Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 31(9): 5250–5266.
DOI: 10.3934/era.2023267
Received: 08 May 2023
Revised: 07 July 2023
Accepted: 17 July 2023
Published: 20 July 2023

Research article

The generalization ability of logistic regression with Markov sampling

Zhiyong Qian, Wangsen Xiao and Shulan Hu*

School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073, China

* **Correspondence:** Email: hu_shulan@zuel.edu.cn.

Abstract: In the case of non-independent and identically distributed samples, we propose a new ueMC algorithm based on uniformly ergodic Markov samples, and study the generalization ability, the learning rate and convergence of the algorithm. We develop the ueMC algorithm to generate samples from given datasets, and present the numerical results for benchmark datasets. The numerical simulation shows that the logistic regression model with Markov sampling has better generalization ability on large training samples, and its performance is also better than that of classical machine learning algorithms, such as random forest and Adaboost.

Keywords: non-independent identically distributed samples; logistic regression model; uniformly ergodic Markov chain algorithm; generalization ability

1. Introduction

The logistic regression model has become one of the most popular machine learning methods in classification [1–4]. Besides the advantages of good performance and strong interpretability in practical applications, the logistic regression model also has a complete theoretical basis in terms of consistency and generalization performance when the training samples are independent and identically distributed (i.i.d) [5–8]. However, the hypothesis of i.i.d is quite hard to be proved in practice, so it is natural to consider the logistic regression model with non-i.i.d samples.

The relaxation of the i.i.d hypothesis has been discussed for a long time in machine learning and statistical literature. For example, Wang et al. [9] pointed out that the statistical learning theory in the case of small samples cannot be directly applied to large samples and proposed the generalization bounds of under-sample models based on strict Bayesian network processing. Sun and Wu [10], Sun and Guo [11] and Chu and Sun [12] used mixed samples to analyze the error of l_2-norm least squares and l_1-norm least squares regression, respectively. Guo and Shi [13] proved that the learning speed of regularized least squares regression was faster than the classical method. Machine learning algorithms in the non-i.i.d case are solved by concentration inequalities because the concentration inequalities can provide probability upper bounds for the deviation [14]. Modha and Masry [15] extended the classical inequalities in the condition from i.i.d to m-correlation and strong mixing, respectively. Merlevède et al. [16] obtained a Bernstein type inequality for a class of weakly dependent and bounded random variables. Fan et al. [17] studied the Hoeffding inequality for general Markov chains and time-dependent functions.

In the paper, we enhance the performance of the logistic regression model from small samples to large samples. Machine learning usually performs well in Markov chain samples [18], so we develop the uniformly ergodic Markov chain algorithm (ueMC algorithm) for the logistic regression model. The ueMC algorithm can identify samples with classification errors and close to the decision boundary, and determine the final Markov samples used in the pre-training models. Compared with the algorithm proposed by Thongkam et al. [19], the ueMC algorithm does not directly eliminate the misclassified samples, as they are based on a random probability. Similar to the method proposed by Miranda et al. [20], the previous eliminated samples may be selected later. Inspired by [21], we study the generalization ability of the ueMC algorithm, and establish the optimal learning rate of the logistic regression classification for ueMC samples. Through numerical study and the simulation, it is verified that the performance of the ueMC algorithm based on ueMC samples is more effective than the algorithm based on other random samples, and the performance of the ueMC algorithm is also better than of the classical machine learning algorithm.

The paper is organized as follows. In Section 2, some definitions and notations are given. In Section 3, we present and prove the main results on the learning rates of the logistic regression model with ueMC samples. In Section 4, we develop a new ueMC algorithm, and present the numerical studies on the generalization performance of the logistic regression model based on Markov sampling for benchmark datasets. Finally, the conclusions are given in Section 5.

2. Preliminaries

In this section we introduce the definitions and notations throughout this paper.

2.1. Logistic regression model

Let $(\mathcal{X}, d)$ be a compact metric space and $\mathcal{Y} = \{0, 1\}$. A binary classifier is a function $\hat{f} : \mathcal{X} \to \mathcal{Y}$ which labels every point $x \in \mathcal{X} \subseteq \mathbb{R}$ with some $y \in \mathcal{Y}$. Let φ be a probability distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and (X, Y) be the corresponding random variable. Given $X_i = (1, x_{i1}, \ldots, x_{ik})^T \subseteq \mathbb{R}^{k+1}$, $i = 1, 2, \ldots, N$, the form of the separating hyperplane is as follows:

$$\hat{f} = W^T X = w_0 + w_1 x_{n1} + w_2 x_{n2} + \ldots + w_k x_{nk} = 0, \tag{2.1}$$

where $W = (w_0, w_1, \ldots, w_k)^T$ is the coefficient of the variable.

Let $P(Y_i \mid X_i, W) = \frac{1}{1+e^{-W^T X_i}} = sigmod(\hat{f})$ represent the probability of the sample being a positive sample ($Y_i = 1$), and $1 - P(Y_i \mid X_i, W)$ represent the probability of the sample being a negative sample ($Y_i = 0$). So we have:

$$\ln \frac{P(Y_i = 1 \mid X_i, W)}{P(Y_i = 0 \mid X_i, W)} = W^T X_i, \tag{2.2}$$

where $P\left(Y_i = 1 \mid X_i, W\right) = \frac{e^{W^T X_i}}{1+e^{W^T X_i}}$.

The log-likelihood function is:

$$g\left(W\right) = \sum_{i=1}^{N} \ln\left(Y_i P\left(Y_i = 1 \mid X_i, W\right) + \left(1 - Y_i\right) P\left(Y_i = 0 \mid X_i, W\right)\right). \tag{2.3}$$

The objective function of the logistic regression model is:

$$L\left(W\right) = \arg\min_{W} \sum_{i=1}^{N} \left(Y_i W^T X_i + \ln\left(1 + e^{W^T X_i}\right)\right). \tag{2.4}$$

Linear models are easy to over-fitting in high dimensions because of the correlation between features. In the paper, the weight term of the model is constrained to alleviate the problem of over-fitting by adding a regular term $c\left(W\right)$. We choose norm l_2 regularization and use BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm [22] to solve the parameters.

$$L\left(W\right) = \arg\min_{W} \sum_{i=1}^{N} \left(Y_i W^T X_i + \ln\left(1 + e^{W^T X_i}\right)\right) + c\left(W\right). \tag{2.5}$$

Let $\mathcal{H}$ be Hilbert space, a set of real function on space $\mathcal{X} \subseteq \mathbb{R}$. If there is a kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying $\forall x \in \mathcal{X}, K\langle\cdot, x\rangle \in \mathbb{R}$, then $\mathcal{H}_K$ is called a reproducing kernel Hilbert space satisfying $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, f(x) = \langle f, K\langle\cdot, x\rangle\rangle$.

