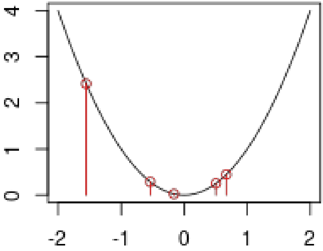
Introduction to Machine Learning

Regression Losses: L2-loss



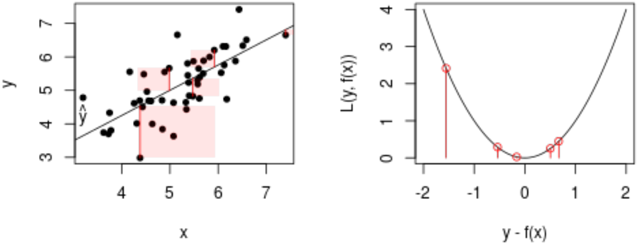
Learning goals

- Derive the risk minimizer of the L2-loss
- Derive the optimal constant model for the L2-loss

L2-LOSS

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

- 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: RISK MINIMIZER

Let us consider the (true) risk for $\mathcal{Y} = \mathbb{R}$ and the *L*2-Loss $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$ with unrestricted $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R}^g\}$.

- By the law of total expectation

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

- Since $\mathcal{H}$ is unrestricted we can choose f as we wish: At any point $\mathbf{x} = \mathbf{x}$ we can predict any value 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: RISK MINIMIZER

$^{(*)}$ follows from:

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

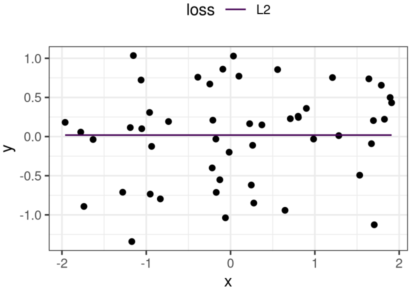
$$= \quad \text{argmin}_c \text{Var}[y] + (\mathbb{E}[y]-c)^2 = \mathbb{E}[y].$$

L2-LOSS: OPTIMAL CONSTANT MODEL

The optimal constant model in terms of the (theoretical) risk for the L2 loss is the expected value over y:

$$f(\mathbf{x}) \quad = \quad \mathbb{E}_{y \mid \mathbf{x}}\left[y \mid \mathbf{x}\right] \overset{\text{drop } \mathbf{x}}{=} \mathbb{E}_y\left[y\right]$$

The optimizer of the empirical risk is $\bar{y}$ (the empirical mean over $y^{(i)}$), which is the empirical estimate for $\mathbb{E}_y\left[y\right]$.



L2-LOSS: OPTIMAL CONSTANT MODEL

Proof:

For the optimal constant model $f(\mathbf{x}) = \theta$ for the L2-loss $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$ we solve the optimization problem

$$\arg\min_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f) = \arg\min_{c \in \mathbb{R}} \sum_{i=1}^{n} (y^{(i)} - \theta)^2.$$

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

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

L2 LOSS MEANS MINIMIZING VARIANCE

Rethinking what we just did: We optimized for the constant, whose squared distance to all data points is minimal (in sum, or on average). This turned out to be the mean.

What happens if we calculcate the incurred loss of $\hat{\theta} = \bar{y}$

Thats obviously $\mathcal{R}_{\text{emp}} = \sum_{i=1}^{n} (y^{(i)} - \bar{y})^2$.

Average this sum by $\frac{1}{n}$ or $\frac{1}{n-1}$, and we get the empirical variance.

The same holds true for the pointwise construction / conditional distribution as considered in the slides before.