Binary classifiers and logistic regression

Omar Hijab HIJAB@TEMPLE.EDU

Department of Mathematics
Temple University
Philadelphia, PA 19122, USA

Editor:

Abstract

Linear separability in a two-class dataset is related to the corresponding logistic regression problem. It is shown logistic regression, in its pure unpenalized form, is trainable exactly when the two classes are not linearly separable. The analogous results for multi-class datasets and soft-class datasets are also derived.

Keywords: binary classifiers, linear separability, logistic regression, trainability, SVM, two-class dataset, multi-class dataset, soft-class dataset.

1 Introduction

A basic problem of machine learning, if not the basic problem, is to train a neural network to replicate a given input-output map.

This is achieved by building a loss function $J(W)$, depending on the input dataset x_1, x_2, …, x_N, the neural network weights W, and the target outputs. Then gradient descent is applied to the loss function, resulting in a weight minimizing the loss function over all weights, or an optimal weight W^*.

The first question that comes to mind is the existence of an optimal weight. Roughly speaking, when this happens, we call the loss function trainable.

Trainability depends on the choice of dataset, targets, and neural network. Even in the simplest cases, the loss function may not be trainable.

For classification of a two-class or a multi-class dataset, the targets are one-hot encoded probability vectors p_1, p_2, …, p_N reflecting the class assignments. Then neural network outputs y_1, y_2, …, y_N are converted to probability vectors q_1, q_2, …, q_N, and the loss function is a sum of information discrepancies $I(p_1, q_1)$, $I(p_2, q_2)$, …, $I(p_N, q_N)$.

Here we focus on classifiers using the simplest neural network, that with no hidden nodes: $y = W^t x$. Then the training problem is standard logistic regression. Even in this simplest case, the loss function is not always trainable. To adjust for this, in actual implementations, trainability is artificially imposed by adding a penalty or regularization term to the loss function.

Surprisingly, a literature search indicates the trainability question, as discussed here, has not been addressed, even in the simplest case of unpenalized logistic regression.

Logistic regression is one approach towards classification of a two-class or a multi-class dataset. A totally different approach towards classification is linear separability of a two-class or a multi-class dataset.

When the classes are linearly separable, SVMs are used to obtain a maximum margin hyperplane. When this is not the case, non-linear separability is obtained by mapping the dataset into a higher dimensional space (Bishop, 2006), (Steinwart and Christmann, 2008).

It turns out the two classification approaches, logistic regression and linear separability, are inextricably tied together: For a two-class dataset, we show logistic regression, in its pure unpenalized form, is trainable exactly when the two classes are not linearly separable.

Subsequently, we derive the analogous results for multi-class datasets and soft-class datasets.

2 Background

Let $x_1, x_2, \dots, x_N$ be the samples of a two-class dataset, and let $p_1, p_2, \dots, p_N$ be the sequence of bits reflecting the class membership of the samples. Then the two classes correspond to $p = 0$ and $p = 1$ respectively.

A hyperplane is specified by the scalar equation

$$m \cdot x + b = 0.$$

Here m is a nonzero vector, b is a scalar and $m \cdot x$ is the dot product. When the dataset is one-dimensional, a hyperplane is simply a threshold.

Let $y = m \cdot x + b$ be the level of a sample x relative to a hyperplane. The hyperplane is separating if

$$\begin{aligned} y &\geq 0, \qquad \text{if } p = 1, \\ y &\leq 0, \qquad \text{if } p = 0, \end{aligned} \qquad \text{for every sample } x. \tag{1}$$

When there is a separating hyperplane, we say the dataset is separable.

If a two-class dataset lies in a hyperplane, then the hyperplane is separating, and the dataset is separable. Thus the question of separability is only interesting when the dataset does not lie in a hyperplane.

When the dataset does not lie in a hyperplane and the dataset is separable, it is easy to see the means of the two classes are distinct.

In regression analysis, given a sample x, one computes a probability q, depending only on the level $y = m \cdot x + b$, interpreted as our confidence that x should be assigned to the class $p = 1$.