For a function $f : \mathcal{X} \to \mathbb{R}$, sign function $\mathrm{sgn}(f) = 1$ if $f(x) \geq 0$ and $\mathrm{sgn}(f) = 0$ if $f(x) < 0$. Then the logistic regression model is defined as $\mathrm{sgn}\left(f_{z,\lambda}\right)$, where $f_{z,\lambda}$ is a minimizer of the following optimization problem involving a set of random sample $S = \{z_i\}_{i=1}^m \in \mathcal{Z}^m$:

$$f_{z,\lambda} := \arg\min_{f \in \mathcal{H}_K} \left\{\lambda \|f\|^2_{\mathcal{H}_K} + \mathcal{E}_z(f)\right\}. \tag{2.6}$$

In Formula (2.6), $\ell(f, z) = -[y \ln(\mathrm{sigmoid}(f(x))) + (1 - y)\ln(1 - \mathrm{sigmoid}(f(x)))]$ is the loss function, $\mathcal{E}_z(f) = \frac{1}{m}\sum_{i=1}^m \ell\left(f, z_i\right)$ is the empirical error, $\mathcal{E}(f) = E[\ell(f, z)]$ denotes the generalization error of the corresponding function f. $\lambda = 1/(2C)$ is the regularization parameter, where C is a constant which depends on $m : C = C(m)$ and often $\lim_{m\to\infty} C(m) = \infty$.

2.2. Uniformly ergodic Markov chains (ueMC)

Suppose a Markov chain on $(\mathcal{Z}, \mathcal{S})$ is a sequence of random variables $\{Z_i\}_{i\geq 1}$ with a set of transition probability measures $P^{(n)}\left(A \mid z_i\right), A \in \mathcal{S}, z_i \in \mathcal{Z}$. It is assumed that

$$P^{(n)}\left(A \mid z_i\right) = P\left\{Z_{n+i} \in A \mid Z_j, j < i, Z_i = z_i\right\}, n > 0. \tag{2.7}$$

For any Markov chain, if the transition probability is independent of time, the Markov chain is stationary [23].

Definition 1. *Given two probabilities ν_1, ν_2 on the measure space $(\mathcal{Z}, \mathcal{S})$, we define $\|\nu_1 - \nu_2\|_{TV} = \sup_{A\in\mathcal{S}} |\nu_1(A) - \nu_2(A)|$ as the total variation distance between the measures ν_1 and ν_2. Meyn and Tweedie [24] pointed out that for a Markov chain $\{Z_i\}_{i\geq 1}$, if there are $0 < \gamma_0 < \infty$ and $0 < \rho_0 < 1$,*

$$\left\|P^k(\cdot \mid z) - \pi(\cdot)\right\|_{TV} \leq \gamma_0 \rho_0^k, \ \forall k \geq 1, \ k \in \mathbb{N}, \tag{2.8}$$

where $\pi(\cdot)$ is the stationary distribution of $\{Z_i\}_{i\geq 1}$, then $\{Z_i\}_{i\geq 1}$ is a ueMC.

Another equivalent definition of ueMC is the following Doeblin condition [25].

Proposition 1. *(Doeblin condition): Suppose $\{Z_i\}_{i\geq 1}$ is a Markov chain with transition probability $P^n(\cdot \mid \cdot)$, and μ is a specific non-negative metric with nonzero mass μ_0. If there is some integer k such that $\forall z \in \mathcal{Z}$ and for all measurable sets A, $P^k(A \mid z) \leq \mu(A)$, then for any integer n, $\forall z_1, z_2 \in \mathcal{Z}$, we have*

$$\left\|P^k(\cdot \mid z_1) - P^k(\cdot \mid z_2)\right\|_{TV} \leq 2\beta_1^{n/k}, \tag{2.9}$$

where $\beta_1 = 1 - \mu_0$.

3. Bounds of generalization ability

In this section, we estimate the bounds on the generalization performance of the logistic regression model based on the ueMC sampling by following the enlightening ideas of [26].

3.1. Bounds of generalization ability

To measure the generalization ability of $f_{z,\lambda}$, we identify how $\operatorname{sgn}(f_{z,\lambda})$ converges (with respect to the misclassification error) to the best classifier as $C(m)$ tends to infinity. Recall the regression function of φ, $f_\varphi = \int_y y d\varphi(y|x)$, $x \in \mathcal{X}$. Then the Bayes rule is given by the sign of the regression function $f_c = \operatorname{sgn}(f_\varphi)$. Referring to Vapnik [27], the speed at which $f_{z,\lambda}$ approaches f_φ is measured by excess generalization error $\mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\varphi)$. Since the minimization (2.6) is taken over the discrete quantity $\mathcal{E}_z(f)$, we have to regulate the capacity of the function set. Here the capacity is measured by the covering number.

We should estimate the excess misclassification error $\mathcal{R}(\operatorname{sgn}(f_{z,\lambda})) - \mathcal{R}(f_c)$ to bound the generalization ability of $f_{z,\lambda}$. The relation between excess misclassification error and excess generalization error $\mathcal{E}(f) - \mathcal{E}\left(f_\varphi\right)$ for convex loss is for $f : \mathcal{X} \to \mathbb{R}$,

$$\mathcal{R}(\operatorname{sgn}(f)) - \mathcal{R}(f_c) \leq \mathcal{E}(f) - \mathcal{E}\left(f_\varphi\right). \tag{3.1}$$

Definition 2. *Let $U > 0$, and $\mathcal{B}_U = \{f : f \in \mathcal{H}_K, \|f\|_{\mathcal{H}_K} \leq U\}$ be the sphere with radius U in $\mathcal{H}_K$, let $\mathcal{N}(\epsilon) = \mathcal{N}(\mathcal{B}_U, \epsilon)$, $\epsilon > 0$ be the covering number of $\mathcal{B}_U$.*

Definition 3. *The complexity index of $\mathcal{H}_K$ is s, if there is some $C_s > 0$ such that $\forall \epsilon > 0$,*

$$\ln \mathcal{N}(\epsilon) \leq C_s (1/\epsilon)^s.$$

Proposition 2. *Let $f_{z,\lambda}$ be the function defined by (2.6), and define $f_\lambda := \arg\min_{f\in\mathcal{H}_K}\{\lambda\|f\|^2_{\mathcal{H}_K} + \mathcal{E}(f)\}$ as a regularizing function, then for any $\lambda > 0$,*

$$\mathcal{E}(f_{z,\lambda}) - \mathcal{E}\left(f_\varphi\right) \leq R_S + R_\lambda, \tag{3.2}$$

where sample error:

$$R_S = \left(\mathcal{E}(f_{z,\lambda}) - \mathcal{E}_z(f_{z,\lambda}) + \mathcal{E}_z(f_\lambda) - \mathcal{E}(f_\lambda)\right),$$

and regularization error:

$$R_\lambda = \left(\mathcal{E}(f_\lambda) - \mathcal{E}\left(f_\varphi\right) + \lambda \|f_\lambda\|^2_{\mathcal{H}_K}\right).$$

