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

Abstract

A large-scale deep model pre-trained on massive labeled or unlabeled data transfers well to downstream tasks. *Linear evaluation* freezes parameters in the pre-trained model and trains a linear classifier separately, which is efficient and attractive for transfer. However, little work has investigated the classifier in linear evaluation except for the default logistic regression. Inspired by the statistical efficiency of naïve Bayes, the paper revisits the classical topic on *discriminative vs. generative classifiers* [2]. Theoretically, the paper considers the surrogate loss instead of the zero-one loss in analyses and generalizes the classical results from binary cases to multiclass ones. We show that, under mild assumptions, multiclass naïve Bayes requires $O(\log n)$ samples to approach its asymptotic error while the corresponding multiclass logistic regression requires $O(n)$ samples, where n is the feature dimension. To establish it, we present a *multiclass $\mathcal{H}$-consistency bound* framework and an explicit bound for logistic loss, which are of independent interests. Simulation results on a mixture of Gaussian validate our theoretical findings. Experiments on various pre-trained deep vision models show that naïve Bayes consistently converges faster as the number of data increases. Besides, naïve Bayes shows promise in few-shot cases and we observe the "two regimes" phenomenon in pre-trained supervised models. Our code is available at *https://github.com/ML-GSAI/Revisiting-Dis-vs-Gen-Classifiers*.

Highlights

- We **challenge the default logistic regression** in the **linear prediction** setting. Specially, we study the comparison between the logistic regression and naive Bayes.
- We consider the practically used **logistic loss** rather than assuming the empirical risk minimization (ERM) can be performed on zero-one loss as [2].
- Technically, we propose a **multiclass $\mathcal{H}$-consistency framework** with **tightness guarantee**. In addition, we obtain an **explicit bound** for the logistic loss and zero-one loss.
- We discuss the empirical implications of our theory in various **deep pre-trained models**.

Notations and definitions

Similarly to [2] and [1], 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 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.
  - We denote by $\boldsymbol{p}(\mathbf{x})$ the conditional distribution of Y given $\mathbf{x}$, i.e., $p_y(\mathbf{x}) = \mathbb{P}(Y = y | X = \mathbf{x})$.
  - Training set S with m *i.i.d.* examples sampled from the distribution $\mathcal{D}$.
- Hypotheses:
  - We mainly consider the constrained linear hypotheses:

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

  - We denote $\mathcal{H}_{\text{all}}$ by the set of all measurable functions.
  - Let $\boldsymbol{h}_{Dis,m}$ and $\boldsymbol{h}_{Gen,m}$ denote the hypothesis returned by logistic regression and naïve Bayes.
  - Let $\boldsymbol{h}_{Dis,\infty}$ and $\boldsymbol{h}_{Gen,\infty}$ be the corresponding asymptotic version.
- Some risks 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: $\ell_{\log}(\boldsymbol{h}(\mathbf{x}), y) = \log_2(1 + e^{-y h(\mathbf{x})})$.
  - Generalization error: $\mathcal{R}_\ell(\boldsymbol{h}) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}}[\ell(\boldsymbol{h}(\mathbf{x}), y)]$, minimal generalization error: $\mathcal{C}^*_{\ell,\mathcal{H}}(\mathbf{x}) = \min_{\boldsymbol{h}} \mathcal{R}_\ell(\boldsymbol{h})$
  - Conditional risk: $\mathcal{C}_\ell(\boldsymbol{h}, \mathbf{x}) = \sum_{y=1}^K p_y(\mathbf{x}) \ell(\boldsymbol{h}(\mathbf{x}), y)$, minimal conditional risk: $\mathcal{C}^*_{\ell,\mathcal{H}}(\mathbf{x}) = \inf_{\boldsymbol{h} \in \mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h}, \mathbf{x})$.
  - Excess conditional risk: $\Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h}, \mathbf{x}) = \mathcal{C}_\ell(\boldsymbol{h}, \mathbf{x}) - \mathcal{C}^*_{\ell,\mathcal{H}}(\mathbf{x})$, Minimizability gap: $M_{\ell,\mathcal{H}} = \mathcal{R}^*_{\ell,\mathcal{H}} - \mathbb{E}_{\mathbf{x}}\left[\mathcal{C}^*_{\ell,\mathcal{H}}(\mathbf{x})\right]$.
- $\mathcal{H}$-consistency bound: it is in the following form that holds for all $\boldsymbol{h} \in \mathcal{H}$, $\mathcal{D} \in \mathcal{P}$ and some non-decreasing function $f : \mathbb{R}_+ \to \mathbb{R}_+$:

$$\mathcal{R}_{\ell_2}(\boldsymbol{h}) - \mathcal{R}^*_{\ell_2,\mathcal{H}} \le f(\mathcal{R}_{\ell_1}(\boldsymbol{h}) - \mathcal{R}^*_{\ell_1,\mathcal{H}}).$$

If $\mathcal{P}$ is composed of all distributions over $\mathcal{X} \times \mathcal{Y}$, we call it a distribution-independent bound.

