Squared loss

Squared loss is a loss function that can be used in the learning setting in which we are predicting a real-valued variable y given an input variable x.

That is, we are given the following scenario: let h be a hypothesis (i.e. a statistical model). Let $S := \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ be our training data where $x_i \in \mathcal{X}$ are the instances ($\mathcal{X}$ is the space of possible instances) and $y_i \in \mathbb{R}$ is a numeric value corresponding to each instance. In this setting, the squared loss for a given item in our training data, (y, x), is given by

$$\ell_{\text{squared}}(x, y, h) := (y - h(x))^2$$

(Definition 1).

Definition 1 *Given a set of possible instances $\mathcal{X}$, an instance $x \in \mathcal{X}$, an associated variable $y \in \mathbb{R}$, and a hypothesis function $h : \mathcal{X} \rightarrow \mathbb{R}$, the **squared loss** of h on (x, y) is given by*

$$\ell_{squared}(x, y, h) := (y - h(x))^2$$

.

The empirical risk function over the training data is then the mean of the individual losses:

$$L_S(h) := \frac{1}{|S|} \sum_{i=1}^{|S|} \ell_{\text{squared}}(x_i, y_i, h)$$

. The empirical risk of the squared error is illustrated geometrically in Figure 1. An empirical risk minimization (ERM) algorithm will then seek an h that minimizes the average area of the squares.

Intuition: maximum likelihood estimation under an implicit Gaussian model

Applying an ERM algorithm over a hypothesis space $\mathcal{H}$ using the least squared loss function is equivalent to finding the maximum likelihood estimate under an implicitly assumed probabilistic model: given an item's value of x, it's value of y is determined by adding Gaussian noise to a deterministic function of x. That is, we assume there exists a "true" function $f \in \mathcal{H}$ such that

$$y_i = f(x_i) + \epsilon_i$$

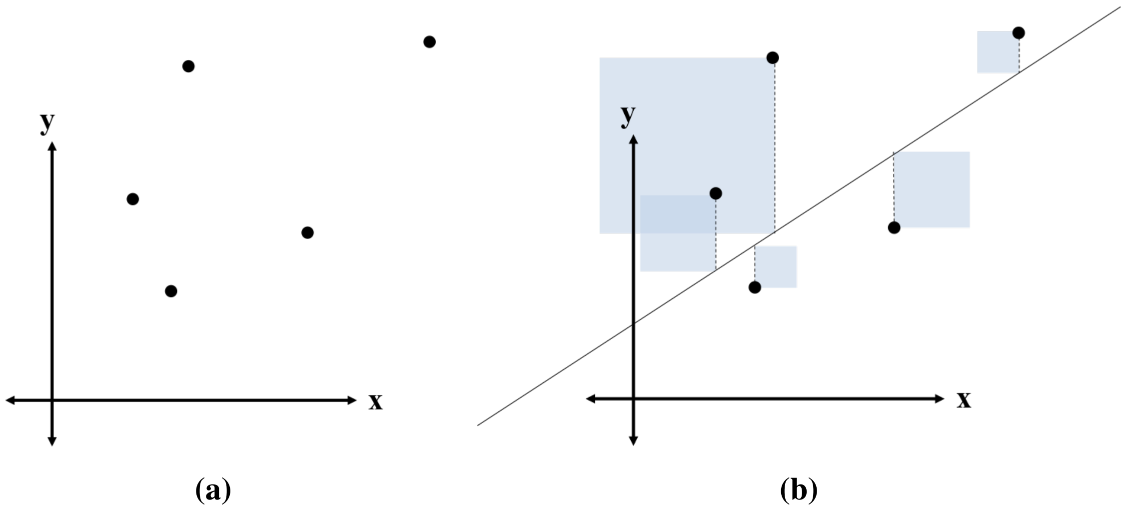

Figure 1: (a) A plot of training set S where $\mathcal{X} := \mathbb{R}$. (b) Fitting the data with a linear hypothesis h. The empirical risk is the average size of the blue squares.

where ϵ_i is Gaussian noise we add to $f(x_i)$. That is,

$$\epsilon_i \sim \text{Normal}(0, \sigma^2)$$

. Stated equivalently, y_i is the outcome of a random variable

$$Y_i \sim \text{Normal}(f(x_i), \sigma^2)$$

. This is proven in Theorem 1.

Theorem 1 *Given a joint distribution over*

$$Y_1, Y_2, \dots, Y_n \mid x_1, x_2, \dots, x_n$$

where

$$Y_i \mid x_i \sim Normal(h(x_i), \sigma^2)$$

and

$$x_i \in \mathcal{X}$$

for a hypothesis $h : \mathcal{X} \to \mathbb{R}$ *in a hypothesis space* $\mathcal{H}$*, the maximum likelihood estimate of* h *over the training data* $S := \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ *(where* y_i *is the realization of* Y_i*) is equal to the ERM estimate using squared loss over* S*.*

Proof:

$$
\begin{aligned}
h_{MLE} &:= \operatorname*{argmax}_{h \in \mathcal{H}} p(S;h) \\
&= \operatorname*{argmax}_{h \in \mathcal{H}} \prod_{i=1}^{|S|} p(y_i, x_i; h) \\
&= \operatorname*{argmax}_{h \in \mathcal{H}} \prod_{i=1}^{|S|} p(y_i \mid x_i; h) p(x_i) \\
&= \operatorname*{argmax}_{h \in \mathcal{H}} \prod_{i=1}^{|S|} p(y_i \mid x_i; h) && h \text{ is only used to explain } y_i \\
&= \operatorname*{argmax}_{h \in \mathcal{H}} \prod_{i=1}^{|S|} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - h(x_i))^2} \\
&= \operatorname*{argmax}_{h \in \mathcal{H}} \sum_{i=1}^{|S|} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - h(x_i))^2} \right) && \log \text{ is monotonic} \\
&= \operatorname*{argmax}_{h \in \mathcal{H}} \sum_{i=1}^{|S|} \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2}(y_i - h(x_i))^2 \right] \\
&= \operatorname*{argmax}_{h \in \mathcal{H}} \sum_{i=1}^{|S|} \left[-\frac{1}{2\sigma^2}(y_i - h(x_i))^2 \right] \\
&= \operatorname*{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{|S|} (y_i - h(x_i))^2 \\
&= \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{|S|} \sum_{i=1}^{|S|} (y_i - h(x_i))^2 \\
&= \operatorname*{argmin}_{h \in \mathcal{H}} L_S(h)
\end{aligned}
$$

□