A Large Scale Analysis of Logistic Regression: Asymptotic Performance and New Insights

Xiaoyi Mai[1,2], Zhenyu Liao[1,2], Romain Couillet[1,2]

[1]LSS, CentraleSupélec, Université ParisSaclay, France.
[2]GIPSA-lab, Université Grenoble Alpes, France.

Abstract

- **Objective**: Quantitative study of logistic regression in modern regime of large dimensional, numerous data.
- **Approach**: Combine flexibility of "leave-one-out" approach for handling implicit solutions and that of random matrix theory (RMT) for structured data models.
- **Results**: Statistical distribution of learned parameters, dependent of training data.

Preliminaries

Logistic Regression:

- Assumption of logic model: for some data vector $\mathbf{x} \in \mathbb{R}^p$ with class label $y = \pm 1$, $\exists \beta_* \in \mathbb{R}^p$ such that

$$P(y|\mathbf{x}) = \sigma(y_i \beta_*^\mathsf{T} \mathbf{x}_i)$$

where $\sigma(t) = \frac{1}{1+e^{-t}}$.

- Method: find estimate $\hat\beta$ of β_* by maximizing posterior probability $P(y|\mathbf{x})$ over training data set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, i.e.,

$$\hat\beta = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \rho(y_i \beta^\mathsf{T} \mathbf{x}_i) \tag{1}$$

with $\rho(t) = \ln(1+e^{-t})$.

- Possibility of ill-defined (1): if $\exists \beta_s$ such that $y_i \beta_s^\mathsf{T} \mathbf{x}_i > 0$ for all i, then $\hat\beta = q\beta_s$ with $q = +\infty$.
- Regularized version:

$$\hat\beta = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \rho(y_i \beta^\mathsf{T} \mathbf{x}_i) + \frac{\lambda}{2}\|\beta\|^2, \quad \lambda > 0. \tag{2}$$

Logic Model under Normality:

- Gaussian mixture: $\mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{C})$ with balanced class priors.
- Verifying assumption:

$$P(y_i|\mathbf{x}_i) = \frac{P(y_i)P(\mathbf{x}_i|y_i)}{P(y_i)P(\mathbf{x}_i|y_i) + P(-y_i)P(\mathbf{x}_i|-y_i)} = \frac{1}{1+e^{2y_i\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\mathbf{x}_i}} = \sigma(y_i\beta_*^\mathsf{T}\mathbf{x}_i)$$

with $\beta_* = 2\mathbf{C}^{-1}\boldsymbol{\mu}$.

High dimensional setting:

- At arbitrarily large p, $n/p \to \xi > 0$.
- Non-trivial regime: $\|\boldsymbol{\mu}\| = O(1)$, $\|\mathbf{C}\| = O(1)$ & $\|\mathbf{C}^{-1}\| = O(1)$ w.r.t. p.

Technical Approach

Objective: asymptotic statistics of implicit solution

$$\lambda\hat\beta = \frac{1}{n}\sum_{i=1}^n c_i y_i \mathbf{x}_i, \quad c_i \equiv \psi(y_i \hat\beta^\mathsf{T}\mathbf{x}_i)$$

where $\psi(t) \equiv -\frac{\partial\rho(t)}{\partial t} = \frac{1}{1+e^t}$.

Main difficulty: intractable statistical behavior of c_i due to implicit dependence between $\hat\beta$ and $\mathbf{x}_i$.

Leave-one-out version:

$$\lambda\hat\beta_{-i} = \frac{1}{n}\sum_{j\neq i} \psi(y_j \hat\beta_{-i}^\mathsf{T}\mathbf{x}_j) y_j \mathbf{x}_j$$

- Tractable leave-one-out error $\hat\beta_{-i}^\mathsf{T}\mathbf{x}_i$ since $\hat\beta_{-i}$ independent of $\mathbf{x}_i$.
- Approximation of $\hat\beta$: $\|\hat\beta_{-i} - \hat\beta\| \to 0$, and $\hat\beta_{-i}^\mathsf{T}\mathbf{x}_j - \hat\beta^\mathsf{T}\mathbf{x}_j \to 0$ for $j \neq i$.