Then, to be as consistent as possible with the dataset, we choose some measure of discrepancy $I(p, q)$ between probabilities p and q, and we minimize, over all (m, b), the sum of the discrepancies between p_k and the probabilities q_k corresponding to x_k, $k = 1, 2, \dots, N$.

Since y is a real number and q is a probability, we use a squashing function $q = \sigma(y)$ to convert real numbers to probabilities.

The standard choices are the sigmoid

$$\sigma(y) = \frac{1}{1 + e^{-y}}$$

and the relative information

$$I(p, q) = p \log(p/q) + (1 - p) \log((1 - p)/(1 - q)).$$

Since $I(p,q) \geq 0$ and $I(p,q) = 0$ only when $q = p$, $I(p,q)$ is a measure of information discrepancy. Since $I(p,q)$ is not symmetric in (p,q), q is thought of as a base probability against which p is compared.

With these choices, logistic regression is the minimization of the loss function

$$J(m,b) = \sum_{k=1}^{N} I(p_k, q_k), \qquad q_k = \sigma(y_k), \qquad y_k = m \cdot x_k + b,$$

over all m and b.

When $m = 0$, a weight (m,b) does not correspond to a hyperplane. When $m \neq 0$, we say (m,b) is a hyperplane weight. Since $(\lambda m, \lambda b)$ is the same hyperplane as (m,b), strictly speaking a hyperplane corresponds to $\lambda = \infty$. From this point of view, a hyperplane weight is a softened hyperplane, in the same way the sigmoid is a softening of the threshold $y = 0$.

A weight (m^*, b^*) is minimizing or optimal if it satisfies $J(m^*, b^*) \leq J(m,b)$ for all (m,b).

3 Properness

We recall the notion of properness of a function, a crucial ingredient in our results. Let $|w|$ denote the absolute value of a scalar weight w or the length of a vector weight w, and let $f(w)$ be a scalar function of weights. We say $f(w)$ is proper if every sublevel set is bounded: for every level c, there is a bound C such that

$$f(w) \leq c \qquad \text{implies} \qquad |w| \leq C. \tag{2}$$

A proper function need not be convex everywhere (Figure 1). If the graph of $f(w)$ is the cross-section of a river, then properness means the river never floods its banks, no matter how much it rains.

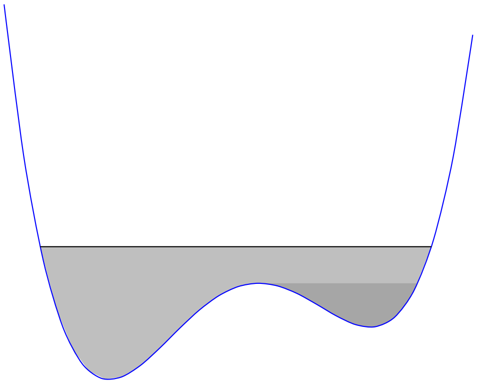

Figure 1: The graph of a proper continuous function.

A weight w^* is minimizing or optimal if it satisfies $f(w^*) \leq f(w)$ for all weights w. It is a standard result that a proper continuous function has an optimal weight.

4 Results

Theorem. *Assume the two-class dataset does not lie in a hyperplane. Then*

1. *the loss function has at most one optimal weight, and*
2. *the means of the classes agree if and only if the loss function has an optimal weight (m, b) with $m = 0$.*

When the loss function is proper and strictly convex, it can be shown gradient descent converges to the unique optimal weight (Hijab, 2025). Because of this, we say logistic regression is trainable if the loss function is proper. The main result for two-class datasets is

Theorem. *Assume neither class lies in a hyperplane. Then logistic regression is trainable if and only if the the two-class dataset is not separable.*

Combining these results, we conclude

Theorem. *If neither class lies in a hyperplane and the two-class dataset is not separable, there is an optimal weight.*

These results are extended below to the multi-class case and the soft-class case.

5 An Example

A simple example of a two-class logistic regression problem, taken from (Wikipedia, 2024), is as follows. A group of students takes an exam. For each student, we know the amount of time x they studied, and the outcome p, whether or not they passed the exam (Table 1).

x	p	x	p	x	p	x	p	x	p
0.5	0	.75	0	1.0	0	1.25	0	1.5	0
1.75	0	1.75	1	2.0	0	2.25	1	2.5	0
2.75	1	3.0	0	3.25	1	3.5	0	4.0	1
4.25	1	4.5	1	4.75	1	5.0	1	5.5	1

Table 1: Hours studied and outcomes.

