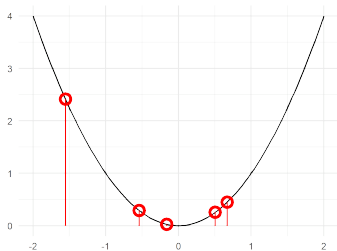
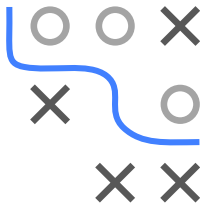


Introduction to Machine Learning

Advanced Risk Minimization

Regression Losses: L2 and L1 loss



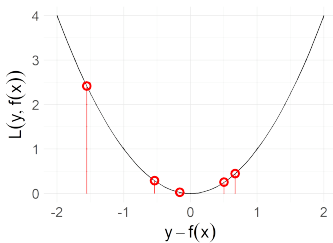
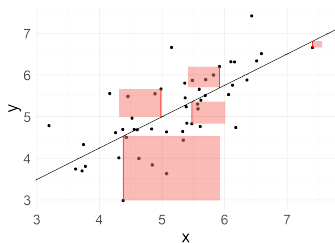
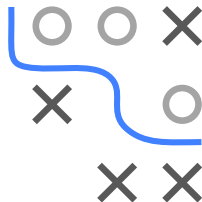
Learning goals

- Derive the risk minimizer of the L2-loss
- Derive the optimal constant model for the L2-loss
- Know risk minimizer and optimal constant model for L1-loss

L2-LOSS

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

- 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 \Rightarrow gradient no problem in loss minimization
- (**Warning:** $\mathcal{R}_{\text{emp}}(f)$ can still be non-smooth/non-convex due to $f(\mathbf{x})$)

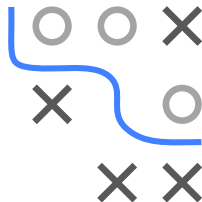


L2-LOSS: OPTIMAL CONSTANT MODEL

Let us consider the (true) risk for $\mathcal{Y} = \mathbb{R}$ and L2-Loss

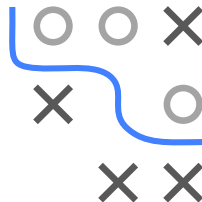
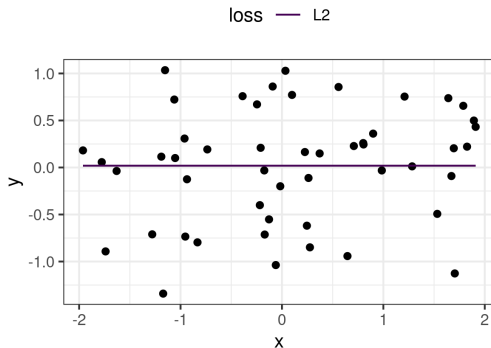
$L(y, f) = (y - f)^2$ with \mathcal{H} restricted to constants. The optimal constant model f_c^* in terms of the theoretical risk is the expected value over y :

$$\begin{aligned} f_c^* &= \arg \min_{c \in \mathbb{R}} \mathbb{E}_{xy} [(y - c)^2] = \arg \min_{c \in \mathbb{R}} \mathbb{E}_y [(y - c)^2] \\ &= \arg \min_{c \in \mathbb{R}} \underbrace{\mathbb{E}_y [(y - c)^2] - (\mathbb{E}_y[y] - c)^2}_{=\text{Var}_y[y-c]=\text{Var}_y[y]} + (\mathbb{E}_y[y] - c)^2 \\ &= \arg \min_{c \in \mathbb{R}} \text{Var}_y[y] + (\mathbb{E}_y[y] - c)^2 \\ &= \mathbb{E}_y[y] \end{aligned}$$



L2-LOSS: OPTIMAL CONSTANT MODEL / 2

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



L2-LOSS: OPTIMAL CONSTANT MODEL / 3

Proof:

For the optimal constant model f_c^* for the L2-loss $L(y, f) = (y - f)^2$ we solve the optimization problem

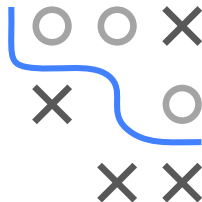
$$\arg \min_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f) = \arg \min_{\theta \in \mathbb{R}} \sum_{i=1}^n (y^{(i)} - \theta)^2.$$

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

$$\frac{\partial \mathcal{R}_{\text{emp}}(\theta)}{\partial \theta} = -2 \sum_{i=1}^n (y^{(i)} - \theta) \stackrel{!}{=} 0$$

$$\sum_{i=1}^n y^{(i)} - n\theta = 0$$

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n y^{(i)} =: \bar{y}.$$



L2-LOSS: RISK MINIMIZER

Let us consider the (true) risk for $\mathcal{Y} = \mathbb{R}$ and the $L2$ -Loss $L(y, f) = (y - f)^2$ with unrestricted $\mathcal{H} = \{f : \mathcal{X} \rightarrow \mathbb{R}^g\}$.

