Lecture 5. Logistic Regression and SGD

Lecturer: Jie Wang Date: March26, 2020

The major references of this lecture are this note by Tom Mitchell and [1].

1 Introduction

Suppose that we are given a set of data $\{(\mathbf{x}_i, y_i)\}_i^n$, where $y_i \in \{0, 1\}$. Clearly, this is a classification problem. As a commonly-used approach for classification, logistic regression aims to learn a mapping $f : X \to Y$, where $X = (X_1, \ldots, X_d)$ and $Y \in \{0, 1\}$.

2 The Probabilistic Approach

Assumption 1.

1. *$Y \sim \text{Bern}(p)$, that is, Y has the Bernoulli distribution with $P(Y = 1) = p$.*
2. *$X = (X_1, \ldots, X_d)$, where each X_j is a continuous random variable.*
3. *For each X_j, $P(X_j|Y = 0) \sim N(\mu_{j,0}, \sigma_j^2)$ and $P(X_j|Y = 1) \sim N(\mu_{j,1}, \sigma_j^2)$.*
4. *For $i \neq j$, X_i and X_j are conditionally independent given Y.*

Notice that, for different values of Y, the conditional distributions of the random variable X_j only differ in the means, while they have the same variance.

The Bayes rule leads to

$$\begin{aligned} P(Y = 0|X) &= \frac{P(X|Y = 0)P(Y = 0)}{P(X|Y = 0)P(Y = 0) + P(X|Y = 1)P(Y = 1)} \\ &= \frac{1}{1 + \frac{P(X|Y=1)P(Y=1)}{P(X|Y=0)P(Y=0)}} \\ &= \frac{1}{1 + \exp\left(\ln \frac{P(X|Y=1)P(Y=1)}{P(X|Y=0)P(Y=0)}\right)} \\ &= \frac{1}{1 + \exp\left(\sum_j \ln \frac{P(X_j|Y=1)}{P(X_j|Y=0)} + \ln \frac{p}{1-p}\right)} \end{aligned} \tag{1}$$

According to 3 of Assumption 1, we have

$$\begin{aligned}\sum_j \ln \frac{P(X_j|Y=1)}{P(X_j|Y=0)} &= \sum_j \ln \frac{\frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(\frac{-(X_j-\mu_{j,1})^2}{2\sigma_j^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(\frac{-(X_j-\mu_{j,0})^2}{2\sigma_j^2}\right)} \\ &= \sum_j \ln \exp\left(\frac{(X_j-\mu_{j,0})^2-(X_j-\mu_{j,1})^2}{2\sigma_j^2}\right) \\ &= \sum_j \left(\frac{(X_j-\mu_{j,0})^2-(X_j-\mu_{j,1})^2}{2\sigma_j^2}\right) \\ &= \sum_j \left(\frac{(X_j^2-2X_j\mu_{j,0}+\mu_{j,0}^2)-(X_j^2-2X_j\mu_{j,1}+\mu_{j,1}^2)}{2\sigma_j^2}\right) \\ &= \sum_j \left(\frac{\mu_{j,1}-\mu_{j,0}}{\sigma_j^2}X_j + \frac{\mu_{j,0}^2-\mu_{j,1}^2}{2\sigma_j^2}\right)\end{aligned} \tag{2}$$

Plugging Eq. (2) into Eq. (1) leads to

$$\begin{aligned}P(Y=0|X) &= \frac{1}{1+\exp\left(\ln\frac{p}{1-p} + \sum_j \left(\frac{\mu_{j,1}-\mu_{j,0}}{\sigma_j^2}X_j + \frac{\mu_{j,0}^2-\mu_{j,1}^2}{2\sigma_j^2}\right)\right)} \\ &= \frac{1}{1+\exp\left(\ln\frac{p}{1-p} + \sum_j \frac{\mu_{j,0}^2-\mu_{j,1}^2}{2\sigma_j^2} + \sum_j \frac{\mu_{j,1}-\mu_{j,0}}{\sigma_j^2}X_j\right)}.\end{aligned} \tag{3}$$

If we let

$$\begin{aligned}w_j &= \frac{\mu_{j,1}-\mu_{j,0}}{\sigma_j^2},\ j=1,\ldots,d, \\ w_0 &= = \frac{p}{1-p} + \sum_j \frac{\mu_{j,0}^2-\mu_{j,1}^2}{2\sigma_j^2},\end{aligned}$$

Eq. (3) takes the form of

$$P(Y=0|X) = \frac{1}{1+\exp(w_0+\sum_{j=1}^d w_j X_j)}. \tag{4}$$

Thus,

$$P(Y=1|X) = 1-P(Y=0|X) = \frac{\exp(w_0+\sum_{j=1}^d w_j X_j)}{1+\exp(w_0+\sum_{j=1}^d w_j X_j)}. \tag{5}$$

Thus, given a data instance $\mathbf{x}$, we compute the conditional probability $P(Y=0|X=\mathbf{x})$ and $P(Y=1|X=\mathbf{x})$, and predict its label as the one which makes the corresponding conditional probability larger.