Technical Approach

Key ideas:

- Express c_i as function fo $\hat\beta_{-i}^\mathsf{T}\mathbf{x}_i$:

$$c_i = \psi\left(y_i\hat\beta^\mathsf{T}\mathbf{x}_i\right) \simeq \psi\left(y_i\hat\beta_{-i}^\mathsf{T}\mathbf{x}_i + \kappa c_i\right) \simeq \psi\left(\operatorname{prox}_\kappa(y_i\hat\beta_{-i}^\mathsf{T}\mathbf{x}_i)\right)$$

where $\operatorname{prox}_\kappa(t) = \operatorname{argmin}_{z\in\mathbb{R}}\{\kappa\rho(z) + (z-t)^2/2\}$ for some scalar κ determined by RMT.

- Demonstrate from

$$\lambda\hat\beta \simeq \frac{1}{n}\sum_{i=1}^n y_i\psi\left(\operatorname{prox}_\kappa(y_i\hat\beta_{-i}^\mathsf{T}\mathbf{x}_i)\right)\mathbf{x}_i \tag{3}$$

the asymptotic normality of $\hat\beta$ by CLT and quasi-independence between $\psi\left(\operatorname{prox}_\kappa(y_i\hat\beta_{-i}^\mathsf{T}\mathbf{x}_i)\right)\mathbf{x}_i$ and $\psi\left(\operatorname{prox}_\kappa(y_j\hat\beta_{-j}^\mathsf{T}\mathbf{x}_j)\right)\mathbf{x}_j$, $i \neq j$.

- Find statistical parameters (i.e., mean, covariance) of $\hat\beta$, by taking corresponding expectation on both sides of (3).

Main Results

Theorem (Distribution of $\hat\beta$) For $\hat\beta$ given by (2)

$$\|\hat\beta - \tilde\beta\| \to 0 \quad \text{where} \quad (\lambda\mathbf{I}_p + \tau\mathbf{C})\tilde\beta \sim \mathcal{N}(\eta\boldsymbol{\mu}, \gamma\mathbf{C}/n)$$

with $(\eta, \gamma, \tau) \in \mathbb{R}_+^3$ given by

$$\eta = \mathbb{E}[\psi(\operatorname{prox}_\kappa(r))], \quad \gamma = \mathbb{E}[\psi^2(\operatorname{prox}_\kappa(r))], \quad \tau = \frac{\mathbb{E}[\psi(\operatorname{prox}_\kappa(r))(m-r)]}{\sigma^2}$$

for some $r \sim \mathcal{N}(m, \sigma^2)$ with

$$m \equiv \eta\boldsymbol{\mu}^\mathsf{T}(\lambda\mathbf{I}_p + \tau\mathbf{C})^{-1}\boldsymbol{\mu}$$

$$\sigma^2 \equiv \eta^2\boldsymbol{\mu}^\mathsf{T}(\lambda\mathbf{I}_p + \tau\mathbf{C})^{-1}\mathbf{C}(\lambda\mathbf{I}_p + \tau\mathbf{C})^{-1}\boldsymbol{\mu} + \gamma\frac{1}{n}\operatorname{tr}\left[(\lambda\mathbf{I}_p + \tau\mathbf{C})^{-1}\mathbf{C}\right]^2.$$

Remarks:

- Test error: $P\left(y_i\beta_{-i}^\mathsf{T}\mathbf{x}_i < 0\right) - P(r < 0) \to 0$.
- Unregularized solutions unbiased in direction, but biased in scale.
- γ/η^2, indicator of variability of $\hat\beta$, is minimized at $\lambda = +\infty$.
- Special case with $\mathbf{C} = \mathbf{I}_p$: classification performance maximized at trivial solution with $\lambda = +\infty$, when $\hat\beta$ proportional to $\frac{1}{n}\sum_{i=1}^n y_i\mathbf{x}_i$.