The samples of this dataset are scalars, and the dataset is one-dimensional (Figure 2).

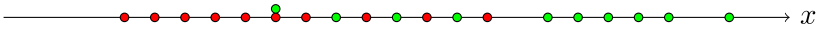

Figure 2: Exam dataset samples.

Plotting the dataset on the (x, p) plane, the goal is to fit a curve

$$p = \sigma(m^* x + b^*) \tag{3}$$

as in Figure 3.

We apply the main result for two-class datasets: The dataset is one-dimensional, so a hyperplane is just a point, a threshold. Neither class lies in a hyperplane, and the dataset is not separable (Figure 2). Hence logistic regression is trainable, and gradient descent is guaranteed to converge to the unique optimal weight

$$m^* = 1.49991537, \qquad b^* = -4.06373862,$$

leading to Figure 3.

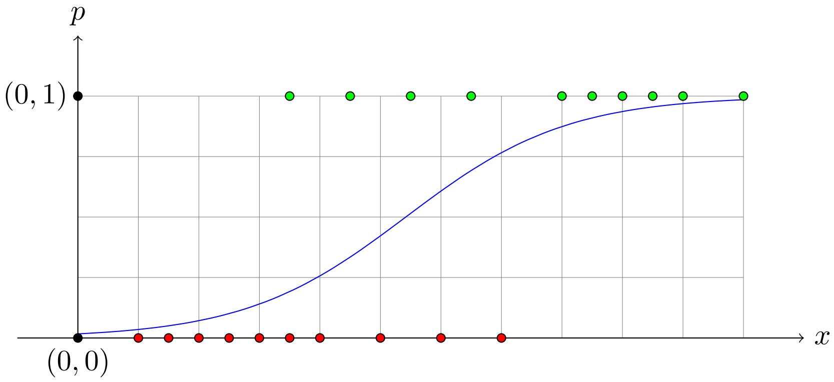

Figure 3: Fitted curve for exam dataset.

6 Proofs

Let $I(p)$ be the absolute information,

$$I(p) = p \log p + (1-p)\log(1-p).$$

Then $I(p) = 0$ when $p = 0, 1$. From this, J is the standard log regression loss function (see the notes at the end).

To compute the derivatives of the loss function, we recall some basic computations. The cumulant generating function of a fair coin, ignoring a constant term, is

$$Z(y) = \log(1 + e^y).$$

Then

$$Z'(y) = q, \qquad Z''(y) = q' = q(1-q), \qquad q = \sigma(y),$$

and

$$I(p,q) = I(p) - py + Z(y), \qquad q = \sigma(y). \tag{4}$$

This last identity, the information error identity, implies $I(p)$ and $Z(y)$ are dual convex functions (Berkovitz, 2003).

From these identities, we have

$$\frac{d}{dy} I(p,q) = q - p, \qquad \frac{d^2}{dy^2} I(p,q) = q(1-q), \qquad q = \sigma(y).$$

Let v be a vector and v_0 a scalar. By the chain rule, the first directional derivative of the loss function is

$$\frac{d}{dt}\bigg|_{t=0} J(m+tv, b+tv_0) = \sum_{k=1}^{N}(q_k - p_k)(v \cdot x_k + v_0), \tag{5}$$

and the second directional derivative of the loss function is

$$\frac{d^2}{dt^2}\bigg|_{t=0} J(m+tv, b+tv_0) = \sum_{k=1}^{N} q_k(1-q_k)(v \cdot x_k + v_0)^2. \tag{6}$$

By (6), the second directional derivative is nonnegative, so the loss function is convex. If (6) vanishes for some m, b, v, and v_0, then the dataset satisfies $v \cdot x_k + v_0 = 0$. Since the dataset does not lie in a hyperplane, $v = v_0 = 0$. This implies strict convexity of J. Since a strictly convex function has at most one minimizer, this establishes *1*.

