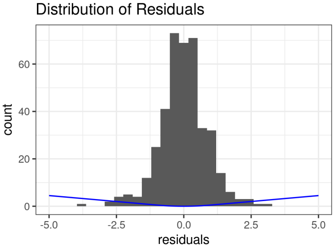
Introduction to Machine Learning

Maximum Likelihood Estimation vs. Empirical Risk Minimization



Learning goals

- Understand the connection between maximum likelihood and risk minimization
- Learn the correspondence of between a Gaussian error distribution and the L2 loss

MAXIMUM LIKELIHOOD

Let us approach regression from a maximum likelihood perspective. In the following we assume that

$$y \mid \mathbf{x} \sim p(y \mid \mathbf{x}, \boldsymbol{\theta}),$$

For example, we commonly consider a true underlying relationship f_{true} with additive noise:

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

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

MAXIMUM LIKELIHOOD

From a statistics / maximum-likelihood perspective, we assume (or we pretend) we know the underlying distribution $p(y \mid \mathbf{x}, \boldsymbol{\theta})$.

- Then, 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 take a machine learning perspective and assume our hypothesis space corresponds to the space of the (parameterized) f_{true}.

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

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

- 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 $y = f_{\text{true}}(\mathbf{x}) + \epsilon$ with additive Gaussian errors are Gaussian, i.e. $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$. Then

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

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 simulate data $y \mid \mathbf{x} \sim \mathcal{N}\left(f_{\text{true}}(\mathbf{x}), 1\right)$ with $f_{\text{true}} = 0.2 \cdot \mathbf{x}$.
- We can plot the empirical error distribution, i.e. the distribution of the residuals after fitting a regression 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.