Solution 1: Equivalent Representation of Separation and Sufficiency

(a) Show the equivalence between

$$\hat{y} \perp\!\!\!\perp \mathbf{A} \mid y$$

and

$$\begin{aligned}
&\mathbb{P}(\hat{y} = 1 \mid y = -1, \mathbf{A} = \mathbf{a}) = \mathbb{P}(\hat{y} = 1 \mid y = -1, \mathbf{A} = \tilde{\mathbf{a}}) && \text{(equal false positive rates)} \\
&\mathbb{P}(\hat{y} = -1 \mid y = 1, \mathbf{A} = \mathbf{a}) = \mathbb{P}(\hat{y} = -1 \mid y = 1, \mathbf{A} = \tilde{\mathbf{a}}) && \text{(equal false negative rates)}
\end{aligned}$$

which holds for all possible realizations $\mathbf{a}, \tilde{\mathbf{a}}$ of $\mathbf{A}$.

Solution:

First, we show that $\hat{y} \perp\!\!\!\perp \mathbf{A} \mid y$ implies equal false positive rates and equal false negative rates. Let $c \in \{-1, 1\}$ and $\neg c := -1c$. For all possible realizations $\mathbf{a}, \tilde{\mathbf{a}}$ of $\mathbf{A}$ it holds that

$$\begin{aligned}
\frac{\mathbb{P}(\hat{y} = c \mid y = \neg c, \mathbf{A} = \mathbf{a})}{\mathbb{P}(\hat{y} = c \mid y = \neg c, \mathbf{A} = \tilde{\mathbf{a}})} &= \frac{\mathbb{P}(\hat{y} = c, y = \neg c, \mathbf{A} = \mathbf{a})\mathbb{P}(y = \neg c, \mathbf{A} = \tilde{\mathbf{a}})}{\mathbb{P}(\hat{y} = c, y = \neg c, \mathbf{A} = \tilde{\mathbf{a}})\mathbb{P}(y = \neg c, \mathbf{A} = \mathbf{a})} \\
&= \frac{\mathbb{P}(\hat{y} = c, y = \neg c, \mathbf{A} = \mathbf{a})\mathbb{P}(y = \neg c, \mathbf{A} = \tilde{\mathbf{a}})\mathbb{P}(y = \neg c)}{\mathbb{P}(\hat{y} = c, y = \neg c, \mathbf{A} = \tilde{\mathbf{a}})\mathbb{P}(y = \neg c, \mathbf{A} = \mathbf{a})\mathbb{P}(y = \neg c)} \\
&= \frac{\mathbb{P}(\hat{y} = c, \mathbf{A} = \mathbf{a} \mid y = \neg c)\mathbb{P}(y = \neg c, \mathbf{A} = \tilde{\mathbf{a}})}{\mathbb{P}(\hat{y} = c, \mathbf{A} = \tilde{\mathbf{a}} \mid y = \neg c)\mathbb{P}(y = \neg c, \mathbf{A} = \mathbf{a})} \\
&\overset{(*)}{=} \frac{\mathbb{P}(\hat{y} = c \mid y = \neg c)\mathbb{P}(\mathbf{A} = \mathbf{a} \mid y = \neg c)\mathbb{P}(y = \neg c, \mathbf{A} = \tilde{\mathbf{a}})}{\mathbb{P}(\hat{y} = c \mid y = \neg c)\mathbb{P}(\mathbf{A} = \tilde{\mathbf{a}} \mid y = \neg c)\mathbb{P}(y = \neg c, \mathbf{A} = \mathbf{a})} \\
&= \frac{\mathbb{P}(\mathbf{A} = \mathbf{a} \mid y = \neg c)\mathbb{P}(y = \neg c, \mathbf{A} = \tilde{\mathbf{a}})}{\mathbb{P}(\mathbf{A} = \tilde{\mathbf{a}} \mid y = \neg c)\mathbb{P}(y = \neg c, \mathbf{A} = \mathbf{a})} \\
&= 1,
\end{aligned}$$

where we used for $(*)$ the conditional independence. Thus, from the latter display we obtain the equality of the false positive and negative rates.

