Model Constraints Independent Optimal Subsampling Probabilities for Softmax Regression

Yaqiong Yao[1], Jiahui Zou[2], and HaiYing Wang[1]

[1]University of Connecticut

[2]Capital University of Economics and Business

Introduction

Softmax regression, which is also called multi-class logistic regression, measuring the association between the multi-level categorical response and the covariates, plays an important role in many fields. Optimal subsampling is an effective way to reduce the computational burden for massive datasets. The goals of our work are to

- construct optimal subsampling probabilities under the summation constraint;
- compare optimal subsampling probabilities under the baseline constraint and the summation constraint;
- derive one kind of optimal subsampling probabilities that are invariant to model constraints.

Softmax Regression

Softmax regression is used as a classification model to measure the relationship between the categorical response with multiple levels and the covariates.

- Consider dataset $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ where
  - $\{\mathbf{x}_i\}_{i=1}^N$ are d dimensional covariates;
  - $\{\mathbf{y}_i\}_{i=1}^N$ are $K+1$ dimensional multivariate responses with $y_{i,k} = 1$ if the k-th category occurs for $k \in \{0, 1, ..., K\}$ and $\sum_{k=0}^K y_{i,k} = 1$.
- Given $\mathbf{x}_i$, suppose y_i follows a softmax regression,

$$\mathbb{P}(y_i = c_k|\mathbf{x}_i) = p_k(\mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k)}{\sum_{l=0}^K \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_l)}, \tag{1}$$

  where $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}, ..., \boldsymbol{\beta}_K^{\mathrm{T}})^{\mathrm{T}}$ is a Kd dimensional unknown parameter. The mean response vector is denoted as $\boldsymbol{p}_i(\boldsymbol{\beta}) = \{p_0(\mathbf{x}_i, \boldsymbol{\beta}), p_1(\mathbf{x}_i, \boldsymbol{\beta}), ..., p_K(\mathbf{x}_i, \boldsymbol{\beta})\}^{\mathrm{T}}$, that is $\mathbb{E}(\mathbf{y}_i|\mathbf{x}_i) = \boldsymbol{p}_i(\boldsymbol{\beta})$.
- Model (1) is not identifiable and extra model constraint should be imposed. Two model constraints are discussed,
  - baseline constraint: assuming that parameter for the baseline category is $\mathbf{0}$, i.e., $\boldsymbol{\beta}_0 = \mathbf{0}$ and
  - summation constraint: assuming that sum of all parameters is $\mathbf{0}$, i.e., $\sum_{k=0}^K \boldsymbol{\beta}_k = \mathbf{0}$.
- No matter which model constraint is used, the mean response vector keeps the same value, and the coefficient estimators obtained under these two model constraints preserve a linear relationship.
- However, these two model constraints lead to different optimal subsampling probabilities and thus produce different results.

Optimal Subsampling under the Baseline Constraint

Under the baseline constraint, the optimal subsampling probabilities for softmax regression was studied in Yao and Wang [2019].

- The unknown parameter under this constraint is a Kd dimensional vector, denoted as $\boldsymbol{\beta}^{\mathrm{b}} = (\boldsymbol{\beta}_1^{\mathrm{b}\mathrm{T}}, ..., \boldsymbol{\beta}_K^{\mathrm{b}\mathrm{T}})^{\mathrm{T}}$.
- The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}_{\mathrm{full}}^{\mathrm{b}}$ can be obtained by maximizing the log-likelihood function

$$\ell(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^N \left[\sum_{k=0}^K y_{i,k}\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k - \log\left\{\sum_{l=0}^K \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_l)\right\}\right] \text{ subject to } \boldsymbol{\beta}_0 = \mathbf{0}.$$

- There is no general closed form solution for the MLE of $\boldsymbol{\beta}^{\mathrm{b}}$. Apply the Newton-Raphson method.
- The subsample estimator is obtained by Algorithm 1.

Algorithm 1 General Subsampling Algorithm

Sampling: Assign subsampling probabilities $\{\pi_i\}_{i=1}^N$ to each data point. Draw n observations from full dataset with replacement. Denote subsample points as $\{\mathbf{x}_i^*, y_i^*\}_{i=1}^n$ and the corresponding subsampling probabilities as $\{\pi_i^*\}_{i=1}^n$.

