Jinchao Xu

Deep Learning Algorithms and Analysis

Summer 2020

Contributors:

This set of notes are based on contributions from many of graduate students, post-doctoral fellows and other collaborators. Here is a partial list:

Juncai He, Qingguo Hong, Li Jiang.....

Contents

1

Logistic Regression

1.1 Introduction to logistic regression

1.1.1 Plain logistic regression

We first introduce the next definition of the set of linearly classifiable weights.

Definition 1 (the set of linearly classifiable weights). *Assume that we are given k linearly separable sets $A_1, A_2, \cdots, A_k \in \mathbb{R}^d$, we define the set of classifiable weights as*

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (W, b) : w_i x + b_i > w_j x + b_j,\ \forall x \in A_i, j \neq i, i = 1, \cdots, k\} \tag{1.1}$$

which means those (W, b) can separate $A_1, A_2, \cdots, A_k$ absolutely correctly.

Our linearly separable assumption implies that $\boldsymbol{\Theta} \neq \emptyset$. Now we know the existence of linearly classifiable weights. But how can we find one element in $\boldsymbol{\Theta}$?

Definition 2 (soft-max). *Given parameter $\boldsymbol{\theta} = (W, b)$, a soft-max mapping $\boldsymbol{p} : \mathbb{R}^d \to \mathbb{R}^k$ is a mapping with the following formulation*

$$\boldsymbol{p}(x; \boldsymbol{\theta}) = \frac{e^{Wx+b}}{e^{Wx+b} \cdot \mathbf{1}} = \frac{1}{\sum_{i=1}^{k} e^{w_i x + b_i}} \begin{pmatrix} e^{w_1 x + b_1} \\ e^{w_2 x + b_2} \\ \vdots \\ e^{w_k x + b_k} \end{pmatrix} = \begin{pmatrix} p_1(x; \boldsymbol{\theta}) \\ p_2(x; \boldsymbol{\theta}) \\ \vdots \\ p_k(x; \boldsymbol{\theta}) \end{pmatrix} \tag{1.2}$$

where $e^{Wx+b} = \begin{pmatrix} e^{w_1 x + b_1} \\ e^{w_2 x + b_2} \\ \vdots \\ e^{w_k x + b_k} \end{pmatrix}$, $\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^k$, *and the i-th component*

$$p_i(x; \boldsymbol{\theta}) = \frac{e^{w_i x + b_i}}{\sum_{i=1}^{k} e^{w_i x + b_i}}. \tag{1.3}$$

The soft-max mapping have several important properties.

1. $0 < p_i(x;\boldsymbol{\theta}) < 1,\ \sum_i p_i(x;\boldsymbol{\theta}) = 1$.
2. $p_i(x;\boldsymbol{\theta}) > p_j(x;\boldsymbol{\theta}) \Leftrightarrow w_i x + b_i > w_j x + b_j$. This implies that the linearly classifiable weights have an equivalent description as

$$\boldsymbol{\Theta} = \left\{\boldsymbol{\theta} : p_i(x;\boldsymbol{\theta}) > p_j(x;\boldsymbol{\theta}),\ \forall x \in A_i, j \neq i, i = 1, \cdots, k\right\} \tag{1.4}$$

3. We usually use the max-out method to do classification. For a given data point x, we first use a soft-max mapping to map it to $\boldsymbol{p}(x;\boldsymbol{\theta})$, then we attached x to the class $i = \arg\max_j p_i(x;\boldsymbol{\theta})$.

Remark 1. The first property implies that $\boldsymbol{p}(x;\boldsymbol{\theta})$ can be regarded as a probability distribution of data points which means given $x \in \mathbb{R}^d$, we have $x \in A_i$ with probability $p_i(x;\boldsymbol{\theta})$, $i = 1, \cdots, k$.

The last properties means we pick the label i as the class of x such that $x \in A_i$ has the biggest probability $p_i(x;\boldsymbol{\theta})$.

More detailed discussion of logistic regression from the probability perspective will be presented in the nearly future.

