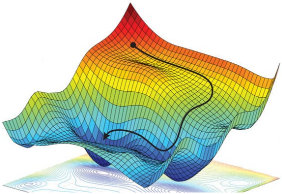
Introduction to Machine Learning

Theoretical Considerations on Regression Losses



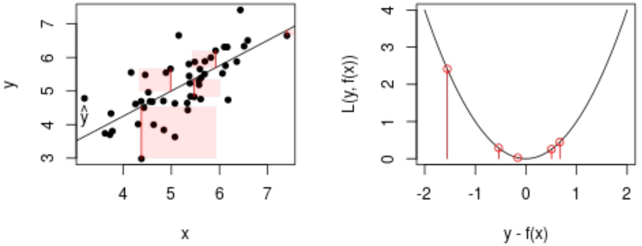
Learning goals

- Understand that an ML model is simply a parametrized curve
- Understand that the hypothesis space lists all admissible models for a learner
- Understand the relationship between the hypothesis space and the parameter space

L2-LOSS

$$L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2 \quad \text{or} \quad L(y, f(\mathbf{x})) = 0.5\,(y - f(\mathbf{x}))^2$$

- Tries to reduce large residuals (if residual is twice as large, loss is 4 times as large), hence outliers in y can become problematic
- Analytic properties: convex, differentiable (gradient no problem in loss minimization)
- Residuals = Pseudo-residuals: $\tilde{r} = -\frac{\partial 0.5(y-f(\mathbf{x}))^2}{\partial f(\mathbf{x})} = y - f(\mathbf{x}) = r$



L2-LOSS: POINT-WISE OPTIMUM

Let us consider the (theoretical) risk for $\mathcal{Y} = \mathbb{R}$ and the $L2$-Loss $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$.

- By the law of total expectation

$$\begin{aligned} \mathcal{R}(f) &= \mathbb{E}_{xy}\left[L(y, f(\mathbf{x}))\right] \\ &= \mathbb{E}_x\left[\mathbb{E}_{y|x}\left[L(y, f(\mathbf{x})) \mid \mathbf{x} = \mathbf{x}\right]\right] \\ &= \mathbb{E}_x\left[\mathbb{E}_{y|x}\left[(y - f(\mathbf{x}))^2 \mid \mathbf{x} = \mathbf{x}\right]\right]. \end{aligned}$$

- Assume we are free to choose f as we wish: At any point $\mathbf{x} = \mathbf{x}$ we can predict any c we want. The best point-wise prediction is the conditional mean

$$\hat{f}(\mathbf{x}) = \text{argmin}_c \mathbb{E}_{y|x}\left[(y - c)^2 \mid \mathbf{x} = \mathbf{x}\right] \overset{(*)}{=} \mathbb{E}_{y|x}\left[y \mid \mathbf{x}\right].$$

L2-LOSS: POINT-WISE OPTIMUM

- $^{(*)}$ follows from:

$$
\begin{aligned}
& \text{argmin}_c \mathbb{E}\left[(y-c)^2\right] \\
= \quad & \text{argmin}_c \underbrace{\mathbb{E}\left[(y-c)^2\right] - (\mathbb{E}[y]-c)^2}_{\text{Var}[y-c]=\text{Var}[y]} + (\mathbb{E}[y]-c)^2 \\
= \quad & \text{argmin}_c \text{Var}[y] + (\mathbb{E}[y]-c)^2 \\
= \quad & \mathbb{E}[y].
\end{aligned}
$$

L2-LOSS: OPTIMAL CONSTANT MODEL

For the sake of simplicity, let us consider the hypothesis space $\mathcal{H}$ of constant models

$$\mathcal{H} = \{f \mid f(\mathbf{x}) = \boldsymbol{\theta}, \boldsymbol{\theta} \in \mathbb{R}\}.$$

Goal: Derive the optimal constant model w.r.t. the $L2$-Loss.