Estimation: Obtain the subsample estimator $\hat{\boldsymbol{\beta}}_{\mathrm{sub}}^{\mathrm{b}}$ by maximizing

$$\ell_s^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \frac{1}{N\pi_i^*}\left[\sum_{k=1}^K y_{i,k}^*\boldsymbol{\beta}_k^{\mathrm{b}\mathrm{T}}\mathbf{x}_i^* - \log\left\{1 + \sum_{l=1}^K \exp(\boldsymbol{\beta}_l^{\mathrm{b}\mathrm{T}}\mathbf{x}_i^*)\right\}\right].$$

- Under mild assumptions, the approximation error $\hat{\boldsymbol{\beta}}_{\mathrm{sub}}^{\mathrm{b}} - \hat{\boldsymbol{\beta}}_{\mathrm{full}}^{\mathrm{b}}$ is asymptotically normal distributed conditionally on the full data and its asymptotic variance-covariance matrix (scaled by n) has the form of

$$\mathbf{V}_N = \mathbf{M}_N^{\circ\,-1}\mathbf{D}_N^{\circ}\mathbf{M}_N^{\circ\,-1},$$

  where

$$\mathbf{M}_N^{\circ} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\phi}_i^{\circ}(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}),$$
$$\mathbf{D}_N^{\circ} = \frac{1}{N^2}\sum_{i=1}^N \frac{\boldsymbol{\psi}_i^{\circ}(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})}{\pi_i},$$
$$\boldsymbol{\phi}_i^{\circ}(\boldsymbol{\beta}) = \mathrm{diag}\{\boldsymbol{p}_i^{\circ}(\boldsymbol{\beta})\} - \{\boldsymbol{p}_i^{\circ}(\boldsymbol{\beta})\}\{\boldsymbol{p}_i^{\circ}(\boldsymbol{\beta})\}^{\mathrm{T}},$$
$$\boldsymbol{\psi}_i^{\circ}(\boldsymbol{\beta}) = \mathbf{s}_i^{\circ}(\boldsymbol{\beta})\mathbf{s}_i^{\circ}(\boldsymbol{\beta})^{\mathrm{T}},$$
$$\mathbf{s}_i^{\circ}(\boldsymbol{\beta}) = \mathbf{y}_i^{\circ} - \boldsymbol{p}_i^{\circ}(\boldsymbol{\beta}),$$

  $\boldsymbol{p}_i^{\circ}(\boldsymbol{\beta}) = \{p_1(\mathbf{x}_i, \boldsymbol{\beta}), p_2(\mathbf{x}_i, \boldsymbol{\beta}), ..., p_K(\mathbf{x}_i, \boldsymbol{\beta})\}^{\mathrm{T}}$ and $\mathbf{y}_i^{\circ} = \{y_{i,1}, y_{i,2}, ..., y_{i,K}\}$
- Under the A-optimality criterion, the optimal subsampling probabilities are to minimize $\mathrm{tr}(\mathbf{V}_N)$ and have expression

$$\pi_i^{\mathrm{b,A}} = \frac{\|\mathbf{M}_N^{\circ\,-1}\{\mathbf{s}_i^{\circ}(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_i\}\|}{\sum_{j=1}^N \|\mathbf{M}_N^{\circ\,-1}\{\mathbf{s}_j^{\circ}(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_j\}\|}, \quad i = 1, 2, ..., N.$$

- We also consider the L-optimality criterion, which is to minimize the trace of the linear transformation of $\mathbf{V}_N$. Here we choose to minimize $\mathrm{tr}(\mathbf{M}_N^{\circ}\mathbf{V}_N\mathbf{M}_N^{\circ}) = \mathrm{tr}(\mathbf{D}_N^{\circ})$ and the expression for optimal subsampling probabilities is

$$\pi_i^{\mathrm{b,L}} = \frac{\|\mathbf{s}_i^{\circ}(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\|\|\mathbf{x}_i\|}{\sum_{j=1}^N \|\mathbf{s}_i^{\circ}(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\|\|\mathbf{x}_i\|}, \quad i = 1, 2, ..., N.$$

We can see that $\pi_i^{\mathrm{b,L}}$ uses $\mathbf{s}_i^{\circ}(\hat{\boldsymbol{\beta}}_{\mathrm{full}})$ who relates to $y_{i,k} - p_k(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{full}})$ only for $k \in \{1, ..., K\}$. The optimal subsampling probabilities treat the baseline category differently, making the importance of the observations in the baseline category be either underestimated or overestimated. This can be revealed by the following example.

