}\left(Y = 1\right) + p(x_1 = \tilde{x}_1 | Y = 0)\mathbb{P}\left(Y = 0\right)} \\
&= \frac{p(x_1 = \tilde{x}_1 | Y = 1)}{p(x_1 = \tilde{x}_1 | Y = 1) + p(x_1 = \tilde{x}_1 | Y = 0)} \qquad \left(\mathbb{P}\left(Y = 1\right) = \mathbb{P}\left(Y = 0\right)\right)
\end{aligned}$$

Question: How to get $p(x_1 = \tilde{x}_1 | Y = 1)$ and other terms?

Exercise 1 (b): Continued

- Marginal over $x_2 \rightsquigarrow p(x_1 = \tilde{x}_1 | Y = 1) = \int p(x_1 = \tilde{x}_1, x_2 = z | Y = 1) \mathrm{d}z$
- Hard to **directly** marginalize because $(x_1, x_2) | Y$ is a mixture of Gaussian components: $p(x_1, x_2 | Y = 1) = \mathbf{0.5}(f(x_1, x_2 | (1, -1)^T) + f(x_1, x_2 | (-1, 1)^T))$
- But it is easy to compute the marginal distribution for a **single Gaussian**.
- Can we first marginalize individual Gaussian components and then mix up them? Yes.
- Marginalize $x_1, x_2 \sim \mathcal{N}((1, -1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(1, 1)$.
- Marginalize $x_1, x_2 \sim \mathcal{N}((-1, 1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(-1, 1)$.
- Let $g_\mu : \mathbb{R} \to [0, 1]$ be the prob. density function of $\mathcal{N}(\mu, 1)$.
- $p(x_1 = \tilde{x}_1 | Y = 1) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$ (Don't forget the weights of each Gaussian component).
- Similarily: $p(x_1 = \tilde{x}_1 | Y = 0) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$.
- So $\mathbb{P}(Y = 1 | x_1 = \tilde{x}_1) = \frac{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))}{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1)) + \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))} = 0.5$ (Same for $\mathbb{P}(Y = 0 | x_1 = \tilde{x}_1)$).

Exercise 1 (b): Continued

- Marginal over $x_2 \rightsquigarrow p(x_1 = \tilde{x}_1 | Y = 1) = \int p(x_1 = \tilde{x}_1, x_2 = z | Y = 1)\mathrm{d}z$
- Hard to **directly** marginalize because $(x_1, x_2)|Y$ is a mixture of Gaussian components: $p(x_1, x_2 | Y = 1) = \mathbf{0.5}(f(x_1, x_2 | (1, -1)^T) + f(x_1, x_2 | (-1, 1)^T))$
- But it is easy to compute the marginal distribution for a **single Gaussian**.
- Can we first marginalize individual Gaussian components and then mix up them? Yes.
- Marginalize $x_1, x_2 \sim \mathcal{N}((1, -1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(1, 1)$.
- Marginalize $x_1, x_2 \sim \mathcal{N}((-1, 1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(-1, 1)$.
- Let $g_\mu : \mathbb{R} \to [0, 1]$ be the prob. density function of $\mathcal{N}(\mu, 1)$.
- $p(x_1 = \tilde{x}_1 | Y = 1) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$ (Don't forget the weights of each Gaussian component).
- Similarily: $p(x_1 = \tilde{x}_1 | Y = 0) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$.
- So $\mathbb{P}(Y = 1 | x_1 = \tilde{x}_1) = \frac{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))}{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1)) + \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))} = 0.5$ (Same for $\mathbb{P}(Y = 0 | x_1 = \tilde{x}_1)$).

