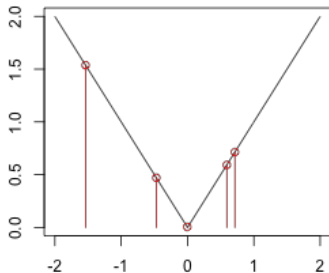


# Introduction to Machine Learning

## Regression Losses: L1-loss



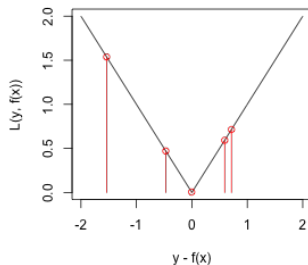
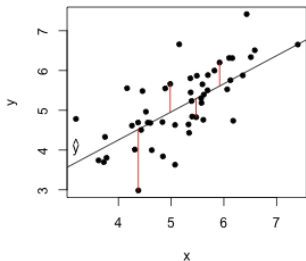
### Learning goals

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

# L1-LOSS

$$L(y, f(\mathbf{x})) = |y - f(\mathbf{x})|$$

- More robust than  $L_2$ , outliers in  $y$  are less problematic.
- Analytical properties: convex, not differentiable for  $y = f(\mathbf{x})$  (optimization becomes harder).



# L1-LOSS: RISK MINIMIZER

We calculate the (true) risk for the L1-Loss  $L(y, f(\mathbf{x})) = |y - f(\mathbf{x})|$  with unrestricted  $\mathcal{H} = \{f : \mathcal{X} \rightarrow \mathcal{Y}\}$ .

- We use the law of total expectation

$$\mathcal{R}(f) = \mathbb{E}_{\mathbf{x}} \left[ \mathbb{E}_{y|\mathbf{x}} [|y - f(\mathbf{x})| \mid \mathbf{x} = \mathbf{x}] \right].$$

- As the functional form of  $f$  is not restricted, we can just optimize point-wise at any point  $\mathbf{x} = \mathbf{x}$ . The best prediction at  $\mathbf{x} = \mathbf{x}$  is then

$$\hat{f}(\mathbf{x}) = \operatorname{argmin}_c \mathbb{E}_{y|\mathbf{x}} [|y - c|] = \operatorname{med}_{y|\mathbf{x}} [y \mid \mathbf{x}].$$

# L1-LOSS: RISK MINIMIZER

**Proof:** Let  $p(y)$  be the density function of  $y$ . Then:

$$\begin{aligned}\operatorname{argmin}_c \mathbb{E}[|y - c|] &= \operatorname{argmin}_c \int_{-\infty}^{\infty} |y - c| p(y) dy \\ &= \operatorname{argmin}_c \int_{-\infty}^c -(y - c) p(y) dy + \int_c^{\infty} (y - c) p(y) dy\end{aligned}$$

We now compute the derivative of the above term and set it to 0

$$\begin{aligned}0 &= \frac{\partial}{\partial c} \left( \int_{-\infty}^c -(y - c) p(y) dy + \int_c^{\infty} (y - c) p(y) dy \right) \\ &\stackrel{* \text{ Leibniz}}{=} \int_{-\infty}^c p(y) dy - \int_c^{\infty} p(y) dy = \mathbb{P}_y(y \leq c) - (1 - \mathbb{P}_y(y \leq c)) \\ &= 2 \cdot \mathbb{P}_y(y \leq c) - 1 \\ \Leftrightarrow 0.5 &= \mathbb{P}_y(y \leq c),\end{aligned}$$

which yields  $c = \operatorname{med}_y(y)$ .

# L1-LOSS: RISK MINIMIZER

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

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

We get

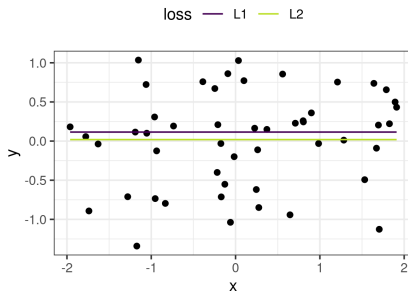
$$\begin{aligned}& \frac{\partial}{\partial c} \left( \int_{-\infty}^c -(y - c) p(y) dy + \int_c^{\infty} (y - c) p(y) dy \right) \\&= \frac{\partial}{\partial c} \left( \int_{-\infty}^c \underbrace{-(y - c) p(y)}_{g_1(c, y)} dy \right) + \frac{\partial}{\partial c} \left( \int_c^{\infty} \underbrace{(y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

# L1-LOSS: OPTIMAL CONSTANT MODEL

The optimal constant model in terms of the theoretical risk for the L1 loss is the median over  $y$ :

$$f(\mathbf{x}) = \text{med}_{y|x} [y | \mathbf{x}] \stackrel{\text{drop } \mathbf{x}}{=} \text{med}_y [y]$$

The optimizer of the empirical risk is  $\text{med}(y^{(i)})$  over  $y^{(i)}$ , which is the empirical estimate for  $\text{med}_y [y]$ .



# L1-LOSS: OPTIMAL CONSTANT MODEL

## Proof:

- Firstly note that for  $n = 1$  the median  $\hat{\theta} = \text{med}(y^{(i)}) = y^{(1)}$  obviously minimizes the empirical risk  $\mathcal{R}_{\text{emp}}$  associated to the  $L1$  loss  $L$ .
- Hence let  $n > 1$  in the following: Let

$$S_{a,b} : \mathbb{R} \rightarrow \mathbb{R}_0^+, \theta \mapsto |a - \theta| + |b - \theta|$$

for  $a, b \in \mathbb{R}$ . It holds that

$$S_{a,b}(\theta) = \begin{cases} |a - b|, & \text{for } \theta \in [a, b] \\ |a - b| + 2 \cdot \min\{|a - \theta|, |b - \theta|\}, & \text{otherwise.} \end{cases}$$

Thus, any  $\hat{\theta} \in [a, b]$  minimizes  $S_{a,b}$ .

# L1-LOSS: OPTIMAL CONSTANT MODEL

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

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

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

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



# L1-LOSS: OPTIMAL CONSTANT MODEL

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 its 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.