A Penalized Method for the Predictive Limit of Learning

Jie Ding, Enmao Diao, Jiawei Zhou and Vahid Tarokh

Duke University, Harvard University

Overview

Machine learning systems learn from and make predictions by building models from observed data. Because large models tend to overfit while small models tend to underfit for a given fixed dataset, a critical challenge is to select an appropriate model (e.g. set of variables/features). Model selection aims to strike a balance between the goodness of fit and model complexity, and thus to gain reliable predictive power. In this paper, we study a penalized model selection technique that asymptotically achieves the optimal expected prediction loss (referred to as the limit of learning) offered by a set of candidate models. We prove that the proposed procedure is both statistically efficient in the sense that it asymptotically approaches the limit of learning, and computationally efficient in the sense that it can be much faster than cross validation methods. The theoretical property of our proposed method is much stronger than what is expected from standard statistical learning theory. Our theory applies for a wide variety of model classes, loss functions, and high dimensions (in the sense that the models' complexity can grow with data size). We released a python package with our proposed method for general usage like logistic regression and neural networks.

Keywords: Cross-validation; Computational efficiency; Ensemble learning; Feature selection; Graph-based expert learning; High dimension; Limit of learning; Model expansion.

Notation

- Data: $Z_1, \ldots, Z_n$ (assumed to be i.i.d.)
- Model class: $\mathcal{A} = \{\alpha\}$
- Parameter space associated with α: $\mathcal{H}_n[\alpha]$
- Loss function: $l_n : \mathcal{Z} \times \mathcal{H}_n[\alpha] \to \mathbb{R}$, usually written as $l_n(\boldsymbol{z}, \boldsymbol{\theta}; \alpha)$
- Minimum loss estimator of α: $\hat{\boldsymbol{\theta}}_n[\alpha] = \arg\min_{\theta \in \mathcal{H}_n[\alpha]} \sum_{i=1}^n l_n(z_i, \boldsymbol{\theta}; \alpha)$
- Gradient: $\boldsymbol{\psi}_n(\boldsymbol{z}, \boldsymbol{\theta}; \alpha) = \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{z}, \boldsymbol{\theta}; \alpha)$
- Empirical measure: $E_n \boldsymbol{l_n}(\cdot, \boldsymbol{\theta}; \alpha) = n^{-1} \sum_{i=1}^n \boldsymbol{l_n}(\boldsymbol{z}_i, \boldsymbol{\theta}; \alpha)$
- In-sample loss: $\hat{\mathcal{L}}_n[\alpha] = E_n l_n\big(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha\big)$
- Empirical Hessian: $\hat{V}_n(\boldsymbol{\theta}; \alpha) = E_n \nabla_{\boldsymbol{\theta}} \boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}; \alpha)$
- Empirical variance of gradient: $\hat{J}_n(\boldsymbol{\theta}; \alpha) = E_n\big\{\boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}; \alpha) \times \boldsymbol{\psi}_n(\cdot, \boldsymbol{\theta}; \alpha)^{\intercal}\big\}$

Limit of Learning

Definition [Predictive Performance]: The expected (or out-sample) prediction loss given by candidate model α, denoted by $\mathcal{L}_n(\alpha)$, is defined by

$$\mathcal{L}_n(\alpha) \triangleq E_* l_n\big(\cdot, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha\big) = \int_{\mathcal{Z}} p(\boldsymbol{z}) l_n\big(\boldsymbol{z}, \hat{\boldsymbol{\theta}}_n[\alpha]; \alpha\big) d\boldsymbol{z}.$$

Here, E_* denotes the expectation with respect to the distribution of a future (unseen) random variable $\boldsymbol{z}$. Typically $\boldsymbol{z}$ consists of response $\boldsymbol{y}$ and covariates $\boldsymbol{x}$, and only the entries of $\boldsymbol{x}$ associated with α are involved in the evaluation of l_n.

Definition [Limit of learning]: For a given data (of size n) and model class $\mathcal{A}$, the limit of learning (LoL) is defined as $\arg\min_{\alpha \in \mathcal{A}} \mathcal{L}_n(\alpha)$, the optimal expected prediction loss offered by candidate models.