Next, we show that equal false positive rates and equal false negative rates imply the conditional independence. For this purpose, note that for any $c, d \in \{-1, 1\}$ and $\tilde{\mathbf{a}}$ it holds that

$$\begin{aligned}
\mathbb{P}(\hat{y} = c \mid y = d) &= \sum_{\mathbf{a}} \mathbb{P}(\hat{y} = c \mid y = d, \mathbf{A} = \mathbf{a})\mathbb{P}(\mathbf{A} = \mathbf{a} \mid y = d) && \text{(Law of total probability)} \\
&= \sum_{\mathbf{a}} \mathbb{P}(\hat{y} = c \mid y = d, \mathbf{A} = \tilde{\mathbf{a}})\mathbb{P}(\mathbf{A} = \mathbf{a} \mid y = d) && \text{(Equal rates)} \\
&= \mathbb{P}(\hat{y} = c \mid y = d, \mathbf{A} = \tilde{\mathbf{a}}) \underbrace{\sum_{\mathbf{a}} \mathbb{P}(\mathbf{A} = \mathbf{a} \mid y = d)}_{=1} \\
&= \mathbb{P}(\hat{y} = c \mid y = d, \mathbf{A} = \tilde{\mathbf{a}}).
\end{aligned}$$

With this, for any $c, d \in \{-1, 1\}$ and $\mathbf{a}$ we obtain

$$\begin{aligned}
\frac{\mathbb{P}(\hat{y} = c, \mathbf{A} = \mathbf{a} \mid y = d)}{\mathbb{P}(\hat{y} = c \mid y = d)\mathbb{P}(\mathbf{A} = \mathbf{a} \mid y = d)} &\overset{(**)}{=} \frac{\mathbb{P}(\hat{y} = c, \mathbf{A} = \mathbf{a} \mid y = d)}{\mathbb{P}(\hat{y} = c \mid y = d, \mathbf{A} = \mathbf{a})\mathbb{P}(\mathbf{A} = \mathbf{a} \mid y = d)} \\
&= \frac{\mathbb{P}(\hat{y} = c, \mathbf{A} = \mathbf{a}, y = d)\mathbb{P}(y = d, \mathbf{A} = \mathbf{a})\mathbb{P}(y = d)}{\mathbb{P}(y = d)\mathbb{P}(\hat{y} = c, y = d, \mathbf{A} = \mathbf{a})\mathbb{P}(\mathbf{A} = \mathbf{a}, y = d)} = 1,
\end{aligned}$$

where we used for $(**)$ the equality we derived before with $\mathbf{a}$ for $\tilde{\mathbf{a}}$.

(b) Show the equivalence between

$$y \perp\!\!\!\perp \mathbf{A} \mid \mathbf{S}$$

and

$$\mathbb{P}(y=1 \mid \mathbf{S}=s, \mathbf{A}=\mathbf{a}) = \mathbb{P}(y=1 \mid \mathbf{S}=s, \mathbf{A}=\tilde{\mathbf{a}})$$

holding for all possible realizations $\mathbf{a}, \tilde{\mathbf{a}}$ of $\mathbf{A}$ and all possible realizations s of $\mathbf{S}$.

Solution:

First, we show that $y \perp\!\!\!\perp \mathbf{A} \mid \mathbf{S}$ implies

$$\mathbb{P}(y=1 \mid \mathbf{S}=s, \mathbf{A}=\mathbf{a}) = \mathbb{P}(y=1 \mid \mathbf{S}=s, \mathbf{A}=\tilde{\mathbf{a}}) \tag{1}$$

for all possible realizations $\mathbf{a}, \tilde{\mathbf{a}}$ of $\mathbf{A}$ and all possible realizations s of $\mathbf{S}$. This can be seen as follows:

$$\begin{aligned}
\frac{\mathbb{P}(y=1 \mid \mathbf{S}=s, \mathbf{A}=\mathbf{a})}{\mathbb{P}(y=1 \mid \mathbf{S}=s, \mathbf{A}=\tilde{\mathbf{a}})} &= \frac{\mathbb{P}(y=1, \mathbf{S}=s, \mathbf{A}=\mathbf{a})\mathbb{P}(\mathbf{S}=s, \mathbf{A}=\tilde{\mathbf{a}})}{\mathbb{P}(\mathbf{S}=s, \mathbf{A}=\mathbf{a})\mathbb{P}(y=1, \mathbf{S}=s, \mathbf{A}=\tilde{\mathbf{a}})} \\
&= \frac{\mathbb{P}(y=1, \mathbf{S}=s, \mathbf{A}=\mathbf{a})\mathbb{P}(\mathbf{S}=s, \mathbf{A}=\tilde{\mathbf{a}})\mathbb{P}(\mathbf{S}=s)}{\mathbb{P}(\mathbf{S}=s, \mathbf{A}=\mathbf{a})\mathbb{P}(y=1, \mathbf{S}=s, \mathbf{A}=\tilde{\mathbf{a}})\mathbb{P}(\mathbf{S}=s)} \\
&= \frac{\mathbb{P}(y=1, \mathbf{A}=\mathbf{a} \mid \mathbf{S}=s)\mathbb{P}(\mathbf{S}=s, \mathbf{A}=\tilde{\mathbf{a}})}{\mathbb{P}(y=1, \mathbf{A}=\tilde{\mathbf{a}} \mid \mathbf{S}=s)\mathbb{P}(\mathbf{S}=s, \mathbf{A}=\mathbf{a})} \\
&\overset{(***)}{=} \frac{\mathbb{P}(y=1 \mid \mathbf{S}=s)\mathbb{P}(\mathbf{A}=\mathbf{a} \mid \mathbf{S}=s)\mathbb{P}(\mathbf{S}=s, \mathbf{A}=\tilde{\mathbf{a}})}{\mathbb{P}(y=1 \mid \mathbf{S}=s)\mathbb{P}(\mathbf{A}=\tilde{\mathbf{a}} \mid \mathbf{S}=s)\mathbb{P}(\mathbf{S}=s, \mathbf{A}=\mathbf{a})} \\
&= 1,
\end{aligned}$$

where we used for (***) the conditional independence.

Next, we show the other direction. For this purpose, note that for any $c \in \{-1, 1\}, s \in [0, 1]$ and $\tilde{\mathbf{a}}$ it holds that

$$\begin{aligned}
\mathbb{P}(y=c \mid \mathbf{S}=s) &= \sum_{\mathbf{a}} \mathbb{P}(y=c \mid \mathbf{S}=s, \mathbf{A}=\mathbf{a})\mathbb{P}(\mathbf{A}=\mathbf{a} \mid \mathbf{S}=s) && \text{(Law of total probability)} \\
&= \sum_{\mathbf{a}} \mathbb{P}(y=c \mid \mathbf{S}=s, \mathbf{A}=\tilde{\mathbf{a}})\mathbb{P}(\mathbf{A}=\mathbf{a} \mid \mathbf{S}=s) && \text{(By (1))} \\
&= \mathbb{P}(y=c \mid \mathbf{S}=s, \mathbf{A}=\tilde{\mathbf{a}}) \underbrace{\sum_{\mathbf{a}} \mathbb{P}(\mathbf{A}=\mathbf{a} \mid \mathbf{S}=s)}_{=1} \\
&= \mathbb{P}(y=c \mid \mathbf{S}=s, \mathbf{A}=\tilde{\mathbf{a}}).
\end{aligned}$$

