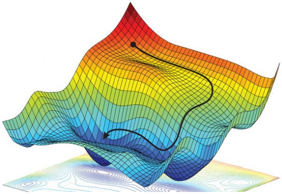
Introduction to Machine Learning

Maximum Likelihood Estimation vs. Empirical Risk Minimization



Learning goals

- Learn the correspondence of between a Laplacian error distributions and the L1 loss
- Learn that there is no error distribution for the Huber loss
- Learn the correpondence between Bernoulli-distributed targets and the Bernoulli loss

LAPLACE ERRORS - L1-LOSS

Let us assume that errors are Laplacian, i.e. ϵ follows a Laplace distribution which has the density

$$\frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right), \sigma > 0.$$

Then

$$y = f_{\text{true}}(\mathbf{x}) + \epsilon$$

follows a Laplace distribution with mean $f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)$ and scale parameter σ.

LAPLACE ERRORS - L1-LOSS

The likelihood is then

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p\left(y^{(i)} \,\middle|\, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right), \sigma\right) \\
&\propto \exp\left(-\frac{1}{\sigma}\sum_{i=1}^{n}\left|y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right|\right).
\end{aligned}
$$

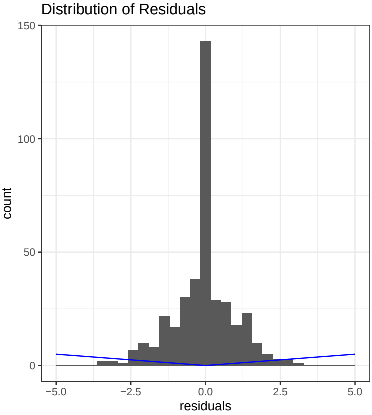
The negative log-likelihood is

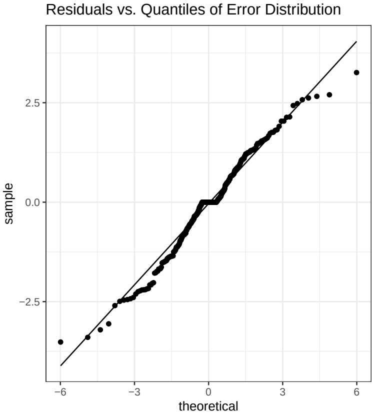
$$
-\ell(\boldsymbol{\theta}) \propto -\sum_{i=1}^{n}\left|y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right|.
$$

Minimizing the negative log-likelihood for Laplacian error terms corresponds to empirical risk minimization with L1-loss.

LAPLACE ERRORS - L1-LOSS

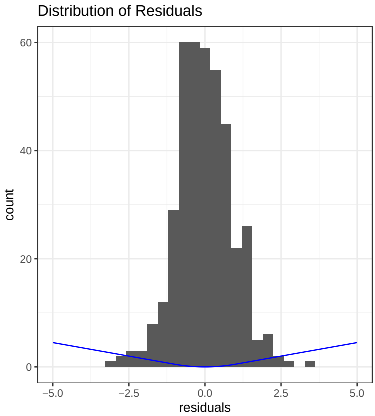
- Distribution of empirical residuals and their comparison to the theoretical quantiles of a Laplace-distribution.

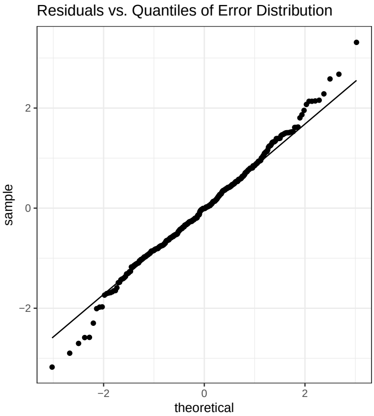




OTHER ERROR DISTRIBUTIONS

- There are losses that do not correspond to “real” error densities, like the Huber loss. (In the QQ-plot below we show residuals against quantiles of a normal.)





OTHER ERROR DISTRIBUTIONS

However, intuitively, we see that a certain type of loss function corresponds to a certain error distribution.

Loss function	Error Distribution
$L2$-Loss	Gaussian Errors
$L1$-Loss	Laplace Errors
Huber Loss	“Huber Errors”

MAXIMUM LIKELIHOOD IN CLASSIFICATION

Let us assume the outputs $y^{(i)}$ to be Bernoulli-distributed, i.e.

$$y^{(i)} \sim \text{Ber}(\pi(\mathbf{x}))$$

with probability $\pi(\mathbf{x})$ that depends on $\mathbf{x}$.

The maximization of the negative log-likelihood is based on

$$\begin{aligned} -\ell(\boldsymbol{\theta}) &= -\sum_{i=1}^{n} \log p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right) \\ &= \sum_{i=1}^{n} -y^{(i)} \log[\pi\left(\mathbf{x}^{(i)}\right)] - \left(1 - y^{(i)}\right) \log[1 - \pi\left(\mathbf{x}^{(i)}\right)]. \end{aligned}$$

MAXIMUM LIKELIHOOD IN CLASSIFICATION

This gives rise to the following loss function

$$L(y, \pi(\mathbf{x})) = -y \ln(\pi(\mathbf{x})) - (1 - y) \ln(1 - \pi(\mathbf{x})), \quad y \in \{0, 1\}$$

which we introduced as **Bernoulli** loss.