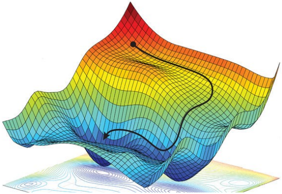
Introduction to Machine Learning

Computational Properties of Loss Functions



Learning goals

- Understand why the choice of the loss function matters
- Know some advanced loss functions

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

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

  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!).

TYPES OF REGRESSION LOSSES

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

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

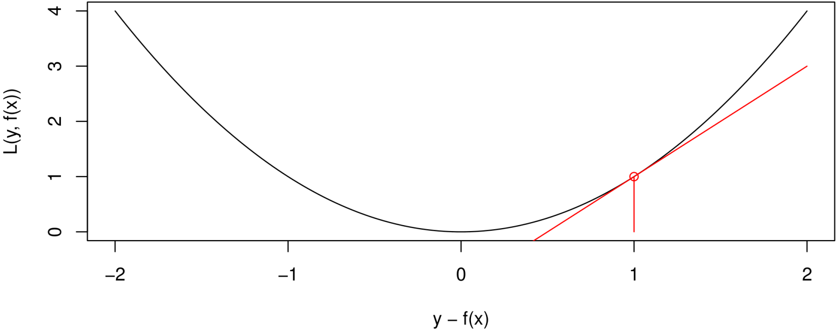
- A loss is called **distance-based** if
  - it can be written in terms of the residual

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

  - $\psi(\epsilon) = 0 \Leftrightarrow \epsilon = 0$.
- 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 $L\left(y, f(\mathbf{x})\right) = L\left(f(\mathbf{x}), y\right)$.

VISULIZING LOSSES VIA LOSS PLOTS

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



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.