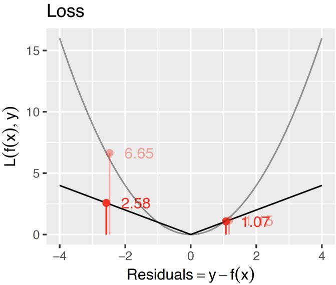
Introduction to Machine Learning

Loss Functions for Regression

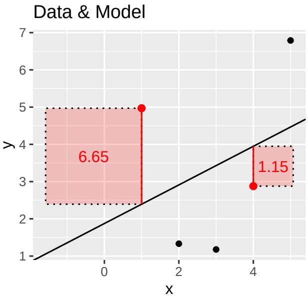


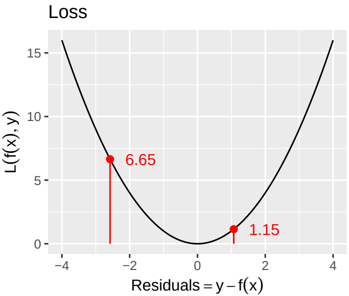
Learning goals

- Know definitions of L1 and L2 loss
- Understand difference between L1 and L2 loss
- Understand why optimization for L1 loss is harder than for L2 loss

REGRESSION LOSSES - L2 / SQUARED ERROR

- $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$ or $L(y, f(\mathbf{x})) = 0.5(y - f(\mathbf{x}))^2$
- Convex
- Differentiable, gradient no problem in loss minimization
- For later: $\frac{\partial 0.5(y - f(\mathbf{x}))^2}{\partial f(\mathbf{x})} = y - f(\mathbf{x}) = \epsilon$, derivative is residual
- Tries to reduce large residuals (if residual is twice as large, loss is 4 times as large), hence outliers in y can become problematic
- Connection to Gaussian distribution (see later)





REGRESSION LOSSES - L2 / SQUARED ERROR

What's the optimal constant prediction c (i.e. the same $\hat{y}$ for all $\mathbf{x}$)?

$$L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2 = (y - c)^2$$

We search for the c that minimizes the empirical risk.

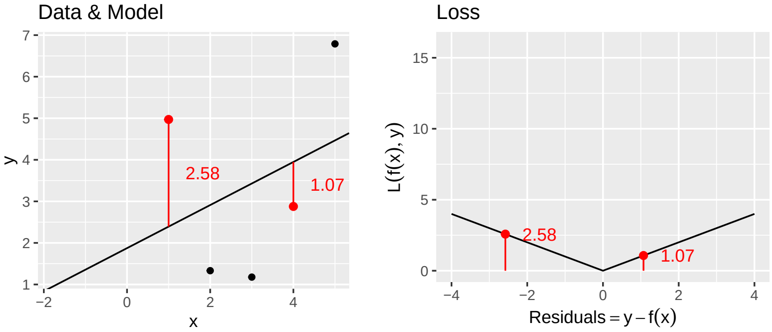
$$\hat{c} = \arg\min_{c \in \mathbb{R}} \mathcal{R}_{\text{emp}}(c) = \arg\min_{c \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - c)^2$$

We set the derivative of the empirical risk to zero and solve for c:

$$\begin{aligned} -\frac{1}{n} \sum_{i=1}^{n} 2(y^{(i)} - c) &= 0 \\ \hat{c} &= \frac{1}{n} \sum_{i=1}^{n} y^{(i)} \end{aligned}$$

REGRESSION LOSSES - L1 / ABSOLUTE ERROR

- $L(y, f(\mathbf{x})) = |y - f(\mathbf{x})|$
- Convex
- No derivatives for $L(y, f(\mathbf{x})) = 0$, optimization becomes harder
- $\hat{f}(\mathbf{x}) = \text{median of } y|\mathbf{x}$



REGRESSION LOSSES - L1 / ABSOLUTE ERROR

- $L(y, f(\mathbf{x})) = |y - f(\mathbf{x})|$
- Convex
- No derivatives for $\epsilon = 0$, $y = f(\mathbf{x})$, optimization becomes harder
- $\hat{f}(\mathbf{x}) =$ median of $y|\mathbf{x}$
- More robust, outliers in y are less influential than for L2