A weight (m, b) is critical if the first directional derivative (5) vanishes for all v and v_0. Since the loss function is strictly convex, (m, b) is optimal if and only if (m, b) is critical.

Assume $(0, b)$ is optimal, and let $q = \sigma(b)$. Then (5) vanishes and $q_k = q$ for all k. Taking $v_0 = 1$ and $v = 0$ in (5), then $v_0 = 0$ and v arbitrary in (5), leads to

$$q \sum_{p_k=0} 1 = (1-q)\sum_{p_k=1} 1, \qquad q \sum_{p_k=0} x_k = (1-q)\sum_{p_k=1} x_k. \tag{7}$$

This implies the sample means of the two classes agree. Conversely, if the means of the classes agree, let q be the proportion of samples satisfying $p_k = 1$. Then (7) holds. It follows (5) vanishes and, with $b = \sigma^{-1}(q)$, $(0, b)$ is critical. This establishes *2*.

To derive the main result, assume there is a separating hyperplane (m, b). Then, by (1) and (4), $I(p_k, q_k) \leq \log 2$ for $k = 1, 2, \ldots, N$, hence $J(m, b) \leq N \log 2$. But $(\lambda m, \lambda b)$ is the same hyperplane, so

$$J(\lambda m, \lambda b) \leq N \log 2, \qquad \lambda > 0.$$

Since this contradicts (2), the loss function is not proper.

Recall a set K has interior if there is a ball B in K. If K_0 and K_1 are compact convex sets having interiors, then the hyperplane separation theorem (Deisenroth et al., 2020) states there is a hyperplane separating K_0 and K_1 if and only if the intersection $K_0 \cap K_1$ has no interior.

Now let K_0 and K_1 be the convex hulls of the two classes. If neither class lies in a hyperplane, both K_0 and K_1 have interiors. If there is no hyperplane separating K_0 and K_1, it follows there is a ball B in the intersection $K_0 \cap K_1$. Let x^* and r be the center and radius of B.

Now suppose $J(m, b) \leq c$ for some level c. We establish properness of the loss function by showing

$$|m| + |b| \leq \frac{c}{r}\left(1 + r + |x^*|\right). \tag{8}$$

If $J(m, b) \leq c$, then $I(p_k, q_k) \leq c$. Since $y < Z(y)$ and $-y < -y + Z(y)$, (4) implies

$$\begin{aligned} y_k &\geq -c, \qquad \text{if } p_k = 1, \\ y_k &\leq c, \qquad \text{if } p_k = 0. \end{aligned}$$

By taking convex combinations,

$$\begin{aligned} y &\geq -c, \qquad \text{for } x \text{ in } K_1, \\ y &\leq c, \qquad \text{for } x \text{ in } K_0. \end{aligned}$$

From this,

$$|m \cdot x + b| \leq c, \qquad \text{for } x \text{ in } K_0 \cap K_1.$$

Let $x^{\pm} = x^* \pm rv$ with v a unit vector, and let $y^{\pm} = m \cdot x^{\pm} + b$. Since $x^{\pm}$ are in $K_0 \cap K_1$,

$$2r|m \cdot v| = |y^+ - y^-| \leq |y^+| + |y^-| \leq 2c.$$

Optimizing over all v, we obtain $r|m| \leq c$. Since $|m \cdot x^* + b| \leq c$,

$$|b| \leq |m \cdot x^* + b| + |m \cdot x^*| \leq c + |m|\,|x^*|.$$

This leads to (8), establishing properness of the loss function.

7 Multi-class Case

Assume the dataset x_1, x_2, ..., x_N is divided into d disjoint classes, denoted $i = 1, 2, \dots, d$. Then the dataset is multi-class.

The multi-class case involves a tuple of hyperplanes (m_1, b_1), (m_2, b_2), ..., (m_d, b_d). Then the levels corresponding to a sample x are

$$\begin{aligned} y_1 &= m_1 \cdot x + b_1, \\ y_2 &= m_2 \cdot x + b_2, \\ \dots &= \dots \\ y_d &= m_d \cdot x + b_d. \end{aligned} \tag{9}$$

Let M be the matrix with columns m_1, m_2, ..., m_d, and let the bias b be the column vector $(b_1, b_2, \dots, b_d)$. With y equal to the column vector $(y_1, y_2, \dots, y_d)$, (9) is equivalent to

$$y = M^t x + b.$$

We say (M, b) is a hyperplane weight if $M \neq 0$.

