

RISK MINIMIZATION

Now, we will discuss the most common loss functions and the optimal solution with respect to

- the theoretical risk

$$\mathcal{R}(f) = \mathbb{E}[L(y, f(\mathbf{x}))] = \int_{\mathcal{X} \times \mathcal{Y}} L(y, f(\mathbf{x})) \, d\mathbb{P}_{xy}$$

and

- the empirical risk

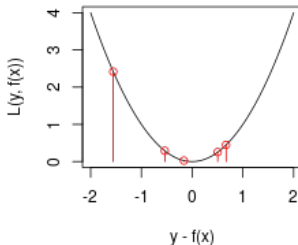
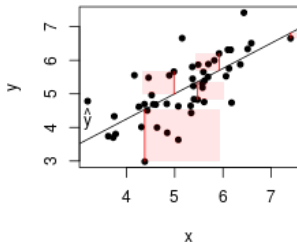
$$\mathcal{R}_{\text{emp}}(f) = \sum_{i=1}^n L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right).$$

L2-Loss

L2-LOSS

$$L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2 \quad \text{or} \quad L(y, f(\mathbf{x})) = 0.5 (y - f(\mathbf{x}))^2$$

- Tries to reduce large residuals (if residual is twice as large, loss is 4 times as large), hence outliers in y can become problematic
- Analytic properties: convex, differentiable (gradient no problem in loss minimization)
- Residuals = Pseudo-residuals: $\tilde{r} = -\frac{\partial 0.5(y-f(\mathbf{x}))^2}{\partial f(\mathbf{x})} = y - f(\mathbf{x}) = r$



L2-LOSS: POINT-WISE OPTIMUM

Let us consider the (theoretical) risk for $\mathcal{Y} = \mathbb{R}$ and the $L2$ -Loss $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$.

- By the law of total expectation

$$\begin{aligned}\mathcal{R}(f) &= \mathbb{E}_{xy} [L(y, f(\mathbf{x}))] \\ &= \mathbb{E}_{\mathbf{x}} [\mathbb{E}_{y|\mathbf{x}} [L(y, f(\mathbf{x})) \mid \mathbf{x} = \mathbf{x}]] \\ &= \mathbb{E}_{\mathbf{x}} [\mathbb{E}_{y|\mathbf{x}} [(y - f(\mathbf{x}))^2 \mid \mathbf{x} = \mathbf{x}]] .\end{aligned}$$

- Assume we are free to choose f as we wish: At any point $\mathbf{x} = \mathbf{x}$ we can predict any c we want. The best point-wise prediction is the conditional mean

$$\hat{f}(\mathbf{x}) = \operatorname{argmin}_c \mathbb{E}_{y|\mathbf{x}} [(y - c)^2 \mid \mathbf{x} = \mathbf{x}] \stackrel{(*)}{=} \mathbb{E}_{y|\mathbf{x}} [y \mid \mathbf{x}] .$$

L2-LOSS: POINT-WISE OPTIMUM

- (*) follows from:

$$\begin{aligned} & \operatorname{argmin}_c \mathbb{E} [(y - c)^2] \\ = & \operatorname{argmin}_c \underbrace{\mathbb{E} [(y - c)^2] - (\mathbb{E}[y] - c)^2}_{\operatorname{Var}[y - c] = \operatorname{Var}[y]} + (\mathbb{E}[y] - c)^2 \\ = & \operatorname{argmin}_c \operatorname{Var}[y] + (\mathbb{E}[y] - c)^2 \\ = & \mathbb{E}[y]. \end{aligned}$$

L2-LOSS: OPTIMAL CONSTANT MODEL

For the sake of simplicity, let us consider the hypothesis space \mathcal{H} of constant models

$$\mathcal{H} = \{f \mid f(\mathbf{x}) = \theta, \theta \in \mathbb{R}\}.$$

Goal: Derive the optimal constant model w.r.t. the $L2$ -Loss.

$$\begin{aligned} f &= \arg \min_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f) = \sum_{i=1}^n L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right) \\ \Leftrightarrow \hat{\theta} &= \arg \min_{\theta \in \mathbb{R}} \sum_{i=1}^n \left(y^{(i)} - \theta\right)^2 \end{aligned}$$

L2-LOSS: OPTIMAL CONSTANT MODEL

We calculate the first derivative of \mathcal{R}_{emp} w.r.t. θ and set it to 0:

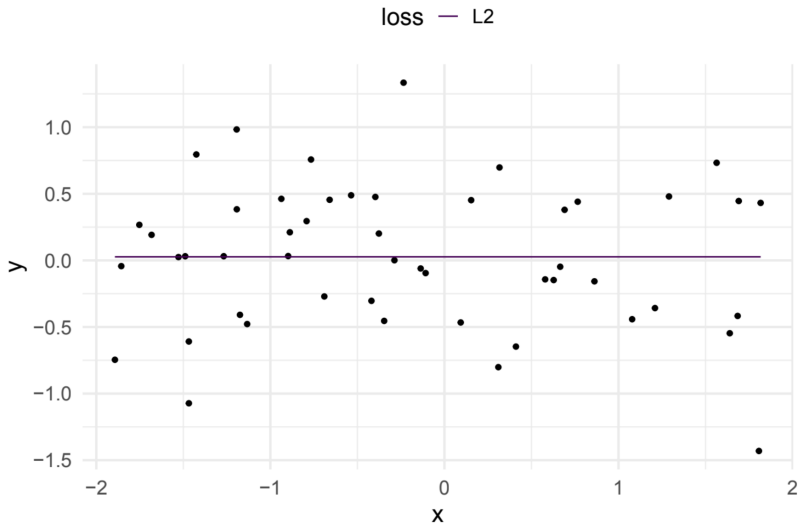
$$\frac{\partial \mathcal{R}_{\text{emp}}(\theta)}{\partial \theta} = 2 \sum_{i=1}^n (y^{(i)} - \theta) \stackrel{!}{=} 0$$

$$\sum_{i=1}^n y^{(i)} - n\theta = 0$$

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n y^{(i)} =: \bar{y}.$$

So the optimal constant model predicts the average of observed outcomes $\hat{f}(\mathbf{x}) = \bar{y}$.

L2-LOSS: OPTIMAL CONSTANT MODEL

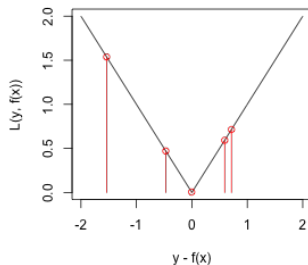
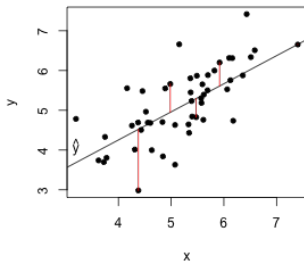


L1-Loss

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: POINT-WISE OPTIMUM

We calculate the (theoretical) risk for the $L1$ -Loss

$L(y, f(\mathbf{x})) = |y - f(\mathbf{x})|$ with unrestricted $\mathcal{H} = \{f : \mathcal{X} \rightarrow \mathcal{Y}\}$.

- Again, 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

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

L1-LOSS: POINT-WISE OPTIMUM

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

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

Setting the derivation w.r.t. c to zero yields:

$$\begin{aligned}0 &= \int_{-\infty}^c p(y) dy - \int_c^{\infty} p(y) dy \\ &= \mathbb{P}_y(y \leq c) - (1 - \mathbb{P}_y(y \leq c)) \\ &= 2 \cdot \mathbb{P}_y(y \leq c) - 1 \\ \Leftrightarrow 0.5 &= \mathbb{P}_y(y \leq c),\end{aligned}$$

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

L1-LOSS: OPTIMAL CONSTANT MODEL

Goal: Derive the optimal constant model

$$f \in \mathcal{H} = \{f(\mathbf{x}) = \theta \mid \theta \in \mathbb{R}\},$$

w.r.t. the $L1$ -Loss.

$$f = \arg \min_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f)$$

$$\Leftrightarrow \hat{\theta} = \arg \min_{\theta \in \mathbb{R}} \sum_{i=1}^n |y^{(i)} - \theta|$$

$$\Leftrightarrow \hat{\theta} = \text{med}(y^{(i)})$$

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}_{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