Exercise 1 (b): Continued

- Marginal over $x_2 \rightsquigarrow p(x_1 = \tilde{x}_1 | Y = 1) = \int p(x_1 = \tilde{x}_1, x_2 = z | Y = 1)\mathrm{d}z$
- Hard to **directly** marginalize because $(x_1, x_2)|Y$ is a mixture of Gaussian components: $p(x_1, x_2 | Y = 1) = \mathbf{0.5}(f(x_1, x_2 | (1, -1)^T) + f(x_1, x_2 | (-1, 1)^T))$
- But it is easy to compute the marginal distribution for a **single Gaussian.**
- Can we first marginalize individual Gaussian components and then mix up them? Yes.
- Marginalize $x_1, x_2 \sim \mathcal{N}((1, -1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(1, 1)$.
- Marginalize $x_1, x_2 \sim \mathcal{N}((-1, 1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(-1, 1)$.
- Let $g_\mu : \mathbb{R} \to [0, 1]$ be the prob. density function of $\mathcal{N}(\mu, 1)$.
- $p(x_1 = \tilde{x}_1 | Y = 1) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$ (Don't forget the weights of each Gaussian component).
- Similarily: $p(x_1 = \tilde{x}_1 | Y = 0) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$.
- So $\mathbb{P}(Y = 1 | x_1 = \tilde{x}_1) = \frac{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))}{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1)) + \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))} = 0.5$ (Same for $\mathbb{P}(Y = 0 | x_1 = \tilde{x}_1)$).

Exercise 1 (b): Continued

- Marginal over $x_2 \rightsquigarrow p(x_1 = \tilde{x}_1 | Y = 1) = \int p(x_1 = \tilde{x}_1, x_2 = z | Y = 1) \mathrm{d}z$
- Hard to **directly** marginalize because $(x_1, x_2)|Y$ is a mixture of Gaussian components: $p(x_1, x_2 | Y = 1) = \mathbf{0.5}(f(x_1, x_2 | (1, -1)^T) + f(x_1, x_2 | (-1, 1)^T))$
- But it is easy to compute the marginal distribution for a **single Gaussian.**
- Can we first marginalize individual Gaussian components and then mix up them? Yes.
- Marginalize $x_1, x_2 \sim \mathcal{N}((1, -1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(1, 1)$.
- Marginalize $x_1, x_2 \sim \mathcal{N}((-1, 1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(-1, 1)$.
- Let $g_\mu : \mathbb{R} \to [0, 1]$ be the prob. density function of $\mathcal{N}(\mu, 1)$.
- $p(x_1 = \tilde{x}_1 | Y = 1) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$ (Don't forget the weights of each Gaussian component).
- Similarily: $p(x_1 = \tilde{x}_1 | Y = 0) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$.
- So $\mathbb{P}(Y = 1 | x_1 = \tilde{x}_1) = \frac{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))}{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1)) + \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))} = 0.5$ (Same for $\mathbb{P}(Y = 0 | x_1 = \tilde{x}_1)$).

Exercise 1 (b): Continued

- Marginal over $x_2 \rightsquigarrow p(x_1 = \tilde{x}_1 | Y = 1) = \int p(x_1 = \tilde{x}_1, x_2 = z | Y = 1)\mathrm{d}z$
- Hard to **directly** marginalize because $(x_1, x_2)|Y$ is a mixture of Gaussian components: $p(x_1, x_2 | Y = 1) = \mathbf{0.5}(f(x_1, x_2|(1, -1)^T) + f(x_1, x_2|(-1, 1)^T))$
- But it is easy to compute the marginal distribution for a **single Gaussian**.
- Can we first marginalize individual Gaussian components and then mix up them? Yes.
- Marginalize $x_1, x_2 \sim \mathcal{N}((1, -1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(1, 1)$.
- Marginalize $x_1, x_2 \sim \mathcal{N}((-1, 1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(-1, 1)$.
- Let $g_\mu : \mathbb{R} \to [0, 1]$ be the prob. density function of $\mathcal{N}(\mu, 1)$.
- $p(x_1 = \tilde{x}_1 | Y = 1) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$ (Don't forget the weights of each Gaussian component).
- Similarily: $p(x_1 = \tilde{x}_1 | Y = 0) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$.
- So $\mathbb{P}(Y = 1 | x_1 = \tilde{x}_1) = \frac{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))}{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1)) + \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))} = 0.5$ (Same for $\mathbb{P}(Y = 0 | x_1 = \tilde{x}_1)$).

