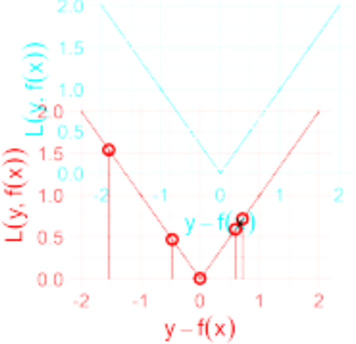
Introduction to Machine Learning

Advanced Risk Minimization

L1 Risk Minimizer (Deep-Dive)

L1 Risk Minimizer (Deep-Dive)



Learning goals

- Derive the risk minimizer of the L1-loss
- Derive the optimal constant model for the L1-loss

Learning goals

- Derive the risk minimizer of the L1-loss
- Derive the optimal constant model for the L1-loss

L1-LOSS: RISK MINIMIZER / 2

* **Note** that since we are computing the derivative w.r.t. the integration boundaries, we need to use Leibniz integration rule

$$\begin{aligned}
\frac{\partial}{\partial c}\left(\int_a^c g(c,y)\, \mathrm{d}y\right) &= g(c,c) + \int_a^c \frac{\partial}{\partial c} g(c,y)\, \mathrm{d}y \\
\frac{\partial}{\partial c}\left(\int_c^a g(c,y)\, \mathrm{d}y\right) &= -g(c,c) + \int_c^a \frac{\partial}{\partial c} g(c,y)\, \mathrm{d}y
\end{aligned}$$

We get

$$\begin{aligned}
& \frac{\partial}{\partial c}\left(\int_{-\infty}^c -(y-c)\, p(y)\, \mathrm{d}y + \int_c^\infty (y-c)\, p(y)\, \mathrm{d}y\right) \\
= & \frac{\partial}{\partial c}\left(\int_{-\infty}^c \underbrace{-(y-c)\, p(y)}_{g_1(c,y)}\, \mathrm{d}y\right) + \frac{\partial}{\partial c}\left(\int_c^\infty \underbrace{(y-c)\, p(y)}_{g_2(c,y)}\, \mathrm{d}y\right) \\
= & \underbrace{g_1(c,c)}_{=0} + \int_{-\infty}^c \frac{\partial}{\partial c}(-(y-c))\, p(y)\, \mathrm{d}y - \underbrace{g_2(c,c)}_{=0} + \int_c^\infty \frac{\partial}{\partial c}(y-c)\, p(y)\, \mathrm{d}y \\
= & \int_{-\infty}^c p(y)\, \mathrm{d}y + \int_c^\infty -p(y)\, \mathrm{d}y.
\end{aligned}$$

L1-LOSS: OPTIMAL CONSTANT MODEL / 2

W.l.o.g. assume now that all $y^{(i)}$ are sorted in increasing order.
Let us define $i_{\max} = n/2$ for n even and $i_{\max} = (n-1)/2$ for n odd and consider the intervals

$$\mathcal{I}_i := [y^{(i)}, y^{(n+1-i)}],\ i \in \{1, ..., i_{\max}\}.$$

By construction $\mathcal{I}_{j+1} \subseteq \mathcal{I}_j$ for $j \in \{1, \dots, i_{\max} - 1\}$ and $\mathcal{I}_{i_{\max}} \subseteq \mathcal{I}_i$.
With this, $\mathcal{R}_{\text{emp}}$ can be expressed as

$$\begin{aligned}
\mathcal{R}_{\text{emp}}(\theta) &= \sum_{i=1}^{n} L(y^{(i)}, \theta) = \sum_{i=1}^{n} |y^{(i)} - \theta| \\
&= \underbrace{|y^{(1)} - \theta| + |y^{(n)} - \theta|}_{=S_{y^{(1)}, y^{(n)}}(\theta)} + \underbrace{|y^{(2)} - \theta| + |y^{(n-1)} - \theta|}_{=S_{y^{(2)}, y^{(n-1)}}(\theta)} + \dots \\
&= \begin{cases} \sum_{i=1}^{i_{\max}} S_{y^{(i)}, y^{(n+1-i)}}(\theta) & \text{for } n \text{ is even} \\ \sum_{i=1}^{i_{\max}} \left(S_{y^{(i)}, y^{(n+1-i)}}(\theta)\right) + |y^{((n+1)/2)} - \theta| & \text{for } n \text{ is odd.} \end{cases}
\end{aligned}$$

L1-LOSS: OPTIMAL CONSTANT MODEL / 3

From this follows that

- for "n is even": $\hat{\theta} \in \mathcal{I}_{i_{\max}} = [y^{(n/2)}, y^{(n/2+1)}]$ minimizes S_i for all $i \in \{1, \dots, i_{\max}\}$ $\Rightarrow$ it minimizes $\mathcal{R}_{\text{emp}}$,
- for "n is odd": $\hat{\theta} = y^{(n+1)/2} \in \mathcal{I}_{i_{\max}}$ minimizes S_i for all $i \in \{1, \dots, i_{\max}\}$ and it's minimal for $|y^{((n+1)/2)} - \theta|$ $\Rightarrow$ it minimizes $\mathcal{R}_{\text{emp}}$.

Since the median fulfills these conditions, we can conclude that it minimizes the $L1$ loss.