Introduction to Machine Learning

Chapter 4: Loss minimization

Bernd Bischl, Christoph Molnar, Daniel Schalk, Fabian Scheipl

Department of Statistics – LMU Munich

WHY DO WE CARE ABOUT LOSSES?

- Assume we trained a model to predict flat rent based on some features (size, location, age, ...).
- The real rent of a new flat, that the model never saw before, is EUR 1600, our model predicts EUR 1300.
- How do we measure the performance of our model?
- We calculate the prediction error and therefore need a suitable error measure, aka a loss function such as:
  - Absolute loss:
    $L(y = 1600, \hat{y} = 1300) = |1600 - 1300| = 300$
  - Squared loss:
    $L(y = 1600, \hat{y} = 1300) = (1600 - 1300)^2 = 90000$
    (puts more emphasis on predictions that are far off the mark)
- The choice of the loss has a major influence on the final model.

LOSS MINIMIZATION

- The "goodness" of a prediction $f(x)$ is measured by a **loss function** $L(y, f(x))$.
- The ability of a model f to reproduce the association between x and y that is present in the data $\mathcal{D}$ can be measured by the average loss: the **empirical risk**

$$\mathcal{R}_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^{n} L\left(y^{(i)}, f\left(x^{(i)}\right)\right).$$

- Learning then amounts to **empirical risk minimization** – figuring out which model f has the smallest average loss:

$$\hat{f} = \arg\min_{f \in H} \mathcal{R}_{\text{emp}}(f).$$

LOSS MINIMIZATION

Since the model f is usually controlled by **parameters** θ in a parameter space Θ, this becomes:

$$
\begin{aligned}
\mathcal{R}_{\text{emp}}(\theta) &= \frac{1}{n}\sum_{i=1}^{n} L\left(y^{(i)}, f\left(x^{(i)} | \theta\right)\right) \\
\hat{\theta} &= \arg\min_{\theta \in \Theta} \mathcal{R}_{\text{emp}}(\theta)
\end{aligned}
$$

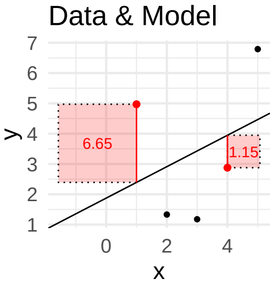
Most learners in ML try to solve the above *optimization problem*, which implies a tight connection between ML and optimization.

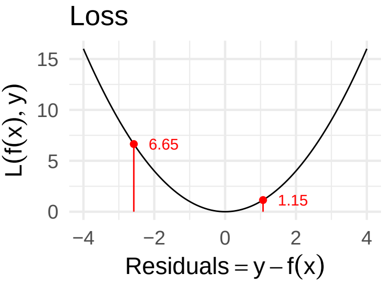
LOSS MINIMIZATION

- For regression tasks, the loss usually only depends on residual $L(y, f(x)) = L(y - f(x)) = L(\epsilon)$.
- Since learning can be re-phrased as minimizing the loss, the choice of loss strongly affects the computational difficulty of learning:
  - smoothness of $\mathcal{R}_{\text{emp}}(\theta)$ in θ
  - can gradient-based methods be applied to minimize $\mathcal{R}_{\text{emp}}(\theta)$?
  - uni- or multimodality $\mathcal{R}_{\text{emp}}(\theta)$ over Θ.
- The choice of loss implies which kinds of errors are important or not – need *domain knowledge*!
- For learners that correspond to probabilistic models, the loss determines / is equivalent to distributional assumptions.

REGRESSION LOSSES - L2 SQUARED LOSS

- $L(y, f(x)) = (y - f(x))^2$ or $L(y, f(x)) = 0.5(y - f(x))^2$
- Convex
- Differentiable, gradient no problem in loss minimization
- For latter: $\frac{\partial 0.5(y-f(x))^2}{\partial f(x)} = y - f(x) = \epsilon$, 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 LOSS

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

$$L(y, f(x)) = (y - f(x))^2 = (y - c)^2$$

We search the c that minimizes the empirical risk.

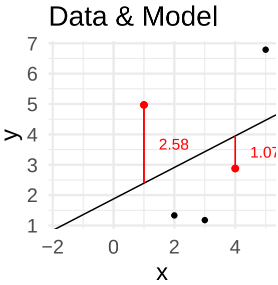
$$\hat{c} = \underset{c \in \mathbb{R}}{\arg\min}\, \mathcal{R}_{\text{emp}}(c) = \underset{c \in \mathbb{R}}{\arg\min}\, \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - c)^2$$

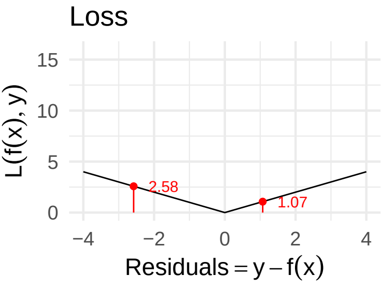
We set the derivative of the empirical risk to zero and solve for c:

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

REGRESSION LOSSES - L1 ABSOLUTE LOSS

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





REGRESSION LOSSES - L1 ABSOLUTE LOSS

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