MATH 462 LECTURE NOTES: BINARY CLASSIFICATION ANALYSIS

ADAM M. OBERMAN

1. Classification loss analysis

In this section we perform further analysis of the classification losses, which allow us to interpret them.

1.1. **Setup.** The setup is as follows.

In this example, we consider a dataset of scores and labels.

$$S^m = \{(x_1, y_1), \dots, (x_m, y_m)\}, \qquad x_i \in \mathbb{R} \tag{1}$$

Define the threshold model and threshold classifier, respectively, by

$$h_w(x) = x - w \qquad c_w(x) = \operatorname{sgn}(h_w(x)) = \begin{cases} +1, & x \geq w \\ -1, & x < w \end{cases} \tag{2}$$

For a given score based loss, $\ell(h, y)$, we consider

$$\widehat{L}(w) = \widehat{L}(w, S^m) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i) \tag{3}$$

A critical point of (3) is given by

$$\frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial h} \ell(h(s_i), y_i) = 0 \tag{4}$$

(since $\partial h / \partial w = -1$)

1.2. **Analysis of the margin loss.** First note that, for the standard margin loss

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

So using (4), we have the following

$$\ell'_{margin}(s, y) = \frac{\partial}{\partial s} \ell_{margin}(s, y) = \begin{cases} -1, & y = +1, s < 1 \\ +1, & y = -1, s > -1 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

We extend the notion of error types to the margin loss, as follows. Define the following classification pairs, see Figure 1.

Definition 1.1. Given the pair (s, y), where $y \in \mathcal{Y}_{\pm}$, and $s \in \mathbb{R}$. Define the pair (y, s) to be

- false: $c(s) \neq y$

Date: November 2, 2022.

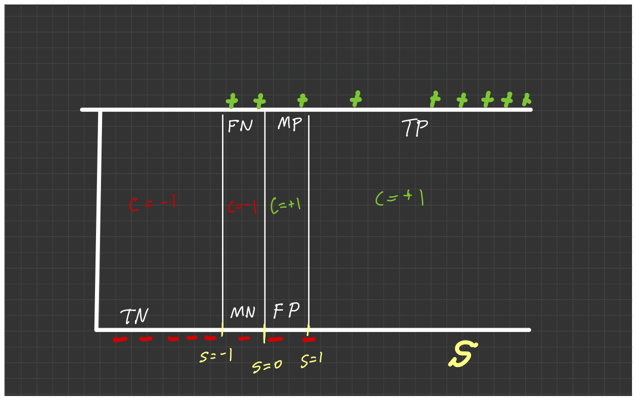

FIGURE 1. Illustration of margin classifier. The classification boundary is at $s = 0$. Correctly classified examples still incur a nonzero loss if they are within distance 1 of the boundary.

  - false positive (FP) $y = -1$, $s > 0$
  - false negative (FN) $y = 1, s < 0$
- marginal: if $y = c(s)$ and $|s| \leq 1$
  - marginal positive (MP) $y = 1$, $0 \leq s \leq 1$
  - marginal negative (MN) $y = -1$, $-1 \leq s \leq 0$
- non-marginal: $c(s) = y$ and $|s| \geq 1$

Note the definitions are overlapping when $s = 0, -1, +1$, which does not pose a problem in the sequel.

Theorem 1.2. *Consider minimizing* (3) *over the dataset* (1) *with the threshold model* $s(x) = x - w$ *and the classifier* $c(s) = \operatorname{sgn}(s)$. *Let* E_p *be the number of false or marginal positives. Let* E_n *be the number of false or marginal negatives. A sufficient condition for a minimizer* w^* *is that*

$$E_n = E_p$$

Proof. 1. Use (4).

2. Each term in the derivative is either (i) zero (for a confident correct) or (ii) equal to ± 1, depending on cases of false/marginal positive or false/marginal negative.

[[details to be filled in]] This leads to

$$\sum_{i \in \{FP, MN\}} 1 = \sum_{i \in \{FN, MP\}} 1$$

So the w^* is the threshold which $E_n = E_p$. □

1.3. **Analysis of log loss.** In this setting we consider the same threshold model and the classifier, (2)

$$h_w(x) = x - w \qquad c_w(x) = \operatorname{sgn}(h_w(x)) = \begin{cases} +1, & x \geq w \\ -1, & x < w \end{cases}$$

First we establish the following result

$$\frac{\partial}{\partial s} \ell_{\log}(\sigma(s), y) = e(\sigma(s), y) \tag{6}$$

where we define

$$e(p,y) = \begin{cases} 1-p & y=1 \\ p & y=-1 \end{cases}$$

to be the error from the optimal probability for the label.

Theorem 1.3. *Consider minimizing* (3) *over the dataset* (1) *with the threshold model and classifier.*

Define, using S_m*,* $J^+ = \{j \in 1,\dots,m \mid y_j = 1\}$ *and* $J^- = \{j \in 1,\dots,m \mid y_j = -1\}$

A sufficient condition for a minimizer w^* *is that*

$$\sum_{j\in J^+}(1-p_j) = \sum_{j\in J^-} p_j$$

Interpretation: class balance of probabilities

The sum of the over the positive examples of the probability gap, is equal to the sum over the negative examples of the probability gap.

Example 1.4. For example, if the probabilities are $.1, .2, .8, .9$ then the classes balance, there are two in each class, and the $e(p) = 1-.9, 1-.9$ and $.1, .2$ which balance.

Proof. 1. Apply (4) with (6) to obtain the result. □

1.4. **Exercises.**

Exercise 1.1. *Verify* (5).

Exercise 1.2. *Verify* (6). *(Hint: use* $\sigma' = \sigma(1-\sigma)$*)*

Exercise 1.3. *Consider the dataset* (1) *with the threshold model* (2). *Show that the 0-1 loss becomes a step function*

$$\ell_{0-1}(c_w(s), y) = 1_{\{\mathrm{sgn}(w-s)=y\}}$$

and that the empirical loss becomes

$$\widehat{L}(w) = \frac{1}{m}\sum_{i=1}^{m} 1_{\{\mathrm{sgn}(w-s)=y\}}$$

Exercise 1.4. *Use the chain rule to prove* (4).

Answer:

$$\frac{\partial}{\partial w}\ell(h(s_i,y_i)) = \frac{\partial}{\partial h}\ell(h(s_i,y_i))\frac{\partial}{\partial w}h_w$$

and $\frac{\partial}{\partial w}h_w = 1$.

Exercise 1.5. *Fill in the details of the proof of Theorem 1.2*

Exercise 1.6. *Complete the details of the proof of Theorem 1.3.*