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}_{emp} can be expressed as

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

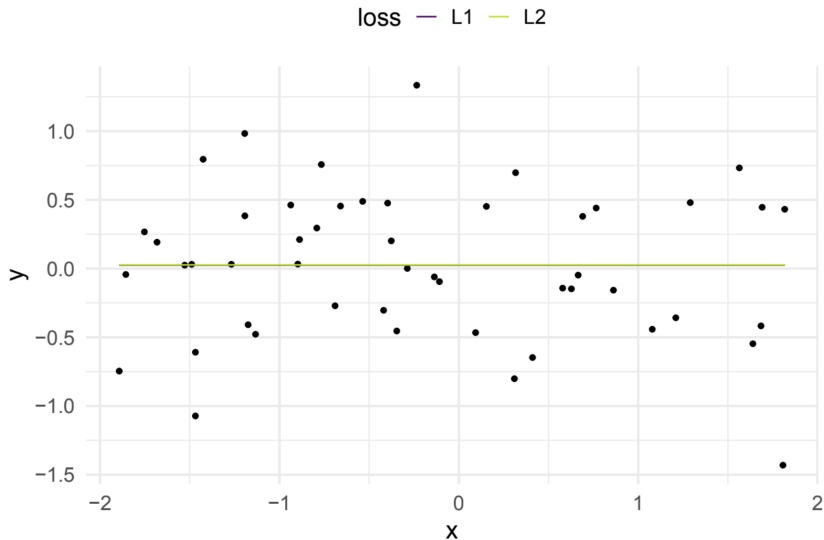
L1-LOSS: OPTIMAL CONSTANT MODEL

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 $i \in \{1, \dots, i_{\max}\} \Rightarrow \mathcal{R}_{\text{emp}}$ reaches its global minimum at $\hat{\theta}$,
- for “ n is odd”: $\hat{\theta} = y^{(n+1)/2} \in \mathcal{I}_{i_{\max}}$ minimizes S_i for all $i \in \{1, \dots, i_{\max}\} \Rightarrow \mathcal{R}_{\text{emp}}$ reaches its global minimum at $\hat{\theta}$.

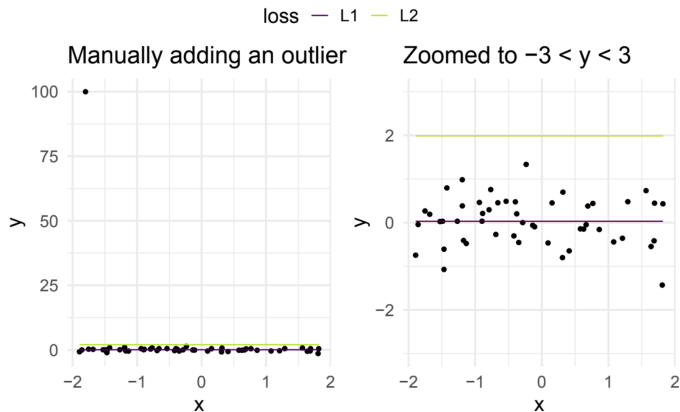
Since the median fulfills these conditions, we can conclude that it minimizes the $L1$ loss.

L1-LOSS: OPTIMAL CONSTANT MODEL



L1-LOSS: OPTIMAL CONSTANT MODEL

We see that the $L1$ -Loss is more robust w.r.t. outliers than the $L2$ -Loss.

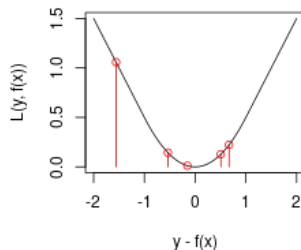
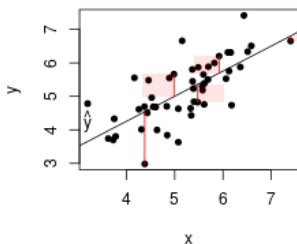


