Revisiting Discriminative vs. Generative Classifiers: Theory and Implications

Chenyu Zheng[1] Guoqiang Wu[2] Fan Bao[3] Yue Cao[4] Chongxuan Li[1] Jun Zhu[3]

[1]Renmin University of China, [2]Shandong University

[3]Tsinghua University, [4]Beijing Academy of Artificial Intelligence

June 2023

Table of Contents

Table of Contents

1.1 Deep representation learning & linear evaluation

Deep representation learning has achieved great success in many fields.

- A common learning paradigm is to (pre-)train a large model on massive data and then transfer it to downstream tasks.
- An attractive method for the transfer is linear evaluation, which freezes parameters in the pre-trained model and learns a linear classifier separately.

A natural question

Can we find a better linear classifier than the default (multiclass) logistic regression?

Our attempt

We challenge the default logistic regression by using naïve Bayes in the context of deep representation learning.

1.2 Discriminative vs. generative linear classifiers

Ng and Jordan [1] have studied the **binary** naïve Bayes and logistic regression.

Limitations of [1]

- It is unclear how to extend the analysis to the multiclass setting.
- They assume that ERM can be performed on zero-one loss directly.
- Their experiments are conducted on shallow and low-dimensional features.

Our contributions

- We analyze the multiclass naïve Bayes vs. logistic regression.
- We consider the practically used logistic loss by introducing $\mathcal{H}$-consistency bounds.
- Technically, we propose a multiclass $\mathcal{H}$-consistency framework with tightness guarantee.
- We discuss the empirical implications of our theory with deep representations.

Table of Contents

2.1 Notions and definitions

We introduce some notations and definitions.

- Data:
  - Let $\mathcal{X} \subseteq [0,1]^n$ be the input space and $\mathcal{Y} = \{1, \cdots, K\}$ be the label space.
  - We denote by $\mathcal{D}$ the population distribution over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.
  - Training set S with m $i.i.d.$ examples sampled from the distribution $\mathcal{D}$.
- We mainly consider the constrained linear hypotheses:

$$\mathcal{H}_{\text{lin}} = \{\mathbf{x} \to \boldsymbol{h}(\mathbf{x}) : h_y(\mathbf{x}) = \langle \mathbf{w}_y, \mathbf{x} \rangle + b_y, \|\mathbf{w}_y\|_2 \leq W, |b_y| \leq B, y \in \mathcal{Y}\}.$$

- Errors based on loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$.
  - Zero-one loss: $\ell_{0-1}(\boldsymbol{h}(\mathbf{x}), y) = \mathbb{1}_{h(\mathbf{x}) \neq y}$,
  - Logistic loss: $\ell_{\log}(\boldsymbol{h}(\mathbf{x}), y) = \log_2(1 + e^{-yh(\mathbf{x})})$,
  - Generalization error: $\mathcal{R}_\ell(\boldsymbol{h}) = \mathbb{E}_{\mathbf{z} \sim \mathcal{D}}[\ell(\boldsymbol{h}(\mathbf{x}), y)]$.
  - Minimal generalization error: $\boldsymbol{h}^*_\ell = \text{argmin}_{\boldsymbol{h} \in \mathcal{H}} \mathcal{R}_\ell(\boldsymbol{h})$, $\mathcal{R}^*_{\ell,\mathcal{H}} = \mathcal{R}_\ell(\boldsymbol{h}^*_\ell)$.

2.2 Our target

Let $\boldsymbol{h}_{Dis,m}$ and $\boldsymbol{h}_{Gen,m}$ denote the hypothesis returned by multiclass logistic regression and naïve Bayes. Let $\boldsymbol{h}_{Dis,\infty}$ and $\boldsymbol{h}_{Gen,\infty}$ be the corresponding asymptotic version.

Our goal

We are interested in comparing the statistical efficiency (sample complexity) of naïve Bayes and logistic regression. Formally, we need to bound

$$R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,m}) - R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,\infty})$$

and

$$R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,m}) - R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,\infty}).$$

2.3 Results of naïve Bayes

In terms of naïve Bayes, we extend the definitions and analysis in [1] to the multiclass setting. The core idea is to bound the gap between the parameters in $\boldsymbol{h}_{Gen,m}$ and $\boldsymbol{h}_{Gen,\infty}$.