- Simulate a balanced dataset with $N = 10000$ where
  - responses have 10 distinct outcomes,
  - covariates are generated from $\mathbb{N}_2(\mathbf{0}, \mathbf{I}_2)$ and
  - true coefficient is $(0, 0, 0, ..., 0, 0, 0)^{\mathrm{T}}$.
- Ideally, uniform subsampling is the best way to sample observations.
- 1000 samples are drawn, and the averaged numbers of sampled observations for all categories for 1000 replications are recorded.
- Based on $\pi_i^{\mathrm{b,L}}$, less observations are drawn from 0-th category.

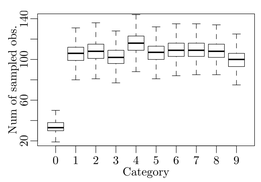

Fig. 1: Compare the average number of sampled observations for 10 categories.

Optimal Subsampling under the Summation Constraint

In this section, we derive the optimal subsampling probabilities under the summation constraint.

- Under this constraint, the unknown parameter is denoted as $\boldsymbol{\beta}^{\mathrm{s}} = (\boldsymbol{\beta}_0^{\mathrm{s}\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{s}\mathrm{T}}, ..., \boldsymbol{\beta}_K^{\mathrm{s}\mathrm{T}})^{\mathrm{T}}$.
- Models under two constraints are equivalent in that $\boldsymbol{p}_i(\boldsymbol{\beta}^{\mathrm{b}}) = \boldsymbol{p}_i(\boldsymbol{\beta}^{\mathrm{s}})$ if

$$\boldsymbol{\beta}_0^{\mathrm{s}} = -\frac{1}{K+1}\sum_{l=1}^K \boldsymbol{\beta}_l^{\mathrm{b}} \quad \text{and} \quad \boldsymbol{\beta}_k^{\mathrm{s}} = \boldsymbol{\beta}_k^{\mathrm{b}} - \frac{1}{K+1}\sum_{l=1}^K \boldsymbol{\beta}_l^{\mathrm{b}}$$

  for $k = 1, 2, ..., K$. In matrix notation,

$$\boldsymbol{\beta}^{\mathrm{s}} = \left\{\begin{pmatrix} -(K+1)^{-1}\mathbf{1}_K^{\mathrm{T}} \\ \mathbf{I}_K - (K+1)^{-1}\mathbf{J}_K \end{pmatrix} \otimes \mathbf{I}_d\right\}\boldsymbol{\beta}^{\mathrm{b}} \equiv \mathbf{G}\boldsymbol{\beta}^{\mathrm{b}}.$$

- Obtain MLE $\hat{\boldsymbol{\beta}}_{\mathrm{full}}^{\mathrm{s}}$ by premultiplying $\mathbf{G}$ to $\hat{\boldsymbol{\beta}}_{\mathrm{full}}^{\mathrm{b}}$. Similarly, obtain the subsample estimator $\hat{\boldsymbol{\beta}}_{\mathrm{sub}}^{\mathrm{s}}$ by $\mathbf{G}\hat{\boldsymbol{\beta}}_{\mathrm{sub}}^{\mathrm{b}}$.

Theorem 1. *Under the same assumptions as the Theorem 1 in Yao and Wang [2019], given the full data $\mathcal{D}_N$, when $N \to \infty$ and $n \to \infty$, $\hat{\boldsymbol{\beta}}_{\mathrm{sub}}^{\mathrm{s}} - \hat{\boldsymbol{\beta}}_{\mathrm{full}}^{\mathrm{s}}$ satisfies*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{sub}}^{\mathrm{s}} - \hat{\boldsymbol{\beta}}_{\mathrm{full}}^{\mathrm{s}}) \overset{a}{\sim} \mathbb{N}(\mathbf{0}, \mathbf{V}_G),$$

where $\overset{a}{\sim}$ means that two quantities have the same asymptotic distribution,

$$\mathbf{V}_G = (\mathbf{M}_N)^+\mathbf{D}_N(\mathbf{M}_N)^+$$
$$\mathbf{M}_N = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\phi}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}),$$
$$\mathbf{D}_N = \frac{1}{N^2}\sum_{i=1}^N \frac{\boldsymbol{\psi}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})}{\pi_i},$$