From the above properties, we can define the next likelihood function to help find elements in $\boldsymbol{\Theta}$:

$$P(\boldsymbol{\theta}) = \prod_{i=1}^{k} \prod_{x \in A_i} p_i(x;\boldsymbol{\theta}) \tag{1.5}$$

where this likelihood function comes from its probabilistic interpretation which we will discuss later. Based on the property that

$$p_i(x;\boldsymbol{\theta}) = \max_{1 \leq j \leq k} p_j(x;\boldsymbol{\theta}), \forall x \in A_i, \tag{1.6}$$

if $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. This somehow means that if

$$P(\boldsymbol{\theta}) = \prod_{i=1}^{k} \prod_{x \in A_i} p_i(x;\boldsymbol{\theta}) = \max, \tag{1.7}$$

if $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Or we say that we may use the next optimization problem

$$\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}). \tag{1.8}$$

to find an element in $\boldsymbol{\Theta}$.

More precisely, let us introduce the next lemmas (properties) of $P(\boldsymbol{\theta})$.

Lemma 1. *Assume that the sets $A_1, A_2, \cdots, A_k$ are linearly separable. Then we have*

$$\left\{\boldsymbol{\theta} : \ P(\boldsymbol{\theta}) > \frac{1}{2}\right\} \subset \boldsymbol{\Theta}. \tag{1.9}$$

Proof. It suffices to show that if $\boldsymbol{\theta} \notin \boldsymbol{\Theta}$, we must have $P(\boldsymbol{\theta}) \leq \frac{1}{2}$. For any $\boldsymbol{\theta} \notin \boldsymbol{\Theta}$, there must exist an i_0 ,an $x_0 \in A_{i_0}$ and a $j_0 \neq i_0$ such that

$$w_{i_0} x_0 + b_{i_0} \leq w_{j_0} x_0 + b_{j_0}. \tag{1.10}$$

Then we have

$$p_{i_0}(x_0; \boldsymbol{\theta}) \leq \frac{e^{w_{i_0} x0 + b_{i_0}}}{e^{w_{i_0} x_0 + b_{i_0}} + e^{w_{j_0} x_0 + b_{j_0}}} \leq \frac{1}{2}. \tag{1.11}$$

Notice that $p_i(x; \boldsymbol{\theta}) < 1$ for all $i = 1, \cdots, k$, $x \in A$. So

$$P(\boldsymbol{\theta}) < p_{i_0}(x_0; \boldsymbol{\theta}) \leq \frac{1}{2}. \tag{1.12}$$

□

Lemma 2. *If $A_1, A_2, \cdots, A_k$ are linearly separable and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we have*

$$\lim_{\alpha \to +\infty} p_i(x; \alpha\boldsymbol{\theta}) = 1 \Leftrightarrow x \in A_i. \tag{1.13}$$

Proof. We first note that if $x \in A_i$,

$$p_i(\theta, x) = \frac{1}{1 + \sum\limits_{j \neq i} e^{\alpha[(w_j x + b_j) - (w_i x + b_i)]}} \to 1, \quad \text{as} \quad \alpha \to \infty. \tag{1.14}$$

On the other hand, if $x \notin A_i$,

$$p_i(x; \boldsymbol{\alpha\theta}) = \frac{1}{1 + \sum\limits_{j \neq i} e^{\alpha[(w_j x + b_j) - (w_i x + b_i)]}} \leq \frac{1}{2}. \tag{1.15}$$

This implies that if $x \notin A_i$, $\lim_{\alpha \to \infty} p_i(x; \alpha\boldsymbol{\theta}) \neq 1$ which is equivalent to the proposition that if $\lim_{\alpha \to \infty} p_i(x; \alpha\boldsymbol{\theta}) = 1$, then $x \in A_i$. □

Lemma 3. *If $A_1, A_2, \cdots, A_k$ are linearly separable,*

$$\boldsymbol{\Theta} = \left\{\boldsymbol{\theta} : \lim_{\alpha \to +\infty} P(\alpha\boldsymbol{\theta}) = 1\right\}. \tag{1.16}$$

