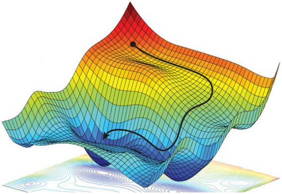
Introduction to Machine Learning

Properties of Loss Functions



Learning goals

- Learn about robust loss functions
- Learn about analytical properties important for optimization

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 (not covered here)
- **Robustness**: Some loss functions are more robust towards outliers than others.
- **Computational** / **Optimization** complexity of the optimization problem: The complexity of the optimization problem

$$\operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$$

is influenced by the choice of the loss function.

TYPES OF REGRESSION LOSSES

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

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

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

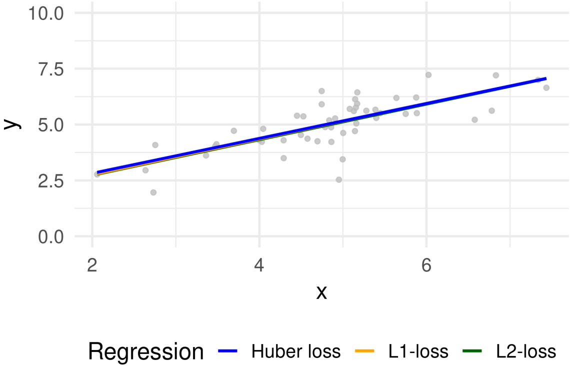
    $$L\left(y, f(\mathbf{x})\right) = \psi(\epsilon) \text{ for some } \psi : \mathbb{R} \to \mathbb{R}$$

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

ROBUSTNESS

Many problems in machine learning require **robustness** – that a model is less influenced by outliers then inliers.

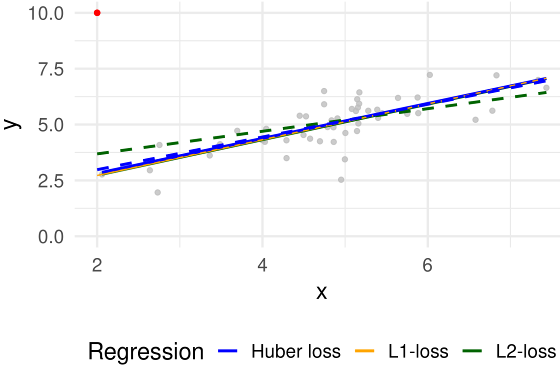
The L2 loss is an example for a loss function that is not very robust towards outliers. It penalizes large residuals more than the L1 or the Huber loss. The L1 and the Huber loss are thus regarded robust.



ROBUSTNESS

Many problems in machine learning require **robustness** – that a model is less influenced by outliers then inliers.

The L2 loss is an example for a loss function that is not very robust towards outliers. It penalizes large residuals more than the L1 or the Huber loss. The L1 and the Huber loss are thus regarded robust.



ANALYTICAL PROPERTIES

- Smoothness of the objective

  Some optimization methods require smoothness (e.g. gradient methods).

- Uni- or multimodality of the problem

  If $L\left(y, f(\mathbf{x})\right)$ 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!).

Summary

SUMMARY OF LOSS FUNCTIONS

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

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

- Quantile-Loss
- ϵ-insensitive-Loss

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