Exercise 1 (b): Continued

- Marginal over $x_2 \rightsquigarrow p(x_1 = \tilde{x}_1 | Y = 1) = \int p(x_1 = \tilde{x}_1, x_2 = z | Y = 1)\mathrm{d}z$
- Hard to **directly** marginalize because $(x_1, x_2)|Y$ is a mixture of Gaussian components: $p(x_1, x_2 | Y = 1) = \mathbf{0.5}(f(x_1, x_2 | (1, -1)^T) + f(x_1, x_2 | (-1, 1)^T))$
- But it is easy to compute the marginal distribution for a **single Gaussian**.
- Can we first marginalize individual Gaussian components and then mix up them? Yes.
- Marginalize $x_1, x_2 \sim \mathcal{N}((1, -1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(1, 1)$.
- Marginalize $x_1, x_2 \sim \mathcal{N}((-1, 1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(-1, 1)$.
- Let $g_\mu : \mathbb{R} \to [0, 1]$ be the prob. density function of $\mathcal{N}(\mu, 1)$.
- $p(x_1 = \tilde{x}_1 | Y = 1) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$ (Don't forget the weights of each Gaussian component).
- Similarily: $p(x_1 = \tilde{x}_1 | Y = 0) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$.
- So $\mathbb{P}(Y = 1 | x_1 = \tilde{x}_1) = \frac{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))}{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1)) + \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))} = 0.5$ (Same for $\mathbb{P}(Y = 0 | x_1 = \tilde{x}_1)$).

Exercise 1 (b): Continued

- Marginal over $x_2 \rightsquigarrow p(x_1 = \tilde{x}_1 | Y = 1) = \int p(x_1 = \tilde{x}_1, x_2 = z | Y = 1)\mathrm{d}z$
- Hard to **directly** marginalize because $(x_1, x_2)|Y$ is a mixture of Gaussian components: $p(x_1, x_2 | Y = 1) = \mathbf{0.5}(f(x_1, x_2 | (1, -1)^T) + f(x_1, x_2 | (-1, 1)^T))$
- But it is easy to compute the marginal distribution for a **single Gaussian.**
- Can we first marginalize individual Gaussian components and then mix up them? Yes.
- Marginalize $x_1, x_2 \sim \mathcal{N}((1, -1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(1, 1)$.
- Marginalize $x_1, x_2 \sim \mathcal{N}((-1, 1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(-1, 1)$.
- Let $g_\mu : \mathbb{R} \to [0, 1]$ be the prob. density function of $\mathcal{N}(\mu, 1)$.
- $p(x_1 = \tilde{x}_1 | Y = 1) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$ (Don't forget the weights of each Gaussian component).
- Similarily: $p(x_1 = \tilde{x}_1 | Y = 0) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$.
- So $\mathbb{P}(Y = 1 | x_1 = \tilde{x}_1) = \frac{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))}{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1)) + \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))} = 0.5$ (Same for $\mathbb{P}(Y = 0 | x_1 = \tilde{x}_1)$).