$(\cdot)^+$ represents the Moore-Penrose inverse,

$$\boldsymbol{\phi}_i(\boldsymbol{\beta}) = \mathrm{diag}\{\boldsymbol{p}_i(\boldsymbol{\beta})\} - \{\boldsymbol{p}_i(\boldsymbol{\beta})\}\{\boldsymbol{p}_i(\boldsymbol{\beta})\}^{\mathrm{T}},$$
$$\boldsymbol{\psi}_i(\boldsymbol{\beta}) = \mathbf{s}_i(\boldsymbol{\beta})\mathbf{s}_i(\boldsymbol{\beta})^{\mathrm{T}}, \text{ and}$$
$$\mathbf{s}_i(\boldsymbol{\beta}) = \mathbf{y}_i - \boldsymbol{p}_i(\boldsymbol{\beta}).$$

- Under A-optimality criterion, the optimal subsampling probabilities are to minimize $\mathrm{tr}(\mathbf{V}_G)$ and have the form of

$$\pi_i^{\mathrm{s,A}} = \frac{\|\mathbf{M}_N^{-1}\{\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_i\}\|}{\sum_{j=1}^N \|\mathbf{M}_N^{-1}\{\mathbf{s}_j(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_j\}\|}, \quad i = 1, 2, ..., N.$$

- Under L-optimality criterion, the optimal subsampling probabilities are to minimize $\mathrm{tr}(\mathbf{D}_N)$ and have the form of

$$\pi_i^{\mathrm{s,L}} = \frac{\|\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\|\|\mathbf{x}_i\|}{\sum_{j=1}^N \|\mathbf{s}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\|\|\mathbf{x}_i\|}, \quad i = 1, 2, ..., N.$$

The optimal subsampling probabilities for summation constraint can treat every category equally. Use the simulated dataset described in last section to illustrate this.

- Based on $\pi_i^{\mathrm{s,L}}$, the average number of sampled observations for all 10 categories are roughly equal.

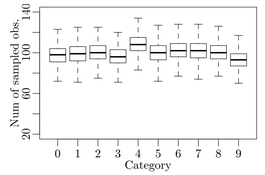

Fig. 2: Compare the average number of sampled observations for 10 categories.

Model Constraints Independent Optimal Subsampling Probabilities

Different model constraints lead to different forms of optimal subsampling probabilities when they are formulated by minimizing asymptotic variance-covariance matrix of the subsample estimators. In this section, we derive the optimal subsampling probabilities that are invariant to the model constraints. Because

- mean response vectors stay invariant to the choice of model constraints, and
- prediction accuracy is an important criterion to assess the quality of the subsample,

we consider to obtain the optimal subsampling probabilities by minimizing the mean squared prediction error

$$\frac{1}{N}\sum_{i=1}^N \left\|\boldsymbol{p}_i(\hat{\boldsymbol{\beta}}_{\mathrm{sub}}) - \boldsymbol{p}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\right\|^2.$$

Theorem 2. *Under the same assumptions as the Theorem 1 of Yao and Wang [2019], given the full data $\mathcal{D}_N$, when $N \to \infty$ and $n \to \infty$,*

$$\frac{n}{N}\sum_{i=1}^N \left\|\boldsymbol{p}_i(\hat{\boldsymbol{\beta}}_{\mathrm{sub}}) - \boldsymbol{p}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\right\|^2 \overset{a}{\sim} \mathbf{z}^{\mathrm{T}}\mathbf{V}_N^{1/2}\boldsymbol{\Omega}_N\mathbf{V}_N^{1/2}\mathbf{z},$$

where $\overset{a}{\sim}$ means that the two quantities have the same asymptotic distribution, $\mathbf{z} \sim \mathbb{N}(\mathbf{0}, \mathbf{I})$,

