Cramér-Rao inequality and machine learning

Hông Vân Lê
Institute of Mathematics, CAS

WS2017, Srni, January 17

OUTLINE

1. Machine learning and statistical learning.

2. Learning algorithms and estimators.

3. Cramér-Rao inequality and efficient estimators.

4. Future directions.

1. Machine learning and statistical learning

• Machine learning is a part of statistical learning which can be translated in algorithms.

• Mathematical foundation of machine learning is statistical learning theory (Vladimir Vapnik).

• Machine learning/statistical learning theory $\sim$ Quantum field theory /theoretical physics.

• David Mumford (statistical pattern theory) and Steve Smale (mathematical foundation of learning theory) are among first movers.

- “Learning is a problem of function estimation on the basis of empirical data” (Vapnik).

- Machine learning is also related to Riemannian geometry, algebraic geometry, topology, non-linear functional analysis, measure theory, category theory, logic.

- A mathematical model of statistical learning consists of a learning machine, observables which are also called empirical data, and a learning algorithm.

1. A learning machine has to estimate an information source - a probability distribution- by observation. A probability distribution is a probability measure on $(\Omega, \mathcal{A})$. A learning machine, also called a statistical model, is a family of probability measures on $(\Omega, \mathcal{A}, P)$, to which we believe/or know that the true probability distribution belongs.

2. Observables in statistical learning theory represent "outcomes" of experiments.
The outcomes are subjected to a true probability distribution which we dont know and we need

to estimate. More often than in physics, we have to repeat our experiments very long, and we observe a sequence of usually i.i.d. (independent identically distributed)

$$D_n = \{X_1, \cdots, X_n | \, X_i \in \Omega\}.$$

Statistical learning: construct a method to estimate the true probability distribution from the set D_n, using machine learning.

Definition(classical math. statistics, AJLS 2015, AJLS 2016, AJLS2017)

- Ω - a measurable space,

- $\mathcal{S}(\Omega)$ - the Banach space of finite signed measure with total variation norm,

- $(M, \Omega, \mathbf{p})$ - a parametrized measure model, *where* M is a Banach manifold, $\mathbf{p} : M \to \mathcal{M}(\Omega) \subset \mathcal{S}(\Omega)$ is a Frechét-C^1-map.

- $(M, \Omega, \mathbf{p})$ is called a statistical model if it consists only of probability measures.

2. Learning algorithms and estimators

Given a sequence $D_n = (X_1, \cdots, X_n) \in (\Omega)^n$ and a statistical model $(P, \Omega, \mathbf{p})$, we need to find out the true distribution p_{true} probably in P using D_n.

Definition. An estimator is a map $\Omega \to P$.

- p_{true} may not belong to $(P, \Omega, \mathbf{p})$.

The most popular criterion used in mathematical statistics for rating the efficiency of an estimator is the Cramér-Rao criterion.

Definition. The Fisher metric $\mathfrak{g}$ is a quadratic form on TM of a parametrized measure model $(M, \Omega, \mathbf{p})$ that is defined by

$$\mathfrak{g}_\xi(V, W) := ||\partial_V \log \mathbf{p} \cdot \partial_{\mathbf{W}} \log \mathbf{p}||_{\mathbf{L^1}(\boldsymbol{\Omega}, \mathbf{p}(\xi))}$$

for $V, W \in T_\xi M$.

This formula has been derived in AJLS 2016 to generalize the classical case when $\mathbf{p}(\xi) = \mathbf{p}(\xi, \omega) \cdot \mu_{\mathbf{0}}$ and

$$\partial_V \log \mathbf{p} = \frac{\partial_{\mathbf{V}} \log \mathbf{p}(\xi, \omega)}{\mathbf{p}(\xi, \omega)}.$$

$$\Longrightarrow \mathfrak{g}(V, V) = \int_\Omega \left(\frac{\partial_V \log p(\xi, \omega)}{p(\xi, \omega)}\right)^2 p(\xi, \omega) d\mu_0(\omega).$$

A necessary condition for the existence of the Fisher metric in the classical case is that $p(\xi, \omega) \in L^2(\Omega, \mathbf{p}(\xi))$. But we also need the continuity of the Fisher metric, when the base measure $\mathbf{p}(\xi)$ varies. We develop a new theory of the Banach spaces of roots of measures.

Definition. (*simplified version*) A statistical model is called 2-integrable if the Fisher metric exists and continuous. A 2-integrable statistical model is called singular, if the Fisher metric degenerate at some point.

Remark. The Fisher metric has been invented by Fisher to quantify "information" of a statistical model. (close to Shannon's information: the same concept of entropy). It has been used first by Rao as a Riemannian metric .Almost all statistical models contain singularity and we cannot ignore them.

People therefore until [JLS2017] assume some extra conditions for estimation problem on singular statistical models.

When is the Fisher metric degenerate?

$$\ker \mathfrak{g} = \ker \mathbf{p} : \mathbf{P} \to \mathcal{M}_+(\boldsymbol{\Omega}) \subset \mathcal{S}(\boldsymbol{\Omega}).$$

Definition. [JLS2017] A reduced tangent space $\hat{T}_\xi P = T_\xi P / ker \mathbf{p}$. The Fisher reduced metric is the metric on the Fisher reduced tangent space whose pull back is the Fisher metric.

