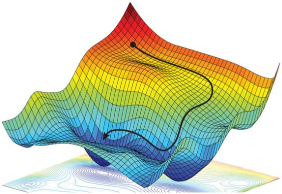
Introduction to Machine Learning

Advanced Regression Losses



Learning goals

- Understand that an ML model is simply a parametrized curve
- Understand that the hypothesis space lists all admissible models for a learner
- Understand the relationship between the hypothesis space and the parameter space

LOSSES AND RESIDUALS

- Regression losses usually only depend on the **residuals**

$$r \quad := \quad y - f(\mathbf{x})$$

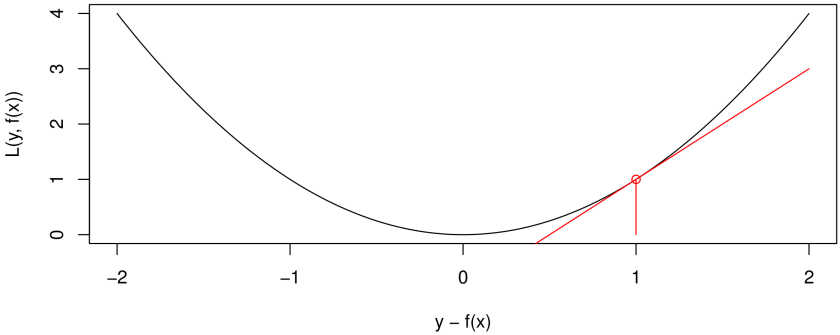
- A loss is called **distance-based** if
  - it can be written in terms of the residual

$$L\left(y, f(\mathbf{x})\right) = \psi(r) \text{ for some } \psi : \mathbb{R} \to \mathbb{R}$$

  - $\psi(r) = 0 \Leftrightarrow r = 0$.
- A loss is **translation-invariant**, if $L(y + a, f(\mathbf{x}) + a) = L\left(y, f(\mathbf{x})\right)$.

LOSS PLOTS

We call the plot that shows the point-wise error, i.e. the loss $L(y, f(\mathbf{x}))$ vs. the **residuals** $r := y - f(\mathbf{x})$ (for regression), **loss plot**. The pseudo-residual corresponds to the slope of the tangent in $(y - f(\mathbf{x}), L(y, f(\mathbf{x})))$.



THE ROLE OF LOSS FUNCTIONS

Why should we care about how to choose the loss function $L(y, f(\mathbf{x}))$?

- **Statistical** properties of f: Choice of loss implies statistical properties of f like robustness and an implicit error distribution.
- **Computational** / **Optimization** complexity of the optimization problem: The complexity of the optimization problem

$$\arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$$

  is influenced by the choice of the loss function, i.e.

  - Smoothness of the objective

    Some optimization methods require smoothness (e.g. gradient methods).
  - Uni- or multimodality of the problem

    If $L(y, f(\mathbf{x}))$ is convex in its second argument, and $f(\mathbf{x} \mid \boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$, then $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ is convex; every local minimum of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ is a global one. If L is not convex, $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ might have multiple local minima (bad!).

RISK MINIMIZATION

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

- the theoretical risk

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

  and

- the empirical risk

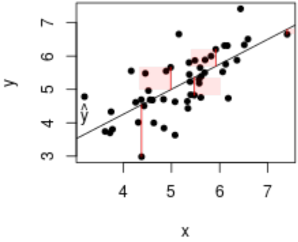
$$\mathcal{R}_{\text{emp}}(f) = \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right).$$

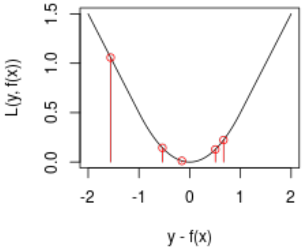
Huber-Loss

HUBER-LOSS

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

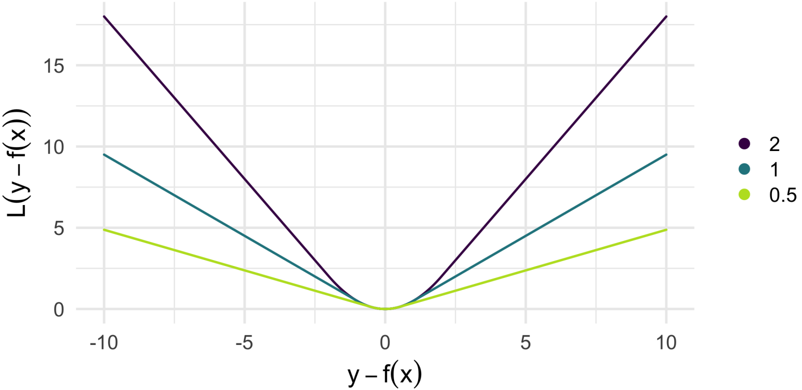
- Piecewise combination of $L1$ and $L2$ loss
- Analytic properties: Convex, differentiable, robust
- Combines advantages of $L1$ and $L2$ loss: differentiable + robust