3 An Alternative Approach to Estimate Parameters

Last section provides an approach to show why logistic regression models the conditional probabilities in the form of Eq. (4) and Eq. (5), and it also derives the value of the parameters. Notice that, all the derivations in last section are based on Assumption 1.

In this section, we describe an alternative approach to estimate the parameters, as Assumption 1 may not hold in many cases. The core idea is MLE, that is

$$\begin{aligned}\hat{\mathbf{w}} &= \operatorname*{\mathbf{argmax}}_{\mathbf{w}\in\mathbb{R}^{d+1}} \prod_i P(y_i|\mathbf{x}_i, \mathbf{w}) \\ &= \operatorname*{\mathbf{argmax}}_{\mathbf{w}\in\mathbb{R}^{d+1}} \sum_i \ln P(y_i|\mathbf{x}_i, \mathbf{w}).\end{aligned}$$

The conditional data log likelihood can be written in a unified form of

$$-L(\mathbf{w}) = \sum_i \left(y_i \ln P(Y=1|\mathbf{x}_i, \mathbf{w}) + (1-y_i)\ln P(Y=0|\mathbf{x}_i, \mathbf{w})\right)$$

Thus, we can see that

$$\hat{\mathbf{w}} = \operatorname*{\mathbf{argmin}}_{\mathbf{w}} L(\mathbf{w}). \tag{6}$$

Moreover,

$$\nabla L(\mathbf{w}) = \mathbf{X}^\top \mathbf{h}(\mathbf{w}),$$

where $\mathbf{X} \in \mathbb{R}^{n\times(d+1)}$ with its i^{th} row being $\bar{\mathbf{x}}_i^\top = (1, \mathbf{x}_i^\top)$ and $\mathbf{h}(\mathbf{w}) = (h_1(\mathbf{w}), \ldots, h_n(\mathbf{w}))^\top$ with

$$h_i(\mathbf{w}) = \frac{1}{1+\exp(-\langle \mathbf{w}, \bar{\mathbf{x}}_i\rangle)} - y_i.$$

We are now ready to apply the Gradient Descent algorithm to find $\hat{\mathbf{w}}$.

Indeed, the problem in (6) is a convex optimization problem. This can be seen from the fact that the hessian of $L(\mathbf{w})$ is positive semidefinite for all $\mathbf{w}$. Specifically, the hessian of $L(\mathbf{w})$ is

$$\nabla^2 L(\mathbf{w}) = \mathbf{X}^\top \mathbf{\Sigma}_{\mathbf{w}} \mathbf{X},$$

where $\mathbf{\Sigma}_{\mathbf{w}}$ is a diagonal matrix with its i^{th} entry on its diagonal being

$$\frac{\exp(-\langle \mathbf{w}, \bar{\mathbf{x}}_i\rangle)}{(1+\exp(-\langle \mathbf{w}, \bar{\mathbf{x}}_i\rangle))^2}.$$

Clearly, the hessian matrix $\nabla^2 L(\mathbf{w}) \succeq 0$.

4 Regularization

$$\hat{\mathbf{w}} = \operatorname*{\mathbf{argmin}}_{\mathbf{w}} L(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{7}$$

Question 1.

1. Without regularization, does the problem in (6) always admit a solution? If (6) admits a solution, is it unique?
2. What about the problem in (7)?

5 Logistic Regression for Multiple Target Values

Previous sections considers the classification problems with two classes. What if there are more than two classes? Can we extend logistic regression to deal with those more general cases?

Suppose that $Y \in \mathcal{C} = \{c_1, \ldots, c_K\}$. Then

$$P(Y = c_k|X) = \frac{\exp(w_{k,0} + \sum_{j=1}^{d} w_{k,j}X_j)}{1 + \sum_{k=1}^{K-1} \exp(w_{k,0} + \sum_{j=1}^{d} w_{k,j}X_j)}, \; k = 1, \ldots, K-1,$$
$$P(Y = c_K|X) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp(w_{k,0} + \sum_{j=1}^{d} w_{k,j}X_j)}, \; k = K.$$

6 Stochastic Gradient Descent

Besides the classic gradient descent, we have another popular suit of methods, called stochastic gradient descent, that are widely used to solve the problems like (6). We first motivate SGD in Section 6.1. Then, we analyze the convergence property of SGD in Section 6.2.