Muticlass $\mathcal{H}$-consistency framework

We propose a **multiclass $\mathcal{H}$-consistency framework** with the **tightness guarantee**, which includes the binary one [1] as a special case.

Definition 1 (Multiclass $\mathcal{H}$-estimation error transformation). The multiclass $\mathcal{H}$-estimation error transformation of a surrogate loss ℓ is defined on $t \in [0,1]$ as $\mathcal{J}_\ell(t) = \inf_{\hat{y} \in \mathcal{Y}, \boldsymbol{p} \in \mathcal{P}_{\hat{y}}(t), \mathbf{x} \in \mathcal{X}, \boldsymbol{h} \in \mathcal{H}_{\hat{y}}(\mathbf{x})} \Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h}, \mathbf{x}, \boldsymbol{p})$. Here $\mathcal{H}_{\hat{y}}(\mathbf{x}) := \{\boldsymbol{h} \in \mathcal{H} : \operatorname{argmax}_{y \in \mathcal{Y}} h_y(\mathbf{x}) = \hat{y}\}$ is a collection of hypotheses that predicts $\mathbf{x}$ as class $\hat{y}$. $\mathcal{P}_{\hat{y}}(t) := \{\boldsymbol{p} \in \Delta_K : \max_y p_y - p_{\hat{y}} = t\}$ is a subset of K-dimensional simplex indexed by classes and the gap between the max component and class-indexed component of $\boldsymbol{p}$.

Theorem 1 (Distribution-independent $\mathcal{H}$-consistency bound). 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 $g(t) \le \mathcal{J}_\ell(t)$. Then it holds for any $\boldsymbol{h} \in \mathcal{H}$ and any distribution $\mathcal{D}$ that

$$g\left(\mathcal{R}_{\ell_{0-1}}(\boldsymbol{h}) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}}\right) \le \mathcal{R}_\ell(\boldsymbol{h}) - \mathcal{R}^*_{\ell,\mathcal{H}} + M_{\ell,\mathcal{H}}.$$

Theorem 2 (Tightness guarantee). If $\mathcal{J}_\ell(t)$ is convex with $\mathcal{J}_\ell(0) = 0$, then for any $t \in [0,1]$ and $\delta > 0$, there exist a distribution $\mathcal{D}$ and a hypothesis $h \in \mathcal{H}$ such that $\mathcal{R}_{\ell_{0-1}}(\boldsymbol{h}) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}} = t$ and $\mathcal{J}_\ell(t) \le \mathcal{R}_\ell(\boldsymbol{h}) - \mathcal{R}^*_{\ell,\mathcal{H}} + M_{\ell,\mathcal{H}} \le \mathcal{J}_\ell(t) + \delta$.

Theoretical results for logistic regression

Proposition 1 (Generalization bound for logistic loss). For any fixed $\delta_0 \in (0,1)$, with probability at least $1 - \delta_0$, the following holds:

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

Assumption 1 (Mild assumption in the context of deep representation learning). The approximate error of the logistic loss is bounded by a small constant $\nu < \frac{1}{2}(\frac{e^{2B}-1}{e^{2B}+K-1})^2$. Namely, $\operatorname{argmin}_{\boldsymbol{h} \in \mathcal{H}_{\text{lin}}} \mathcal{R}_{\ell_{\log}}(\boldsymbol{h}) - \operatorname{argmin}_{\boldsymbol{h} \in \mathcal{H}_{\text{all}}} \mathcal{R}_{\ell_{\log}}(\boldsymbol{h}) \le \nu$, which implies that $M_{\ell_{\log},\mathcal{H}_{\text{lin}}} \le \nu$.