Suppose the dataset lies in a hyperplane (m, b), and let M be the matrix with first column m, second column $-m$, and all other columns zero. Similarly, let b_1 be the column vector $(b, -b, 0, \dots, 0)$, and let $y_k = M^t x_k + b_1$, $k = 1, 2, \dots, N$. Then $y_k = 0$, $k = 1, 2, \dots, N$.

Conversely, if (M, b) is such that $y_k = 0$, $k = 1, 2, \dots, N$, the dataset lies in the intersection of r hyperplanes, where r is the rank of M.

Thus a dataset lies in a hyperplane if and only if there is a hyperplane weight (M, b) with $y_k = 0$, $k = 1, 2, \dots, N$.

There are at least two generalizations of separability to multi-class datasets. They are strong separability ("all-against-all"), and weak separability ("some-against-some"). Let y_1, y_2, ..., y_d be the levels (9).

A dataset is strongly separable if there is a hyperplane (m_i, b_i) separating class i from the rest of the dataset, for every $i = 1, 2, \dots, d$. This is the same as saying there is a weight (M, b) such that

$$\begin{aligned} y_i &\geq 0, && \text{for } x \text{ in class } i, \\ y_i &\leq 0, && \text{for } x \text{ in class } j, \end{aligned} \qquad \text{for every } i = 1, 2, \dots, d \text{ and every } j \neq i.$$

On the other hand, a dataset is weakly separable if there is a hyperplane (m, b) separating some class i and some class $j \neq i$. By setting the i-th entry of (M, b_1) equal to (m, b), the j-th entry equal to $(-m, -b)$, and all other entries equal to zero, we see this is the same as saying there is a weight (M, b_1) such that

$$\begin{aligned} y_i &\geq 0, && \text{for } x \text{ in class } i, \\ y_i &\leq 0, && \text{for } x \text{ in class } j, \end{aligned} \qquad \text{for some } i = 1, 2, \dots, d \text{ and some } j \neq i.$$

Clearly strong separability implies weak separability, and, with $d = 2$ and weight $(m, -m)$, $(b, -b)$, both notions reduce to separability as defined before.

If a dataset lies in a hyperplane, the dataset is separable, in both strong and weak senses. Thus the question of separability is only interesting when the dataset does not lie in a hyperplane.

Recall a set K has interior if there is a ball B in K. In the binary case, if neither class lies in a hyperplane, then the dataset is separable if and only if $K_0 \cap K_1$ has no interior.

For each $i = 1, 2, \dots, d$, let K_i be the convex hull of the samples in class i. Then K_i has interior if and only if class i does not lie in a hyperplane. In the multi-class case, by using the hyperplane separation theorem again, we have

Theorem. *Assume none of the classes lie in a hyperplane. Then the multi-class dataset is weakly separable if and only if $K_i \cap K_j$ has no interior for some i and some $j \neq i$.*

A vector $p = (p_1, p_2, \dots, p_d)$ is a probability vector if $p_j \geq 0$ for $j = 1, 2, \dots, d$, and $p_1 + p_2 + \dots + p_d = 1$. Let $p = (p_1, p_2, \dots, p_d)$ and $q = (q_1, q_2, \dots, q_d)$ be probability vectors, and let $y = (y_1, y_2, \dots, y_d)$ be a vector.

