Research presentation for PhD admits

Edgar Dobriban

March 3, 2022

Overview

Overview

Uncertainty quantification: calibration

Fairness: Bayes-optimal classifiers

High-dim. statistics & deterministic equivalents

Research Interests

- Goal: work on important problems at the interface of statistics and machine learning

Research Interests

- Goal: work on important problems at the interface of statistics and machine learning
- Mainly study theoretical problems, but also methods/algorithms

Research Interests

- Goal: work on important problems at the interface of statistics and machine learning
- Mainly study theoretical problems, but also methods/algorithms
- Work with interdisciplinary team of students/postdocs/collaborators: stats, math/applied math, computer science, electrical engineering, biostats

Research Interests

- Goal: work on important problems at the interface of statistics and machine learning
- Mainly study theoretical problems, but also methods/algorithms
- Work with interdisciplinary team of students/postdocs/collaborators: stats, math/applied math, computer science, electrical engineering, biostats
- Advising style:
  - Open to new problems/directions—we can work on anything!

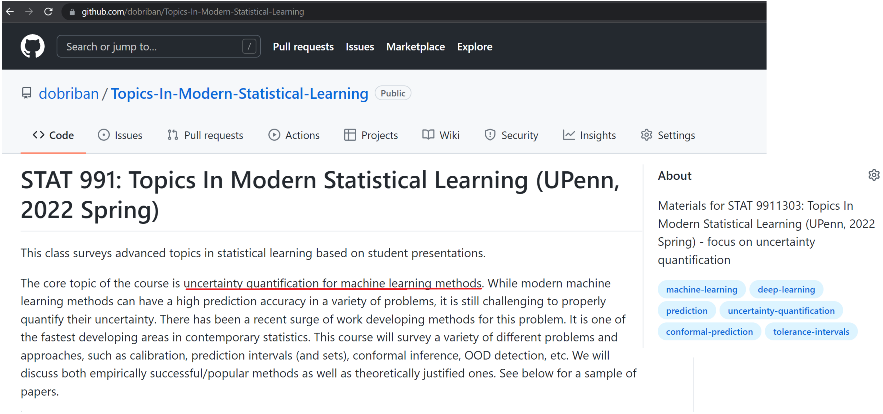
Research Interests

- Goal: work on important problems at the interface of statistics and machine learning
- Mainly study theoretical problems, but also methods/algorithms
- Work with interdisciplinary team of students/postdocs/collaborators: stats, math/applied math, computer science, electrical engineering, biostats
- Advising style:
  - Open to new problems/directions—we can work on anything!
  - Usually work closely on first project, then as hands-on/off as you would like.

Research Interests: see my website for details

- the efficient statistical analysis of "big data" using advanced tools, such as those from random matrix theory
  - dimension reduction, PCA: [1], [2], [3], [4], [5], [6]
  - multiple testing: [1], [2], [3]
  - high-dimensional regression: [1], [2]
  - invariance-based randomization tests
- the theoretical foundations of modern machine learning, including deep learning
  - data augmentation: [1], [2]
  - weight normalization
  - (stochastic) gradient descent and flow: [1], [2]
  - overparametrization
  - sketching and random projections, [1], [2], [3], [4], [5]
  - distributed learning: [1], [2], [3]
  - adversarial robustness: [1], [2]
  - retraining of ML models
  - uncertainty quantification: [1], [2]
  - fairness: [1]
  - reinforcement learning inspired by child-like learning
- in addition, we occasionally work on important applications and methods, such as
  - genomics
  - group testing for COVID-19

Uncertainty quantification for ML - My course at Penn



github.com/dobriban/Topics-In-Modern-Statistical-Learning

Search or jump to... Pull requests Issues Marketplace Explore

dobriban / Topics-In-Modern-Statistical-Learning Public

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

STAT 991: Topics In Modern Statistical Learning (UPenn, 2022 Spring)

This class surveys advanced topics in statistical learning based on student presentations.

The core topic of the course is uncertainty quantification for machine learning methods. While modern machine learning methods can have a high prediction accuracy in a variety of problems, it is still challenging to properly quantify their uncertainty. There has been a recent surge of work developing methods for this problem. It is one of the fastest developing areas in contemporary statistics. This course will survey a variety of different problems and approaches, such as calibration, prediction intervals (and sets), conformal inference, OOD detection, etc. We will discuss both empirically successful/popular methods as well as theoretically justified ones. See below for a sample of papers.

