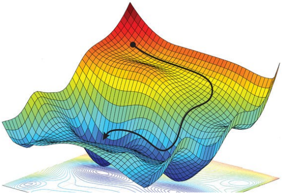
Introduction to Machine Learning

Maximum Likelihood Estimation vs. Empirical Risk Minimization



Learning goals

- Understand the connection between Maximum Likelihood and Risk Minimization
- Learn the correspondence of loss functions and distributions

Regression

MAXIMUM LIKELIHOOD

Let us approach regression from a maximum likelihood perspective.

We assume that

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

where f_{true} is a function that is parameterized by $\boldsymbol{\theta}$ with ϵ being a random variable that follows some distribution $\mathbb{P}_\epsilon$, with $\mathbb{E}[\epsilon] = 0$. Further, we assume ϵ to be independent of $\mathbf{x}$.

It follows that

- $y \mid \mathbf{x}$ follows a distribution with mean $f_{\text{true}}(\mathbf{x})$ and variance $\mathrm{Var}(\epsilon)$.
- We denote the corresponding density function by $p(y \mid \mathbf{x}, \boldsymbol{\theta})$.

MAXIMUM LIKELIHOOD

- Given data

$$\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)} \right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)} \right) \right)$$

the maximum-likelihood principle is to maximize the **likelihood**

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} p \left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta} \right)$$

or to minimize the **negative log-likelihood**:

$$-\ell(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log p \left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta} \right)$$

MAXIMUM LIKELIHOOD

- Let us now simply define the negative log-likelihood as **loss function**

$$L(y, f(\mathbf{x} \mid \boldsymbol{\theta})) := -\log p(y \mid \mathbf{x}, \boldsymbol{\theta})$$

- Maximum-likelihood optimization can be formulated as an empirical risk minimization problem

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

- We can even disregard multiplicative or additive constants in the loss as they do not change the minimizer.

MAXIMUM LIKELIHOOD

- For every error distribution $\mathbb{P}_\epsilon$ we can derive an equivalent loss function, which leads to the same point estimator for the parameter vector $\boldsymbol{\theta}$ as maximum-likelihood.

- **NB**: The other way around does not always work: We cannot derive a probability density function or error distribution corresponding to every loss function – the Hinge loss is a prominent example.

GAUSSIAN ERRORS - L2-LOSS

Let us assume that errors are Gaussian, i.e. $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$. Then

$$y = f_{\text{true}}(\mathbf{x}) + \epsilon \sim N\left(f_{\text{true}}(\mathbf{x}), \sigma^2\right).$$

The likelihood is then

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p\left(y^{(i)} \,\middle|\, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right), \sigma^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2\right).
\end{aligned}$$

GAUSSIAN ERRORS - L2-LOSS

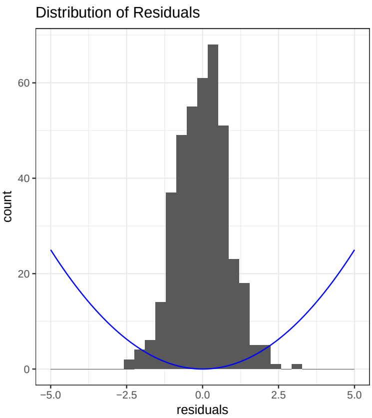
It is easy to see that minimizing the negative log-likelihood is equivalent to the *L*2-loss minimization approach since

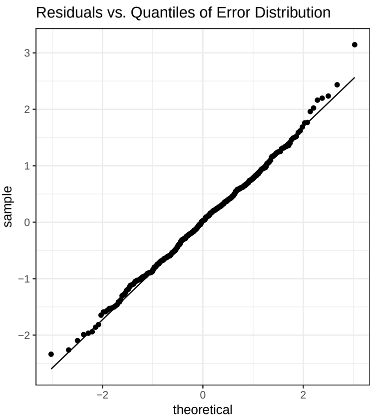
$$
\begin{aligned}
-\ell(\boldsymbol{\theta}) &= -\log\left(\mathcal{L}(\boldsymbol{\theta})\right) \\
&= -\log\left(\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2\right)\right) \\
&\propto \sum_{i=1}^{n}\left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2.
\end{aligned}
$$

Note: We use $\propto$ as “proportional to ... up to multiplicative and additive constants”.

GAUSSIAN ERRORS - L2-LOSS

- We can plot the "empirical" error distribution, i.e. the distribution of the residuals after fitting a model w.r.t. $L2$-loss.
- With the help of a Q-Q-plot we can compare the empirical residuals vs. the theoretical quantiles of a Gaussian distribution.





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

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

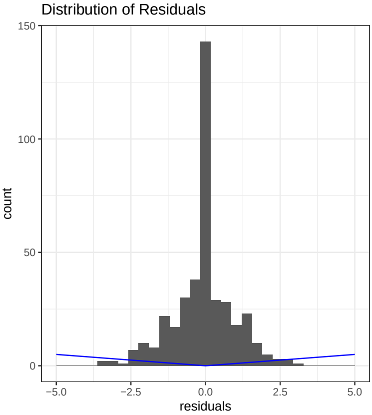
The negative log-likelihood is

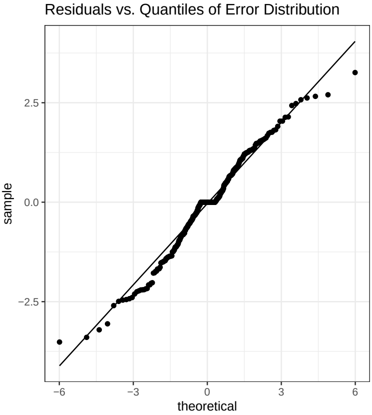
$$
-\ell(\boldsymbol{\theta}) \propto -\sum_{i=1}^{n}\left|y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right|.
$$

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

LAPLACE ERRORS - L1-LOSS

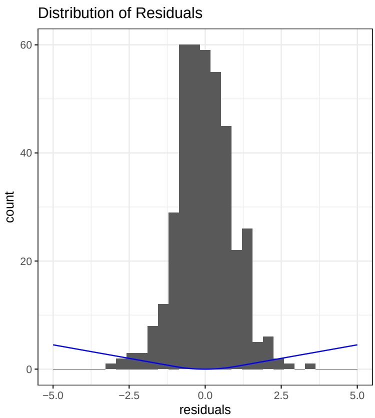
- Distribution of empirical residuals and their comparison to the theoretical quantiles of a Laplace-distribution.

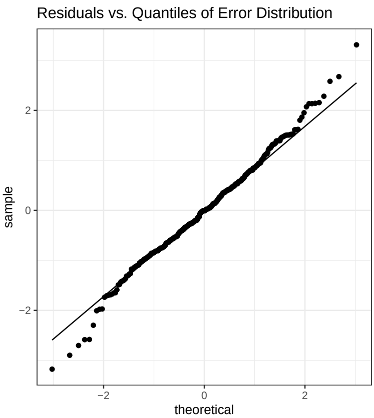




OTHER ERROR DISTRIBUTIONS

- 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"

Classification

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

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

which we introduced as **Bernoulli** loss.