MATH 462 LECTURE NOTES
BINARY CLASSIFICATION

ADAM M. OBERMAN

1. Introduction to binary classification: errors

Reference for this section

- [Mur12, Chapter 8] (mostly the first equation) or [Mur22, Section 5.1.2].
- review earlier notes on logistic and softmax. Will be used in this material.

1.1. **Binary classification setup.** Here we consider the case of binary classification consisting, so there are two labels, which we denote by $-1, +1$, and we write

In the binary classification problem, the target set is

$$\mathcal{Y} = \mathcal{Y}_{\pm} = \{-1, +1\}$$

Remark 1.1. Sometimes the target set will instead be $\mathcal{Y} = \mathcal{Y}_2 = \{0, 1\}$.

We are given a dataset (S) consisting of m pairs of (x_i, y_i), $i = 1, \dots, m$, of data, $x_i \in \mathcal{X}$ and labels, $y_i \in \mathcal{Y}$,

$$S = \{(x_1, y_1), \dots, (x_m, y_m)\} \tag{S}$$

Definition 1.2 (Error)**.** The error, or zero-one loss, is given by $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$,

$$\ell_{0,1}(y_1, y_2) = \begin{cases} 0 & y_1 = y_2 \\ 1 & \text{otherwise} \end{cases}$$

Given a function $c : \mathcal{X} \to \mathcal{Y}$ and the dataset (S), the error of the model on the dataset is given by

$$L_{0-1}(c, S) = \frac{1}{m} \sum_{i=1}^{m} \ell_{0-1}(c(x_i), y_i)$$

We have seen some classifiers which work directly on the zero-one loss. However, for general problems, if we work directly with this loss, the optimization problem can be intractable. So instead we use a score based loss.

2. Binary classification Losses

The approach to (supervised) binary classification we take is score based, differentiable loss. The main advantage of this approach is that a differentiable loss is amenable to optimization

This means, instead of a function whose values are in the target set $\mathcal{Y}_{\pm}$, we define a real-valued function, the score, and then threshold it to determine the classification.

We can think of the score, s as generating a probability, $p = \sigma(s)$ or equivalently, as $s = \text{logit}(p)$, where p is the probability of the positive class. However, this is not needed for score based classification.

Date: October 10, 2023.

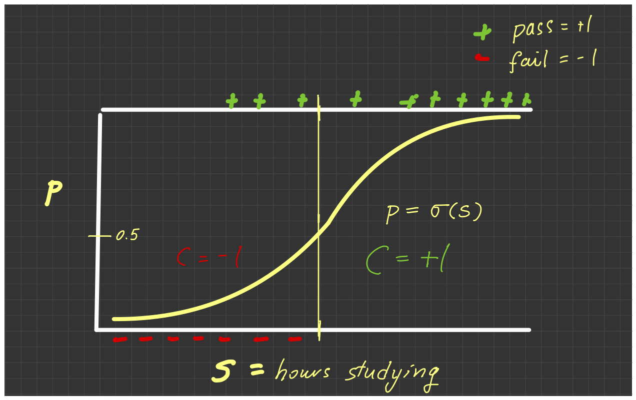

FIGURE 1. Illustration of logistic classifier: probability of passing ($y = +1$) an exam, as a function of hours studying. $p = \sigma(s)$. From wikipedia `https://en.wikipedia.org/wiki/Logistic_regression`. The model is an approximation of the probability of positive class, as a function of s.

2.1. **Standard log-logistic loss.**

In this section we study how to classify using the standard log-logistic loss. The classifier is given by (1). Recall $\sigma(s) = 1/(1 + \exp(s))$.

The loss is defined by setting $p = \sigma(s)$ is the probability based log loss,

$$\ell_{\log}(s, y) = \begin{cases} -\log(\sigma(s)), & y = +1 \\ -\log(1 - \sigma(s)), & y = -1 \end{cases}$$

The loss can be simplified to

$$\ell_{\log}(s, y) = \begin{cases} \log(1 + \exp(-s)), & y = +1 \\ \log(1 + \exp(+s)), & y = -1 \end{cases}$$

2.2. **Loss design.**

We want a loss that encourages more confident (correct) classifications. This leads to the following loss design principle. Here is a general principle which is satisfied by the log loss.