Theorem 1 (Sample complexity of naïve Bayes, informal)

Suppose that data distribution satisfies some mild assumptions. Then, it suffices to pick $m = O(\log n)$ *training samples such that* $R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,m}) \leq R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,\infty}) + \epsilon_0$ *hold with probability* $1 - \delta_0$*, for any fixed* $\epsilon_0 \in (0, 1)$ *and* $\delta_0 \in (0, \frac{\epsilon_0}{K^2}]$.

2.4 Generalization bounds for logistic loss

We can use the classical technique based on Rademacher complexity to obtain a generalization bound with $\ell_{\log}$.

Theorem 2

For any fixed $\delta_0 \in (0,1)$, with probability at least $1-\delta_0$, the following holds:

$$R_{\ell_{\log}}(\boldsymbol{h}_{Dis,m}) \leq R_{\ell_{\log}}(\boldsymbol{h}_{Dis,\infty}) + O(\sqrt{\frac{K^3 n}{m}}).$$

Problem

Our final goal is to establish bounds for $R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,m}) - R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,\infty})$. We need a tool to connect ℓ_{0-1} and $\ell_{\log}$.

2.5 $\mathcal{H}$-consistency bounds

$\mathcal{H}$-consistency bounds try to build the quantitative relationship between $\mathcal{R}_{\ell_1}(h) - \mathcal{R}_{\ell_1}(h^*_{\ell_1})$ and $\mathcal{R}_{\ell_2}(h) - \mathcal{R}_{\ell_2}(h^*_{\ell_2})$.

Definition 3

$\mathcal{H}$-consistency bound is in the following form that holds for all $h \in \mathcal{H}$ and some non-decreasing function $f : \mathbb{R}_+ \to \mathbb{R}_+$:

$$\mathcal{R}_{\ell_2}(h) - \mathcal{R}_{\ell_2}(h^*_{\ell_2}) \leq f\left(\mathcal{R}_{\ell_1}(h) - \mathcal{R}_{\ell_1}(h^*_{\ell_1})\right).$$

Our contributions

[2] constructs a binary $\mathcal{H}$-consistency framework for different hypotheses and losses. We extend it to the multiclass setting with a tightness guarantee.

2.6 Multiclass $\mathcal{H}$-consistency framework

Theorem 4 (Multiclass $\mathcal{H}$-consistency bounds, informal)

Suppose that $\mathcal{H}$ satisfies that $\{\operatorname{argmax}_{y\in\mathcal{Y}} h_y(\mathbf{x}) : \boldsymbol{h} \in \mathcal{H}\} = \{1, \ldots, K\}$ for any $\mathbf{x} \in \mathcal{X}$. If there exists a convex function $g : \mathbb{R}_+ \to \mathbb{R}$ with $g(0) = 0$ and $g(t) \leq \mathcal{J}_\ell(t)$. Then it holds for any $\boldsymbol{h} \in \mathcal{H}$ and any distribution $\mathcal{D}$ that

$$g(R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}}) \leq R_\ell(\boldsymbol{h}) - R^*_{\ell,\mathcal{H}} + M_{\ell,\mathcal{H}},$$

where $\mathcal{J}_\ell(t)$ and $M_{\ell,\mathcal{H}}$ are some well-defined terms (see our paper for details).

Remark

Our multiclass $\mathcal{H}$-consistency framework degenerates to the binary one [2] with $K = 2$. In addition, we note that if $\mathcal{J}_\ell(t)$ is convex and $\mathcal{J}_\ell(0) = 0$, then it leads to the tightest multiclass $\mathcal{H}$-consistency bound.

2.7 Results of logistic regression

Theorem 5 (Explicit $\mathcal{H}$-consistency bound for logistic regression)

If $R_{\ell_{\log}}(\boldsymbol{h}) - R^*_{\ell_{\log},\mathcal{H}_{\text{lin}}} + M_{\ell_{\log},\mathcal{H}_{\text{lin}}} \leq \frac{1}{2}(\frac{e^{2B}-1}{e^{2B}+K-1})^2$, *then for any distribution satisfiying* $\max_y p_y(\mathbf{x}) - \min_y p_y(\mathbf{x}) \leq \frac{e^{2B}-1}{e^{2B}+K-1}$ *for all* $\mathbf{x}$, *it holds that*

$$R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}_{\text{lin}}} + M_{\ell_{0-1},\mathcal{H}_{\text{lin}}} \leq \sqrt{2}(R_{\ell_{\log}}(\boldsymbol{h}) - R^*_{\ell_{\log},\mathcal{H}_{\text{lin}}} + M_{\ell_{\log},\mathcal{H}_{\text{lin}}})^{\frac{1}{2}}.$$