Proof. We first note that if $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we have $\lim_{\alpha \to +\infty} p_i(x; \alpha\boldsymbol{\theta}) = 1$ for all $x \in A_i$. So

$$\lim_{\alpha\to+\infty} H(\alpha\boldsymbol{\theta}) = \lim_{\alpha\to+\infty} \prod_{i=1}^{k} \prod_{x\in A_i} p_i(x;\alpha\boldsymbol{\theta}) = \prod_{i=1}^{k} \prod_{x\in A_i} \lim_{\alpha\to+\infty} p_i(x;\alpha\boldsymbol{\theta}) = 1. \tag{1.17}$$

On the other hand, if $\lim_{\alpha\to+\infty} P(\alpha\boldsymbol{\theta}) = 1$, there must exist one $\alpha_0 > 0$ such that $P(\alpha_0\boldsymbol{\theta}) > \frac{1}{2}$. From Lemma 1, we have $\alpha_0\boldsymbol{\theta} \in \boldsymbol{\Theta}$, which means $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. □

These properties above imply that if we can obtain a classifiable weight through maximizing $P(\boldsymbol{\theta})$, while lemma 3 tells us that $P(\boldsymbol{\theta})$ will not have a minimum actually.

More specifically, we just need to find some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that

$$P(\boldsymbol{\Theta}) > \frac{1}{2} \Leftrightarrow L(\boldsymbol{\theta}) := -\log P(\boldsymbol{\theta}) < \log(2). \tag{1.18}$$

Question: how to find these element?

1.1.2 Regularized logistic regression

Here, we start from the regularization term $e^{-\lambda R(\|\boldsymbol{\theta}\|)}$ with these next properties:

1. $\lambda > 0$.
2. $R(t)$ is a strictly increasing function on $\mathbb{R}^+$ with $R(0) = 0$, $\lim_{t\to+\infty} R(t) = +\infty$. For example, $R(t) = t^2$.
3. $\|\cdot\|$ is a norm on $R^{k\times(d+1)}$, a commonly used norm is the following Frobenius norm:

$$\|\boldsymbol{\theta}\|_F = \sqrt{\sum_{i,j} W_{ij}^2 + \sum_i b_i^2}. \tag{1.19}$$

Based on this regularization term, we may consider the following regularized likelihood function $P_\lambda(\boldsymbol{\theta})$ as

$$P_\lambda(\boldsymbol{\theta}) = P(\boldsymbol{\theta}) e^{-\lambda R(\|\boldsymbol{\theta}\|)}. \tag{1.20}$$

Here, let us define

$$\boldsymbol{\Theta}_\lambda = \arg\max_{\boldsymbol{\theta}} P_\lambda(\boldsymbol{\theta}), \tag{1.21}$$

where we have the next definition of arg max

$$\arg\max_{\boldsymbol{\theta}} P_\lambda(\boldsymbol{\theta}) = \left\{\boldsymbol{\theta} \ : \ P_\lambda(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} P_\lambda(\boldsymbol{\theta})\right\}. \tag{1.22}$$

The next lemma show that the maximal set of modified objective is not empty.

Lemma 4. *Suppose that $A_1, A_2, \cdots, A_k$ are linearly separable, then*

1. *if $\lambda = 0$, $\boldsymbol{\Theta}_0 = \emptyset$,*
2. *$\boldsymbol{\Theta}_\lambda$ must be nonempty for all $\lambda > 0$.*

Proof. Lemma 3 shows the first proposition. For the second proposition, we notice that

1. $P_\lambda(\mathbf{0}) = \frac{1}{k^N}$.
2. $\exists\, M_\lambda > 0$ such that $e^{-\lambda R(\|\boldsymbol{\theta}\|)} < \frac{1}{k^N}$ whenever $\|\boldsymbol{\theta}\| > M_\lambda$ because of the properties of $R(\|\boldsymbol{\theta}\|)$.