Exercise 1 (b): Continued

- Marginal over $x_2 \rightsquigarrow p(x_1 = \tilde{x}_1 | Y = 1) = \int p(x_1 = \tilde{x}_1, x_2 = z | Y = 1) \mathrm{d}z$
- Hard to **directly** marginalize because $(x_1, x_2) | Y$ is a mixture of Gaussian components: $p(x_1, x_2 | Y = 1) = \mathbf{0.5}(f(x_1, x_2 | (1, -1)^T) + f(x_1, x_2 | (-1, 1)^T))$
- But it is easy to compute the marginal distribution for a **single Gaussian**.
- Can we first marginalize individual Gaussian components and then mix up them? Yes.
- Marginalize $x_1, x_2 \sim \mathcal{N}((1, -1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(1, 1)$.
- Marginalize $x_1, x_2 \sim \mathcal{N}((-1, 1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(-1, 1)$.
- Let $g_\mu : \mathbb{R} \to [0, 1]$ be the prob. density function of $\mathcal{N}(\mu, 1)$.
- $p(x_1 = \tilde{x}_1 | Y = 1) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$ (Don't forget the weights of each Gaussian component).
- Similarily: $p(x_1 = \tilde{x}_1 | Y = 0) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$.
- So $\mathbb{P}(Y = 1 | x_1 = \tilde{x}_1) = \frac{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))}{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1)) + \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))} = 0.5$ (Same for $\mathbb{P}(Y = 0 | x_1 = \tilde{x}_1)$).

Exercise 1 (b): Continued

- Marginal over $x_2 \rightsquigarrow p(x_1 = \tilde{x}_1 | Y = 1) = \int p(x_1 = \tilde{x}_1, x_2 = z | Y = 1)\mathrm{d}z$
- Hard to **directly** marginalize because $(x_1, x_2)|Y$ is a mixture of Gaussian components: $p(x_1, x_2 | Y = 1) = \mathbf{0.5}(f(x_1, x_2|(1, -1)^T) + f(x_1, x_2|(-1, 1)^T))$
- But it is easy to compute the marginal distribution for a **single Gaussian**.
- Can we first marginalize individual Gaussian components and then mix up them? Yes.
- Marginalize $x_1, x_2 \sim \mathcal{N}((1, -1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(1, 1)$.
- Marginalize $x_1, x_2 \sim \mathcal{N}((-1, 1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(-1, 1)$.
- Let $g_\mu : \mathbb{R} \to [0, 1]$ be the prob. density function of $\mathcal{N}(\mu, 1)$.
- $p(x_1 = \tilde{x}_1 | Y = 1) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$ (Don't forget the weights of each Gaussian component).
- Similarily: $p(x_1 = \tilde{x}_1 | Y = 0) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$.
- So $\mathbb{P}(Y = 1 | x_1 = \tilde{x}_1) = \frac{\mathbf{0.5}(g_1(\tilde{x}_1)+g_{-1}(\tilde{x}_1))}{\mathbf{0.5}(g_1(\tilde{x}_1)+g_{-1}(\tilde{x}_1))+\mathbf{0.5}(g_1(\tilde{x}_1)+g_{-1}(\tilde{x}_1))} = 0.5$ (Same for $\mathbb{P}(Y = 0 | x_1 = \tilde{x}_1)$).

Exercise 1 (b): Continued

- Marginal over $x_2 \rightsquigarrow p(x_1 = \tilde{x}_1 | Y = 1) = \int p(x_1 = \tilde{x}_1, x_2 = z | Y = 1)\mathrm{d}z$
- Hard to **directly** marginalize because $(x_1, x_2)|Y$ is a mixture of Gaussian components: $p(x_1, x_2 | Y = 1) = \mathbf{0.5}(f(x_1, x_2 | (1, -1)^T) + f(x_1, x_2 | (-1, 1)^T))$
- But it is easy to compute the marginal distribution for a **single Gaussian**.
- Can we first marginalize individual Gaussian components and then mix up them? Yes.
- Marginalize $x_1, x_2 \sim \mathcal{N}((1, -1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(1, 1)$.
- Marginalize $x_1, x_2 \sim \mathcal{N}((-1, 1)^T, I_2)$ over $x_2 \rightsquigarrow x_1 \sim \mathcal{N}(-1, 1)$.
- Let $g_\mu : \mathbb{R} \to [0, 1]$ be the prob. density function of $\mathcal{N}(\mu, 1)$.
- $p(x_1 = \tilde{x}_1 | Y = 1) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$ (Don't forget the weights of each Gaussian component).
- Similarily: $p(x_1 = \tilde{x}_1 | Y = 0) = \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))$.
- So $\mathbb{P}(Y = 1 | x_1 = \tilde{x}_1) = \frac{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))}{\mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1)) + \mathbf{0.5}(g_1(\tilde{x}_1) + g_{-1}(\tilde{x}_1))} = 0.5$ (Same for $\mathbb{P}(Y = 0 | x_1 = \tilde{x}_1)$).