6.1 Motivation

We have introduced three models in this class so far: linear regression, naive Bayes, and logistic regression. Recall that, both linear regression and logistic regression take the form of

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\mathbf{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i, \mathbf{w}), y_i), \tag{8}$$

where $h(\cdot; \mathbf{w})$ is the parameterized model and $\ell(\cdot, \cdot)$ measures the prediction error (loss) at the i^{th} data instance. The objective function is the average of the sample losses, which is also known as the *empirical risk*

$$R_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i; \mathbf{w}), y_i). \tag{9}$$

To save notations, let

$$f_i(\mathbf{w}) = \ell(h(\mathbf{x}_i; \mathbf{w}), y_i).$$

Thus, in this section, we consider the following optimization problem:

$$\min_{\mathbf{w}} R_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{w}). \tag{10}$$

Remark 1. If we assume that each single sample $(\mathbf{x}_i, y_i)$ is the realization of a random vector $\xi \in \mathbb{R}^{d+1}$ with an unknown distribution $\mathcal{D}$, the objective function we would like to minimize is indeed the *expected risk*

$$R(\mathbf{w}) = \mathbb{E}_{\xi}[f(\xi; \mathbf{w})], \tag{11}$$

where

$$f(\xi; \mathbf{w}) = \ell(h(\xi_{[1:n]}; \mathbf{w}), \xi_{[n+1]}),$$

with $\xi_{[1:n]}$ being the vector consisting of the first n components of ξ, and $\xi_{[n+1]}$ the last entry of ξ.

To solve the problem (10), we can apply the gradient descent algorithm introduced in Lecture 3. This requires a scan of the entire data set in each iteration to compute the full gradient

$$\nabla R_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{w}),$$

which can be quite time-consuming if the training set contains a huge amount of data instances. Moreover, in many real applications, we have no access to the full gradient, as the data instance comes in only one at a time. This motives the popular *stochastic gradient descent* (SGD) method.

Let us consider the k^{th} iteration. Instead of computing the full gradient $\nabla R_n(\mathbf{w}_k)$, SGD aims to find an update direction $\mathbf{g}_k$ to approximate the full gradient. The only requirement is that the expectation of this approximate update direction equals to the full gradient, i.e.,

$$\mathbb{E}[\mathbf{g}_k] = \nabla R_n(\mathbf{w}_k).$$

A simple choice of $\mathbf{g}_k$ is to uniformly sample a data instance $\xi_k = (\mathbf{x}_{i_k}, y_{i_k})$ from $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and set

$$\mathbf{g}_k(\xi_k) = \nabla f_{i_k}(\mathbf{w}_k) = \nabla \ell(h(\mathbf{x}_{i_k}; \mathbf{w}), y_{i_k}). \tag{12}$$

It is easy to see that

$$\mathbb{E}_{\xi_k}[\mathbf{g}_k(\xi_k)] = \nabla R_n(\mathbf{w}_k).$$

We summarize the SGD algorithm as follows.

Algorithm 1 Stochastic Gradient Descent Algorithm

Input: an initial point $\mathbf{w}_0$, the number of iterations K, stepsize $\alpha > 0$, $k = 0$

Output: $\mathbf{w}_K$

1: **repeat**
2: choose update direction $\mathbf{g}_k(\xi_k)$ by Eq. (12)
3: $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k - \alpha \mathbf{g}_k(\xi_k)$
4: $k \leftarrow k + 1$
5: **until** $k \geq K$

6.2 Convergence Analysis

To keep the notation simple, we rewrite the problem (10) as follows.

$$\min_{\mathbf{w}} F(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{w}). \tag{13}$$

Similar to the analysis of GD, we first make some assumptions on the problem (13).

Assumption 2.

1. *The objective function F is convex and continuously differentiable, which implies that*

$$F(\mathbf{w}_1) \geq F(\mathbf{w}_2) + \langle \nabla F(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2 \rangle. \tag{14}$$

2. *The gradient of function F is Lipschitz continuous, i.e., $\exists\, L > 0$, such that*

$$\|\nabla F(\mathbf{w}_2) - \nabla F(\mathbf{w}_1)\| \leq L\|\mathbf{w}_2 - \mathbf{w}_1\|. \tag{15}$$

3. *The function F attains its minimum at $\mathbf{w}^*$, i.e.,*

$$F(\mathbf{w}^*) = F^* = \min_{\mathbf{w}} F(\mathbf{w}). \tag{16}$$

Recall that, for GD, the above assumptions imply the descent lemma as follows.