Proof. According to the definition of $f_{z,\lambda}$,

$$\lambda \left\|f_{z,\lambda}\right\|_{\mathcal{H}_K}^2 + \mathcal{E}_z\left(f_{z,\lambda}\right) \leq \lambda \left\|f_\lambda\right\|_{\mathcal{H}_K}^2 + \mathcal{E}_z\left(f_\lambda\right),$$

we have

$$\mathcal{E}_z\left(f_\lambda\right) - \mathcal{E}_z\left(f_{z,\lambda}\right) + \lambda \left\|f_\lambda\right\|_{H_K}^2 \geq \lambda \left\|f_{z,\lambda}\right\|_{\mathcal{H}_K}^2 \geq 0.$$

So

$$\begin{aligned}&\mathcal{E}\left(f_{z,\lambda}\right) - \mathcal{E}\left(f_\varphi\right)\\ \leq&\mathcal{E}\left(f_{z,\lambda}\right) - \mathcal{E}\left(f_\varphi\right) + \mathcal{E}_z\left(f_\lambda\right) - \mathcal{E}_z\left(f_{z,\lambda}\right) + \lambda \left\|f_\lambda\right\|_{\mathcal{H}_K}^2\\ =&\left(\mathcal{E}\left(f_{z,\lambda}\right) - \mathcal{E}_z\left(f_{z,\lambda}\right) + \mathcal{E}_z\left(f_\lambda\right) - \mathcal{E}\left(f_\lambda\right)\right) + \left(\mathcal{E}\left(f_\lambda\right) - \mathcal{E}\left(f_\varphi\right) + \lambda \left\|f_\lambda\right\|_{\mathcal{H}_K}^2\right).\end{aligned}$$

In Proposition 2, sample error

$$R_S = R_{S_1} + R_{S_2} = \left\{E\xi_1 - \frac{1}{m}\sum_{i=1}^m \xi_1\left(z_i\right)\right\} + \left\{\frac{1}{m}\sum_{i=1}^m \xi_2\left(z_i\right) - E\xi_2\right\},$$

where $\xi_1 = \ell\left(f_{z,\lambda}, z\right) - \ell(f_\varphi, z)$ and $\xi_2 = \ell\left(f_\lambda, z\right) - \ell(f_\varphi, z)$. □

Definition 4. *We say the function f_φ can be approximated by $\mathcal{H}_K$ with exponent $0 < \beta \leq 1$, if there exists a constant C_β such that for any $\lambda > 0, R_\lambda \leq C_\beta \lambda^\beta$. In the paper, we assume that there is a constant B such that $\left|f_\varphi\right| \leq B$.*

3.2. Main tools

To prove the main results, there are four lemmas.

Lemma 1. *Let f be a continuous function defined on $\mathcal{X}$ and $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. Let $\kappa = \sup_{x\in\mathcal{X}} \sqrt{K\langle x, x\rangle}$, then $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)| \leq \kappa\|f\|_{\mathcal{H}_K}, \forall f \in \mathcal{H}_K$.*

Proof. According to the Cauchy-Schwartz inequality of PSD kernel and the reproducing property of reproducing kernel Hilbert space, we have:

$$f^2(x) \leq \langle f, f\rangle \cdot K\langle x, x\rangle,$$

it follows that

$$\sup_{x\in\mathcal{X}} |f(x)| \leq \kappa\|f\|_{\mathcal{H}_K}, \forall f \in \mathcal{H}_K.$$

□

Lemma 2. *$\xi_2 = \ell\left(f_\lambda, z\right) - \ell(f_\varphi, z)$, $|f_\varphi| \leq B$, we have*

$$|\xi_2| \leq d := \kappa\sqrt{R_\lambda/\lambda} + B + \ln\left(1 + e^{\kappa\sqrt{R_\lambda/\lambda}}\right) + \ln(1 + e^B). \tag{3.3}$$

Proof. Due to

$$R_\lambda = \mathcal{E}\left(f_\lambda\right) - \mathcal{E}\left(f_\varphi\right) + \lambda \left\|f_\lambda\right\|_{\mathcal{H}_K}^2 \geq \lambda \left\|f_\lambda\right\|_{\mathcal{H}_K}^2,$$

then $\|f_\lambda\|_{\mathcal{H}_K} \le \sqrt{R_\lambda/\lambda}$. By applying Lemma 1, there holds $\|f_\lambda\|_\infty \le \kappa\sqrt{R_\lambda/\lambda}$ as $\xi_2 = \ell(f_\lambda, z) - \ell(f_\varphi, z)$ and $|f_\varphi| \le B$, and when $y = 1$,

$$\xi_2 = \ln(\text{sigmod}(f_\lambda)) - \ln(\text{sigmod}(f_\varphi)) = f_\varphi - f_\lambda + \ln(1 + e^{f_\lambda}) - \ln(1 + e^{f_\varphi});$$

when $y = 0$,

$$\xi_2 = \ln(1 - \text{sigmod}(f_\lambda)) - \ln(1 - \text{sigmod}(f_\varphi)) = \ln(1 + e^{f_\lambda}) - \ln(1 + e^{f_\varphi}).$$

Then, the proof ends with $|\xi_2| = |\ell(f_\lambda, z) - \ell(f_\varphi, z)| \le |f_\lambda| + |f_\varphi| + |\ln(1 + e^{f_\varphi})| + |\ln(1 + e^{f_\lambda})|$. □

Lemma 3. *(proved in [21]): Hoeffding's inequality: Let $Z = \{Z_i\}_{i=1}^m$ be a ueMC sample and $\mathcal{F}$ be a set of bounded and measurable functions, i.e. there is a constant C, such that $\forall z \in \mathcal{Z}$, $\forall g \in \mathcal{F}$, $0 \le g(z) \le C$. For any $\varepsilon > 0$, we have*

$$\Pr\left\{\left|\frac{1}{m}\sum_{i=1}^m g(z_i) - E(g)\right| \ge \varepsilon\right\} \le 2\exp\left\{\frac{-m\varepsilon^2}{56C\,\|\Gamma_0\|^2\,E(g)}\right\}, \tag{3.4}$$

where $\Gamma_0 = \sqrt{2}/(1-\beta^{1/2t})$ and $\beta = 1 - \mu_0$. Here μ_0 and t are from Doeblin condition [25].