$$\begin{aligned} f &= \underset{f \in \mathcal{H}}{\arg\min}\, \mathcal{R}_{\text{emp}}(f) = \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right) \\ \Leftrightarrow \quad \hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta} \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}\right)^2 \end{aligned}$$

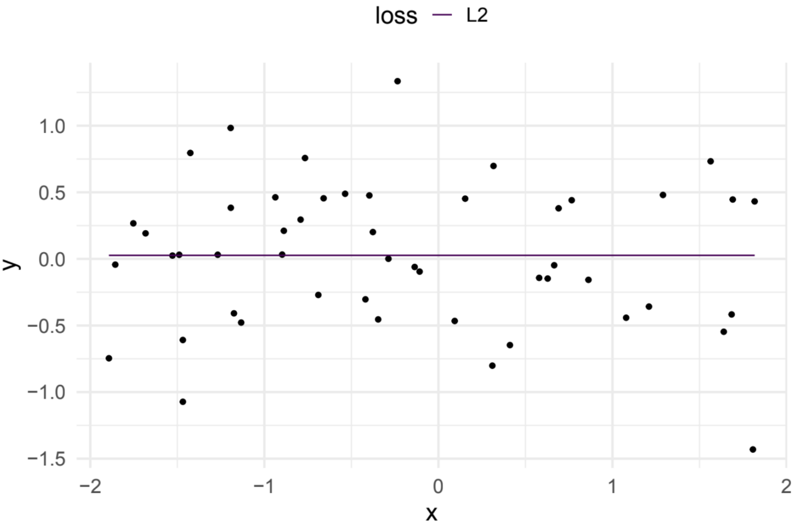
L2-LOSS: OPTIMAL CONSTANT MODEL

We calculate the first derivative of $\mathcal{R}_{\text{emp}}$ w.r.t. $\boldsymbol{\theta}$ and set it to 0:

$$
\begin{aligned}
\frac{\partial \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 2 \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta} \right) &\overset{!}{=} 0 \\
\sum_{i=1}^{n} y^{(i)} - n\boldsymbol{\theta} &= 0 \\
\hat{\boldsymbol{\theta}} &= \frac{1}{n} \sum_{i=1}^{n} y^{(i)} =: \bar{y}.
\end{aligned}
$$

So the optimal constant model predicts the average of observed outcomes $\hat{f}(\mathbf{x}) = \bar{y}$.

L2-LOSS: OPTIMAL CONSTANT MODEL

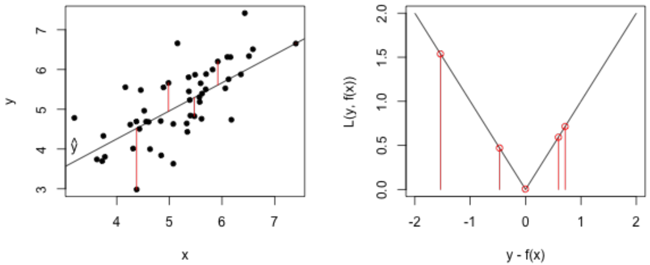


L1-Loss

L1-LOSS

$$L(y, f(\mathbf{x})) = |y - f(\mathbf{x})|$$

- More robust than $L2$, outliers in y are less problematic.
- Analytical properties: convex, not differentiable for $y = f(\boldsymbol{x})$ (optimization becomes harder).



L1-LOSS: POINT-WISE OPTIMUM

We calculate the (theoretical) risk for the $L1$-Loss $L(y, f(\mathbf{x})) = |y - f(\mathbf{x})|$ with unrestricted $\mathcal{H} = \{f : \mathcal{X} \to \mathcal{Y}\}$.