Theorem 6 (Sample complexity of logistic regression)

Suppose that $M_{\ell_{\log},\mathcal{H}_{\text{lin}}} \leq \nu$. *Then, it suffices to pick* $m = O(n)$ *training samples such that* $R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,m}) \leq R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,\infty}) + \epsilon_0$ *hold with probability* $1-\delta_0$, *for any fixed* $\epsilon_0 \in [\sqrt{2\nu}, \frac{e^{2B}-1}{e^{2B}+K-1}]$ *and* $\delta_0 \in (0,1)$.

2.8 Theoretical conclusion

Theoretical conclusion

Theorem 1 and Theorem 6 show that the $O(n)$ vs. $O(\log n)$ result [1] still holds in multiclass cases with weaker assumptions, which suggests that naïve Bayes is possibly better than logistic regression when the sample size is limited.

Table of Contents

3.1 Simulations

We validate our theory on a mixture of Gaussian distribution. For a fixed feature dimension n, we increase the number of samples m until the two models approach the corresponding asymptotic error, which is approximately tractable in the experiment.

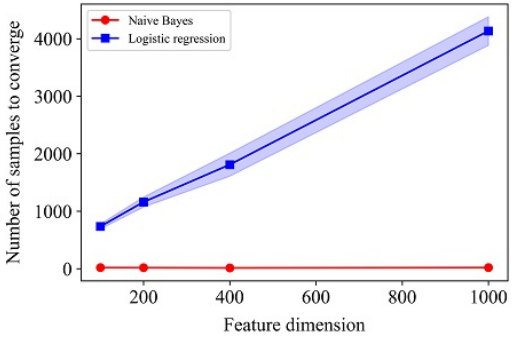
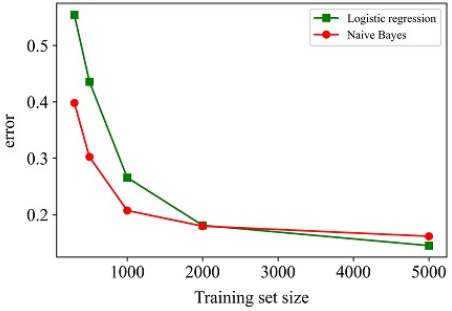


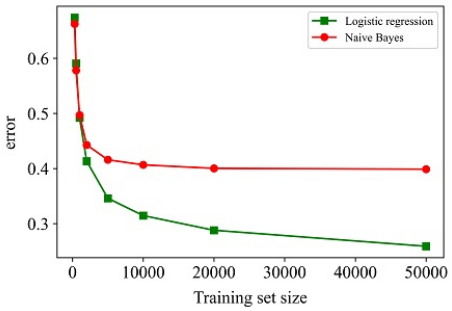
Figure: Multiclass ($K = 5$) simulation results.

3.2 Deep learning results

We systematically compare logistic regression and naïve Bayes based on various pre-trained vision models. Notably, naïve Bayes approaches its asymptotic error much faster than logistic regression in all settings, which is consistent with our theoretical results. We also observe the "two regimes" phenomenon [1], which shows the promise of naïve Bayes with limited data.



(a) ViT-B16



(b) ResNet-50

Table of Contents

Conclusion

Our contributions

- We challenge the default logistic regression in the linear prediction setting.
- We consider the practically used logistic loss rather than the zero-one loss.
- Technically, we propose a multiclass $\mathcal{H}$-consistency framework with tightness guarantee.
- We discuss the empirical implications of our theory with deep pre-trained models.

References I

[1] Andrew Y. Ng and Michael I. Jordan. "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes". In: *Advances in Neural Information Processing Systems*. 2001, pp. 841–848.

[2] Pranjal Awasthi et al. "$\mathcal{H}$-Consistency Bounds for Surrogate Loss Minimizers". In: *International Conference on Machine Learning*. Vol. 162. 2022, pp. 1117–1174.