According to Lemma 3, for any $\varepsilon > 0$,

$$\Pr\left\{\frac{\frac{1}{m}\sum_{i=1}^m \xi_2(z_i) - E(\xi_2)}{\sqrt{E(\xi_2) + \varepsilon}} \ge \sqrt{\varepsilon}\right\} \le \exp\left\{\frac{-\varepsilon m}{56\,\|\Gamma_0\|^2\,d}\right\}. \tag{3.5}$$

Let $\delta_1 = \exp\left\{\frac{-\varepsilon m}{56\|\Gamma_0\|^2 d}\right\}$, so $\varepsilon = \frac{-56\|\Gamma_0\|^2 d\ln(\delta_1)}{m}$ and $\sqrt{\varepsilon}\sqrt{\varepsilon(f) + \varepsilon} \le \frac{1}{2}\varepsilon(f) + \varepsilon$. Then for any $0 < \delta_1 < 1$, we obtain

$$R_{S_2} = \frac{1}{m}\sum_{i=1}^m \xi_2(z_i) - E\xi_2 \le \frac{1}{2}R_\lambda - \frac{56\ln(\delta_1)\,d\,\|\Gamma_0\|^2}{m}, \tag{3.6}$$

with probability at least $1 - \delta_1$.

Let $U > 0$ and $\mathcal{F}_U = \{g = \ell(f, z) - \ell(f_\varphi, z), f \in \mathcal{B}_U\}$, we have

$$E(g) = \mathcal{E}(f) - \mathcal{E}(f_\varphi) \ge 0, \quad \frac{1}{m}\sum_{i=1}^m g(z_i) = \mathcal{E}_z(f) - \mathcal{E}_z(f_\varphi).$$

From the definition of $\mathcal{B}_U$, it can be seen that for any $f \in \mathcal{B}_U$, there are $\|f\|_\infty \le \kappa\|f\|_{\mathcal{H}_K} \le \kappa U$ and then $\left|f_\varphi\right| \le B$. It follows that

$$|g(z)| \le b := \kappa U + B + \ln(1 + e^{\kappa U}) + \ln(1 + e^B).$$

Under the condition of Lemma 3, for any $\epsilon > 0$

$$\begin{aligned} &\Pr\left\{\sup_{f\in\mathcal{B}_U}\frac{(\mathcal{E}(f) - \mathcal{E}_z(f)) - (\mathcal{E}(f_\varphi) - \mathcal{E}_z(f_\varphi))}{\sqrt{\mathcal{E}(f) - \mathcal{E}(f_\varphi) + \varepsilon}} \ge \sqrt{\varepsilon}\right\} \\ =&\Pr\left\{\sup_{g\in\mathcal{F}_U}\frac{E(g) - \frac{1}{m}\sum_{i=1}^m g(z_i)}{\sqrt{E(g) + \varepsilon}} \ge \sqrt{\varepsilon}\right\} \\ \le&\mathcal{N}(\mathcal{F}_U, \epsilon)\exp\left\{\frac{-\varepsilon m}{56b\,\|\Gamma_0\|^2}\right\}. \end{aligned} \tag{3.7}$$

For any $g_1, g_2 \in \mathcal{F}_U$, $|\ln(1+e^{x_1}) - \ln(1+e^{x_2})| \leq |x_1 - x_2|$, it follows that $|g_1(x) - g_2(x)| \leq \|f_1 - f_2\|_\infty$.

According to inequality (3.7), for any $\varepsilon > 0$

$$\Pr\left\{\sup_{f\in\mathcal{B}_U} \frac{(\mathcal{E}(f) - \mathcal{E}_z(f)) - (\mathcal{E}(f_\varphi) - \mathcal{E}_z(f_\varphi))}{\sqrt{\mathcal{E}(f) - \mathcal{E}(f_\varphi) + \varepsilon}} \geq \sqrt{\varepsilon}\right\} \leq \mathcal{N}\left(\frac{\varepsilon}{U}\right)\exp\left\{\frac{-\varepsilon m}{56b\,\|\Gamma_0\|^2}\right\}. \tag{3.8}$$

Therefore, according to $\mathcal{N}\left(\frac{\varepsilon}{U}\right) \leq \exp\left\{C_s\left(\frac{U}{\varepsilon}\right)^s\right\}$ by Definition 3, there holds, for $f_{z,\lambda}$,

$$\Pr\left\{\sup_{f\in\mathcal{B}_U} \frac{(\mathcal{E}(f_{z,\lambda}) - \mathcal{E}_z(f_{z,\lambda})) - (\mathcal{E}(f_\varphi) - \mathcal{E}_z(f_\varphi))}{\sqrt{\mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\varphi) + \varepsilon}} \geq \sqrt{\varepsilon}\right\} \leq \exp\left\{C_s\left(\frac{U}{\varepsilon}\right)^s - \frac{\varepsilon m}{56b\,\|\Gamma_0\|^2}\right\}. \tag{3.9}$$

In the (3.9), let

$$\exp\left\{C_s\left(\frac{U}{\varepsilon}\right)^s - \frac{\varepsilon m}{56b\,\|\Gamma_0\|^2}\right\} = \delta_2,$$

so

$$\frac{\varepsilon m}{56b\,\|\Gamma_0\|^2} - C_s\left(\frac{U}{\varepsilon}\right)^s = -\ln(\delta_2),$$

it is obtained that

$$\varepsilon^{s+1} - \frac{C_s U^s 56b\,\|\Gamma_0\|^2}{m} + \frac{\varepsilon^s \ln(\delta_2)\,56b\,\|\Gamma_0\|^2}{m} = 0. \tag{3.10}$$

Lemma 4. *(Lemma 4 in [28]) Let $c_1, c_2 > 0$, and $p_1 > p_2 > 0$, then the equation $x^{p_1} - c_1 x^{p_2} - c_2 = 0$ has a unique positive zero x^*. In addition, $x^* \leq \max\left\{(2c_1)^{1/(p_1-p_2)},\ (2c_2)^{1/p_1}\right\}$.*

3.3. Main results

Through the above inference process, we form the following main results.

Theorem 1. *Let $Z = \{Z_i\}_{i=1}^m$ be a ueMC sample, then for any $0 < \delta < 1$,*

$$\mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\varphi) + 2\lambda\left\|f_{z,\lambda}\right\|^2_{\mathcal{H}_K} \leq 3R_\lambda + 2\hat{\mathcal{E}} + \frac{112d\,\|\Gamma_0\|^2\ln(1/\delta)}{m} \tag{3.11}$$

holds true with probability at least $1-\delta$, where

$$\hat{\varepsilon} \leq \max\{\varepsilon_1, \varepsilon_2\},\ \|\Gamma_0\|^2 = \sqrt{2}/\left(1 - \beta_1^{1/2t}\right),$$

$$\varepsilon_1 = \frac{112b\,\|\Gamma_0\|^2\ln(1/\delta)}{m},\quad \varepsilon_2 = \left[\frac{112C_sU^sb\,\|\Gamma_0\|^2}{m}\right]^{1/(1+s)},$$

here β_1 and t are defined as that in Proposition 1.