About

Materials for STAT 9911303: Topics In Modern Statistical Learning (UPenn, 2022 Spring) - focus on uncertainty quantification

machine-learning deep-learning prediction uncertainty-quantification conformal-prediction tolerance-intervals

Topics in Deep Learning - My course at Penn

STAT 991: Topics in deep learning (UPenn)

STAT 991: Topics in Deep Learning is a seminar class at UPenn started in 2018. It surveys advanced topics in deep learning based on student presentations.

Fall 2019

- Syllabus.
- Lecture notes. (~170 pages, file size ~30 MB, mostly covering notes from previous semesters.)

Lectures

Lectures 1 and 2: Introduction and uncertainty quantification (jackknife+, and Pearce at al, 2018), presented by Edgar Dobriban.

Lecture 3: NTK by Jiayao Zhang. Blog post on the off-convex blog.

Lecture 4: Adversarial robustness by Yinjun Wu.

Lecture 5: ELMo and BERT by Dan Deutsch.

Lecture 6: TCAV by Ben Auerbach (adapted from Been Kim's slides).

Lecture 7: Spherical CNN by Arjun Guru and Claudia Zhu.

Lecture 8: DNNs and approximation by Yebiao Jin.

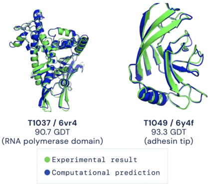
Overview

Context

- Prediction accuracy of machine learning methods is steadily increasing

Context

- Prediction accuracy of machine learning methods is steadily increasing
- Success stories: AlphaFold, cancer tissue image classification, computer vision, NLP ...

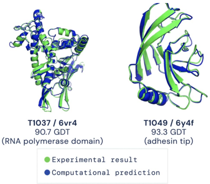


Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: **After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.**

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

Context

- Prediction accuracy of machine learning methods is steadily increasing
- Success stories: AlphaFold, cancer tissue image classification, computer vision, NLP ...



Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: **After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.**

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

- Meanwhile, growing concerns: safety, ethics, energy- and sample-efficiency, uncertainty

Calibration

- Calibration: construct probability predictions that reflect true probabilities. For binary classification, for all appropriate z,

$$P(y = 1 | f(x) = z) \approx z$$

Calibration

- Calibration: construct probability predictions that reflect true probabilities. For binary classification, for all appropriate z,

$$P(y = 1 | f(x) = z) \approx z$$

- Modern finding: powerful ML methods (e.g., deep CNNs) are *over-confident* and *mis-calibrated*

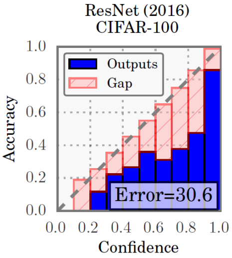
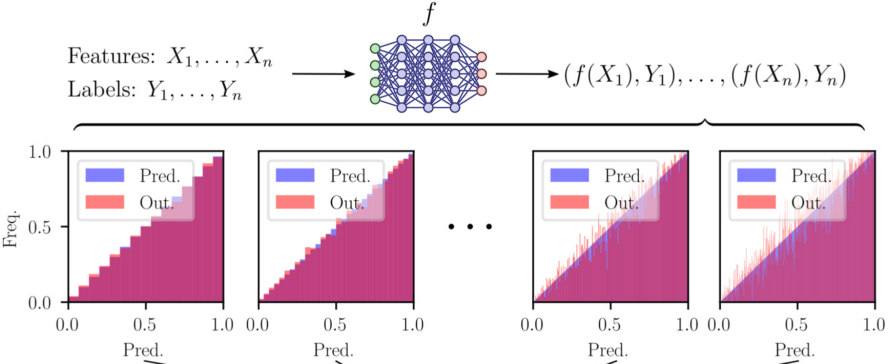


Figure: Guo et al, 2017

T-Cal: An optimal test of calibration



$$T^{\mathrm{d}}_{m,n} = \sum_{\substack{1 \le i \le m \\ |\mathcal{I}_{m,i}| \ge 1}} \frac{1}{n|\mathcal{I}_{m,i}|} \sum_{j_1 \neq j_2 \in \mathcal{I}_{m,i}} [Y_{j_1} - f(X_{j_1})][Y_{j_2} - f(X_{j_2})]$$

-7.65×10^{-4} ✓ 1.60×10^{-5} ✓ $\cdots$ 4.51×10^{-4} ✗ 2.67×10^{-4} ✓

✗ (f is mis-calibrated)

T-Cal

- Theoretical result: minimax optimal under Hölder smoothness
- Empirical results: large power in simulations; can use it to detect mis-calibration of state-of-the-art deep networks

Overview

Motivation

- Machine learning algorithms are becoming integrated into more and more high-stakes decision-making processes.
- Algorithm-based decision-making systems could retain or even amplify historical unfairness in data.