In multi-class logistic regression, the basic objects are the cumulant generating function

$$Z(y) = \log\left(e^{y_1} + e^{y_2} + \dots + e^{y_d}\right),$$

the softmax function

$$\sigma(y) = e^{-Z}\left(e^{y_1}, e^{y_2}, \dots, e^{y_d}\right),$$

the relative information

$$I(p, q) = p_1 \log(p_1/q_1) + p_2 \log(p_2/q_2) + \dots + p_d \log(p_d/q_d),$$

the absolute information

$$I(p) = p_1 \log p_1 + p_2 \log p_2 + \dots + p_d \log p_d,$$

and the information error identity

$$I(p, q) = I(p) - p \cdot y + Z(y), \qquad q = \sigma(y). \tag{10}$$

Then $I(p,q)$ is well-defined for any p and q, $I(p,q)$ is nonnegative, and $I(p,q) = 0$ only when $p = q$. Moreover $I(p)$ is well-defined for any p, and $0 \geq I(p) \geq -\log d$.

Let i be an integer satisfying $1 \leq i \leq d$. Then, as usual, a probability vector $p = (p_1, p_2, \dots, p_d)$ is one-hot encoded at slot i if $p_i = 1$. When p is one-hot encoded at slot i, $I(p) = 0$.

Let $\max y = \max_j y_j$. Then, by definition of $Z(y)$,

$$y_j \leq Z(y) \leq \max y + \log d, \qquad j = 1, 2, \dots, d.$$

Let p be one-hot encoded at slot i and let $q = \sigma(y)$. Then (10) implies

$$y_j - y_i \leq I(p,q) \leq \max y - y_i + \log d, \qquad j = 1, 2, \dots, d. \tag{11}$$

For x in class i, let $p(x)$ be the probability vector that is one-hot encoded at slot i. For each sample x_k, let

$$p_k = p(x_k), \qquad y_k = M^t x_k + b, \qquad q_k = \sigma(y_k), \qquad k = 1, 2, \dots, N.$$

Then the the logistic loss function is

$$J(M,b) = \sum_{k=1}^{N} I(p_k, q_k).$$

Here p_k and q_k are probability vectors, not scalars.

Let $\mathbf{1}$ be the column vector $(1, 1, \dots, 1)$. A matrix M is centered if $M\mathbf{1} = 0$, and a vector b is centered if $b \cdot \mathbf{1} = 0$. A weight (M,b) is centered if both M and b are centered. When (M,b) is centered, $y = M^t x + b$ is centered. A centered weight (M^*, b^*) is optimal if $J(M^*, b^*) \leq J(M,b)$ for all centered (M,b).

Let $\bar{m}$ be the mean of the columns of M, let $\bar{b}$ be the mean of the components of b, and let $\bar{y} = \bar{m} \cdot x + \bar{b}$. If (M_1, b_1) be obtained from (M,b) by subtracting $\bar{m}$ from the columns of M, and subtracting $\bar{b}$ from the components of b, then (M_1, b_1) is centered, with corresponding layers $y_1 = y - \bar{y}\mathbf{1}$.

Since

$$\sigma(y) = \sigma(y + \lambda\mathbf{1}),$$

for every scalar λ,

$$\sigma(y_1) = \sigma(y - \bar{y}\mathbf{1}) = \sigma(y), \qquad \text{hence} \qquad J(M_1, b_1) = J(M,b).$$

Thus there is no harm in restricting the loss function to centered weights.

We say logistic regression is trainable if the loss function is proper on centered weights. The main result for multi-class datasets is

Theorem. *If the multi-class dataset is strongly separable, logistic regression is not trainable. If none of the classes lie in a hyperplane and the multi-class dataset is not weakly separable, logistic regression is trainable.*

Combining the results, we obtain

Theorem. *If none of the classes lie in a hyperplane and the multi-class dataset is not weakly separable, there is an optimal centered weight.*

To derive the main result, suppose (M, b) is strongly separating. Since strong separability is equivalent to

$$\begin{aligned} &y_i \geq 0, \qquad \text{for } x \text{ in } K_i, \\ &y_j \leq 0, \qquad \text{for } x \text{ in } K_i \text{ and every } j \neq i, \end{aligned} \qquad \text{for every } i = 1, 2, \dots, d,$$

by (11), strong separability implies $I(p_k, q_k) \leq \log d$ for each x_k in class i, for every i. From this follows $J(M, b) \leq N \log d$. Since for $\lambda > 0$, $(\lambda M, \lambda b)$ is also strongly separating, it follows the loss function is not proper.