Theorem 3 ($\mathcal{H}$-consistency bound connects ℓ_{0-1} and $\ell_{\log}$). If $\mathcal{R}_{\ell_{\log}}(\boldsymbol{h}) - \mathcal{R}^*_{\ell_{\log},\mathcal{H}_{\text{lin}}} + M_{\ell_{\log},\mathcal{H}_{\text{lin}}} \le \frac{1}{2}(\frac{e^{2B}-1}{e^{2B}+K-1})^2$, then for any distribution satisfying $\max_y p_y(\mathbf{x}) - \min_y p_y(\mathbf{x}) \le \frac{e^{2B}-1}{e^{2B}+K-1}$ for all $\mathbf{x}$, it holds that $\mathcal{R}_{\ell_{0-1}}(\boldsymbol{h}) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\text{lin}}} + M_{\ell_{0-1},\mathcal{H}_{\text{lin}}} \le \sqrt{2}(\mathcal{R}_{\ell_{\log}}(\boldsymbol{h}) - \mathcal{R}^*_{\ell_{\log},\mathcal{H}_{\text{lin}}} + M_{\ell_{\log},\mathcal{H}_{\text{lin}}})^{\frac{1}{2}}$.

Corollary 1 (Sample complexity of logistic regression). Suppose that Assumption 1 holds. Then, it suffices to pick $m = O(n)$ training samples such that $\mathcal{R}_{\ell_{0-1}}(\boldsymbol{h}_{Dis,m}) \le \mathcal{R}_{\ell_{0-1}}(\boldsymbol{h}_{Dis,\infty}) + \epsilon_0$ hold with probability $1 - \delta_0$, for any fixed $\epsilon_0 \in [\sqrt{2\nu}, \frac{e^{2B}-1}{e^{2B}+K-1}]$ and $\delta_0 \in (0,1)$.

Theoretical results for naive Bayes

Theorem 4 (Generalization bound and sample complexity of naive Bayes, **informal**). Under some mild assumptions that are similar to [2], with probability at least $1 - \delta_0$:

$$\mathcal{R}_{\ell_{0-1}}(\boldsymbol{h}_{Gen,m}) \le \mathcal{R}_{\ell_{0-1}}(\boldsymbol{h}_{Gen,\infty}) + \frac{K(K-1)}{2}\left(\widetilde{G}\big(O(\sqrt{\frac{1}{m}\log(\frac{n}{\delta_0})})\big) + \delta\right),$$

where $\widetilde{G}(\tau)$ is polynomially small in n. Then, it suffices to pick $m = O(\log n)$ training samples such that $\mathcal{R}_{\ell_{0-1}}(\boldsymbol{h}_{Gen,m}) \le \mathcal{R}_{\ell_{0-1}}(\boldsymbol{h}_{Gen,\infty}) + \epsilon_0$ hold with probability $1 - \delta_0$, for any $\epsilon_0 \in (0,1)$ and $\delta_0 \in (0, \frac{\epsilon_0}{K^2}]$.

Experimental results

Simulations. We validate our theory on a **mixture of Gaussian distribution**. For a fixed feature dimension n, we increase the number of samples m and record the performance.

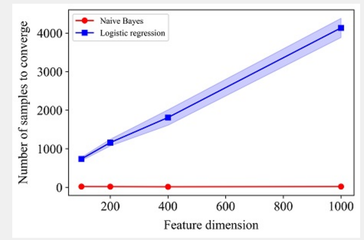
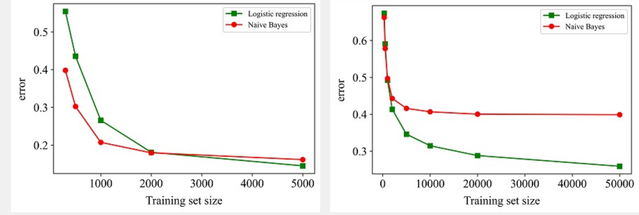


Figure 1. Multiclass ($K = 5$) simulation results. Empirically, logistic regression and naïve Bayes require $O(n)$ and $O(\log n)$ samples to approach the corresponding asymptotic error respectively.

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 [2] when model is pre-trained in a supervised manner**, which shows the promise of naïve Bayes when training data is limited.



(a) ViT-B16 (b) ResNet-50

Table 1. Convergence comparison between multiclass logistic regression and naïve Bayes. "NB faster" means naïve Bayes approaches its asymptotic error faster.

Method	NB faster/ CIFAR10	Two regimes CIFAR100
ViT	√ / √	√ / √
ResNet	√ / √	√ / √
CLIP	√ / √	√ / √
MoCov2	√ / ×	√ / ×
SimCLRv2	√ / ×	√ / √
MAE	√ / √	√ / ×

References

[1] Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong.
$\mathcal{H}$-consistency bounds for surrogate loss minimizers.
In *International Conference on Machine Learning*, volume 162, pages 1117–1174, 2022.

[2] 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*, pages 841–848, 2001.

Emails: chongxuanli@ruc.edu.cn; chenyu.zheng666@gmail.com