- Again, we use the law of total expectation

$$\mathcal{R}(f) = \mathbb{E}_x \left[\mathbb{E}_{y|x} \left[|y - f(\mathbf{x})| \; |\mathbf{x} = \mathbf{x}\right]\right].$$

- As the functional form of f is not restricted, we can just optimize point-wise at any point $\mathbf{x} = \mathbf{x}$. The best prediction at $\mathbf{x} = \mathbf{x}$ is then

$$\hat{f}(\mathbf{x}) = \text{argmin}_c \mathbb{E}_{y|x} \left[|y - c|\right] \overset{(*)}{=} \text{med}_{y|x} \left[y \mid \mathbf{x}\right].$$

L1-LOSS: POINT-WISE OPTIMUM

- $^{(*)}$ Let $p(y)$ be the density function of y. Then:

$$
\begin{aligned}
\text{argmin}_c \mathbb{E}\left[|y - c|\right] &= \text{argmin}_c \int_{-\infty}^{\infty} |y - c|\; p(y) \mathrm{d}y \\
= \quad & \text{argmin}_c \int_{-\infty}^{c} -(y - c)\; p(y)\; \mathrm{d}y + \int_{c}^{\infty} (y - c)\; p(y)\; \mathrm{d}y
\end{aligned}
$$

Setting the derivation w.r.t. c to zero yields:

$$
\begin{aligned}
0 &= \int_{-\infty}^{c} p(y)\; \mathrm{d}y - \int_{c}^{\infty} p(y)\; \mathrm{d}y \\
&= \mathbb{P}_y(y \leq c) - (1 - \mathbb{P}_y(y \leq c)) \\
&= 2 \cdot \mathbb{P}_y(y \leq c) - 1 \\
\Leftrightarrow 0.5 &= \mathbb{P}_y(y \leq c),
\end{aligned}
$$

which yields $c = \text{med}_y(y)$.

L1-LOSS: OPTIMAL CONSTANT MODEL

Goal: Derive the optimal constant model

$$f \in \mathcal{H} = \left\{ f(\mathbf{x}) = \boldsymbol{\theta} \mid \boldsymbol{\theta} \in \mathbb{R} \right\},$$

w.r.t. the $L1$-Loss.

$$\begin{aligned} f &= \underset{f \in \mathcal{H}}{\arg\min}\, \mathcal{R}_{\text{emp}}(f) \\ \Leftrightarrow \quad \hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta} \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} \left| y^{(i)} - \boldsymbol{\theta} \right| \\ \Leftrightarrow \quad \hat{\boldsymbol{\theta}} &= \text{med}(y^{(i)}) \end{aligned}$$

L1-LOSS: OPTIMAL CONSTANT MODEL

Proof:

- Firstly note that for $n = 1$ the median $\hat{\theta} = \text{med}(y^{(i)}) = y^{(1)}$ obviously minimizes the empirical risk $\mathcal{R}_{\text{emp}}$ associated to the $L1$ loss L.
- Hence let $n > 1$ in the following: Let

$$S_{a,b} : \mathbb{R} \to \mathbb{R}_0^+, \boldsymbol{\theta} \mapsto |a - \boldsymbol{\theta}| + |b - \boldsymbol{\theta}|$$

for $a, b \in \mathbb{R}$. It holds that

$$S_{a,b}(\boldsymbol{\theta}) = \begin{cases} |a - b|, & \text{for } \boldsymbol{\theta} \in [a, b] \\ |a - b| + 2 \cdot \min\{|a - \boldsymbol{\theta}|, |b - \boldsymbol{\theta}|\}, & \text{otherwise.} \end{cases}$$

Thus, any $\hat{\boldsymbol{\theta}} \in [a, b]$ minimizes $S_{a,b}$.