Huber-Loss

HUBER-LOSS

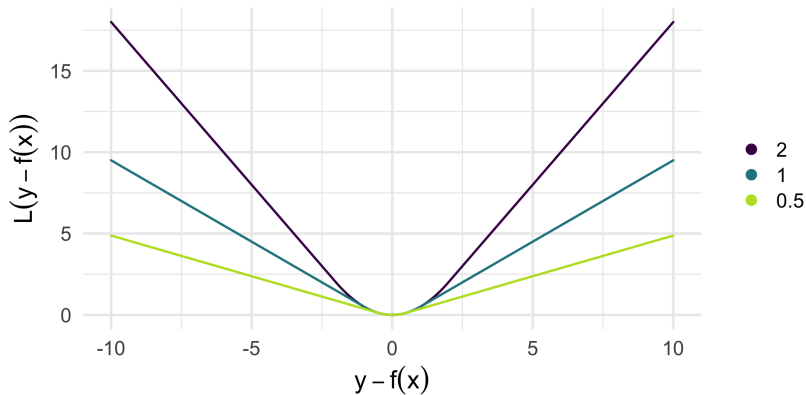
$$L(y, f(\mathbf{x})) = \begin{cases} \frac{1}{2}(y - f(\mathbf{x}))^2 & \text{if } |y - f(\mathbf{x})| \leq \delta \\ \delta|y - f(\mathbf{x})| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}, \delta > 0$$

- Piecewise combination of $L1$ and $L2$ loss
- Analytic properties: Convex, differentiable, robust
- Combines advantages of $L1$ and $L2$ loss: differentiable + robust



HUBER-LOSS

The following plot shows the Huber loss for different values of δ .



HUBER LOSS: OPTIMAL CONSTANT MODEL

What is the optimal constant model $f(\mathbf{x}) = \theta$ w.r.t. the Huber loss?

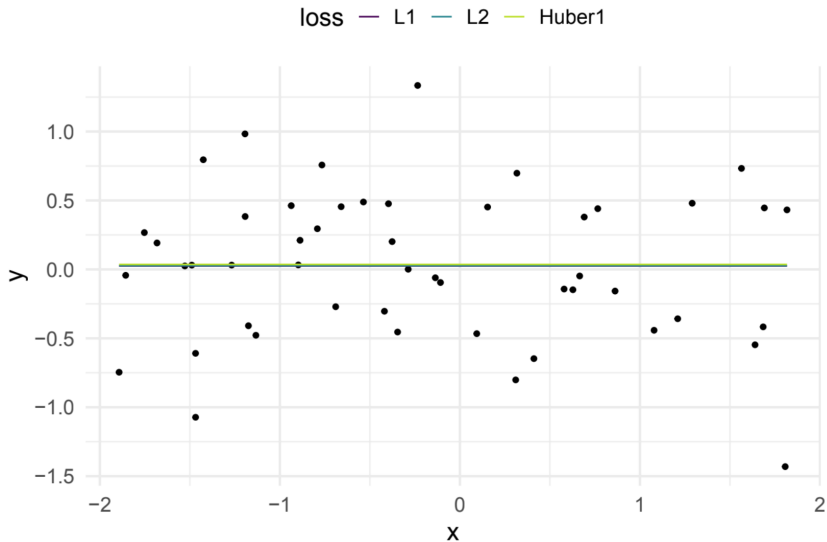
$$f = \arg \min_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f)$$

$$\Leftrightarrow \hat{\theta} = \arg \min_{\theta \in \mathbb{R}} \sum_{i=1}^n L(y, \theta)$$

$$\text{with } L(y, \theta) = \begin{cases} \frac{1}{2}(y - \theta)^2 & \text{if } |y - \theta| \leq \delta \\ \delta|y - \theta| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}.$$

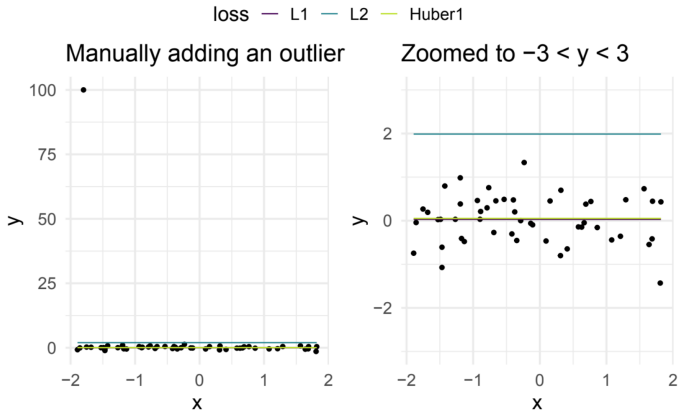
- There is no closed-form solution.
- Numerical optimization methods are necessary.
- \rightarrow the “optimal” solution can only be approached to a certain degree of accuracy via iterative optimization.

HUBER LOSS: OPTIMAL CONSTANT MODEL



HUBER LOSS: OPTIMAL CONSTANT MODEL

And when adding an outlier:



Summary