COMPAS Algorithm

Amazon Recruitment System

RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Group Fairness

- Consider a classification problem with two types of feature: the usual feature $X \in \mathcal{X}$, and the protected (or, sensitive) feature $A \in \mathcal{A} = \{0, 1\}$.
- Binary labels in $\mathcal{Y} = \{0, 1\}$, prediction $\hat{Y}$.

Fair Bayes-optimal Classifier under Demographic Parity

Several group fairness measures have been proposed. Measure "unfairness" by *Difference in demographic parity*:

$$DDP = P(\hat{Y} = 1|A = 1) - P(\hat{Y} = 1|A = 0).$$

For input x, let $f(x) := P(\hat{Y} = 1|X = x)$.
Goal: Find δ-fair Bayes-optimal classifier with respect to demographic parity; defined as

$$f^{\star}_{D,\delta} \in \operatorname*{argmin}_{f:|\mathrm{DDP}(f)|\leqslant\delta} [P(Y \neq \widehat{Y})].$$

Main Theorem

Denote

$$p_a := P(A = a)$$
$$\eta_a(x) := P(Y = 1 | A = a, X = x)$$
$$S_a(t) := P(\eta_a(X) > t | A = a)$$

Theorem (Fair Bayes-optimal Classifier under Demographic Parity)

Let $D^\star = \mathrm{DDP}(f^\star)$, where $f^\star$ is unconstrained Bayes-optimal classifier. For any $\delta > 0$, all δ-fair Bayes optimal classifiers $f^\star_{D,\delta}$ have the following form:

- *When $|D^\star| \le \delta$, $f^\star_{D,\delta} = f^\star$.*
- *When $|D^\star| > \delta$, for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$,*

$$f^\star_{D,\delta}(x,a) = I\left(\eta_a(x) > \frac{1}{2} + \frac{(2a-1)t^\star_{D,\delta}}{2p_a}\right) + a t^\star_{D,\delta} I\left(\eta_a(x) = \frac{1}{2} + \frac{(2a-1)t^\star_{D,\delta}}{2p_a}\right), \tag{1}$$

Main Theorem

Theorem (continued)

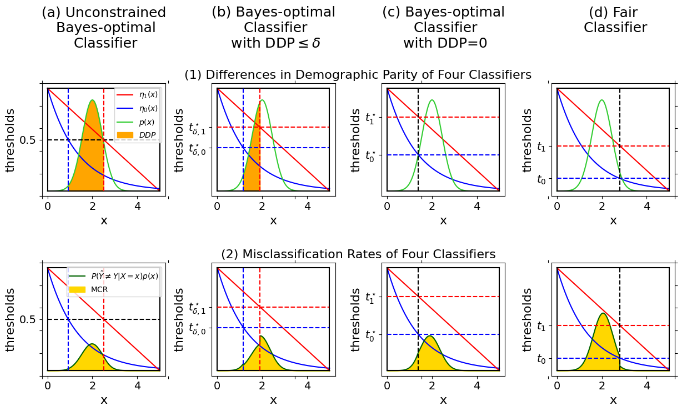
where $t^\star_{D,\delta}$ is defined as

$$t^\star_{D,\delta} = \sup\left\{ t : S_1\left(\frac{1}{2} + \frac{t}{2p_1}\right) > S_0\left(\frac{1}{2} - \frac{t}{2p_0}\right) + \frac{D^\star}{|D^\star|}\delta \right\}. \tag{2}$$

Here, $\tau^\star_{D,\delta} \in [0,1]$ can be an arbitrary constant if $P_{X|A=1}(\eta_1(X) = \frac{1}{2} + \frac{t}{2p_1}) = 0$, and otherwise

$$\tau^\star_{D,\delta} = \frac{S_1\left(\frac{1}{2} + \frac{t}{2p_1}\right) - S_0\left(\frac{1}{2} - \frac{t}{2p_0}\right) - \frac{D^\star}{|D^\star|}}{P_{X|A=1}(\eta_1(X) = \frac{1}{2} + \frac{t}{2p_1})}. \tag{3}$$

Illustration of Theorem



Proof Sketch

Lemma (Generalized Neyman-Pearson lemma)

Let $f_0, f_1, ..., f_m$ be $m+1$ real-valued functions defined on a Euclidean space $\mathcal{X}$. Assume they are ν-integrable for a σ-finite measure ν. Let ϕ_0 be any function of the form

$$\phi_0(x) = \begin{cases} 1, & f_0(x) > \sum_{i=1}^m c_i f_i(x); \\ \gamma(x) & f_0(x) = \sum_{i=1}^m c_i f_i(x); \\ 0, & f_0(x) < \sum_{i=1}^m c_i f_i(x), \end{cases} \tag{4}$$

where $0 \leqslant \gamma(x) \leqslant 1$ for all $x \in \mathcal{X}$.