Definition 2.1 (Loss design principle). Given $s \in \mathbb{R}$ and $y \in \mathcal{Y}_\pm = \{-1, +1\}$, a non-negative function $\ell(s, y)$ is called a score-based binary classification loss. Given s, define the classification of s to be

$$c_{\mathrm{sgn}}(s) = \mathrm{sgn}(s) = \begin{cases} +1, & s \geq 0 \\ -1 & s < 0 \end{cases} \tag{1}$$

We say the loss is:

(1) balanced if $\ell(s, +1) = \ell(-s, -1)$, for all $s \in \mathbb{R}$
(2) normalized if $\ell(0, -1) = \ell(0, +1) = 1$
(3) monotone if $\ell(s, +1)$ is increasing in s and if $\ell(s, -1)$ is decreasing in s
(4) convex if $\ell(s, y)$ is convex as a function of s for each y.

Remark 2.2. Often in the definition of the losses, the normalization property is dropped, since it does not affect the minimizer.

On a dataset (S), given a score based loss, ℓ and function $h : \mathcal{X} \to \mathbb{R}$, the average loss of h on S is given by

$$L(h, S) = \frac{1}{m}\sum_{i=1}^{m} \ell(h(x_i), y_i)$$

2.3. **Standard margin loss.** In this section we study the standard margin (or hinge) loss,

$$\ell_{margin}(s, y) = \begin{cases} \max(0, 1 - s), & y = +1 \\ \max(0, 1 + s), & y = -1 \end{cases} \tag{2}$$

This loss is designed to score which lead to incorrect classification, as well as marginal scores. See Figure 2. See also Figure 4

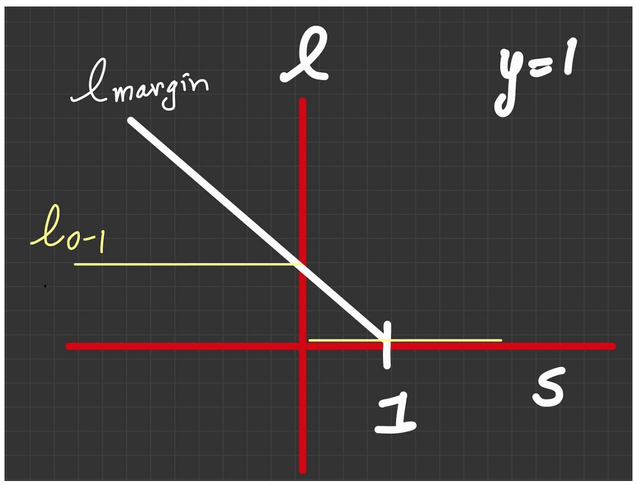

FIGURE 2. Margin loss, this loss is differentiable except at the corner, and lies above the 0-1 loss

3. Error bounds from the loss

We defined the classification task to be (in-distribution) generalization. For this purpose, both the losses work equally well. So does any abstract loss which satisfies the properties above.

When we use the score-based approach, we need to check that our loss minimization (which is defined $h \in \mathbb{R}$) results in an effective *classification*. In other words we care about the average classification error (the 0-1 loss, defined below).

Theorem 3.1. *The score-based classification loss is an upper bound for the error if*

$$\ell_{score}(s, y) \geq \ell_{0-1}(\mathrm{sgn}(s), y), \quad \text{for all } s \in \mathbb{R},\ y \in \mathcal{Y}_{\pm} \tag{LvE}$$

Suppose (LvE) *holds for ℓ_{score}. Then, for any function $h : \mathcal{X} \to \mathbb{R}$, and any dataset S*

$$L_{0-1}(\mathrm{sgn}(h), S) \leq L_{score}(h, S)$$

In particular, the bound above holds for a minimizer h^* of the loss.

