Yeheng GE

My notes series

Paper Reading

Do not be distracted

Contents

Chapter 1

ROBUST NONPARAMETRIC REGRESSION WITH DEEP NEURAL NETWORKS

Overview

Use Deep neural network do regression. Main contributions lies in the theoretical part.

ROBUST NONPARAMETRIC REGRESSION WITH DEEP NEURAL NETWORKS

- excess risk grows with d, the dimension of data in sublinear form
- only require that the Y has finite p order moment. it is the key point of the ROBUST in the title.
- loose the assumption of exact manifold suppurt assumption
- limitation: X have bounded support. The condition is needed in the approximation theory. The condition appears in page 10. (Writting skills.)

f^0 is the true function we want. $Z = (X, Y)$ is random vector independent of f. So no post selection problem here.? why emphisize it. $Y = f^0(X) + \eta$, where η is the noise vector independent with X. L is the loss function that is Lipschitz and continuous.
So we can define the risk we interested in :

$$R(f) = E_z L\{f(x), Y\}$$

and we define the target here is $f^* = argmin_f \, R(f) = argmin_f \, E_z L\{f(x), Y\}$
The target of optimization and what we optimize is in the same page with the definition of f^*.
We denote $\mathcal{S} = \{X_i, Y_i\}_{i=1}^n$ is the data set with sample size n.
Then we can define the empirical risk on the data set is :

$$R_n(f) = \frac{1}{n} \sum_i L\{f(X_i), Y_i\}$$

.
We define the estimated function as

$$\hat{f}_n = \underset{f \in \mathcal{F}_n}{argmin} \, R_n(f)$$

The estimated function must be defined in a proper function class $\mathcal{F}_n$. The subscripts n says the depedency of the function class with the sample size n.
So we have the excess risk:

$$R(\hat{f}_n) - R(f^*) = E_z L\{\hat{f}_n(X), Y\} - E_z L\{f^*(X), Y\}$$

The excess risk is the key point in machine learning We think a "better" f_n should have smaller excess risk. However, the excess risk is still unobeservable.

Chapter 2

Cube Root Asymptotics

Overview

This paper gives a functional central limit theory for empirical process.

- many convergence rate $n^{-1/3}$
- Key point is: continuous mapping theorem for the location of maximum point.
-

The paper is a very nice material for understanding the core idea and techniques of empirical process.
Main reference of this paper is in the lecture notes of Pollard "Empirical process: Theory and applications" 1990 version.

0.1 The Mode Estimation Problem

The paper first intuitively gives an example of "mode estimation" and gives its convergence rate $n^{-1/3}$
Suppose $\hat{\theta}_n$ is chosen to maximize

$$\Gamma_n(\theta) = P_n[\theta - 1, \theta + 1] \tag{2.1}$$

is the proportion of observations in an interval of length 2.
If P has a smooth density $p(\dot{})$, the function Γ is approximately parabolic of its optimal value θ_0, which means that

$$\Gamma(\theta) - \Gamma\left(\theta_0\right) = \int_{1+\theta_0}^{1+\theta} p(x)dx - \int_{-1+\theta_0}^{-1+\theta} p(x)dx \approx -C\left(\theta - \theta_0\right)^2$$

> Take care that $\Gamma(\theta)$ is the expectation of $\Gamma_n(\theta)$. The term above is the "bias" caused by the departure of θ from θ_0.

Then we consider the stochastic term,

$$D_n(\theta) = \left[\Gamma_n(\theta) - \Gamma_n\left(\theta_0\right)\right] - \left[\Gamma(\theta) - \Gamma\left(\theta_0\right)\right].$$

For fixed θ, the $D_n(\theta)$ is approximately $N(0, \sigma_\theta^2/n)$ where

$$\sigma_\theta^2 \approx \int_{1+\theta_0}^{1+\theta} p(x)dx + \int_{-1+\theta_0}^{-1+\theta} p(x)dx \approx C\,|\theta - \theta_0|$$

Intuitively, when the bias term $C|\theta - \theta_0|^2$ is large comparing with the stochastic term $C|\theta - \theta_0|$, the θ is far away from the true value θ_0. Thus not maximize the $\Gamma_n(\theta)$.
So the θ could be the solution of $\Gamma_n(\theta)$ if the bias term is the same order or smaller than the stochastic term. It means

The "add" and "minus" produces different order here. The order produced by "add" is due to the finite density of $p()$. Thus the density integral is proportional to the length of θ

$$\begin{aligned} C|\theta - \theta_0|^2 &< Cn^{-1/2}|\theta - \theta_0|^{1/2} \\ C|\theta - \theta_0|^{3/2} &< Cn^{-1/2} \\ C|\theta - \theta_0| &< Cn^{-1/3} \end{aligned}$$

> However, it is just an intuitive explaination. theoretically, we need build error bound uniformly in θ and the normal approximation must hold uniformly over θ.