Proof Sketch

Lemma (continued)

For given constants $t_1, ..., t_m \in \mathbb{R}$*, let* $\mathcal{T}$ *be the class of Borel functions* $\phi : \mathcal{X} \mapsto \mathbb{R}$ *satisfying*

$$\int_{\mathcal{X}} \phi f_i d\nu \leq t_i, \quad i = 1, 2, ..., m. \tag{5}$$

and $\mathcal{T}_0$ *be the set of* ϕ*s in* $\mathcal{T}$ *satisfying* (5) *with all inequalities replaced by equalities. If* $\phi_0 \in \mathcal{T}_0$*, then* $\phi_0 \in \underset{\phi \in \mathcal{T}_0}{\text{argmax}} \int_{\mathcal{X}} \phi f_0 d\nu$*. Moreover, if* $c_i \geqslant 0$ *for all* $i = 1, \ldots, m$*, then* $\phi_0 \in \underset{\phi \in \mathcal{T}}{\text{argmax}} \int_{\mathcal{X}} \phi f_0 d\nu$*.*

Overview

Overview

Uncertainty quantification: calibration

Fairness: Bayes-optimal classifiers

High-dim. statistics & deterministic equivalents

Motivation

- Standard linear model $Y = X\beta + \varepsilon$, where
  1. Y is $n \times 1$ outcome, X is $n \times p$ feature matrix.
  2. β is p-dim parameter
- Ordinary least squares

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

- Mean squared error of OLS, assuming $\mathbb{E}\varepsilon = 0$, $\mathrm{cov}(\varepsilon) = \sigma^2 I_n$

$$\mathbb{E}\|\hat{\beta} - \beta\|^2 = \sigma^2 \operatorname{tr}[(X^\top X)^{-1}]$$

- How large is this? (How hard? How much error?)

Motivation ctd

- When $X_{ij} \sim \mathcal{N}(0,1)$ are iid standard normal,

$$\mathbb{E}\,\mathrm{tr}[(X^\top X)^{-1}] = \frac{p}{n-p-1}.$$

- More general data distributions? There are only approximate expressions.

Deterministic equivalents

- We have sequences of (not necessarily symmetric) $p_n \times p_n$ random matrices A_n and deterministic matrices B_n of growing dimensions
- **Definition**: B_n is a *deterministic equivalent* for A_n,

$$A_n \asymp B_n$$

if

$$\lim_{n \to \infty} |\mathrm{tr}(C_n A_n) - \mathrm{tr}(C_n B_n)| = 0$$

almost surely, for any $p_n \times p_n$ sequence C_n of (not necessarily symmetric) deterministic real matrices with bounded trace norm, i.e.,

$$\lim\sup_{n \to \infty} \|C_n\|_{tr} = \lim\sup_{n \to \infty} \sum_i \sigma_i(C_n) < \infty.$$

e.g, $C_n = c_n c_n^\top$, $\|c_n\|_2$ bounded

Sample covariance matrices

Example (Mestre et al., 2011)

Let $\hat{\Sigma} = X^\top X/n$, where $X = Z\Sigma^{1/2}$ and Z is an $n \times p$ random matrix with iid entries of zero mean, unit variance and finite $8 + \eta$ moment. Also, $\Sigma^{1/2}$ is any sequence of $p \times p$ positive semi-definite matrices satisfying $\sup \|\Sigma\|_2 < \infty$. As $n, p \to \infty$ proportionally, for any $\lambda > 0$

$$(\widehat{\Sigma} + \lambda I_p)^{-1} \asymp (q_p \Sigma + \lambda I_p)^{-1} \,,$$

where q_p is the solution of a fixed point equation.

- This is the simplest way I know how to think of a broad class of results in random matrix theory.

Distributed linear regression

- Standard linear model $Y = X\beta + \varepsilon$
- Data distributed across k machines. The i-th machine has matrix X_i ($n_i \times p$) and outcomes Y_i.

$$X = \begin{bmatrix} X_1 \\ \dots \\ X_k \end{bmatrix}, \; Y = \begin{bmatrix} Y_1 \\ \dots \\ Y_k \end{bmatrix}$$

- Global least squares - infeasible
- *Local* least squares estimator $\hat{\beta}_i = (X_i^\top X_i)^{-1} X_i^\top Y_i$ (assume $n_i > p$)
- Send to parameter server, average
- How does this compare to OLS on full data?