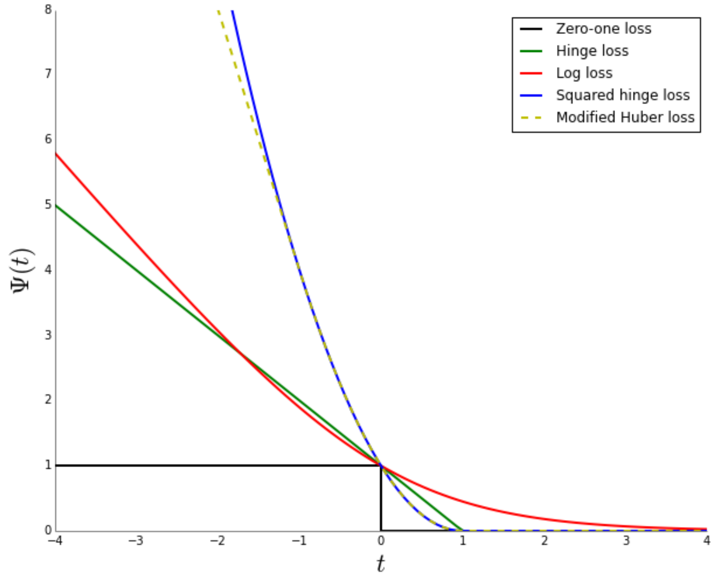

FIGURE 3. Plot of loss functions, image takes from `https://fa.bianp.net/blog/2014/surrogate-loss-functions-in-machine-learning/`. Here the log loss is normalized.

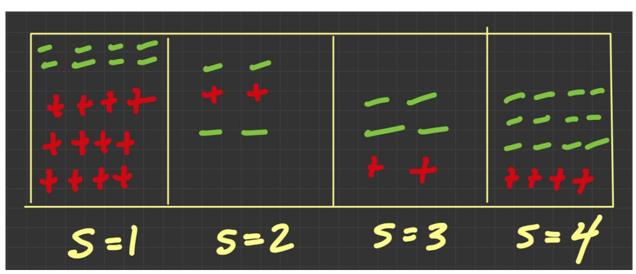

FIGURE 4. Score based classification example

4. Additional Exercises

Exercise 4.1. *In this exercise, use c_{sgn}. (i) Is the loss $\ell(h, y) = (h - y)^2$ an upper bound for the zero one loss? If so, what is the best (largest) constant for which* (LvE) *holds.*

(ii) Show that $\ell(h, y) = |h + y|$ is not an upper bound for the zero one loss.

(iii) Given the function $\ell(h, y)$, suppose there is an $h < 0$ with $\ell(h, 1) = 0$. Show this function cannot be an upper bound for the zero one loss.

(iv) Converse. Given $\ell(h, y)$, suppose (1) $\ell(h, y) \geq 0$ for all h, y, (2) there is some $c > 0$ such that $\ell(h, y) \geq c$ for all h with $(h) \neq y$. Prove that there is a C_{class} which makes ℓ and upper bound for the zero one loss. What is the best value of C_{class}?

Exercise 4.2. *Prove Theorem 3.1.*

Exercise 4.3. *Consider the example of score-based classification illustrated in Figure 4. Find the minimizer of the empirical loss using the score-based absolute value loss*

$$\ell_{abs}(s,y) = \begin{cases} \max(s,0) & y=-1 \\ \max(-s,0) & y=+1 \end{cases}$$

the threshold model $s_w(x) = x - w$, *and the sign classifier* $c(s) = \operatorname{sgn}(s)$. *Compare to the majority classifier which chooses the most popular class in each bin. Show that in Figure 4, if we relabel the scores from* $1,2,3,4$ *to any other non-decreasing values (e.g. try* $10,15,20,25$*), and use the absolute value loss, we get the same classifier. (Hint: can check this directly or use the condition for a minimizer).*

Exercise 4.4. *Show that with the margin loss* (2)*, the cases in* **??** *correspond to*

$$\ell_{margin}(s,y) \begin{cases} [1,\infty) & \textit{incorrect} \\ \in [0,1] & \textit{marginal} \\ = 0 & \textit{confident} \end{cases}$$

Exercise 4.5. *Show that* (LvE) *holds for the* $\ell_{margin-t}$ *with* $C_{class} = 1$ *and the* $c = \operatorname{sgn}$ *classifier. Justify* (**??**).

Definition 4.1. Given a threshold $t > 0$. Define the t-margin loss,

$$\ell_{margin,t}(s,y) = \begin{cases} \max(0, 1 - s/t) & y = 1 \\ \max(0, 1 + s/t) & y = -1 \end{cases} \tag{3}$$

Exercise 4.6. *(i) Show that setting* $t = 1$ *in* (3) *recovers that standard margin loss. (ii) Generalize the definitions of the error types* **??**.

Exercise 4.7. *Plot the loss* (3) *for* $y = 1$ *and* $t > 1$. *Show symmetry of loss* $\ell_{margin,t}(-s,-y) = \ell(s,y)$. *Use this to plot loss for* $y = -1$.

References

[Mur12] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
[Mur22] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2022.