Note that the variance term σ_θ decreases with $|\theta - \theta_0|$.
If the loss function $g(\theta, \dot{})$ is differentiable, σ_θ decreases with $|\theta - \theta_0|^2$.
Thus

$$\begin{aligned} C|\theta - \theta_0|^2 &< Cn^{-1/2}|\theta - \theta_0| \\ C|\theta - \theta_0| &< Cn^{-1/2} \end{aligned}$$

It produces the common $n^{-1/2}$ rate.
Thus the variance term σ_θ decreases with $|\theta - \theta_0|$, the non-standard case is a consequence of "shape-edge effect"

0.2 Convergence in distribution and the argmax functional

Chapter 3

Distribution-Invariant Differential Privacy

This paper has very wierd organization. Key point: the trade-off between privacy protection and statistical accuracy

The key point of the paper is that one can reconcile both accuracy and privacy, which we achieve by preserving the original data's distribution. it is believed that there is a trade-off between statistical accuracy and differential privacy.

It transforms and perturbs the data and employs a suitable transformation to recover the original distribution.

The first achieves privacy protection by either a privatization mechanism or a privatized sampling method, including the Laplace mechanism [19, 20], the exponential mechanism [43], the minimax optimal procedures [15], among others.

The second achieves differential privacy via privatization for a category of models or algorithms, such as deep learning [1], boosting [21], stochastic gradient descent [2], risk minimization [7], random graphs [38], func- tion estimation [30], parametric estimation [4], regression diagnostics [8], and top-k selection [16].

One main challenge is that existing privatization mechanisms protect data privacy at the expense of altering a sample's distribution

DIP approximately maintains statistical accuracy even with strict privacy protection in that it does not suffer from the trade-off between accuracy and privacy strictness

These characteristics enable us to perform data analysis without sacrificing statistical accuracy, as in regression, classification, graphical models, clustering, among other statistical and machine learning tasks

DIP's privatization process consists of three steps.

- First, DIP splits the original sample randomly into two independent subsamples, hold-out and to-be-privatized samples, both are fixed after the split.
- Second, it estimates an unknown data distribution by, say, the empirical distribution on the hold-out sample, which is referred to as a reference distribution.
- Third, we privatize the to-be-privatized sample through data perturbation, which (i) satisfies the requirement of differen- tial privacy, and (ii) preserves the reference

distribution approximating the original distribution. As a result, DIP is differential private on the to-be-privatized sample while retaining the original distribution asymptotically, c.f., Theorem 2.

need to estimate the empirical distribution.

For univariate data, (1)do probability-integral transformation

(2)random Laplace noise to perturb and mask the data

(3) we design a new function transforming the obfuscated data to follow the reference distribution approximating the original data distribution

For multivariate data, we propose to apply the probability chain rule [51], in place of privatizing each variable independently

Detail methodology

First, we focus on the case where the underlying distribution is known. The cumulative distribution function is F.

For continuous variable,

- apply F on the random sample Z_i and get $F(Z_i)$ which follows uniform distribution.
- add independent noise e_i to the $F(Z_i)$ where e_i follows Laplace distribution $Laplace(0, 1/\epsilon)$
- Finally, we apply a nonlinear transformation H to produce a privatized sample that follows the original distribution F.

The H dependes on the data type. For continuous variable, G converges $F(Z_i) + e_i$ to uniform distribution.

Then F^{-1} converges $G(F(Z_i) + e_i)$ to original function. Then $H(\dot{}) = F^{-1} \circ G(\dot{})$

Details of G in Appendix S1.1.

□

Chapter 4

Property of Schur Complement

Overview

In this part we give some basic result for the Schur Complement.
Consider a matrix M,

$$M = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}$$

, we define the Schur Complement of P in the matrix M as

$$(M/P) = S - RP^{-1}Q \tag{4.1}$$

Main result in this part comes from the book "The Schur Complement and its applications"

Lemma 0.1. *Schur determinant formula*
P,Q,S,R all $n \times n$ matrix, then we have
$\det(M) = \det(PS - RQ)$
and
$\det M = \det(P)\det(S - RP^{-1}Q)$

Definition 1. *The inertia of hermitian matrix H is the triple tuple $In(H) = \{\pi, \nu, \delta\}$ is the number of positive, negative and zero eigenvalue.*

and we have the result

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{12}^* & H_{22} \end{pmatrix}$$

then
$In(H) = In(H_{11}) + In(H/H_{11})$
For matrix M,

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

we have

$$M/D = A - BD^{-1}C$$

and

$$M/A = D - BA^{-1}C$$

Lemma 0.2. *Schur formula if M square, and A nonsingular.* $\det(M/A) = \det(M)/\det(A)$

