Exercise 1: Equivalent Representation of Separation and Sufficiency

(a) Show the equivalence between

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

and

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

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

(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 $\mathbf{S}$.

Exercise 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?

Exercise 3: Calibration

Assume we have a set

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

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

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

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 $\mathcal{X}_g$.

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

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

(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 $\mathcal{X}$?

(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).