WIKIPEDIA
The Free Encyclopedia

WIKIPEDIA

Huber loss

In statistics, the **Huber loss** is a loss function used in robust regression, that is less sensitive to outliers in data than the squared error loss. A variant for classification is also sometimes used.

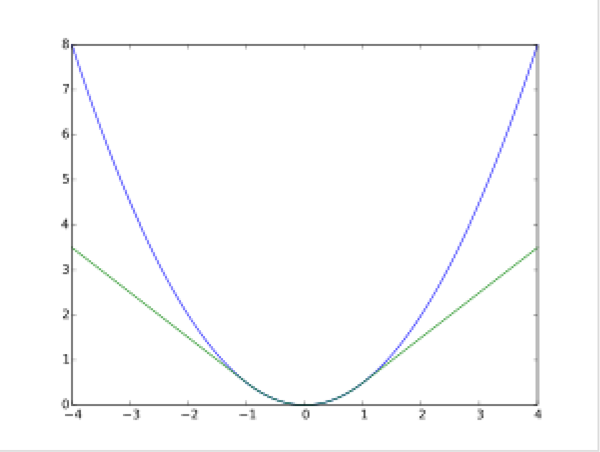
Definition

The Huber loss function describes the penalty incurred by an estimation procedure f. Huber (1964) defines the loss function piecewise by[1]

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \le \delta, \\ \delta \cdot \left(|a| - \frac{1}{2}\delta\right), & \text{otherwise.} \end{cases}$$

This function is quadratic for small values of a, and linear for large values, with equal values and slopes of the different sections at the two points where $|a| = \delta$. The variable a often refers to the residuals, that is to the difference between the observed and predicted values $a = y - f(x)$, so the former can be expanded to[2]

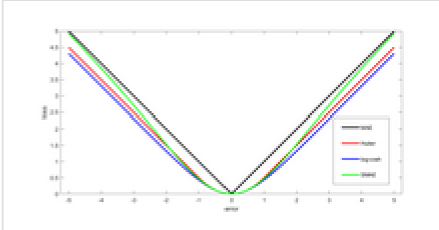
$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \le \delta, \\ \delta \ \cdot \left(|y - f(x)| - \frac{1}{2}\delta\right), & \text{otherwise.} \end{cases}$$



Huber loss (green, $\delta = 1$) and squared error loss (blue) as a function of $y - f(x)$

The Huber loss is the convolution of the absolute value function with the rectangular function, scaled and translated. Thus it "smoothens out" the former's corner at the origin.

Motivation



Comparison of Huber loss with other loss functions used for robust regression.

Two very commonly used loss functions are the squared loss, $L(a) = a^2$, and the absolute loss, $L(a) = |a|$. The squared loss function results in an arithmetic mean-unbiased estimator, and the absolute-value loss function results in a median-unbiased estimator (in the one-dimensional case, and a geometric median-unbiased estimator for the multi-dimensional case). The squared loss has the disadvantage that it has the tendency to be dominated by outliers—when summing over a set of a's (as in $\sum_{i=1}^{n} L(a_i)$), the sample mean is influenced too much by a few particularly large a-values when the distribution is heavy tailed: in terms of estimation theory, the asymptotic relative efficiency of the mean is poor for heavy-tailed distributions.

As defined above, the Huber loss function is strongly convex in a uniform neighborhood of its minimum $a = 0$; at the boundary of this uniform neighborhood, the Huber loss function has a differentiable extension to an affine function at points $a = -\delta$ and $a = \delta$. These properties allow it to combine much of the sensitivity of the mean-unbiased, minimum-variance estimator of the mean (using the quadratic loss function) and the robustness of the median-unbiased estimator (using the absolute value function).