With this, for any $c \in \{-1, 1\}, s \in [0, 1]$ and $\mathbf{a}$ we obtain

$$\begin{aligned}
\frac{\mathbb{P}(y=c, \mathbf{A}=\mathbf{a} \mid \mathbf{S}=s)}{\mathbb{P}(y=c \mid \mathbf{S}=s)\mathbb{P}(\mathbf{A}=\mathbf{a} \mid \mathbf{S}=s)} &\overset{(****)}{=} \frac{\mathbb{P}(y=c, \mathbf{A}=\mathbf{a} \mid \mathbf{S}=s)}{\mathbb{P}(y=c \mid \mathbf{S}=s, \mathbf{A}=\mathbf{a})\mathbb{P}(\mathbf{A}=\mathbf{a} \mid \mathbf{S}=s)} \\
&= \frac{\mathbb{P}(y=c, \mathbf{A}=\mathbf{a}, \mathbf{S}=s)\mathbb{P}(\mathbf{S}=s, \mathbf{A}=\mathbf{a})\mathbb{P}(\mathbf{S}=s)}{\mathbb{P}(\mathbf{S}=s)\mathbb{P}(y=c, \mathbf{S}=s, \mathbf{A}=\mathbf{a})\mathbb{P}(\mathbf{A}=\mathbf{a}, \mathbf{S}=s)} \\
&= 1,
\end{aligned}$$

where we used for (****) the equality we derived before (with $\mathbf{a}$ for $\tilde{\mathbf{a}}$).

Solution 2: Fairness Measure

Assume that the sensitive attribute $\mathbf{A}$ has only two possible realizations, say $\mathbf{a}$ and $\tilde{\mathbf{a}}$, and $\mathbf{a}$ corresponds to a privileged group and $\tilde{\mathbf{a}}$ to an unprivileged group. Let $d_{\max} = \min\left(\frac{\mathbb{P}(\hat{y}=1)}{\mathbb{P}(\mathbf{A}=\mathbf{a})}, \frac{\mathbb{P}(\hat{y}=-1)}{\mathbb{P}(\mathbf{A}=\tilde{\mathbf{a}})}\right)$ and consider the following measure for fairness

$$\delta = \frac{\mathbb{P}(\hat{y}=1 \mid \mathbf{A}=\mathbf{a}) - \mathbb{P}(\hat{y}=1 \mid \mathbf{A}=\tilde{\mathbf{a}})}{d_{\max}}.$$

Which values can this measure realize and how can we interpret the extreme values from a fairness perspective?

Solution:

The enumerator $d = \mathbb{P}(\hat{y}=1 \mid \mathbf{A}=\mathbf{a}) - \mathbb{P}(\hat{y}=1 \mid \mathbf{A}=\tilde{\mathbf{a}})$ measures discrimination in terms of differences of the acceptance rates between the privileged group $\mathbf{a}$ and the unprivileged group $\tilde{\mathbf{a}}$:

- The larger the value, the stronger the discrimination against the unprivileged group $\tilde{\mathbf{a}}$.
- The smaller the value, the stronger the discrimination against the privileged group $\mathbf{a}$.
- If the value is zero, there is no discrimination.

The maximum value of d is $d_{\max} = \min\left(\frac{\mathbb{P}(\hat{y}=1)}{\mathbb{P}(\mathbf{A}=\mathbf{a})}, \frac{\mathbb{P}(\hat{y}=-1)}{\mathbb{P}(\mathbf{A}=\tilde{\mathbf{a}})}\right)$, which can be seen as follows: First, we can write

$$\begin{aligned}
d &= \mathbb{P}(\hat{y}=1 \mid \mathbf{A}=\mathbf{a}) - \mathbb{P}(\hat{y}=1 \mid \mathbf{A}=\tilde{\mathbf{a}}) \\
&= \frac{\mathbb{P}(\hat{y}=1, \mathbf{A}=\mathbf{a})}{\mathbb{P}(\mathbf{A}=\mathbf{a})} - \frac{\mathbb{P}(\hat{y}=1, \mathbf{A}=\tilde{\mathbf{a}})}{\mathbb{P}(\mathbf{A}=\tilde{\mathbf{a}})} \\
&= \frac{\mathbb{P}(\hat{y}=1) - \mathbb{P}(\hat{y}=1, \mathbf{A}=\tilde{\mathbf{a}})}{\mathbb{P}(\mathbf{A}=\mathbf{a})} - \frac{\mathbb{P}(\hat{y}=1, \mathbf{A}=\tilde{\mathbf{a}})}{\mathbb{P}(\mathbf{A}=\tilde{\mathbf{a}})} \qquad (\text{Since } \mathbb{P}(\hat{y}=1, \mathbf{A}=\tilde{\mathbf{a}}) + \mathbb{P}(\hat{y}=1, \mathbf{A}=\mathbf{a}) = \mathbb{P}(\hat{y}=1)) \\
&= \frac{\mathbb{P}(\hat{y}=1)}{\mathbb{P}(\mathbf{A}=\mathbf{a})} - \mathbb{P}(\hat{y}=1, \mathbf{A}=\tilde{\mathbf{a}}) \left(\frac{1}{\mathbb{P}(\mathbf{A}=\tilde{\mathbf{a}})} + \frac{1}{\mathbb{P}(\mathbf{A}=\mathbf{a})}\right) =: (A).
\end{aligned}$$