$$\boldsymbol{\Omega}_N = \frac{1}{N}\sum_{i=1}^N \left\{\mathbf{B}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\mathbf{B}_i(\hat{\boldsymbol{\beta}}_{\mathrm{full}})\right\},$$
$$\mathbf{B}_i(\boldsymbol{\beta}) = \begin{pmatrix} -p_0(\mathbf{x}_i, \boldsymbol{\beta})p_1(\mathbf{x}_i, \boldsymbol{\beta}) & ... & -p_0(\mathbf{x}_i, \boldsymbol{\beta})p_K(\mathbf{x}_i, \boldsymbol{\beta}) \\ p_1(\mathbf{x}_i, \boldsymbol{\beta}) - p_1^2(\mathbf{x}_i, \boldsymbol{\beta}) & ... & -p_1(\mathbf{x}_i, \boldsymbol{\beta})p_K(\mathbf{x}_i, \boldsymbol{\beta}) \\ ... & ... & ... \\ -p_1(\mathbf{x}_i, \boldsymbol{\beta})p_K(\mathbf{x}_i, \boldsymbol{\beta}) & ... & p_K(\mathbf{x}_i, \boldsymbol{\beta}) - p_K^2(\mathbf{x}_i, \boldsymbol{\beta}) \end{pmatrix} \otimes \mathbf{x}_i^{\mathrm{T}}.$$

- The asymptotic mean of the mean squared prediction error is

$$\mathbb{E}\{\mathbf{z}^{\mathrm{T}}\mathbf{V}_N^{1/2}\boldsymbol{\Omega}_N\mathbf{V}_N^{1/2}\mathbf{z}|\mathcal{D}_N\} = \mathrm{tr}\left(\mathbf{V}_N\boldsymbol{\Omega}_N\right).$$

- The optimal subsampling probabilities minimizing the asymptotic mean of the mean squared prediction error are

$$\pi_i^{\mathrm{P}} = \frac{\|\boldsymbol{\Omega}_N^{1/2}\mathbf{M}_N^{\circ\,-1}\{\mathbf{s}_i^{\circ}(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_i\}\|}{\sum_{i=1}^N \|\boldsymbol{\Omega}_N^{1/2}\mathbf{M}_N^{\circ\,-1}\{\mathbf{s}_i^{\circ}(\hat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_i\}\|}.$$

Practical Implementation

The optimal subsampling probabilities depend on the full data MLE, which is an unknown quantity. To solve this problem, an adaptive algorithm is used where we take a pilot sample and use the pilot sample estimator to substitute the full data MLE to approximate the optimal subsampling probabilities.

Algorithm 2 Two-Stage Adaptive Optimal Subsampling Algorithm

First Stage Sampling: Run Algorithm 1 with sample size n_1 and subsampling probabilities $\{n_1\pi_i^{\mathrm{prop}}\}_{i=1}^N$ to obtain the pilot sample estimator $\hat{\boldsymbol{\beta}}_1$. Substitute $\hat{\boldsymbol{\beta}}_{\mathrm{full}}$ with $\hat{\boldsymbol{\beta}}_1$ when calculating the optimal subsampling probabilities.

Second stage sampling: Run Algorithm 1 with sample size n_2 and approximate optimal subsampling probabilities to obtain a second stage sample.

Combining: Implement Newton-Raphson method to the combined samples of two stages. Obtain the final estimator $\bar{\boldsymbol{\beta}}^{\mathrm{OS}}$.

Real Data Example

- Forest cover type dataset [Dheeru and Karra Taniskidou, 2019]
- Total number of observations : 581012
- Response variable has 7 categories corresponding to 7 forest types:
  Spruce/Fir (36.46%), Lodgepole Pine (48.76%),
  Ponderosa Pine (6.15%), Cottonwood/Willow (0.427%),
  Aspen (1.63%), Douglas-fir (2.99%),
  Krummholz (3.53%).
- Use 10 quantitive covariates which are measuring geographical location and lighting condtions.

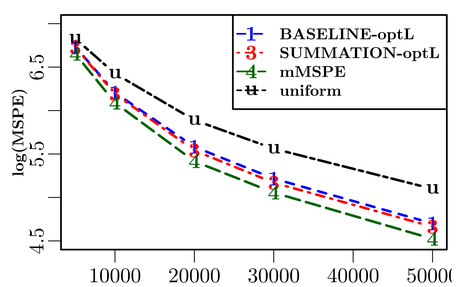

Fig. 3: Compare Empirical mean squared prediction errors among different n_2 for cover type dataset when $n_1 = 5000$ for 1000 replicates.

Fig. 3 shows that using π_i^{P} gives us the best prediction accuracy among the three kinds of optimal subsampling probabilities.

References

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2019. URL `http://archive.ics.uci.edu/ml`.

Yaqiong Yao and HaiYing Wang. Optimal subsampling for softmax regression. *Statistical Papers*, 60(2):585–599, 2019.