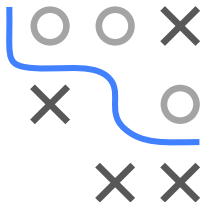
- By the law of total expectation

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

- Since \mathcal{H} is unrestricted, at any point $\mathbf{x} = \mathbf{x}$, we can predict any value c we want. The best point-wise prediction is the cond. mean

$$f^*(\mathbf{x}) = \arg \min_c \mathbb{E}_{y|\mathbf{x}} [(y - c)^2 \mid \mathbf{x} = \mathbf{x}] \stackrel{(*)}{=} \mathbb{E}_{y|\mathbf{x}} [y \mid \mathbf{x}] .$$

^(*) follows from the derivation of f_c^*

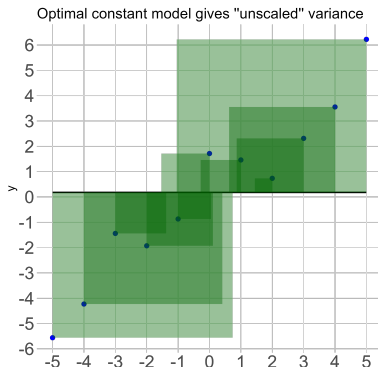


L2 LOSS MEANS MINIMIZING VARIANCE

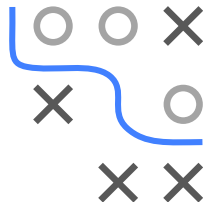
Rethinking what we did in the opt. constant model: We optimized for a constant whose squared distance to all data points is minimal (in sum, or on average). This turned out to be the mean.

What if we calculate the loss of $\hat{\theta} = \bar{y}$? That's $\mathcal{R}_{\text{emp}} = \sum_{i=1}^n (y^{(i)} - \bar{y})^2$.

Average this by $\frac{1}{n}$ or $\frac{1}{n-1}$ to obtain variance.



- Generally, if model yields unbiased predictions, $\mathbb{E}_{y | \mathbf{x}} [y - f(\mathbf{x}) | \mathbf{x}] = 0$, using L_2 -loss means minimizing variance of model residuals
- Same holds for the pointwise construction / conditional distribution considered before

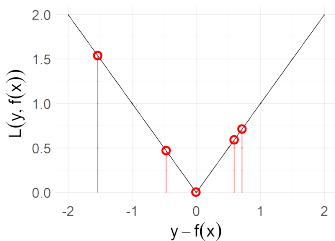
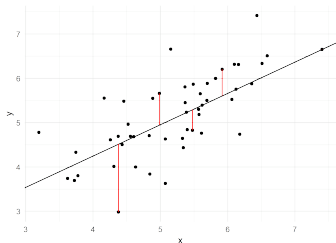
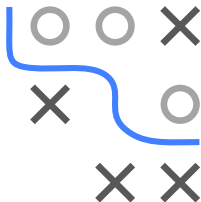


L1-LOSS

The L1 loss is defined as

$$L(y, f) = |y - f|$$

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



L1-LOSS: RISK MINIMIZER

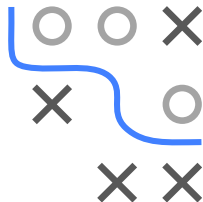
We calculate the (true) risk for the L1-Loss $L(y, f) = |y - f|$ with unrestricted $\mathcal{H} = \{f : \mathcal{X} \rightarrow \mathcal{Y}\}$.

- We use the law of total expectation

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

- 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

$$f^*(\mathbf{x}) = \arg \min_c \mathbb{E}_{y|\mathbf{x}} [|y - c|] = \text{med}_{y|\mathbf{x}} [y \mid \mathbf{x}] .$$



L1-LOSS: OPTIMAL CONSTANT MODEL

The optimal constant model in terms of the theoretical risk for the L1 loss is the median over y :

$$f_c^* = \text{med}_{y|x} [y | \mathbf{x}] \stackrel{\text{drop } \mathbf{x}}{=} \text{med}_y [y]$$

The optimizer \hat{f}_c of the empirical risk is $\text{med}(y^{(i)})$ over $y^{(i)}$, which is the empirical estimate for $\text{med}_y [y]$.