L1-LOSS: OPTIMAL CONSTANT MODEL

Let us define $i_{\max} = n/2$ for n even and $i_{\max} = (n-1)/2$ for n odd and consider the intervals

$$\mathcal{I}_i := [y^{(i)}, y^{(n+1-i)}], i \in \{1, ..., i_{\max}\}.$$

By construction $\mathcal{I}_{j+1} \subseteq \mathcal{I}_j$ for $j \in \{1, \ldots, i_{\max} - 1\}$ and $\mathcal{I}_{i_{\max}} \subseteq \mathcal{I}_i$. With this, $\mathcal{R}_{\text{emp}}$ can be expressed as

$$\begin{aligned}
\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} L(y^{(i)}, \boldsymbol{\theta}) = \sum_{i=1}^{n} |y^{(i)} - \boldsymbol{\theta}| \\
&= \underbrace{|y^{(1)} - \boldsymbol{\theta}| + |y^{(n)} - \boldsymbol{\theta}|}_{= S_{y^{(1)}, y^{(n)}}(\boldsymbol{\theta})} + \underbrace{|y^{(2)} - \boldsymbol{\theta}| + |y^{(n-1)} - \boldsymbol{\theta}|}_{= S_{y^{(2)}, y^{(n-1)}}(\boldsymbol{\theta})} + ... \\
&= \begin{cases} \sum\limits_{i=1}^{i_{\max}} S_{y^{(i)}, y^{(n+1-i)}}(\boldsymbol{\theta}) & \text{for } n \text{ is even} \\ \sum\limits_{i=1}^{i_{\max}} \left(S_{y^{(i)}, y^{(n+1-i)}}(\boldsymbol{\theta})\right) + |y^{(n+1)} - \boldsymbol{\theta}| & \text{for } n \text{ is odd.} \end{cases}
\end{aligned}$$

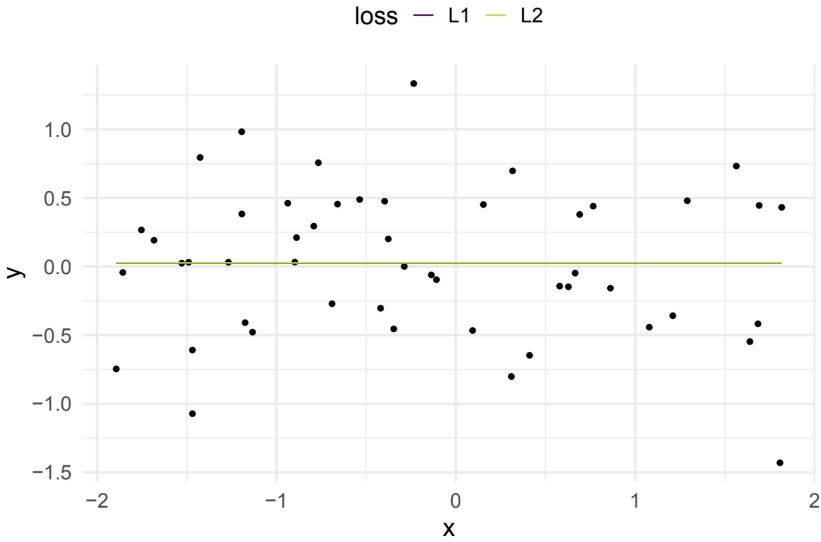
L1-LOSS: OPTIMAL CONSTANT MODEL

From this follows that

- for "n is even": $\hat{\boldsymbol{\theta}} \in \mathcal{I}_{i_{\max}} = [y^{(n/2)}, y^{(n/2+1)}]$ minimizes S_i for all $i \in \{1, \dots, i_{\max}\} \;\Rightarrow \mathcal{R}_{\text{emp}}$ reaches its global minimum at $\hat{\boldsymbol{\theta}}$,
- for "n is odd": $\hat{\boldsymbol{\theta}} = y^{(n+1)/2} \in \mathcal{I}_{i_{\max}}$ minimizes S_i for all $i \in \{1, \dots, i_{\max}\} \;\Rightarrow \mathcal{R}_{\text{emp}}$ reaches its global minimum at $\hat{\boldsymbol{\theta}}$.

Since the median fulfills these conditions, we can conclude that it minimizes the $L1$ loss.

L1-LOSS: OPTIMAL CONSTANT MODEL



L1-LOSS: OPTIMAL CONSTANT MODEL

We see that the $L1$-Loss is more robust w.r.t. outliers than the $L2$-Loss.