So a maxima on $\{\boldsymbol{\theta} : \|\boldsymbol{\theta}\| < M_\lambda\}$ must be a global maxima. Then we can easily obtain the result in the lemma from the boundedness and closeness of $\{\boldsymbol{\theta} : \|\boldsymbol{\theta}\| < M_\lambda\}$. □

Furthermore, we have the next theorem which shows that we can indeed get $\boldsymbol{\Theta}$ by maximizing $P_\lambda(\boldsymbol{\theta})$.

Theorem 1. *If $A_1, A_2, \cdots, A_k$ are linearly separable,*

$$\boldsymbol{\Theta}_\lambda \subset \boldsymbol{\Theta}, \tag{1.23}$$

when $\lambda > 0$ and sufficiently small.

Proof. By Lemma 1, we can take $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ such that $P(\boldsymbol{\theta}_0) > \frac{3}{4}$. Then, for any $\lambda < \frac{\log \frac{3}{2}}{R(\|\boldsymbol{\theta}_0\|)}$, $\boldsymbol{\theta}_\lambda \in \boldsymbol{\Theta}_\lambda$, we have

$$P(\boldsymbol{\theta}_\lambda) \geq P_\lambda(\boldsymbol{\theta}_\lambda) \geq P_\lambda(\boldsymbol{\theta}_0) = P(\boldsymbol{\theta}_0)e^{-\lambda R(\|\boldsymbol{\theta}_0\|)} > \frac{3}{4} \cdot \frac{2}{3} = \frac{1}{2},$$

which implies that $\boldsymbol{\theta}_\lambda \in \Theta$. Thus, for any $0 < \lambda < \frac{\log \frac{3}{2}}{R(\|\boldsymbol{\theta}_0\|)}$, $\boldsymbol{\Theta}_\lambda \subset \boldsymbol{\Theta}$.

□

The design of logistic regression is that maximize $P_\lambda(\boldsymbol{\theta})$ is equivalent to minimize $-\log P_\lambda(\boldsymbol{\theta})$, i.e.,

$$\max_{\boldsymbol{\theta}} \{P_\lambda(\boldsymbol{\theta})\} \Leftrightarrow \min_{\boldsymbol{\theta}} \{-\log P_\lambda(\boldsymbol{\theta})\}, \tag{1.24}$$

while the second one is more convenient to evaluate the gradient. Meanwhile, we add a regularization term $R(\boldsymbol{\theta})$ to the objective function which makes the optimization problem has a unique solution.

Mathematically, we can formulate Logistic regression as

$$\min_{\boldsymbol{\theta}} L_\lambda(\boldsymbol{\theta}), \tag{1.25}$$

where

$$L_\lambda(\boldsymbol{\theta}) := -\log P_\lambda(\boldsymbol{\theta}) = -\log P(\boldsymbol{\theta}) + \lambda R(\|\boldsymbol{\theta}\|) = L(\boldsymbol{\theta}) + \lambda R(\|\boldsymbol{\theta}\|), \tag{1.26}$$

with

$$L(\boldsymbol{\theta}) = -\sum_{i=1}^{k} \sum_{x \in A_i} \log p_i(x; \boldsymbol{\theta}). \tag{1.27}$$

Then we have the next logistic regression algorithm.

Algorithm 1 Logistic Regression

Given data $A_1, A_2, \cdots, A_k$, find

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} L_\lambda(\boldsymbol{\theta}), \tag{1.28}$$

for some sufficient small $\lambda > 0$.

Remark 2. Here

$$L(\boldsymbol{\theta}) = -\log P(\boldsymbol{\theta}), \tag{1.29}$$

is known as the loss function of logistic regression model. The next reasons may show that why $L(\boldsymbol{\theta})$ is popular.

1. It is more convenient to take gradient for $L(\boldsymbol{\theta})$ than $P(\boldsymbol{\theta})$.
2. $L(\boldsymbol{\theta})$ is related the so-called cross-entropy loss function which will be discussed in the next section.
3. $L(\boldsymbol{\theta})$ is a convex function which will be discussed later.