Pseudo-Huber loss function

The **Pseudo-Huber loss function** can be used as a smooth approximation of the Huber loss function. It combines the best properties of **L2** squared loss and **L1** absolute loss by being strongly convex when close to the target/minimum and less steep for extreme values. The scale at which the Pseudo-Huber loss function transitions from **L2** loss for values close to the minimum to **L1** loss for extreme values and the steepness at extreme values can be controlled by the δ value. The **Pseudo-Huber loss function** ensures that derivatives are continuous for all degrees. It is defined as[3][4]

$$L_\delta(a) = \delta^2 \left(\sqrt{1 + (a/\delta)^2} - 1 \right).$$

As such, this function approximates $a^2/2$ for small values of a, and approximates a straight line with slope δ for large values of a.

While the above is the most common form, other smooth approximations of the Huber loss function also exist.[5]

Variant for classification

For classification purposes, a variant of the Huber loss called *modified Huber* is sometimes used. Given a prediction $f(x)$ (a real-valued classifier score) and a true binary class label $y \in \{+1, -1\}$, the modified Huber loss is defined as[6]

$$L(y, f(x)) = \begin{cases} \max(0, 1 - y\,f(x))^2 & \text{for } y\,f(x) > -1, \\ -4y\,f(x) & \text{otherwise.} \end{cases}$$

The term $\max(0, 1 - y\,f(x))$ is the hinge loss used by support vector machines; the quadratically smoothed hinge loss is a generalization of L.[6]

Applications

The Huber loss function is used in robust statistics, M-estimation and additive modelling.[7]

See also

- Winsorizing
- Robust regression
- M-estimator
- Visual comparison of different M-estimators

References

1. Huber, Peter J. (1964). "Robust Estimation of a Location Parameter" (https://doi.org/10.1214%2Faoms%2F1177703732). *Annals of Statistics*. **53** (1): 73–101. doi:10.1214/aoms/1177703732 (https://doi.org/10.1214%2Faoms%2F1177703732). JSTOR 2238020 (https://www.jstor.org/stable/2238020).
2. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). *The Elements of Statistical Learning* (https://web.archive.org/web/20150126123924/http://statweb.stanford.edu/~tibs/ElemStatLearn/). p. 349. Archived from the original (http://statweb.stanford.edu/~tibs/ElemStatLearn/) on 2015-01-26. Compared to Hastie *et al.*, the loss is scaled by a factor of ½, to be consistent with Huber's original definition given earlier.
3. Charbonnier, P.; Blanc-Féraud, L.; Aubert, G.; Barlaud, M. (1997). "Deterministic edge-preserving regularization in computed imaging". *IEEE Trans. Image Processing*. **6** (2): 298–311. Bibcode:1997ITIP....6..298C (https://ui.adsabs.harvard.edu/abs/1997ITIP....6..298C). CiteSeerX 10.1.1.64.7521 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.64.7521). doi:10.1109/83.551699 (https://doi.org/10.1109%2F83.551699). PMID 18282924 (https://pubmed.ncbi.nlm.nih.gov/18282924).
4. Hartley, R.; Zisserman, A. (2003). *Multiple View Geometry in Computer Vision* (https://archive.org/details/multipleviewgeom00hart_833) (2nd ed.). Cambridge University Press. p. 619 (https://archive.org/details/multipleviewgeom00hart_833/page/n634). ISBN 978-0-521-54051-3.
5. Lange, K. (1990). "Convergence of Image Reconstruction Algorithms with Gibbs Smoothing". *IEEE Trans. Med. Imaging*. **9** (4): 439–446. doi:10.1109/42.61759 (https://doi.org/10.1109%2F42.61759). PMID 18222791 (https://pubmed.ncbi.nlm.nih.gov/18222791).
6. Zhang, Tong (2004). *Solving large scale linear prediction problems using stochastic gradient descent algorithms* (https://dl.acm.org/citation.cfm?id=1015332). ICML.
7. Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine" (https://doi.org/10.1214%2Faos%2F1013203451). *Annals of Statistics*. **26** (5): 1189–1232. doi:10.1214/aos/1013203451 (https://doi.org/10.1214%2Faos%2F1013203451). JSTOR 2699986 (https://www.jstor.org/stable/2699986).

-