EDAN96
Applied Machine Learning
Lecture 9: Complement on the Logistic Loss

Pierre Nugues

Pierre.Nugues@cs.lth.se

November 28, 2022

Logistic Loss

1. For one observation, logistic regression yields the predicted probabilities of the classes
2. The logistic loss is defined as the opposite of the logarithm of predicted probability of the true class.
3. For a dataset, it can be reformulated as a cross entropy:

$$-\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i \cdot \log \hat{\mathbf{y}}_i,$$

where $\mathbf{y}_i$ is a one-hot vector giving the position of the true value:

$$\mathbf{y}_i = (0, 0, ..., 0, \mathbf{1}, 0, ...0)$$

and $\hat{\mathbf{y}}_i$ the vector of estimated probabilities of the observations for all the classes

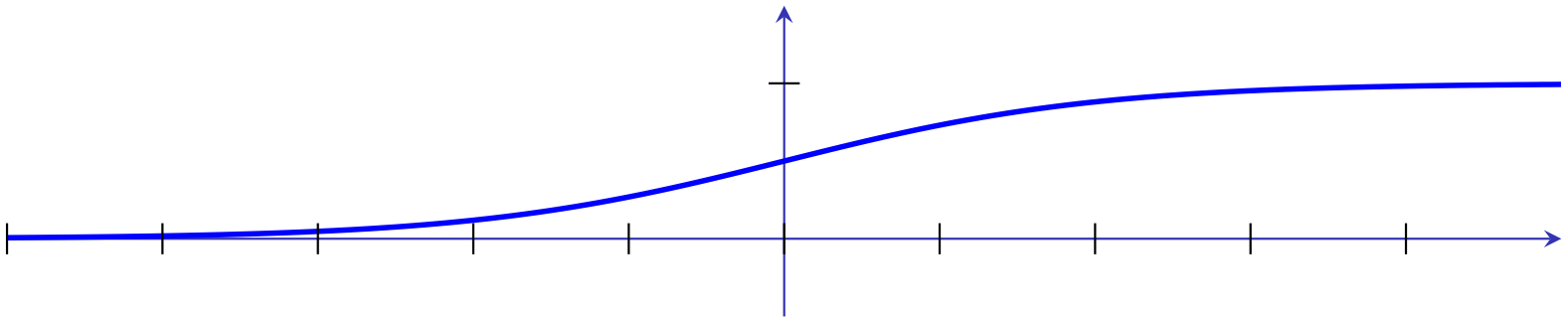
$$\hat{\mathbf{y}}_i = (0.01, 0.005, ..., \mathbf{0.70}, 0.10, ...0.001)$$

4. For this term: $-1 \times \log 0.7 = -0.36$

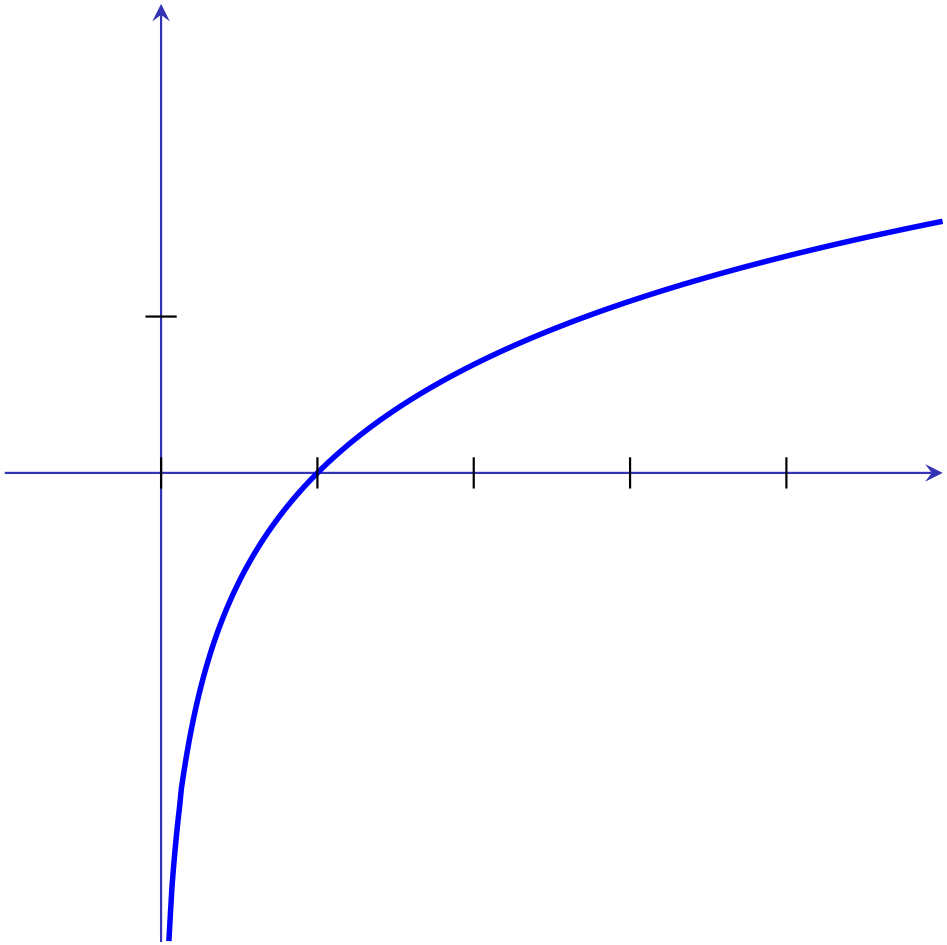
Logistic Curve

The logistic function:

$$f(x) = \frac{1}{1 + e^{-x}}$$



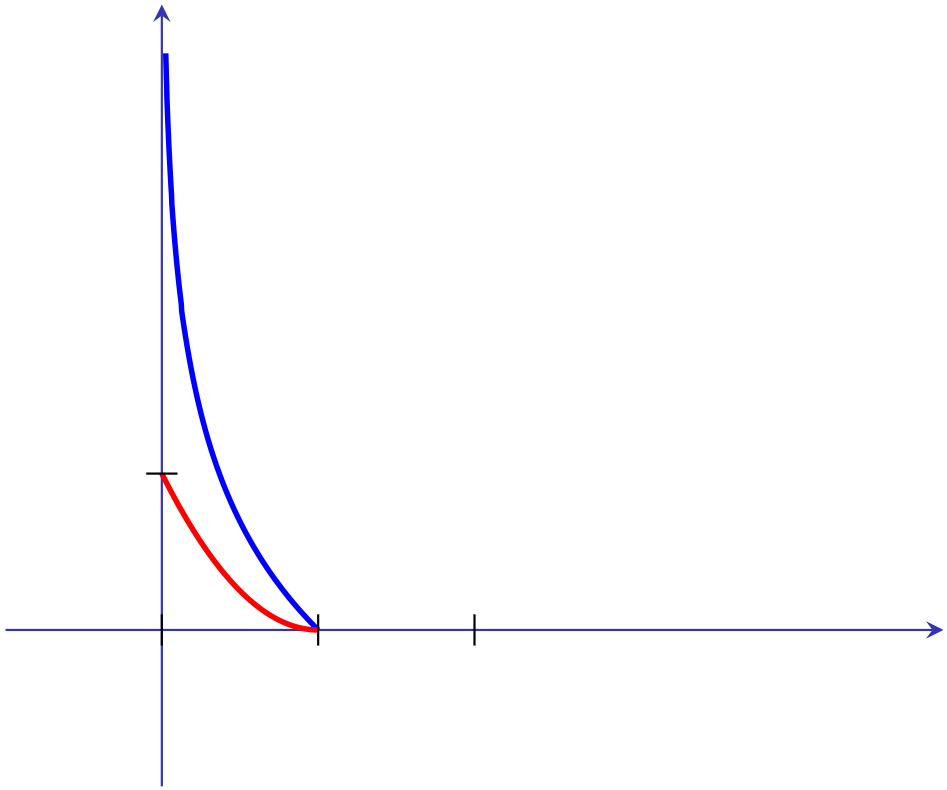
Logarithm Function



Loss

The logistic loss is defined as: $L(\hat{y}) = -\log \hat{y}$ in blue compared with the squared error loss in red $(1-\hat{y})^2$, where

$$\hat{y} = \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}}.$$



Example

Guess the language!

	Catalan	**French**	**Italian**	**Occitan**	**Portuguese**	**Spanish**
Buenos dias	0.01	0.01	0.01	0.01	0.01	0.01
Buongiorno	0.01	0.01	0.01	0.01	0.01	0.01
Bon jorn	0.01	0.01	0.01	0.01	0.01	0.01
Bom dia	0.01	0.01	0.01	0.01	0.01	0.01
Bon dia	0.01	0.01	0.01	0.01	0.01	0.01
Bonjour	0.01	0.01	0.01	0.01	0.01	0.01

How Berkson Solved it (I)

Excerpt from Joseph Berkson, Application of the Logistic Function to Bio-Assay, *Journal of the American Statistical Association*, Vol. 39, No. 227 (Sep., 1944), pp. 357-365

For the application of the logistic function, the question arose as to what method of fitting to utilize, and in particular whether to attempt to fit by the method of maximum likelihood. The principle of this method has been employed by many workers in particular situations, but under the name of maximum likelihood it has been advocated for general application by Professor R. A. Fisher. The method seems to be favored also by other mathematical authorities including, at least for the present application, Professor E. B. Wilson [12].

In spite of earnest prayer and the greatest desire to adhere to proper statistical behavior, I have not been able to see why the method of maximum likelihood is to be preferred over other methods, particularly the method of least squares. In the logistic function (1) there are two parameters α and β which if known determine the effect at any dose.

How Berkson Solved it (II)

Excerpt from Joseph Berkson, Application of the Logistic Function to Bio-Assay, *Journal of the American Statistical Association*, Vol. 39, No. 227 (Sep., 1944), pp. 357-365

probability are those of maximum likelihood. The maximum likelihood values are determined by obtaining the expression for the probability of all the observations occurring together, or its logarithm, differentiating with respect to the parameters and solving for maximum values. The method has considerable immediate plausibility. It employs a principle used in inverse probability and it has a generality which is attractive. However, the results that it gives in some cases conflict with other principles that seem equally well or better established. It is

Your Textbook in AI

Excerpt from Russell and Norvig, *Artificial Intelligence: A Modern Approach*, 4th edition, 2021, pp. 703-704 (EDAP01 course textbook)

> *The process of fitting the weights of this model to minimize loss on a data set is called* **logistic regression**. *There is no easy closed-form solution to find the optimal value of* $\mathbf{w}$ *with this model, but the gradient descent computation is straightforward. Because our hypotheses no longer output 0 or 1, we will use the* L_2 *loss function;*

They derive the update rule:

$$w_i \leftarrow w_i - \alpha\left(\frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}} - y\right) \times \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}}\left(1 - \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}}\right) \times x_i$$

With the logistic loss, you would find:

$$w_i \leftarrow w_i - \alpha \cdot \left(\frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}} - y_i\right) \cdot x_i;$$

The update rules are equal to a factor.