Numerical Validation

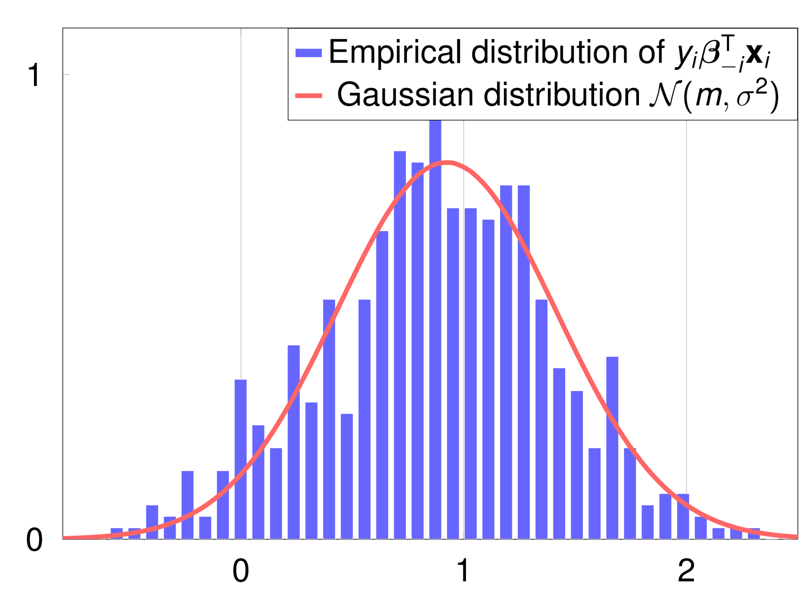

Figure: Comparison between $y_i\beta_{-i}^\mathsf{T}\mathbf{x}_i$ and a Gaussian distribution $\mathcal{N}(m, \sigma^2)$ with $\boldsymbol{\mu} = [2,\ \mathbf{0}_{p-1}]$, $\mathbf{C} = \mathbf{I}_p$, for $\lambda = 1$, $p = 256$ and $n = 512$.

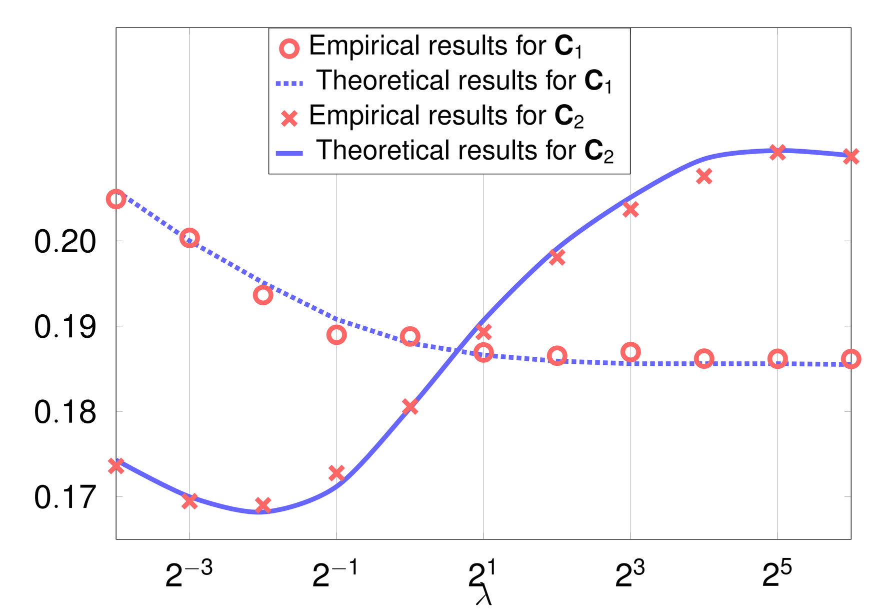

Figure: Misclassification error as a function of λ, with $\boldsymbol{\mu} = [1,\ 1,\ \mathbf{0}_{p-2}]$, $\mathbf{C}_1 = 2\mathbf{I}_p$ and $\mathbf{C}_2 = \operatorname{diag}[1,\ 5,\ \mathbf{1}_{p-2}]$, where $p = 128$, $n = 512$ and with number of test samples $n_{\text{test}} = 512$. Empirical results obtained by averaging over 500 runs.

Email: {xiaoyi.mai,zhenyu.liao,romain.couillet}@l2s.centralesupelec.fr