Lemma 1. *Suppose that a function F is continuously differentiable and its gradient is Lipschitz continuous with constant $L > 0$. Then, for the sequence generated by GD algorithm, we have*

$$F(\mathbf{w}_{k+1}) \leq F(\mathbf{w}_k) - \alpha(1 - \frac{L}{2}\alpha)\|\nabla F(\mathbf{w}_k)\|^2. \tag{17}$$

Thus, with stepsize $0 < \alpha < \frac{2}{L}$, the sequence $(F(\mathbf{w}_k))$ decreases monotonously.

Though there is a *descent* in SGD, it is not a descent algorithm. Due to the stochastic nature of the update direction $\mathbf{g}_k(\xi_k)$, the function values may even go up in some iterations. Can we show a *descent* property for SGD in terms of the *expectation*? The answer is still no, due to the nonnegative variance of the update direction.

Lemma 2. *Suppose that F is continuously differentiable and its gradient is Lipschitz continuous with constant $L > 0$. Then, for the sequence generated by SGD, we have*

$$\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F(\mathbf{w}_k)] \leq -\alpha(1 - \frac{L}{2}\alpha)\|\nabla F(\mathbf{w}_k)\|^2 + \frac{L}{2}\alpha^2 \mathbb{V}_{\xi_k}[\mathbf{g}_k]. \tag{18}$$

Proof. Let us consider the k^{th} iteration of SGD.

Lipschitz continuity of ∇F implies that

$$F(\mathbf{w}_{k+1}) - F(\mathbf{w}_k) \leq \langle \nabla F(\mathbf{w}_k), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle + \frac{L}{2}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2. \tag{19}$$

Noting the update rule in Algorithm 1 and taking expectation with respect to ξ_k of both sides of (19) lead to

$$\begin{aligned} \mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F(\mathbf{w}_k)] \leq & \langle \nabla F(\mathbf{w}_k), -\alpha \mathbb{E}[\mathbf{g}_k(\xi_k)] \rangle + \frac{L}{2}\alpha^2 \mathbb{E}_{\xi_k}[\|\mathbf{g}_k(\xi_k)\|^2] \\ = & -\alpha\|\nabla F(\mathbf{w}_k)\|^2 + \frac{L}{2}\alpha^2 \mathbb{E}_{\xi_k}[\|\mathbf{g}_k(\xi_k)\|^2]. \end{aligned} \tag{20}$$

Moreover, we have

$$\begin{aligned} \mathbb{V}_{\xi_k}[\mathbf{g}_k(\xi_k)] &= \mathbb{E}_{\xi_k}[\|\mathbf{g}_k(\xi_k)\|^2] - \|\mathbb{E}_{\xi_k}[\mathbf{g}_k(\xi_k)]\|^2 \\ &= \mathbb{E}_{\xi_k}[\|\mathbf{g}_k(\xi_k)\|^2] - \|\nabla F(\mathbf{w}_k)\|^2. \end{aligned} \tag{21}$$

Thus, the inequality (18) follows immediately by plugging Eq. (21) into (20). □

This lemma shows that the expected difference of two successive function values consists of two terms: the descent term and the variance term. We can see that the $\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1})]$ decrease with respect to $F(\mathbf{w}_k)$ only if the descent term dominates the variance term.

Notice that, even if $\nabla F(\mathbf{w}_k) = \mathbf{0}$—that is, $\mathbf{w}_k$ is one of the optimum of F—the variance $\mathbb{V}_{\xi_k}[\mathbf{g}_k(\xi_k)]$ can still be positive. This suggests that the $(\mathbb{E}[f(x_k)])$ can not converge to F^* with a fixed stepsize $\alpha > 0$. Besides, if we set $\mathbf{g}_k(\xi_k) = \nabla F(\mathbf{w}_k)$, then we have $\mathbb{V}_{\xi_k}[g_k] = 0$ and we recover the descent lemma in GD immediately.

For SGD, we cannot expect that $\mathbb{V}_{\xi_k}[\mathbf{g}_k(\xi_k)] = 0$, and we cannot even expect that it is bounded. However, we can make the following reasonable assumption for the objective function F for SGD.