Proof. According to Lemma 4, Eq (3.10) has a solution $\hat{\varepsilon} = \max\{\varepsilon_1, \varepsilon_2\}$ with

$$\varepsilon_1 = -\frac{112b\,\|\Gamma_0\|^2\ln(\delta_2)}{m},\quad \varepsilon_2 = \left[\frac{112C_sU^sb\,\|\Gamma_0\|^2}{m}\right]^{1/(1+s)}.$$

Since $\sqrt{\varepsilon}\sqrt{\varepsilon(f)+\varepsilon} \leq \frac{1}{2}\varepsilon(f)+\varepsilon$, in combination with inequality (3.9), inequality (3.12) holds at least with a probability of $1-\delta_2$,

$$R_{S_1} = E\xi_1 - \frac{1}{m}\sum_{i=1}^{m}\xi_1(z_i) \leq \frac{1}{2}\left(\mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\varphi)\right) + \hat{\varepsilon}. \tag{3.12}$$

Combining inequality (3.6) with inequality (3.12), inequality (3.13) holds at least $1-\delta$ probability for any $0<\delta<1$,

$$\mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\varphi) \leq 3R_\lambda + 2\hat{\varepsilon} - 2\lambda\left\|f_{z,\lambda}\right\|_{\mathcal{H}_K}^2 - \frac{112d\,\|\Gamma_0\|^2\ln(\delta)}{m}. \tag{3.13}$$

□

Theorem 2. *Let $Z=\{Z_i\}_{i=1}^m$ be a ueMC sample, Taking $\lambda=(1/m)^{\vartheta}$. For any $\epsilon>0$ and $0<\delta<1$, there exists a constant $\hat{C}$ independent of m such that*

$$\mathcal{R}(\text{sgn}(f_{z,\lambda})) - \mathcal{R}(f_c) \leq \hat{C}(1/m)^{\theta}, \tag{3.14}$$

holds true with probability at least $1-\delta$, providing $m \geq 112b\,\|\Gamma_0\|^2\ln(1/\delta)\left(\ln(1/\delta)/C_s\right)^{1/s}/U$, where

$$\vartheta = \min\left\{\frac{2}{\beta+1}, \frac{2}{(1+\beta)(1+s)}\right\},$$
$$\theta = \min\left\{\frac{2\beta}{\beta+1}, \frac{2\beta}{(1+\beta)(1+s)} - \epsilon\right\}.$$

Proof. The proof is easily obtained from Theorem 1 with the proof of Theorem 2 in [21]. □

4. Markov sampling and numerical studies

In this section, we introduce a ueMC algorithm to generate the samples from a given dataset. We give numerical studies on the learning performance of the logistic regression model and present some useful discussions.

4.1. Markov sampling algorithm

Here are the notations:

- S_T: the initial training set
- S_{iid}: a i.i.d sample from S_T
- S_{mkv}: a ueMC sample from S_T
- N_T: the number of positive samples in S_{mkv}
- N_F: the number of negative samples in S_{mkv}
- k, m, N, n, a, $max_$iteration: super-parameters

The pseudocode of the ueMC algorithm is as follows:

```
The ueMC algorithm: Markov sampling for Logistic Regression
Input: $S_T, m, k, N, n$, max_iteration, $a$.
Output: $f_k$.
Step 1: Set $i = 0$, draw randomly $m$ samples from $S_T$ called $S_{iid}$, and train $S_{iid}$, get the premilinary mode $f_i$.
Step 2: Let $S_{mkv} = \emptyset$, $N_T = 0$, $N_F = 0$, $t := 1$.
Step 3: Draw randomly a sample $z_t$ from $S_T$, let $S_{mkv} = S_{mkv} \cup z_t$, $N_T = N_T + 1$, if $y_t = 1$; $N_F = N_F + 1$, if $y_t = 0$. set $j := 0$.
Step 4: Draw randomly another sample $z_{\text{candidate}}$ from $S_T$, calculate $p_1 = e^{-\ell(f_i,\ z_{\text{cantididete}})}/e^{-\ell(f_i,z_t)}$, $j := j + 1$.
Step 5: Draw randomly a number $p_{\text{random}}$ from $U(0, 1)$.
Step 6: If $1 > p_1 > p_{\text{random}}$, accept $z_{\text{candidate}}$ with $p_1$; If $p_1 \geq 1$ and $y_t y_{\text{candidate}} = 1$, calculate $p_2 = e^{-y_{\text{candidate}} f_i}/e^{-y_t f_i}$, if $p_2 > p_{\text{random}}$, accept $z_{\text{candidate}}$ with $p_2$; If $p_1 \geq 1$ and $y_t y_{\text{candidate}} = -1$, accept $z_{\text{candidate}}$ with $p_1$; If $n$ samples are rejected continuously, accept $z_{\text{candidate}}$ with $p_3 = ap_1$.
Step 7: If $y_{\text{candidate}} = 1$ and $N_T < \frac{N}{2}$, let $N_T = N_T + 1, S_{mkv} = S_{mkv} \cup z_{\text{candidate}}$; If $y_{\text{candidate}} = -1$ and $N_F < \frac{N}{2}$, let $N_F = N_F + 1, S_{mkv} = S_{mkv} \cup z_{\text{candidate}}$.
Step 8: If $j >$ max_iteration or $N_T + N_F > N$, train $S_{mkv}$ to get $f_{i+1}$, else go to Step 4.
Step 9: If $i < k$, let $i := i + 1$ and go to Step 2, else output $f_{i+1}$.
```

Compared with data noise, the over-fitting problem caused by small sample size has a stronger impact on the generalization performance of the classifier ([29]). Different from the methods from [11, 19] based on threshold to eliminate noise data, the ueMC is constructed to avoid the problem of over-fitting caused by small training samples.

The ueMC algorithm uses the initial model f_0 of S_{iid} and the loss function $\ell(f, z)$ to construct the transition probability p_1, p_2, p_3 of ueMC. Since p_1, p_2, p_3 is greater than 0 and the sample size in S_T is limited, S_{mkv} obtained by the ueMC algorithm is a ueMC, according to the research conclusion of [30].

Different from the MCMC algorithm proposed by [31, 32], the ueMC algorithm does not need to know the probability distribution information of the training set samples. In addition, when $k = 0$, the algorithm degenerates into the classical logistic regression model. In order to make the following experimental results without loss of generality, we take $k = 2$ and $a = 1.2$.

4.2. Experiment results

In this subsection, we present the numerical study on the learning performance of the logistic regression model based on linear prediction models for 9 real-world datasets (from http://archive.ics.uci.edu/ml/datasets and https://www.fml.tuebingen.mpg.de/). The information of these data sets is summarized in Table 1, and all these datasets are 2-classes realworld datasets. The samples in these datasets obey non-independent identical distribution. We use the SMOTE algorithm to deal with unbalanced data. For each data set, it is randomly divided into two parts: the training set and the test set.

Table 1. General Information about 9 Real-world Datasets.

