

# Lecture 3: Logistic Regression

Week 3

Lecturer: Tianyu Wang

## 1 Surrogate Functions for the Classification Objective

Recall the goal of classification is to find a function  $f$  in a hypothesis class  $\mathcal{H}$  to fit a given dataset  $\{(x_i, y_i)\}_{i=1}^n$  ( $x_i \in \mathbb{R}^d, y_i \in \{+1, -1\}$ ). The optimization objective (empirical risk) for this problem is

$$\min_{f \in \mathcal{H}} \sum_{i=1}^n \mathbb{I}_{[f(x_i) \neq y_i]},$$

which can be written as

$$\min_{f \in \mathcal{H}} \sum_{i=1}^n \mathbb{I}_{[y_i f(x_i) < 0]}.$$

Minimizing this empirical risk is not easy, since it is highly discontinuous and we cannot do much better than trial-and-error. Luckily, we can use surrogate functions, that are more friendly to work with, to approximate this objective. See Figure 1 for an illustration.

## 2 Logistic Regression

### 2.1 The surrogate objective

Consider a linear classifier  $f(x) = \begin{cases} +1, & \text{if } w^\top x + b \geq 0 \\ -1, & \text{otherwise.} \end{cases}$  Here we use a positive tie-breaking.

The empirical risk for this classifier is

$$\sum_{i=1}^n \mathbb{I}_{[y_i(w^\top x_i + b) < 0]}.$$

If we use the logistic loss, the surrogate objective is

$$l(w, b) = \sum_{i=1}^n \log(1 + \exp(-y_i(w^\top x_i + b))).$$

With this surrogate objective, one can solve for  $w$  and  $b$  using gradient-based methods, such as gradient descent. We will briefly discuss some optimization methods later in this course.

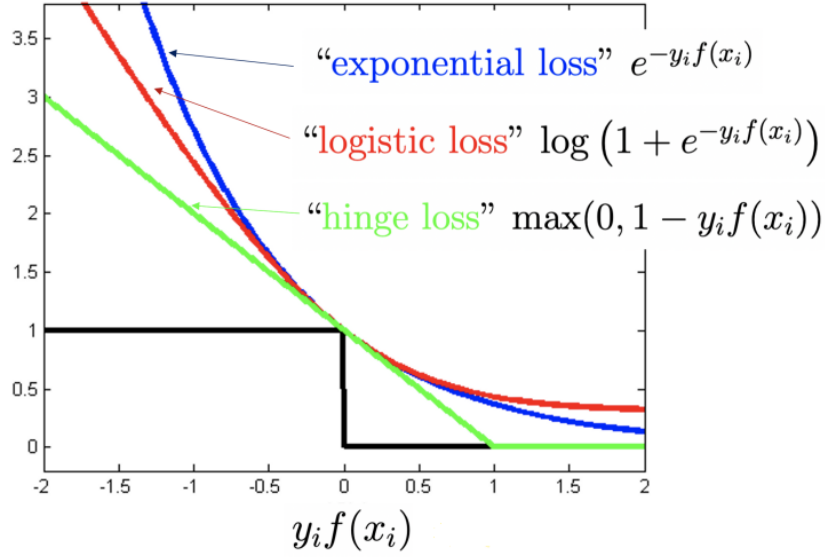


Figure 1: Surrogate functions for the objective. Picture Credit: C. Rudin.

## 2.2 A probabilistic perspective

Let's say the labels ( $y_i$ 's) are generated from a Bernoulli distribution:

$$y \sim \text{Bernoulli}(\mathbb{P}(y = 1|x, w, b)).$$

The Bernoulli distribution is fully determined by the success probability, which is a number in  $[0, 1]$ . This means we cannot directly impose a linear model on the success probability, since a non-trivial linear function takes values outside of  $[0, 1]$ . To handle this, we consider the following equation

$$\log \frac{\mathbb{P}(y = 1|x, w, b)}{1 - \mathbb{P}(y = 1|x, w, b)} = w^\top x + b.$$

Note that both sides of the above equation can take value in  $(-\infty, \infty)$ . Solve for  $\mathbb{P}(y = 1|x, w, b)$  in the above equation gives

$$\mathbb{P}(y = 1|x, w, b) = \frac{\exp(w^\top x + b)}{1 + \exp(w^\top x + b)}.$$

For both  $y_i = 1$  and  $y_i = -1$ , the probability can be expressed as

$$\mathbb{P}(y|x, w, b) = \frac{1}{1 + \exp(-y(w^\top x + b))}.$$

*Exercise.* Verify the above fact.

With this probability, the likelihood is

$$L(w, b | \{(x_i, y_i)\}_{i=1}^n) = \prod_{i=1}^n \mathbb{P}(y_i | x_i, w, b) = \prod_{i=1}^n \frac{1}{1 + \exp(-y_i(w^\top x_i + b))}.$$

The log likelihood is

$$\log L(w, b | \{(x_i, y_i)\}_{i=1}^n) = - \sum_{i=1}^n \log(1 + \exp(-y_i(w^\top x_i + b))).$$

Again, maximizing the log-likelihood is equivalent to minimizing the logistic loss.

### 3 ROC Curve and Area Under Curve

We will use slides by Rudin for this part. Recall the confusion matrix.

#### Acknowledgement

TW used lecture notes by Cynthia Rudin to compile this notes.