Assumption 3. *We assume that the upper bound of $\mathbb{V}_{\xi_k}[\mathbf{g}_k(\xi_k)]$ takes the form of*

$$\mathbb{V}_{\xi_k}[\mathbf{g}_k(\xi_k)] \leq M + M_V \|\nabla F(\mathbf{w}_k)\|^2, \tag{22}$$

where M and M_V are positive constants.

Lemma 3. *Let $M_G = M_V + 1$. Assumptions 2 and 3 imply that*

$$\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F(\mathbf{w}_k)] \leq -\alpha(1 - \frac{L}{2} M_G \alpha)\|\nabla F(\mathbf{w}_k)\|^2 + \frac{L}{2} M \alpha^2. \tag{23}$$

We are now ready to analyze the convergence property of SGD. To light the burden, we consider the strongly convex objective functions, i.e.,

$$F(\mathbf{w}_2) \geq F(\mathbf{w}_1) + \langle \nabla F(\mathbf{w}_1), \mathbf{w}_2 - \mathbf{w}_1 \rangle + \frac{\mu}{2}\|\mathbf{w}_2 - \mathbf{w}_1\|^2, \forall\, \mathbf{w}_1, \mathbf{w}_2 \in \mathbf{dom}\ F, \tag{24}$$

where $\mu > 0$. Recall that, we have following result for strongly convex functions.

Lemma 4. *Suppose that F is strongly convex with convexity parameter $\mu > 0$ and continuously differentiable. Then,*

$$F(\mathbf{w}) - F^* \leq \frac{1}{2\mu}\|\nabla F(\mathbf{w})\|^2, \forall\, \mathbf{w} \in \mathbf{dom}\ F. \tag{25}$$

The following result shows a linear convergence rate of SGD for strongly convex objective functions.

Theorem 1. ***(Strongly Convex Objective, Fixed Stepsize)*** *Suppose that Assumptions 2 and 3 hold and $0 < \alpha < \frac{1}{LM_G}$. Then, the sequence $(\mathbb{E}[F(\mathbf{w}_k)])$ generated by SGD converges to a neighborhood of F^* with a linear rate. Specifically, we have*

$$\begin{aligned} \mathbb{E}_{\xi_0:\xi_{k-1}}[F(\mathbf{w}_k) - F^*] &\leq \frac{LM}{2\mu}\alpha + (1 - \mu\alpha)^k (F(x_0) - F^* - \frac{LM}{2\mu}\alpha) \\ &\xrightarrow{linear} \frac{LM}{2\mu}\alpha. \end{aligned} \tag{26}$$

Proof. Subtracting F^* from both sides of (23) and rearranging the terms yield

$$\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F^*] \leq F(\mathbf{w}_k) - F^* - \alpha(1 - \frac{L}{2} M_G \alpha)\|\nabla F(\mathbf{w}_k)\|^2 + \frac{L}{2} M \alpha^2. \tag{27}$$

As $0 < \alpha < \frac{1}{LM_G}$, we have

$$\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F^*] \leq F(\mathbf{w}_k) - F^* - \frac{\alpha}{2}\|\nabla F(\mathbf{w}_k)\|^2 + \frac{LM}{2}\alpha^2. \tag{28}$$

Combining (25) and (28) leads to

$$\begin{aligned}\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F^*] &\leq F(\mathbf{w}_k) - F^* - \mu\alpha(F(\mathbf{w}_k) - F^*) + \frac{LM}{2}\alpha^2 \\ &= (1 - \mu\alpha)(F(\mathbf{w}_k) - F^*) + \frac{LM}{2}\alpha^2,\end{aligned} \tag{29}$$

which is equivalent to

$$\mathbb{E}_{\xi_k}\left[F(\mathbf{w}_{k+1}) - F^* - \frac{LM}{2\mu}\alpha\right] \leq (1 - \mu\alpha)\left(F(\mathbf{w}_k) - F^* - \frac{LM}{2\mu}\alpha\right).$$

Now take the expectation with respect to $\xi_0, \ldots, \xi_{k-1}$ of both sides of the above inequaltiy, we have

$$\mathbb{E}_{\xi_0:\xi_k}\left[F(\mathbf{w}_{k+1}) - F^* - \frac{LM}{2\mu}\alpha\right] \leq (1 - \mu\alpha)^{k+1}\left[F(\mathbf{w}_0) - F^* - \frac{LM}{2\mu}\alpha\right]. \tag{30}$$

The claim follows immediately. □

References

[1] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 2018.