On the other hand, we have also that

$$\begin{aligned}
d &= \mathbb{P}(\hat{y}=1 \mid \mathbf{A}=\mathbf{a}) - \mathbb{P}(\hat{y}=1 \mid \mathbf{A}=\tilde{\mathbf{a}}) \\
&= 1 - \mathbb{P}(\hat{y}=-1 \mid \mathbf{A}=\mathbf{a}) - (1 - \mathbb{P}(\hat{y}=-1 \mid \mathbf{A}=\tilde{\mathbf{a}})) \\
&= \mathbb{P}(\hat{y}=-1 \mid \mathbf{A}=\tilde{\mathbf{a}}) - \mathbb{P}(\hat{y}=-1 \mid \mathbf{A}=\mathbf{a}) \\
&= \frac{\mathbb{P}(\hat{y}=-1, \mathbf{A}=\tilde{\mathbf{a}})}{\mathbb{P}(\mathbf{A}=\tilde{\mathbf{a}})} - \frac{\mathbb{P}(\hat{y}=-1, \mathbf{A}=\mathbf{a})}{\mathbb{P}(\mathbf{A}=\mathbf{a})} \\
&= \frac{\mathbb{P}(\hat{y}=-1) - \mathbb{P}(\hat{y}=-1, \mathbf{A}=\mathbf{a})}{\mathbb{P}(\mathbf{A}=\tilde{\mathbf{a}})} - \frac{\mathbb{P}(\hat{y}=-1, \mathbf{A}=\mathbf{a})}{\mathbb{P}(\mathbf{A}=\mathbf{a})} \\
&\qquad\qquad (\text{Since } \mathbb{P}(\hat{y}=-1, \mathbf{A}=\tilde{\mathbf{a}}) + \mathbb{P}(\hat{y}=-1, \mathbf{A}=\mathbf{a}) = \mathbb{P}(\hat{y}=-1)) \\
&= \frac{\mathbb{P}(\hat{y}=-1)}{\mathbb{P}(\mathbf{A}=\tilde{\mathbf{a}})} - \mathbb{P}(\hat{y}=-1, \mathbf{A}=\mathbf{a}) \left(\frac{1}{\mathbb{P}(\mathbf{A}=\mathbf{a})} + \frac{1}{\mathbb{P}(\mathbf{A}=\tilde{\mathbf{a}})}\right) =: (B).
\end{aligned}$$

Note that in the (A) representation of d, we substract from $\frac{\mathbb{P}(\hat{y}=1)}{\mathbb{P}(\mathbf{A}=\mathbf{a})}$ a non-negative term, so that we can infer that $d \leq \frac{\mathbb{P}(\hat{y}=1)}{\mathbb{P}(\mathbf{A}=\mathbf{a})}$. In the (B) representation of d, we substract from $\frac{\mathbb{P}(\hat{y}=-1)}{\mathbb{P}(\mathbf{A}=\tilde{\mathbf{a}})}$ a non-negative term as well, so that we can infer that $d \leq \frac{\mathbb{P}(\hat{y}=-1)}{\mathbb{P}(\mathbf{A}=\tilde{\mathbf{a}})}$. Thus, $d \leq \min\left(\frac{\mathbb{P}(\hat{y}=1)}{\mathbb{P}(\mathbf{A}=\mathbf{a})}, \frac{\mathbb{P}(\hat{y}=-1)}{\mathbb{P}(\mathbf{A}=\tilde{\mathbf{a}})}\right) = d_{\max}$. The terms which are substracted in (A) and (B) can be zero, i.e., it can hold that $\mathbb{P}(\hat{y}=1, \mathbf{A}=\tilde{\mathbf{a}}) = 0 = \mathbb{P}(\hat{y}=-1, \mathbf{A}=\mathbf{a})$, which both corresponds to maximal discrimination against the $\tilde{\mathbf{a}}$ group. Hence, $d_{\max}$ can be attained by d.