HUBER-LOSS

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



HUBER LOSS: OPTIMAL CONSTANT MODEL

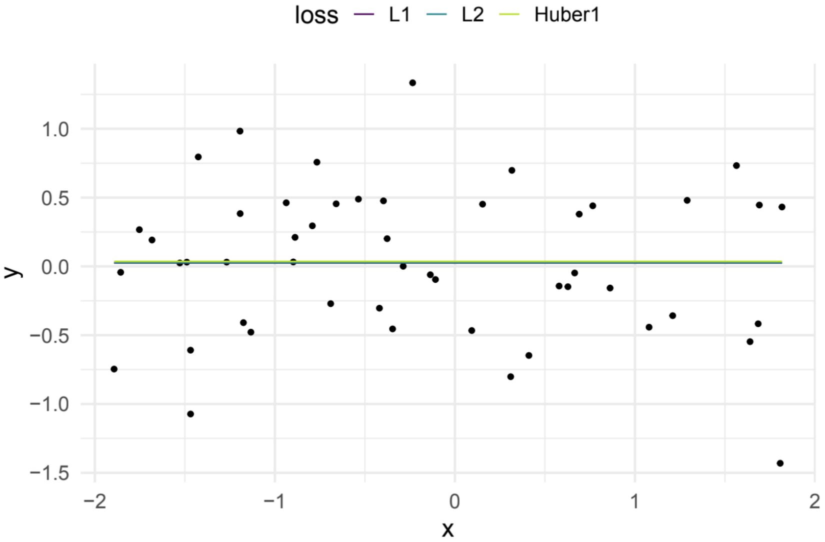
What is the optimal constant model $f(\mathbf{x}) = \boldsymbol{\theta}$ w.r.t. the Huber loss?

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

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

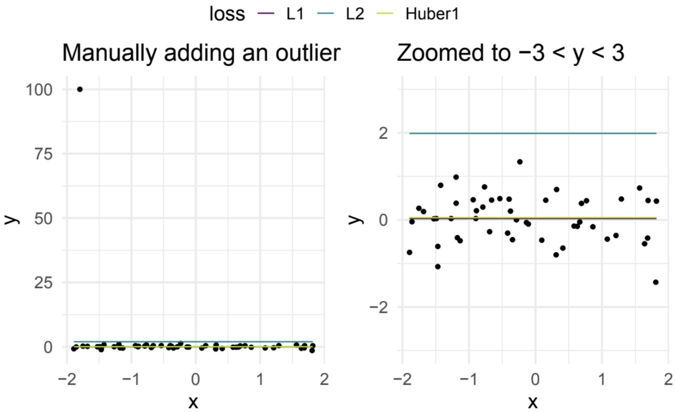
- There is no closed-form solution.
- Numerical optimization methods are necessary.
- $\rightarrow$ the “optimal” solution can only be approached to a certain degree of accuracy via iterative optimization.

HUBER LOSS: OPTIMAL CONSTANT MODEL



HUBER LOSS: OPTIMAL CONSTANT MODEL

And when adding an outlier:



Summary

SUMMARY OF LOSS FUNCTIONS

	$L2$	$L1$	Huber
Point-wise optimum	$\mathbb{E}_{y\|x}\left[y \mid \mathbf{x}\right]$	$\text{med}_{y\|x}\left[y \mid \mathbf{x}\right]$	n.a.
Best constant	$\frac{1}{n}\sum_{i=1}^{n} y^{(i)}$	$\text{med}\left(y^{(i)}\right)$	n.a.
Differentiable	✓	✗	✓
Convex	✓	✓	✓
Robust	✗	✓	✓

There are many other loss functions for regression tasks, for example:

- Quantile-Loss
- ϵ-insensitive-Loss
- Log-Barrier-Loss

Loss functions might also be customized to an objective that is defined by an application.

Log-Barrier Loss