A general framework

- Important to study not only estimation, but also prediction/test error, residual error, confidence intervals etc
- Predict the linear functional

$$L_A = A\beta + Z$$

- Using the plug-in estimator

$$\hat{L}_A(\hat{\beta}_0) = A\hat{\beta}_0$$

- A - fixed $d \times p$ matrix; mean and covariance of Z has the structure: $Z \sim (0, h\sigma^2 I_d)$, $h \geqslant 0$
- The noise can be correlated with ε: $\mathrm{Cov}[\varepsilon, Z] = N$ (e.g., to study residuals)
- Relative efficiency:

$$E(A; X_1, \ldots, X_k) := \frac{\mathbb{E}\|L_A - \hat{L}_A(\hat{\beta})\|^2}{\mathbb{E}\|L_A - \hat{L}_A(\hat{\beta}_{dist})\|^2}.$$

Examples: Predict $L_A = A\beta + Z$ by $\hat{L}_A(\hat{\beta}_0) = A\hat{\beta}_0$

Statistical learning problem	L_A	$\hat{L}_A$	A	h	N
Parameter estimation	β	$\hat{\beta}$	I_p	0	0
Regression function estimation	$X\beta$	$X\hat{\beta}$	X	0	0
Confidence interval for marginal effect	β_j	$\hat{\beta}_j$	$E_j^\top$	0	0
Test error	$x_t^\top \beta + \varepsilon_t$	$x_t^\top \hat{\beta}$	$x_t^\top$	1	0
Training error/Residual	$X\beta + \varepsilon$	$X\hat{\beta}$	X	1	$\sigma^2 I_n$

Finite sample results

- When $h = 0$ (no noise), the MSE of estimating $L_A = A\beta$ by OLS $\hat{L}_A = A\hat{\beta} = A(X^\top X)^{-1}X^\top Y$ is

$$M(\hat{\beta}) = \sigma^2 \cdot \mathrm{tr}\left[(X^\top X)^{-1}A^\top A\right].$$

- For the distributed estimator $\hat{\beta}_{dist}(w) = \sum_i w_i \hat{\beta}_i$, $\sum_i w_i = 1$

$$M(\hat{\beta}_{dist}) = \sigma^2 \cdot \sum_{i=1}^{k} w_i^2 \cdot \mathrm{tr}\left[(X_i^\top X_i)^{-1}A^\top A\right].$$

- So optimal efficiency is

$$E(A; X_1, \ldots, X_k) = \mathrm{tr}\left[(X^\top X)^{-1}A^\top A\right] \cdot \sum_{i=1}^{k} \frac{1}{\mathrm{tr}\left[(X_i^\top X_i)^{-1}A^\top A\right]}.$$

CDE: $\mathrm{tr}[(X_i^\top X_i)^{-1}A^\top A] \asymp \frac{p}{n_i - p} \cdot \mathrm{tr}[\Sigma^{-1}A^\top A]/p$.

Plot efficiencies

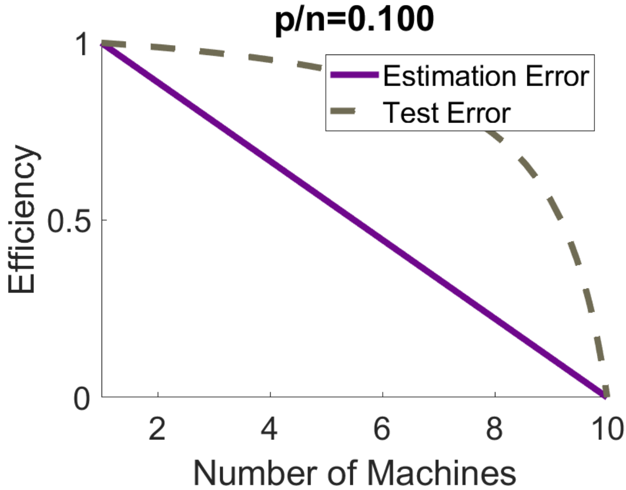


Figure: The loss of efficiency is much worse for estimation ($\frac{\mathbb{E}\|\hat{\beta}-\beta\|^2}{\mathbb{E}\|\hat{\beta}_{dist}-\beta\|^2}$) than for test error ($\frac{\mathbb{E}(x_t^\top\hat{\beta}-y_t)^2}{\mathbb{E}(x_t^\top\hat{\beta}_{dist}-y_t)^2}$).