Exercise 1 (c)

(c) Compute $\mathbb{P}\left(Y = 1 | x_1 = 1, x_2 = 1\right)$.

$$\begin{aligned}
\mathbb{P}\left(Y = 1 | x_1 = 1, x_2 = 1\right) &= \frac{p((1,1)|Y=1)\mathbb{P}\left(Y=1\right)}{p((1,1)|Y=1)\mathbb{P}\left(Y=1\right) + p((1,1)|Y=0)\mathbb{P}\left(Y=0\right)} \\
&= \frac{1}{1 + \frac{p((1,1)|Y=1)}{p((1,1)|Y=1)}} \\
&= \frac{1}{1 + \frac{\exp(0)+\exp(-0.5(-2,-2)^T(-2,-2))}{2\exp(-0.5(0,-2)^T(0,-2))}} \qquad \text{(Use the given density functions)} \\
&\approx 0.21
\end{aligned}$$

Exercise 1 (d)

(d) Explain what happens if we apply mutual information as filter in this scenario.

- From (b): $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 1 | x_1 = \tilde{x}_1) = \mathbb{P}(Y = 1 | x_2 = \tilde{x}_2) = 0.5$.
- So x_1 is independent from Y, and the same hold for x_2.
- Mutual information between x_i and Y will be 0 for $i = 1, 2$.
- Any feature will be more preferred over them.
- But Y is clearly jointly dependent on x_1 and x_2, as shown in (c), $\mathbb{P}(Y = 1 | x_1 = 1, x_2 = 1) \neq \mathbb{P}(Y = 1)$.

Exercise 1 (d)

(d) Explain what happens if we apply mutual information as filter in this scenario.

- From (b): $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 1 | x_1 = \tilde{x}_1) = \mathbb{P}(Y = 1 | x_2 = \tilde{x}_2) = 0.5$.
- So x_1 is independent from Y, and the same hold for x_2.
- Mutual information between x_i and Y will be 0 for $i = 1, 2$.
- Any feature will be more preferred over them.
- But Y is clearly jointly dependent on x_1 and x_2, as shown in (c), $\mathbb{P}(Y = 1 | x_1 = 1, x_2 = 1) \neq \mathbb{P}(Y = 1)$.

Exercise 1 (d)

(d) Explain what happens if we apply mutual information as filter in this scenario.

- From (b): $\mathbb{P}\left(Y = 1\right) = \mathbb{P}\left(Y = 1 | x_1 = \tilde{x}_1\right) = \mathbb{P}\left(Y = 1 | x_2 = \tilde{x}_2\right) = 0.5$.
- So x_1 is independent from Y, and the same hold for x_2.
- Mutual information between x_i and Y will be 0 for $i = 1, 2$.
- Any feature will be more preferred over them.
- But Y is clearly jointly dependent on x_1 and x_2, as shown in (c), $\mathbb{P}\left(Y = 1 | x_1 = 1, x_2 = 1\right) \neq \mathbb{P}\left(Y = 1\right)$.

Exercise 1 (d)

(d) Explain what happens if we apply mutual information as filter in this scenario.

