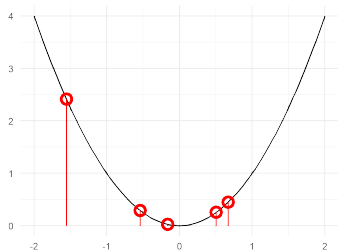


Introduction to Machine Learning

Advanced Risk Minimization

Regression Losses: L2 and L1 loss



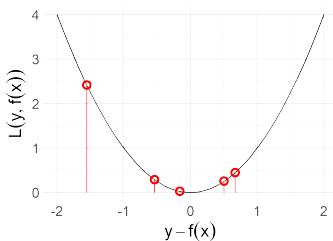
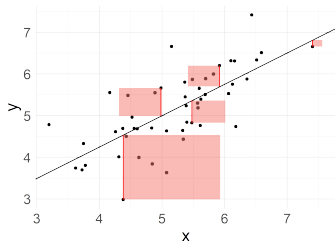
Learning goals

- L2 loss and risk minimizers
- L1 loss and risk minimizers

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, sensitive to outliers in y
- Analytic properties: convex, differentiable



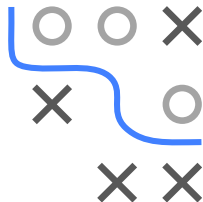
L2: OPTIMAL VALUE IS EXPECTATION

- Can derive a general result now for any $z \sim Q$
- Consider

$$\arg \min_{c \in \mathbb{R}} \mathbb{E}_z[L(z, c)] = \arg \min_{c \in \mathbb{R}} \mathbb{E}[(z - c)^2]$$

$$\mathbb{E}[(z - c)^2] = \mathbb{E}[z^2 - 2zc + c^2] = \mathbb{E}[z^2] - 2c\mathbb{E}[z] + c^2$$

- The RHS is minimized by $c = \mathbb{E}[z]$
(simple quadratic, or take derivative and set to 0)



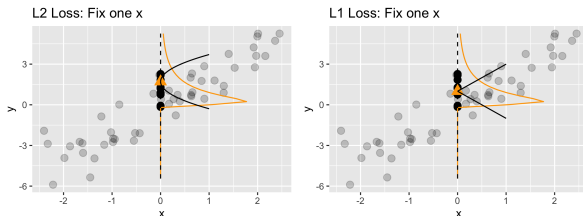
L2-LOSS: RISK MINIMIZER

- Let's minimize true risk for unrestricted hypothesis space and L_2
- We know: At any point $\mathbf{x} = \tilde{\mathbf{x}}$, our loss-optimal prediction is

$$f^*(\tilde{\mathbf{x}}) = \arg \min_{c \in \mathbb{R}} \mathbb{E}_{y|x} [L(y, c) \mid \mathbf{x} = \tilde{\mathbf{x}}]$$

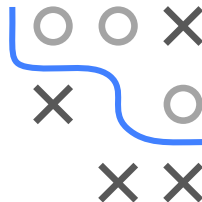
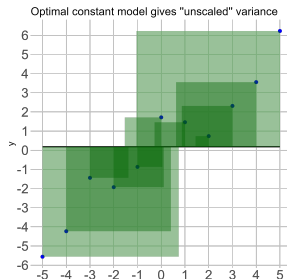
- Again from the previous,
we know the minimizer for L_2 is the conditional expectation

$$f^*(\tilde{\mathbf{x}}) = \mathbb{E}_{y|x} [y \mid \mathbf{x} = \tilde{\mathbf{x}}] .$$



L2 LOSS MEANS MINIMIZING VARIANCE

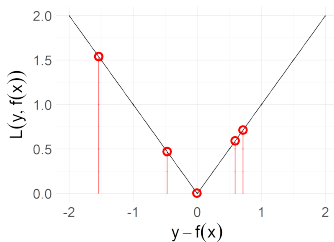
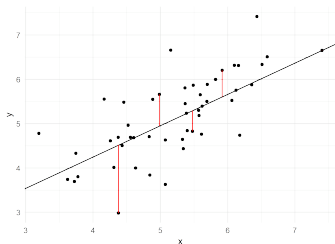
- Let's reconsider the previous
- Optimized for const whose squared dist to points is minimal (on avg)
- Result: $\hat{\theta} = \bar{y}$
- What is the associated risk?
 $\mathcal{R}(\hat{\theta}) = \sum_{i=1}^n (y_i - \bar{y})^2$
- Average this by $\frac{1}{n}$ or $\frac{1}{n-1}$ to obtain variance
- Same holds for the pointwise construction / conditional distribution considered before



L1-LOSS

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

- More robust than L_2 , outliers in y are less problematic
- Analytical properties: convex, not differentiable for $y = f(\mathbf{x})$ (optimization becomes harder)



L1-LOSS: OPTIMAL PREDICTIONS

- Optimal constant model is median: $f_c^* = \text{med}[y]$
- Empirical version: $\hat{f}_c = \text{med}(y^{(1)}, \dots, y^{(n)})$
- Derivations slightly harder and in deep-dive
- Risk minimizer / optimal conditional prediction:

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