3. Cramér-Rao inequality and efficient estimation

Given a statistical model $(P, \Omega, \mathbf{p})$, we set

$$L^2_P(\Omega) := \{\psi \in L^2(\Omega, \mathbf{p}(\xi)) \text{ for all } \xi \in P\}.$$

- V - a topological vector space.
- V^P the vector space of all V-valued functions on P.

For an estimator $\hat{\sigma} : \Omega \to P$ we set

$$L^2_{\hat{\sigma}}(P, V) := \{\varphi \in V^P | \, l \circ \varphi \circ \hat{\sigma} \in L^2_P(\Omega) \forall l \in V^*\},$$

$$\langle \varphi_{\widehat{\sigma}}(\xi), l \rangle := \mathbb{E}_{\mathbf{p}(\xi)}(l \circ \varphi \circ \widehat{\sigma}) = \int_{\Omega} l \circ \varphi \circ \widehat{\sigma} \, d\mathbf{p}(\xi) \forall l \in V^*.$$

Definition The difference

$$b_{\widehat{\sigma}}^{\varphi} := \varphi_{\widehat{\sigma}} - \varphi \in (V^{**})^P$$

will be called the bias of the estimator $\widehat{\sigma}$ w.r.t. the map φ.

Definition Given $\varphi \in L^2_{\widehat{\sigma}}(P, V)$ the estimator $\widehat{\sigma}$ is called φ-unbiased, if $\varphi_{\widehat{\sigma}} = \varphi$, equivalently, $b_{\widehat{\sigma}}^{\varphi} = 0$.

For $l, k \in V^*$ we set

$$V^{\varphi}_{\mathbf{p}(\xi)}[\hat{\sigma}](l,k) :=$$

$$E_{\mathbf{p}(\xi)}[(\varphi^l \circ \hat{\sigma} - E_{\mathbf{p}(\xi)}(\varphi^l \circ \hat{\sigma})) \cdot (\varphi^k \circ \hat{\sigma} - E_{\mathbf{p}(\xi)}(\varphi^k \circ \hat{\sigma}))],$$

$$(\mathfrak{g}^{\varphi}_{\hat{\sigma}})^{-1}(\xi)(l,k) := \langle d\varphi^l_{\hat{\sigma}}, d\varphi^k_{\hat{\sigma}} \rangle_{\mathfrak{g}^{-1}}(\xi).$$

If $P \subset \mathbf{R}^n$ and $\varphi : P \to \mathbf{R}^n$ defines coordinates of P, if $\hat{\sigma}$ is an biased estimator and, then $(\mathfrak{g}^{\varphi}_{\hat{\sigma}})^{-1}$ is just the inverse of the Fisher metric.

Theorem(JLS2017) Let $(P, \Omega, \mathbf{p})$ be a finite dimensional 2-integrable statistical model, $\hat{\sigma}$: $\Omega \to P$ an estimator and $\varphi \in L^2_{\hat{\sigma}}(P, V)$. Then the difference $V^{\varphi}_{\mathbf{p}(\xi)}[\hat{\sigma}] - (\mathfrak{g}^{\varphi}_{\hat{\sigma}})^{-1}(\xi)$ is a positive semi-definite quadratic form on V' for any $\xi \in P$.

Assume that V is finite dimensional and φ is a coordinate mapping and $\hat{\sigma}$ is φ-unbiased. We get from the general Cramér-Rao inequality the well-known Cramér-Rao inequality

$$V_{\xi}[\hat{\sigma}] \geq \mathfrak{g}^{-1}(\xi).$$

This classical Cramer-Rao inequality is often compared with the Heisenberg uncertainty in physics !

Further results: - First examples of singular statistical models that admits efficient estimation. (Complicated technique using resolution of singularity in algebraic geometry). This example show that our extension of the Cramér-Rao inequality to singular spaces is optimal.

- First examples of infinite dimensional statistical model that admits efficient biased estimations. (complicated technique using RKHS).

Future directions

• Define the Fisher gradient flow on singular statistical model and understand its convergences.

• Prove that biased estimator under certain condition converges to the true estimation.

- Develop geometric methods for infinite dimensional exponential models: they are used in big data theory.

- Improve the gradient flow method and sell it to Google, Facebook, Elon Musk, etc, .

- Develop geometric theory of unsupervised learning.

REFERENCES - N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer, Information geometry, Ergebnisse der Mathematik und ihre Grenzgebiete, Springer 2017.
- J. Jost, H. V. Lê, and L. Schwachhöfer, Cramér-Rao inequality on singular statistical models (in preparation).
- D. Geiger, C. Meek, B. Sturmfels, On the toric algebra of graphical models, The Annals of Statistics 34 (5) (2006) 1463–1492.
- H. Hironaka, Resolution of singularities of an algebraic variety over a field of

characteristic zero. Annals of Math. 79 (1964), 109-326.
- S. Lang, Fundamentals of differential geometry, Springer 1999.
- H. V. Lê, P. Somberg and J. Vanžura, Poisson smooth structures on stratified symplectic spaces, Springer Proceeding in Mathematics and Statistics, Volume 98, (2015), chapter 7, p. 181-204.
- K. Muandet, K. Fukumizu, B. Sriperumbudur and B. Schölkopf, Kernel Mean Embedding of Distributions: A Review and Beyonds, arXiv:1605.09522.

- V. N. Vapnik, The nature of statistical learning theory, Springer, 1999.
- S. Watanabe, Algebraic Geometry and Statistical Learning Theory, Cambridge University Press, 2009.

THANK YOU!