Dataset	Training Size	Test Size	Input Dimension
UCI- Heart Disease	800	225	13
UCI-Skin	160000	85057	3
UCI-HTRU2	20000	12518	8
UCI-Wine	1200	399	11
UCI-Diabetes	600	168	8
UCI-Waveform	4000	1000	21
MNIST	60000	10000	784
ELEC2	35000	10312	6
Splice	20000	43500	60

In order to simplify the experimental process, we take $N = m$ in the ueMC algorithm , and carry out 50 repeated experiments for each dataset. The experimental results are shown in Table 2, where "MR (i.i.d.)" and "MR (Markov)" denote the misclassification rates of the logistic regression model based on random sampling and Markov sampling, respectively. In the following, we also discuss the experimental results based on $N < m$.

Table 2. Misclassification Rates (%) for 500 Training Samples.

Dataset	MR (i.i.d.)	MR (Markov)
UCI- Heart Disease	20.94 ± 0.98	**19.90** ± 1.07
UCI-Skin	8.92 ± 0.48	**8.86** ± 0.54
UCI-HTRU2	17.38 ± 1.71	**14.65** ± **1.45**
UCI-Wine	24.91 ± 1.40	**24.34** ± **1.38**
UCI-Diabetes	35.25 ± 1.61	**30.00** ± **1.22**
UCI-Waveform	17.42 ± 2.19	**12.48** ± **1.67**
MNIST	0.51 ± 0.10	**0.49** ± **0.09**
ELEC2	26.83 ± 0.68	**25.89** ± **0.36**
Splice	12.85 ± 5.12	**9.30** ± **3.77**

From Table 2, we can find that for 500 training samples, the standard deviations and means of average misclassification rates of the logistic regression model based on Markov sampling are smaller than that of random sampling except UCI-Heart Disease and UCI-Skin, and the means of average misclassification rates based on Markov sampling in UCI-Heart Disease and UCI-Skin datasets are still smaller than that of random samples. To show the learning performance of the logistic regression model based on Markov sampling, we present the average misclassification rates for 50 experimental results of the logistic regression model based on Markov sampling (non-i.i.d) and random sampling (i.i.d.) for different training sizes and four datasets in Figures 1–4.

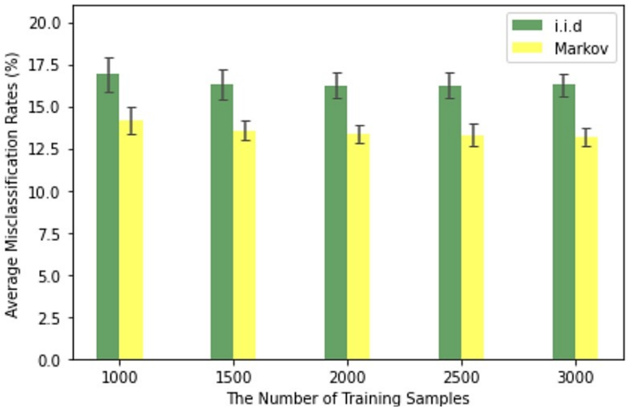

Figure 1. Average misclassification rates for UCI-HTRU2 and m = 1000, 1500, 2000, 2500, 3000.

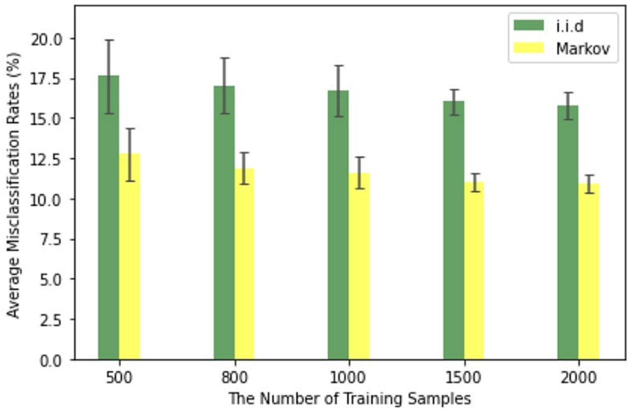

Figure 2. Average misclassification rates for UCI-Waveform and m = 500, 800, 1000, 1500, 2000.

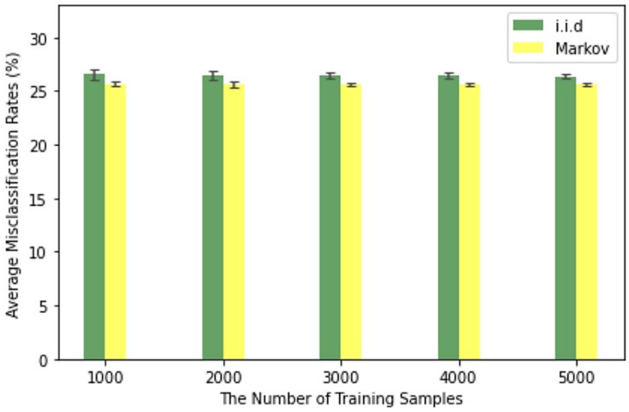

Figure 3. Average misclassification rates for ELEC2 and m = 1000, 2000, 3000, 4000, 5000.

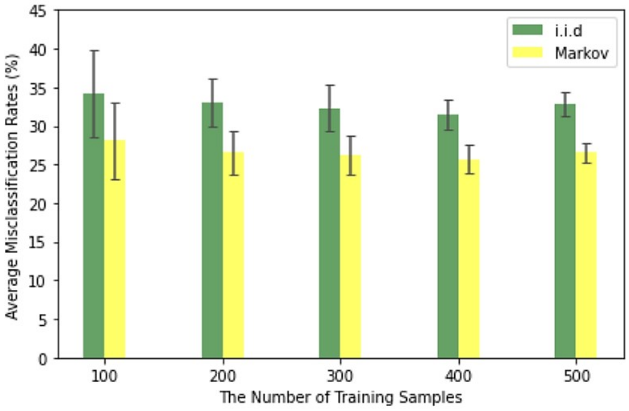

Figure 4. Average misclassification rates for UCI-Diabetes and m = 100, 200, 300, 400, 500.

To have a better understanding of learning performance of the logistic regression model based on Markov sampling, the following figures are presented to show the 50 experimental misclassification rates of the logistic regression model based on Markov sampling.

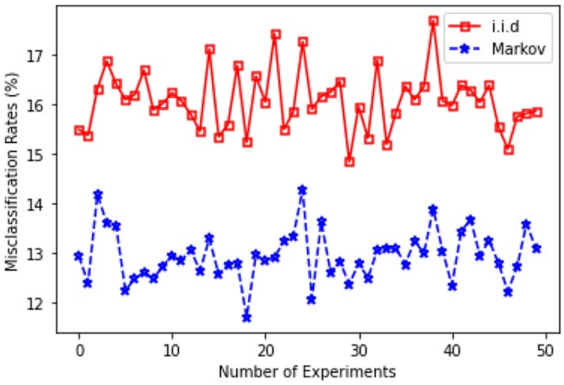

Figure 5. UCI-HTRU2, m = 3000.