To establish properness of the loss function, suppose none of the classes lie in a hyperplane and the dataset is not weakly separable. Then $K_i \cap K_j$ has interior for all i and all $j \neq i$. Let x^*_{ij} be the centers of balls in $K_i \cap K_j$ for each $i \neq j$. By making the balls small enough, we may assume the radii of the balls equal the same $r > 0$.

Suppose $J(M, b) \leq c$ for some level c, with $M = (m_1, m_2, \dots, m_d)$, $b = (b_1, b_2, \dots, b_d)$ centered. We establish properness of the loss function by showing

$$|m_i| + |b_i| \leq \frac{c}{r} \left(1 + r + \frac{1}{d-1} \sum_{j \neq i} |x^*_{ij}| \right), \qquad i = 1, 2, \dots, d. \tag{12}$$

If $J(M, b) \leq c$, then $I(p, q) \leq c$ for each sample x. Then (11) implies

$$y_j - y_i \leq c, \qquad \text{for } x \text{ in class } i \text{ and } j \neq i.$$

By taking convex combinations,

$$y_j - y_i \leq c, \qquad \text{for } x \text{ in } K_i \text{ and } j \neq i.$$

Interchanging i and j,

$$y_i - y_j \leq c, \qquad \text{for } x \text{ in } K_j \text{ and } j \neq i.$$

Hence

$$|y_i - y_j| \leq c, \qquad \text{for } x \text{ in } K_i \cap K_j. \tag{13}$$

Let v be a unit vector, and let

$$x^{\pm} = x^*_{ij} \pm rv, \qquad y^{\pm}_i = m_i \cdot x^{\pm} + b_i, \qquad y^{\pm}_j = m_j \cdot x^{\pm} + b_j.$$

Since $x^{\pm}$ are in $K_i \cap K_j$, by (13),

$$2r|(m_i - m_j) \cdot v| = |(y^+_i - y^+_j) - (y^-_i - y^-_j)| \leq 2c.$$

Optimizing over all v, we obtain

$$r|m_i - m_j| \leq c.$$

Let

$$y_i = m_i \cdot x^*_{ij} + b_i, \qquad y_j = m_j \cdot x^*_{ij} + b_j.$$

Since x^*_{ij} is in $K_i \cap K_j$, by (13),

$$|b_i - b_j| \le |y_i - y_j| + |(m_i - m_j) \cdot x^*_{ij}| \le c + |m_i - m_j|\, |x^*_{ij}| \le c \left(1 + \frac{1}{r}|x^*_{ij}|\right).$$

Hence

$$|m_i - m_j| + |b_i - b_j| \le \frac{c}{r}(1 + r + |x^*_{ij}|).$$

Since M is centered,

$$dm_i = (d-1)m_i + m_i = (d-1)m_i - \sum_{j \ne i} m_j = \sum_{j \ne i}(m_i - m_j).$$

Similarly, since b is centered,

$$db_i = \sum_{j \ne i}(b_i - b_j).$$

Hence

$$|m_i| + |b_i| \le \frac{1}{d} \sum_{j \ne i} |m_i - m_j| + |b_i - b_j| \le \frac{1}{d-1} \sum_{j \ne i} \frac{c}{r}(1 + r + |x^*_{ij}|),$$

resulting in (12), and establishing properness of the loss function.

8 Soft-class Case

A soft-class dataset is a dataset $x_1, x_2, \dots, x_N$ with corresponding target probability vectors $p_1, p_2, \dots, p_N$. Here the targets need not be one-hot encoded.

For each $i = 1, 2, \dots, d$, let class i be the samples x whose corresponding targets $p = (p_1, p_2, \dots, p_d)$ satisfy $p_i > 0$.