- From (b): $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 1 | x_1 = \tilde{x}_1) = \mathbb{P}(Y = 1 | x_2 = \tilde{x}_2) = 0.5$.
- So x_1 is independent from Y, and the same hold for x_2.
- Mutual information between x_i and Y will be 0 for $i = 1, 2$.
- Any feature will be more preferred over them.
- But Y is clearly jointly dependent on x_1 and x_2, as shown in (c), $\mathbb{P}(Y = 1 | x_1 = 1, x_2 = 1) \neq \mathbb{P}(Y = 1)$.

Exercise 1 (d)

(d) Explain what happens if we apply mutual information as filter in this scenario.

- From (b): $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 1 | x_1 = \tilde{x}_1) = \mathbb{P}(Y = 1 | x_2 = \tilde{x}_2) = 0.5$.
- So x_1 is independent from Y, and the same hold for x_2.
- Mutual information between x_i and Y will be 0 for $i = 1, 2$.
- Any feature will be more preferred over them.
- But Y is clearly jointly dependent on x_1 and x_2, as shown in (c), $\mathbb{P}(Y = 1 | x_1 = 1, x_2 = 1) \neq \mathbb{P}(Y = 1)$.

Exercise 1 (d)

(d) Explain what happens if we apply mutual information as filter in this scenario.

- From (b): $\mathbb{P}\left(Y = 1\right) = \mathbb{P}\left(Y = 1 | x_1 = \tilde{x}_1\right) = \mathbb{P}\left(Y = 1 | x_2 = \tilde{x}_2\right) = 0.5$.
- So x_1 is independent from Y, and the same hold for x_2.
- Mutual information between x_i and Y will be 0 for $i = 1, 2$.
- Any feature will be more preferred over them.
- But Y is clearly jointly dependent on x_1 and x_2, as shown in (c), $\mathbb{P}\left(Y = 1 | x_1 = 1, x_2 = 1\right) \neq \mathbb{P}\left(Y = 1\right)$.

Exercise 2: Filter simulation study

Show the standard solution.

Exercise 3: Wrappers

You are given the following features and their respective BICs. BIC_i with $i \in \{\{A\}, \{B\}, \{C\}, \{D\}, \{A, B\}, \{A, C\}\{A, D\}, \{B, C\}, \{B, D\}, \{C, D\}, \{A, B, C\}, \{A, B, D\}, \{B, C, D\}, \{A, B, C, D\}\}$.

Features	BIC_i
$\{A\}$	0.9
$\{B\}$	0.8
$\{C\}$	1.0
$\{D\}$	1.0
$\{A, B\}$	0.8
$\{A, C\}$	0.7
$\{A, D\}$	0.8

Features	BIC_i
$\{B, C\}$	0.7
$\{B, D\}$	0.6
$\{C, D\}$	0.9
$\{A, B, C\}$	0.6
$\{A, B, D\}$	0.8
$\{B, C, D\}$	0.5
$\{A, B, C, D\}$	0.6

(a) Do forward search and note down each iteration.

(b) Do backward search and note down each iteration.

Exercise 3 (a)

Features	BIC_i
$\{A\}$	0.9
$\{B\}$	0.8
$\{C\}$	1.0
$\{D\}$	1.0
$\{A, B\}$	0.8
$\{A, C\}$	0.7
$\{A, D\}$	0.8

Features	BIC_i
$\{B, C\}$	0.7
$\{B, D\}$	0.6
$\{C, D\}$	0.9
$\{A, B, C\}$	0.6
$\{A, B, D\}$	0.8
$\{B, C, D\}$	0.5
$\{A, B, C, D\}$	0.6

(a) Do forward search and note down each iteration.

1. $\{B\}$ since $\mathrm{BIC}_{\{B\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{A\}, \{C\}, \{D\}\}$.
2. $\{B, D\}$ since $\mathrm{BIC}_{\{B,D\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{A, B\}, \{B, C\}\}$
3. $\{B, C, D\}$ since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{A,B,D\}}$.
4. $\{B, C, D\}$ and terminate since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{A,C,B,D\}}$.