For a Matrix A of size $n \times n$, we define the index set α and β are subsets of $\{1, \ldots, n\}$. Then we define $A[\alpha, \beta]$ is the submatrix with row index α and column index β. And we say $A[\alpha] = A[\alpha, \alpha]$. Then $A/A[\alpha, \beta]$ is the schur complement of $A[\alpha, \beta]$ in the matrix A.

$$A/A[\alpha, \beta] = A[\alpha^c, \beta^c] - A[\alpha^c, \beta](A[\alpha, \beta]^{-1})A[\alpha, \beta^c]$$

And we denote $A/A[\alpha]$ as A/α.
For matrix A,

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^* & A_{22} \end{pmatrix}$$

$A > 0$ if and only if $A_{11} > 0$ and $A/A_{11} > 0$
$A \geq 0$ if and only if $A_{11} > 0$ and $A/A_{11} \geq 0$
if $A \geq 0$ and $A_{11} > 0$, then $A/A_{11} = A_{22} - A_{12}A_{11}^{-1}a_{12} \geq 0$ so $A_{22} \geq A/A_{11} \geq 0$, $\det(A_{22}) \geq 0$

Lemma 0.3. *A,B are $n \times n$ positive matrices, then* $\det(A + B) \geq \det(A) + \det(B)$

Eigenvalue and singular value of schur complement

$\mathcal{H}_n$ is $n \times n$ Hermitian matrix sets.
For $A \in \mathcal{H}_n$ we define eigenvalues $\lambda_1(A) \geq \lambda_2(A) \geq \ldots, \geq \lambda_n(A)$.
For $A \in \mathcal{C}^{m \times n}$ we define singular value $\sigma_1(A) \geq \sigma_2(A) \geq \ldots, \geq \sigma_n(A)$.

Lemma 0.4. *Cauthy eigenvalue interlacing theorem*
for

$$H = \begin{pmatrix} A & B \\ B^* & D \end{pmatrix}$$

where A is $r \times r$ and H is $n \times n$.
we have

$$\lambda_i(H) \geq \lambda_i(A) \geq \lambda_{i+n-r}(H)$$

Lemma 0.5. *$H \in \mathcal{H}_n$, α is a index set of size k, $1 \leq k < n$, if $H[\alpha]$ positive definite, then*

$$\lambda_i(H) \geq \lambda_i(H/\alpha \bigoplus \mathbf{0}) \geq \lambda_{i+k}(H)$$

Corollary 1. *H is a $n \times n$ positive semidefinite matrix. $H[\alpha]$ is $k \times k$. then*

$$\lambda_i(H) \geq \lambda_i(H/\alpha) \geq \lambda_{i+k}(H)$$

$$\lambda_i(H) \geq \lambda_i(H[\alpha^c]) \geq \lambda_{i+k}(H/\alpha) \geq \lambda_{i+k}(H)$$

Corollary 2. *H is a $n \times n$ positive semidefinite matrix. α and α' are to nonnull index set. $\alpha' \subset \alpha \subset \{1, 2, \ldots, n\}$*
if $H[\alpha]$ non-singular, for every $i = 1, 2, \ldots, n - |\alpha|$

$$\lambda_i(H/\alpha') \geq \lambda_i(H[\alpha' \cap \alpha^c]/\alpha') \geq \lambda_i(H/\alpha) \geq \lambda_{i+|\alpha|-|\alpha'|}(H/\alpha')$$

Chapter 5

An Error Analysis of Generative Adversarial Networks for Learning Distributions

Overview

However, theoretical explanations for their empirical success are not well established. More specifically, to estimate a target distribution μ, one chooses an easy-to-sample source distribution v (for example, uniform or Gaussian distribution) and find the generator by solving the following minimax optimization problem, at the population level,

$$\min_{g \in \mathcal{G}} \max_{f \in \mathcal{F}} \mathbb{E}_{x \sim \mu}[f(x)] - \mathbb{E}_{z \sim \nu}[f(g(z))]$$

max f to increase the margin. so f is the discriminator.
min g to decrease the margin ,so g is the generator.
We show that, if the generator and discriminator network architectures are properly chosen, GANs are able to learn any distributions with bounded support

Huang, J., Jiao, Y., Li, Z., Liu, S., Wang, Y., Yang, Y. (2022). An error analysis of generative adversarial networks for learning distributions. Journal of Machine Learning Research, 23(116), 1-43.

$$\underset{g \in \mathcal{G}}{\operatorname{argmin}} d_{\mathcal{F}}\left(\widehat{\mu}_n, g_{\#} \nu\right) = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f\left(X_i\right) - \mathbb{E}_\nu[f \circ g] \right\}$$

$$\underset{g \in \mathcal{G}}{\operatorname{argmin}} d_{\mathcal{F}}\left(\widehat{\mu}_n, g_{\#} \widehat{\nu}_m\right) = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f\left(X_i\right) - \frac{1}{m} \sum_{j=1}^{m} f\left(g\left(Z_i\right)\right) \right\},$$