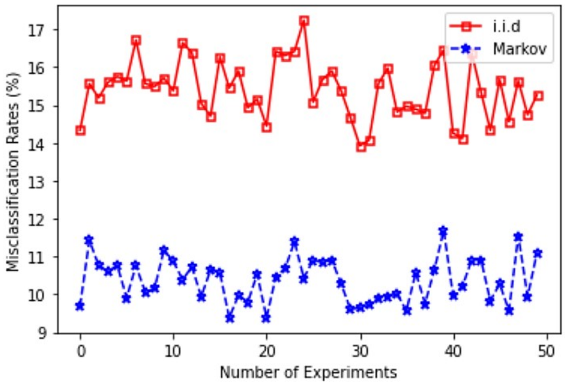

Figure 6. UCI-Waveform, m = 2000.

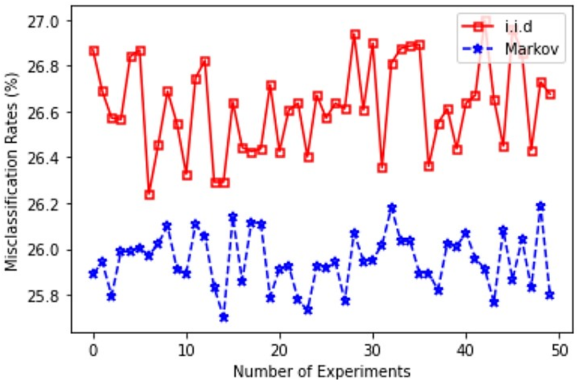

Figure 7. ELEC2, m = 5000.

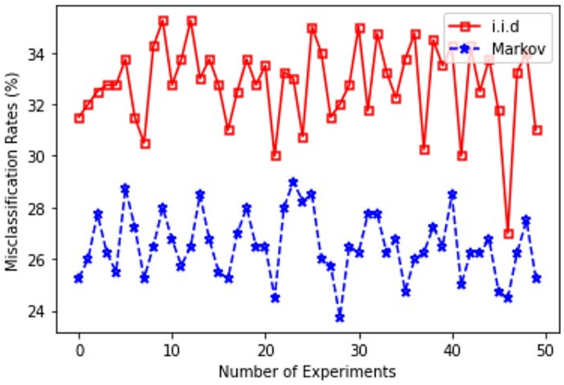

Figure 8. UCI-Diabetes, m = 500.

Figures 1–8 show that for UCI-HTRU2, UCI-Waveform, ELEC2 and UCI-Diabetes, the logistic regression model based on Markov sampling would have better learning performance than that of random sampling as the number of training samples is large. For other datasets, since the experimental results are similar, we do not present all of them here.

Table 3. Misclassification Rates for Different Training Sizes.

Dataset	i.i.d. (2000)	Markov (600)	Markov (800)	Markov (1000)
UCI-HTRU2	16.18 ± 0.94	9.54 ± 0.55	10.23 ± 0.47	11.10 ± 0.65
ELEC2	26.94 ± 0.28	25.53 ± 0.21	25.55 ± 0.20	25.49 ± 0.25
UCI-Waveform	15.41 ± 0.76	8.48 ± 0.42	8.42 ± 0.42	8.28 ± 0.33

For the case of $N < m$, Table 3 shows that for the datasets of UCI-HTRU2, ELEC2 and UCI-Waveform, the logistic regression model based on smaller Markov chain samples (600 samples for UCI-HTRU2, ELEC2 and UCI-Waveform) can present smaller misclassification rates compared to more i.i.d. samples (2000 samples) .

We compare the performance of the logistic regression model with the ueMC samples with the classical logistic regression model, SVMC, Adaboost, RandomForest and other classical machine learning models on UCI-HTRU2, UCI-Skin and MNIST. Tables 4 and 5 show the average classification error rates and Wilkeson signed rank test results of 50 experiments.

Table 4. Classification Error Rates Comparison Results.

Dataset	Markov_Logistic	Classical_Logistic	SVM	RandomForest	Adaboost
MNIST	**0.4054**	0.5201	0.6374	0.7292	0.4657
SKIN	6.6451	9.2786	7.6901	3.1789	4.4347
HTRU2	**10.2353**	14.5536	13.5935	11.6205	11.9470

Table 5. Wilkerson signed rank test results of classification error rate.

Comparison	Statistic		P-value	Optimal
Markov_Logistic vs. Classical_Logistic	MNIST	109.0000	6.1×10^{-04}	Markov_Logistic
	SKIN	0.0000	7.5×10^{-10}	Markov_Logistic
	HTRU2	0.0000	7.5×10^{-10}	Markov_Logistic
Markov_Logistic vs. RandomForest	MNIST	0.0000	7.5×10^{-10}	Markov_Logistic
	SKIN	0.0000	7.5×10^{-10}	RandomForest
	HTRU2	0.0000	7.5×10^{-10}	Markov_Logistic
Markov_Logistic vs. Adaboost	MNIST	302.5101	5.3×10^{-04}	Markov_Logistic
	SKIN	0.0000	7.5×10^{-10}	Adaboost
	HTRU2	0.0000	7.5×10^{-10}	Markov_Logistic

On the three real data sets, the difference between the logistic regression model based on the ueMC sampling and classical logistic regression models in classification error rates is statistically significant. It can be concluded that: 1) The generalization ability of the logistic regression model based on the ueMC sampling is better than that of the classical logistic regression model classifier, and it is robust. 2) The generalization ability of the logistic regression model with the ueMC samples is comparable to that of complex classifiers, such as random forest and Adaboost.

4.3. Explanation of learning performance

The ueMC algorithm divides the samples into three categories according to the model pre-trained in the previous step. The first one is the samples with correct classification and close to the decision boundary, the second one is the samples with correct classification but far away from the decision boundary, and the third one is the samples with wrong classification.

According to the definition of the loss function, the distance between the sample and the decision boundary determines the size of the loss. Samples close to the decision boundary would train a better decision boundary for classification. The ueMC algorithm designs the acceptance probability p_1 and p_2 which ensure that the samples obtained according to the acceptance probability are close to the decision boundary. Therefore, when the initial logistic regression model can better fit the data set, the ueMC algorithm can ensure that the ueMC samples are excellent samples. In addition, we do not directly eliminate the samples with classification errors or far away from the decision boundary, but accept them with a small probability, which not only ensures the excellent properties of the training set, but also maintains the diversity of the training set samples to a certain extent, and can reduce the error caused by the misjudgment of the pre-training model. Therefore, the learning performance of the logistic regression model based on Markov sampling is better than that of random sampling, and the classifier based on Markov sampling is more sparse compared to random sampling.