Definition [Efficient learning]: Our goal is to select $\hat{\alpha}_n \in \mathcal{A}$ that is asymptotically efficient, in the sense that $\mathcal{L}_n[\hat{\alpha}_n] / \min_{\alpha \in \mathcal{A}} \mathcal{L}_n[\alpha] \to_p 1$ as $n \to \infty$.

Note that this requirement is weaker than selecting the exact optimal model $\arg\min_{\alpha \in \mathcal{A}} \mathcal{L}_n[\alpha]$, but can be much stronger than typical bounds in the context of statistical learning theory.

We propose to use the following penalized model selection procedure, which generalizes Takeuchi's and Akaike's information criteria from negative log-likelihood to general loss functions.

Our learning procedure: Given data $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ and a specified model class $\mathcal{A}$. We first estimate each model and then select the $\hat{\alpha} \in \mathcal{A}$ in the following way: 1) for each $\alpha \in \mathcal{A}$, find the minimal loss estimator $\hat{\boldsymbol{\theta}}_n[\alpha]$ defined in (**??**), and record the minimum as $\hat{\mathcal{L}}_n[\alpha]$; 2) select $\hat{\alpha} = \arg\min_{\alpha \in \mathcal{A}} \mathcal{L}_t^c[\alpha]$, where

$$\mathcal{L}_t^c[\alpha] \triangleq \hat{\mathcal{L}}_n[\alpha] + n^{-1} tr\big\{\hat{V}_n(\hat{\boldsymbol{\theta}}_n[\alpha]; \alpha)^{-1} \hat{J}_n(\hat{\boldsymbol{\theta}}_n[\alpha]; \alpha)\big\}.$$

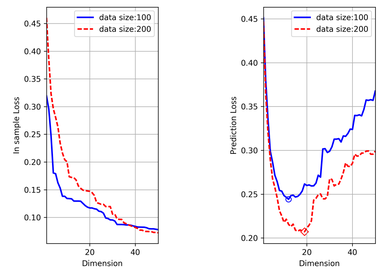
Theoretical Guarantee

Under some regularity assumptions, the $\hat{\alpha}_n$ selected by GTIC procedure is asymptotically efficient.

Technical challenge: Classical asymptotic analysis typically relies on a type of uniform convergence of empirical process around $\boldsymbol{\theta}_n^*[\alpha]$. Because our functions are vector valued with dimension depending on data size, we cannot directly use state-of-art technical tools. The classical proof by White in proving asymptotic normality in mis-specified class cannot be directly adapted, either, for parameter spaces that depend on n.

Sequential Model Expansion

In terms of predictive power, a model in a mis-specified model class could be determined to be unnecessarily large, suitable, or inadequately small, depending on specific data size (see Fig. 1). A realistic learning procedure thus requires models of different complexity levels as more data become available.



(a) The "goodness of fit" (b) The predictive performance

Figure: Experiment showing the "bigger models for bigger data" phenomena, under $n = 100$ (solid blue) and $n = 200$ (dash red).

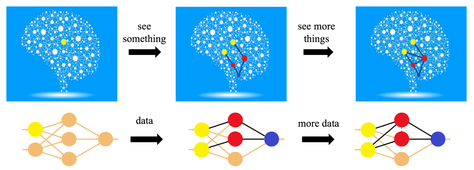


Figure: Illustration of neurons being activated to adapt to the currently observed data in an online fashion, similar to human brains.

To address the above challenge, we first propose a concept referred to as the *graph-based* expert tracking, which extends some classical online learning techniques based on expert tracking. Motivated by the particular path graph $1 \to 2 \to \cdots N$, where $1, 2, \ldots, N$ index the model class, we further propose a model expansion strategy, where each candidate model and its corrected prediction loss can be regarded respectively as an expert and loss at each time. The proposed algorithm can be used for online prediction, which ensures not only statistically reliable results but also simple computation.

Reference: Jie Ding, Enmao Diao, Jiawei Zhou and Vahid Tarokh, "Approaching the Predictive Limit of Learning," preprint, 2017.

Acknowledgement: This work is supported by Defense Advanced Research Projects Agency (DARPA) grant numbers N66001-15-C-4028 and W911NF-16-1-0561.