Exercise 3 (a)

Features	BIC_i
$\{A\}$	0.9
$\{B\}$	0.8
$\{C\}$	1.0
$\{D\}$	1.0
$\{A, B\}$	0.8
$\{A, C\}$	0.7
$\{A, D\}$	0.8

Features	BIC_i
$\{B, C\}$	0.7
$\{B, D\}$	0.6
$\{C, D\}$	0.9
$\{A, B, C\}$	0.6
$\{A, B, D\}$	0.8
$\{B, C, D\}$	0.5
$\{A, B, C, D\}$	0.6

(a) Do forward search and note down each iteration.

1. $\{B\}$ since $\mathrm{BIC}_{\{B\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{A\}, \{C\}, \{D\}\}$.
2. $\{B, D\}$ since $\mathrm{BIC}_{\{B,D\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{A, B\}, \{B, C\}\}$
3. $\{B, C, D\}$ since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{A,B,D\}}$.
4. $\{B, C, D\}$ and terminate since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{A,C,B,D\}}$.

Exercise 3 (a)

Features	BIC_i
$\{A\}$	0.9
$\{B\}$	0.8
$\{C\}$	1.0
$\{D\}$	1.0
$\{A, B\}$	0.8
$\{A, C\}$	0.7
$\{A, D\}$	0.8

Features	BIC_i
$\{B, C\}$	0.7
$\{B, D\}$	0.6
$\{C, D\}$	0.9
$\{A, B, C\}$	0.6
$\{A, B, D\}$	0.8
$\{B, C, D\}$	0.5
$\{A, B, C, D\}$	0.6

(a) Do forward search and note down each iteration.

1. $\{B\}$ since $\mathrm{BIC}_{\{B\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{A\}, \{C\}, \{D\}\}$.
2. $\{B, D\}$ since $\mathrm{BIC}_{\{B,D\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{A, B\}, \{B, C\}\}$
3. $\{B, C, D\}$ since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{A,B,D\}}$.
4. $\{B, C, D\}$ and terminate since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{A,C,B,D\}}$.

Exercise 3 (a)

Features	BIC_i
$\{A\}$	0.9
$\{B\}$	0.8
$\{C\}$	1.0
$\{D\}$	1.0
$\{A, B\}$	0.8
$\{A, C\}$	0.7
$\{A, D\}$	0.8

Features	BIC_i
$\{B, C\}$	0.7
$\{B, D\}$	0.6
$\{C, D\}$	0.9
$\{A, B, C\}$	0.6
$\{A, B, D\}$	0.8
$\{B, C, D\}$	0.5
$\{A, B, C, D\}$	0.6

(a) Do forward search and note down each iteration.

1. $\{B\}$ since $\mathrm{BIC}_{\{B\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{A\}, \{C\}, \{D\}\}$.
2. $\{B, D\}$ since $\mathrm{BIC}_{\{B,D\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{A, B\}, \{B, C\}\}$
3. $\{B, C, D\}$ since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{A,B,D\}}$.
4. $\{B, C, D\}$ and terminate since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{A,C,B,D\}}$.

Exercise 3 (a)

Features	BIC_i
$\{A\}$	0.9
$\{B\}$	0.8
$\{C\}$	1.0
$\{D\}$	1.0
$\{A, B\}$	0.8
$\{A, C\}$	0.7
$\{A, D\}$	0.8

Features	BIC_i
$\{B, C\}$	0.7
$\{B, D\}$	0.6
$\{C, D\}$	0.9
$\{A, B, C\}$	0.6
$\{A, B, D\}$	0.8
$\{B, C, D\}$	0.5
$\{A, B, C, D\}$	0.6

(a) Do forward search and note down each iteration.

1. $\{B\}$ since $\mathrm{BIC}_{\{B\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{A\}, \{C\}, \{D\}\}$.
2. $\{B, D\}$ since $\mathrm{BIC}_{\{B,D\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{A, B\}, \{B, C\}\}$
3. $\{B, C, D\}$ since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{A,B,D\}}$.
4. $\{B, C, D\}$ and terminate since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{A,C,B,D\}}$.