Because we do not assume the targets are one-hot encoded, the classes may overlap. The main result for soft-class datasets is

Theorem. *If the soft-class dataset is strongly separable, logistic regression is not trainable. If none of the classes lie in a hyperplane and the soft-class dataset is not weakly separable, logistic regression is trainable.*

When the dataset is strongly separable, the same argument as before implies logistic regression is not trainable. Assume none of the classes lie in a hyperplane and the soft-class dataset is not weakly separable, and let ϵ_i be the minimum of p_i over all p corresponding to x in class i. Let ϵ be the least of $\epsilon_1, \epsilon_2, \dots, \epsilon_d$. If $j \ne i$, then

$$\epsilon(y_j - y_i) \le \epsilon(Z(y) - y_i) \le p_i(Z(y) - y_i) \le \sum_{k=1}^{d} p_k(Z(y) - y_k) = Z(y) - p \cdot y.$$

To establish properness, suppose $J(M, b) \le c$. Then $I(p, q) \le c$. By (11),

$$Z(y) - p \cdot y = I(p, q) - I(p) \le c + \log d, \qquad q = \sigma(y).$$

Combining the last two inequalities,

$$\epsilon(y_j - y_i) \le c + \log d, \qquad \text{for } x \text{ in class } i \text{ and } j \ne i.$$

The rest of the proof is as before, leading to (12) with c replaced by $(c + \log d)/\epsilon$.

A probability vector $p = (p_1, p_2, \dots, p_d)$ is strict if $p_i > 0$ for every slot i. As a special case of this generalization,

Theorem. *In a soft-class dataset, let K be the convex hull of the samples having strict corresponding targets, and suppose K has interior. Then logistic regression is trainable, and there is an optimal centered weight.*

9 Notes

1. Suppose the soft-class dataset does not lie in a hyperplane. It can be shown the loss function has at most one optimal weight.

2. Suppose the multi-class dataset does not lie in a hyperplane. It can be shown the means of the classes agree if and only if there is an optimal weight (M, b) with $M = 0$.

3. Since the bound (12) does not depend on the number of features (the sample dimension), a version of the main result should hold when the sample space is a Hilbert space.

4. Since the bound (12) is independent of N, a version of the main result should hold for continuous data sets, with an integral instead of a sum in the definition of the loss function.

5. Let i be an integer satisfying $1 \le i \le d$. Soft-class i may be defined as *any* collection of samples x whose targets p satisfy $p_i > 0$. In particular, the union of the classes need not be the whole dataset.

   For example, if $p = (p_1, p_2, \dots, p_d)$ and $\max p = \max_j p_j$, we may define class i to be the samples x whose targets p satisfy $p_i = \max p$. Then the results remain valid.

6. The Iris dataset is weakly separable. The two-dimensional projection of the MNIST dataset is not weakly separable.

7. In the literature, $I(p, q)$ is often called the "Kullback-Liebler divergence" or "relative entropy". We prefer the term relative information because this terminology is more descriptive and consistent with the absolute information $I(p)$, and because $I(p, q)$ is convex (see below).

8. Let $q = \sigma(y)$. In the literature, $I(p, q) - I(p) = -p \cdot y + Z(y)$ is often called the "cross-entropy". We prefer the term cross-information because this terminology is consistent with the terminology for $I(p)$ and $I(p, q)$, and because $I(p, q) - I(p)$ is convex (see below).

9. In the literature, the logistic regression loss function is a sum over samples of $I(p, q) - I(p)$. We work with $I(p, q)$ because the targets p for soft-class datasets are not one-hot encoded, so $I(p)$ need not be zero.

10. To further support our terminology choices, we note entropy is the negative of information, entropy is concave, information is convex, and loss functions are minimized, not maximized.

    This suggests that the loss function should be a sum of information terms, not entropy terms. Of course, the issue is the choice of terminology, not the choice of loss function: the loss function in the literature is identical with the loss function here.

10 Disclosure of Funding

This research was not supported by third-party funding.

References

L. D. Berkovitz. *Convexity and Optimization in* $\mathbf{R}^n$. Wiley, Hoboken, 2003.

C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, Heidelberg, 2006.

M. P. Deisenroth, A. A. Faisal, and C. S. Ong. *Mathematics for Machine Learning.* Cambridge University Press, Cambridge, 2020.

O. Hijab. *Math for Data Science.* To appear, 2025.

I. Steinwart and A. Christmann. *Support Vector Machines.* Springer, Heidelberg, 2008.

Wikipedia. *Logistic Regression.* URL: `https://en.wikipedia.org/wiki/Logisticregression`.