5. Conclusions

To study the generalization performance of the logistic regression model based on the ueMC sampling, inspired by the idea from [21], we estimate first the generalization error of logistic model algorithm based on the ueMC sampling. The generalization error is deconstructed into sample error and regularization error by error decomposition, and the convergence of the algorithm is proved. In addition, we also generate the samples from given dataset by the ueMC algorithm. The numerical studies show that as the number of training samples is large, the learning performance of the logistic regression model based on Markov sampling is better than that of random sampling, and its performance is also better than that of classical machine model algorithms, such as random forest and Adaboost. In other words, the ueMC algorithm significantly improves the learning performance of the logistic regression model. To our knowledge, this study is the first on this topic.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by the Innovation and Talent Base for Digital Technology and Finance (B21038) and "the Fundamental Research Funds for the Central Universities", Zhongnan University of Economics and Law (2722023EJ002).

Conflict of interest

The authors declare there is no conflicts of interest.

References

1. A. Bayaga, Multinomial logistic regression: Usage and application in risk analysis, *J. Appl. Quant. Methods*, **5** (2010), 288–297.
2. A. Selmoune, Z. Liu, J. Lee, To pay or not to pay? Understanding public acceptance of congestion pricing: A case study of Nanjing, *Electron. Res. Arch*, **30** (2022), 4136–4156. https://doi.org/10.3934/era.2022209
3. Z. Ahmad, Z. Almaspoor, F. Khan, S. E. Alhazmi, M. El-Morshedy, O. Y. Ababneh, et al., On fitting and forecasting the log-returns of cryptocurrency exchange rates using a new logistic model and machine learning algorithms, *AIMS Math.*, **7** (2022), 18031–18049. https://doi.org/10.3934/math.2022993

4. N. Dwarika, Asset pricing models in South Africa: A comparative of regression analysis and the Bayesian approach, *Data Sci. Financ. Econ.*, **3** (2023), 55–75. https://doi.org/10.3934/DSFE.2023004

5. D. McAllester, Generalization bounds and consistency, *Predicting Struct. Data*, 2007. https://doi.org/10.7551/mitpress/7443.003.0015

6. N. Kordzakhia, G. D. Mishra, L. Reiersølmoen, Robust estimation in the logistic regression model, *J. Stat. Plan. Infer.*, **98** (2001), 211–223. https://doi.org/10.1016/S0378-3758(00)00312-8

7. M. Rashid, *Inference on Logistic Regression Models*, Ph.D thesis, Bowling Green State University, 2008.

8. D. Dai, D. Wang, A generalized Liu-type estimator for logistic partial linear regression model with multicollinearity, *AIMS Math.*, **8** (2023), 11851–11874. https://doi.org/10.3934/math.2023600

9. Z. Wang, Z. Wang, B. Fu, Learning restricted bayesian network classifiers with mixed non-i.i.d. sampling, in *2010 IEEE International Conference on Data Mining Workshops*, (2010), 899–904. https://doi.org/10.1109/ICDMW.2010.199

10. H. Sun, Q. Wu, Least square regression with indefinite kernels and coefficient regularization, *Appl. Comput. Harmon A*, **30** (2011), 96–109 https://doi.org/10.1016/j.acha.2010.04.001

11. H. Sun, Q. Guo, Coefficient regularized regression with non-iid sampling, *Int. J. Comput. Math.*, **88** (2011), 3113–3124. https://doi.org/10.1080/00207160.2011.587511

12. X. Chu, H. Sun, Regularized least square regression with unbounded and dependent sampling, *Abstr. Appl. Anal.*, **2013** (2013), 900–914. https://doi.org/10.1155/2013/139318 .

13. Z. C. Guo, L. Shi, Learning with coefficient-based regularization and l1-penalty, *Adv. Comput. Math.*, **39** (2013), 493–510. https://doi.org/10.1007/s10444-012-9288-6

14. B. Jiang, Q. Sun, J. Q. Fan, Bernstein's inequality for general Markov chains, preprint, arXiv: 1805.10721.

15. D. S. Modha, E. Masry, Minimum complexity regression estimation with weakly dependent observations, *IEEE Trans. Inf. Theory*, **42** (1996), 2133–2145. https://doi.org/10.1109/18.556602

16. F. Merlevède, M. Peligrad, E. Rio, Bernstein inequality and moderate deviations under strong mixing conditions, *Inst. Math. Stat. (IMS) Collect.*, **2009** (2009), 273–292. https://doi.org/10.1214/09-IMSCOLL518

17. J. Q. Fan, B. Jiang, Q. Sun, Hoeffding's lemma for Markov Chains and its applications to statistical learning, preprint, arXiv:1802.00211.

18. P. J. M. Laarhoven, E. H. L. Aarts, *Simulated Annealing: Theory and Applications*, Springer, Dordrecht, 1987.

19. J. Thongkam, G. Xu, Y. Zhang, et.al., Support vector machine for outlier detection in breast cancer survivability prediction, in *Asia-Pacific Web Conference*, Springer, (2008), 99–109. https://doi.org/10.1007/978-3-540-89376-9_10

20. A. L. B. Miranda, L. P. F. Garcia, A. C. P. L. F. Carvalho, A. C. Lorena, Use of classification algorithms in noise detection and elimination, in *International Conference on Hybrid Artificial Intelligence Systems*, Springer, (2009), 417–424. https://doi.org/10.1007/978-3-642-02319-4_50

21. J. Xu, Y. Y. Tang, B. Zou, Z. Xu, L. Li, Y. Lu, et al., The generalization ability of SVM classification based on Markov sampling, *IEEE Trans. Cybern.*, **45** (2014), 1169–1179. https://doi.org/10.1109/TCYB.2014.2346536

22. J. D. Head, M. C. Zerner, A Broyden—Fletcher—Goldfarb—Shanno optimization procedure for molecular geometries, *Chem. Phys. Lett.*, **122** (1985), 264–270. https://doi.org/10.1016/0009-2614(85)80574-1

23. M. Vidyasagar, *Learning and Generalization: With Applications to Neural Networks*, Springer, London, 2003.

24. S. P. Meyn, R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer, Berlin, 2012.

25. P. Doukhan, *Mixing: Properties and Examples*, Springer, Berlin, 2012.

26. P. Zhang, N. Riedel, Discriminant analysis: A unified approach, in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, 2005. https://doi.org/10.1109/ICDM.2005.51

27. V. N. Vapnik, An overview of statistical learning theory, *IEEE T. Neur. Net. Lear.*, **10** (1999), 988–999. https://doi.org/10.1109/72.788640

28. F. Cucker, S. Smale, Best choices for regularization parameters in learning theory: On the bias-variance problem, *Found. Comput. Math.*, **2** (2002), 413–428. https://doi.org/10.1007/s102080010030

29. G. Stempfel, L. Ralaivola, Learning SVMs from sloppily labeled data, in *Lecture Notes in Computer Science*, Springer, 2009. http://dx.doi.org/10.1007/978-3-642-04274-4_91

30. M. P. Qian, G. L. Gong, *Applied random processes*, Peking University Press, Beijing, 1998.

31. W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, **57** (1970), 97–109. https://doi.org/10.1093/biomet/57.1.97

32. S. Geman S, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Mach. Intell.*, **6** (1984), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596