Exercise 3 (b)

Features	BIC_i
$\{A\}$	0.9
$\{B\}$	0.8
$\{C\}$	1.0
$\{D\}$	1.0
$\{A, B\}$	0.8
$\{A, C\}$	0.7
$\{A, D\}$	0.8

Features	BIC_i
$\{B, C\}$	0.7
$\{B, D\}$	0.6
$\{C, D\}$	0.9
$\{A, B, C\}$	0.6
$\{A, B, D\}$	0.8
$\{B, C, D\}$	0.5
$\{A, B, C, D\}$	0.6

(b) Do backward search and note down each iteration.

1. Start with all features $\{A, B, C, D\}$.
2. $\{B, C, D\}$ and since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{A, B, C\}, \{A, C, D\}\}$.
3. $\{B, C, D\}$ and terminate since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{B, C\}, \{B, D\}, \{C, D\}\}$.

Exercise 3 (b)

Features	BIC_i
$\{A\}$	0.9
$\{B\}$	0.8
$\{C\}$	1.0
$\{D\}$	1.0
$\{A, B\}$	0.8
$\{A, C\}$	0.7
$\{A, D\}$	0.8

Features	BIC_i
$\{B, C\}$	0.7
$\{B, D\}$	0.6
$\{C, D\}$	0.9
$\{A, B, C\}$	0.6
$\{A, B, D\}$	0.8
$\{B, C, D\}$	0.5
$\{A, B, C, D\}$	0.6

(b) Do backward search and note down each iteration.

1. Start with all features $\{A, B, C, D\}$.
2. $\{B, C, D\}$ and since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{A, B, C\}, \{A, C, D\}\}$.
3. $\{B, C, D\}$ and terminate since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{B, C\}, \{B, D\}, \{C, D\}\}$.

Exercise 3 (b)

Features	BIC_i
$\{A\}$	0.9
$\{B\}$	0.8
$\{C\}$	1.0
$\{D\}$	1.0
$\{A, B\}$	0.8
$\{A, C\}$	0.7
$\{A, D\}$	0.8

Features	BIC_i
$\{B, C\}$	0.7
$\{B, D\}$	0.6
$\{C, D\}$	0.9
$\{A, B, C\}$	0.6
$\{A, B, D\}$	0.8
$\{B, C, D\}$	0.5
$\{A, B, C, D\}$	0.6

(b) Do backward search and note down each iteration.

1. Start with all features $\{A, B, C, D\}$.
2. $\{B, C, D\}$ and since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{A, B, C\}, \{A, C, D\}\}$.
3. $\{B, C, D\}$ and terminate since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{B, C\}, \{B, D\}, \{C, D\}\}$.

Exercise 3 (b)

Features	BIC_i
$\{A\}$	0.9
$\{B\}$	0.8
$\{C\}$	1.0
$\{D\}$	1.0
$\{A, B\}$	0.8
$\{A, C\}$	0.7
$\{A, D\}$	0.8

Features	BIC_i
$\{B, C\}$	0.7
$\{B, D\}$	0.6
$\{C, D\}$	0.9
$\{A, B, C\}$	0.6
$\{A, B, D\}$	0.8
$\{B, C, D\}$	0.5
$\{A, B, C, D\}$	0.6

(b) Do backward search and note down each iteration.

1. Start with all features $\{A, B, C, D\}$.
2. $\{B, C, D\}$ and since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{A, B, C\}, \{A, C, D\}\}$.
3. $\{B, C, D\}$ and terminate since $\mathrm{BIC}_{\{B,C,D\}} < \mathrm{BIC}_{\{X\}} \quad \forall X \in \{\{B, C\}, \{B, D\}, \{C, D\}\}$.