In summary, dividing $d = \mathbb{P}(\hat{y}=1 \mid \mathbf{A}=\mathbf{a}) - \mathbb{P}(\hat{y}=1 \mid \mathbf{A}=\tilde{\mathbf{a}})$ by $d_{\max}$ normalizes the measure of fairness

$$\delta = \frac{\mathbb{P}(\hat{y}=1 \mid \mathbf{A}=\mathbf{a}) - \mathbb{P}(\hat{y}=1 \mid \mathbf{A}=\tilde{\mathbf{a}})}{d_{\max}}.$$

Thus, a maximal value of $\delta = 1$ corresponds to maximal discrimination of the unprivileged group $\mathbf{A}=\tilde{\mathbf{a}}$.

Solution 3: Calibration

Assume we have a set

$$\mathcal{D} = \left\{ (\mathbf{x}^{(i)}, y^{(i)}) \right\}_{i=1}^{N} \in (\mathcal{X} \times \mathcal{Y})^N$$

of training examples $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$. Recall that in logistic regression the probability $p(y=+1 \,|\, \mathbf{x})$ is modeled as

$$\begin{aligned}
\pi_{\boldsymbol{\theta}} : \; \mathcal{X} &\to [0,1] \\
\mathbf{x} &\mapsto \frac{1}{1+\exp\left(-\langle \boldsymbol{\theta}, \mathbf{x} \rangle\right)},
\end{aligned}$$

with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top \in \mathbb{R}^d$ is a parameter vector.

Assume we have a partition of $\mathcal{X}$ into $G \in \mathbb{N}$ groups[1], say $\mathcal{X}_1, \ldots, \mathcal{X}_G$. Consider the following quantity

$$H_{\boldsymbol{\theta}}(G) = \sum_{g=1}^{G} \frac{(O_{g,+1} - E_{g,+1|\boldsymbol{\theta}})^2}{E_{g,+1|\boldsymbol{\theta}}} + \frac{(O_{g,-1} - E_{g,-1|\boldsymbol{\theta}})^2}{E_{g,-1|\boldsymbol{\theta}}},$$

[1]Here, we mean disjoint subsets of $\mathbb{R}^d$ whose union is $\mathcal{X} = \mathbb{R}^d$.

where $O_{g,\pm 1}$ is the number of *observed* $y's$ which are ± 1 and the corresponding $\mathbf{x}$ is an element of $\mathcal{X}_g$, and $E_{g,\pm 1|\boldsymbol{\theta}}$ is the number of *expected* $y's$ which are ± 1 under the model $\pi_{\boldsymbol{\theta}}$ and the corresponding $\mathbf{x}$ is an element of $\mathcal{X}_g$.

(a) Give a mathematical definition of $O_{g,+1}$, $O_{g,-1}$, $E_{g,+1|\boldsymbol{\theta}}$ and $E_{g,-1|\boldsymbol{\theta}}$.

Solution: Recall the interpretation of $\pi_{\boldsymbol{\theta}}(\mathbf{x})$: It is giving the (fitted) probability that $y = +1$ for given $\mathbf{x}$. In other words, we expect that $y = +1$ given $\mathbf{x}$ with probability $\pi_{\boldsymbol{\theta}}(\mathbf{x})$. Similarly, we expect that $y = -1$ given $\mathbf{x}$ with probability $(1 - \pi_{\boldsymbol{\theta}}(\mathbf{x}))$. With this, we can write the quantities as follows:

$$
\begin{aligned}
O_{g,+1} &= \sum_{i=1}^{N} \mathbb{1}_{[y^{(i)}=+1]} \mathbb{1}_{[\mathbf{x}^{(i)} \in \mathcal{X}_g]}, \\
O_{g,-1} &= \sum_{i=1}^{N} \mathbb{1}_{[y^{(i)}=-1]} \mathbb{1}_{[\mathbf{x}^{(i)} \in \mathcal{X}_g]}, \\
E_{g,+1|\boldsymbol{\theta}} &= \sum_{i=1}^{N} \pi_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) \mathbb{1}_{[\mathbf{x}^{(i)} \in \mathcal{X}_g]}, \\
E_{g,-1|\boldsymbol{\theta}} &= \sum_{i=1}^{N} (1 - \pi_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})) \mathbb{1}_{[\mathbf{x}^{(i)} \in \mathcal{X}_g]}.
\end{aligned}
$$

(b) If the model $\pi_{\boldsymbol{\theta}}$ is (approximately) well-calibrated, what values should $H_{\boldsymbol{\theta}}(G)$ take? What is a desirable property of the partition $\mathcal{X}_1, \ldots, \mathcal{X}_G$ of $\mathcal{X}$?

Solution:

If the model is approximately well-calibrated, then it should hold for any $s \in [0, 1]$

$$
\mathbb{P}(y = +1 \,|\, \pi_{\boldsymbol{\theta}}(\mathbf{x}) = s) \approx s, \quad \forall \mathbf{x} \in \mathcal{X}.
$$

In words, if the logistic model predicts that $y = +1$ will occur with probability $s = \pi_{\boldsymbol{\theta}}(\mathbf{x})$, then $y = +1$ should occur with probability (approximately) s. In particular, the frequency with which $y = +1$ is observed for some particular $\mathbf{x}$ should match the expected frequency under $\pi_{\boldsymbol{\theta}}(\mathbf{x})$.

Thus, we would expect that $O_{g,+1} \approx E_{g,+1|\boldsymbol{\theta}}$ and $O_{g,-1} \approx E_{g,-1|\boldsymbol{\theta}}$ for every group $g \in \{1, \ldots, G\}$, if the model $\pi_{\boldsymbol{\theta}}$ is approximately well-calibrated. Hence, $H_{\boldsymbol{\theta}}(G)$ should be close to 0 or not "too large".

However, as we rarely observe the same exact feature vector $\mathbf{x}$ we could group some of them together. If we assume that the probability of occurrence for $y = +1$ is similar if $\mathbf{x}$ and $\mathbf{x}'$ are close (with respect to some metric on $\mathbb{R}^d$), then it would be sensible to use a group $\mathcal{X}_g$ which is a connected subset of $\mathbb{R}^d$, e.g., a hypercube. Moreover, it would be desirable if the number of $\mathbf{x}^{(i)}$'s in $\mathcal{X}_g$ is roughly the same among all groups. This would make sure that the frequencies considered by $O_{g,+1}$ (or $O_{g,-1}$) are balanced.

(c) Generate a data set $\mathcal{D}$ with $\mathcal{X} = \mathbb{R}$ of size $N = 100$ in the following way:

- Sample each x_i according to a standard normal distribution;
- Sample u_i uniformly at random from the unit interval;
- Set $y^{(i)} = 2 \cdot \mathbb{1}_{[u_i < \exp(x_i)/(1+\exp(x_i))]} - 1$

Fit a logistic regression model π_θ to the data and visualize whether the model is calibrated in a suitable way. Next, compute $H_\theta(G)$ for different values of G, say $G \in \{5, \ldots, 15\}$, where you use a suitable partition $\mathcal{X}_1, \ldots, \mathcal{X}_G$ of $\mathcal{X}$. Repeat this whole procedure (of computing $H_\theta(G)$) for 1000 times and compute the average over the computed $H_\theta(G)$ values and plot these averages as a function of G into one figure.

(d) Generate your data set $\mathcal{D}$ of size $N = 100$ in the following way:

- Sample each x_i according to a standard normal distribution;
- Sample u_i uniformly at random from the unit interval;
- Set $y^{(i)} = 2 \cdot \mathbb{1}_{[u_i < \exp(x_i^2)/(1+\exp(x_i^2))]} - 1$

Repeat (c) for this data generating process and include the resulting average curve into the figure of (c).