CSIE5002 Prediction, learning, and games

Lecture 1: Learning with & without probability

Yen-Huan Li (yenhuan.li@csie.ntu.edu.tw)

18.02.2019

Department of Computer Science and Information Engineering
National Taiwan University

Abstract

This lecture consists of two parts. The first part is about the course logistics. The second part discusses the tentative topics to be covered in this course.

Table of contents

1. Logistics
2. What is this course about?
3. Case study: Logistic regression
4. Case study: Calibrated forecasting
5. Conclusions

Logistics

Caveat

- This is a *theory course*.
- There will be *no programming & no deep learning*.
- We will focus on the *ideas* instead of implementation details.
- You are expected to read and work out *rigorous mathematical proofs*.
- # students who withdrawn last semester: 17/41.
- There may be no TA...

Prerequisites

Calculus, linear algebra, probability, and *mathematical maturity*.

Take a look at Homework #0 on the course website.

- *To be graded!* Due at 2pm, March 4, 2019.

There will not be any review of required mathematical knowledge.

Grading

- 40%: Homework/quiz.
  - Only the *three highest grades* are counted.
- 20%: Midterm exam.
  - One and only one *hand-written double-sided A4-sized cheatsheet* is allowed.
- 40%: Final project report.
  - Review of a research topic (instead of a paper!), report of a novel research result, etc. Must be *theory-oriented*.
  - One double-column *report* of at most eight pages & one *oral presentation*.
  - Details TBD.

Late submission & cheating

A late submission results in a discounted score.

- Late by h hours $\Rightarrow$ Score scaled by $(1 - 0.05h)$.

Cheating, plagiarism, and/or self-plagiarism are not tolerated.

- First time $\Rightarrow$ 0 point for the homework/exam/project.
- Second time $\Rightarrow$ 0 point for the whole course.

Course attendance & auditing

Course attendance is not counted for grades.

Auditing is welcome, as long as there are seats available.

Schedule

There are 18 weeks in this semester.

No courses in the weeks officially for midterm and final exams.

Tentative: The last two weeks (excluding the week for final exam) will be for project presentations.

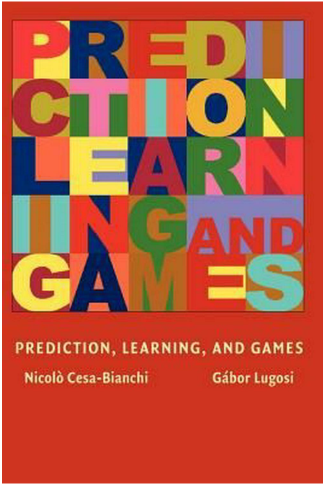
Course website

`https://cool.ntu.edu.tw/courses/233`

- Any announcement.
- Lecture slides/notes. (Released before the course.)
- Homework problems.
- Submission of homework and project reports.

What is this course about?

What is this course about? (1/2)


PREDICTION LEARNING AND GAMES
PREDICTION, LEARNING, AND GAMES
Nicolò Cesa-Bianchi
Gábor Lugosi

N. Cesa-Bianchi *et al. Prediction, Learning, and Games*. 2006.
- A classic reference but not the textbook of this course.

What is this course about? (2/2)

Online learning @ machine learning

Individual sequence prediction @ information theory

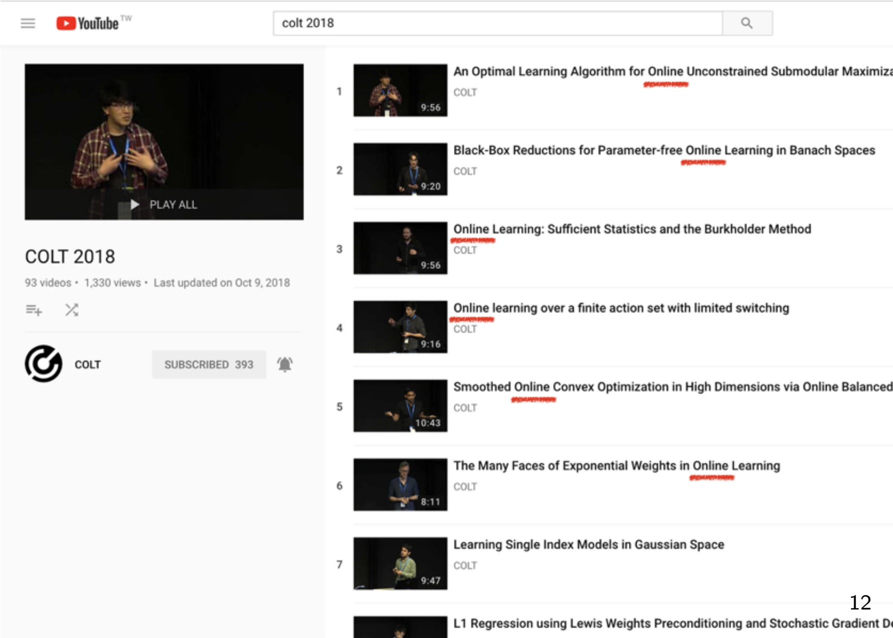
Learning in games @ game theory

Prequential analysis @ statistics

Competitive analysis @ theoretical computer science

Closely related to certain algorithms @ optimization theory

A popular topic in learning theory



colt 2018

COLT 2018

93 videos • 1,330 views • Last updated on Oct 9, 2018

COLT SUBSCRIBED 393

1 An Optimal Learning Algorithm for Online Unconstrained Submodular Maximiza
COLT
9:56

2 Black-Box Reductions for Parameter-free Online Learning in Banach Spaces
COLT
9:20

3 Online Learning: Sufficient Statistics and the Burkholder Method
COLT
9:56

4 Online learning over a finite action set with limited switching
COLT
9:16

5 Smoothed Online Convex Optimization in High Dimensions via Online Balanced
COLT
10:43

6 The Many Faces of Exponential Weights in Online Learning
COLT
8:11

7 Learning Single Index Models in Gaussian Space
COLT
9:47

L1 Regression using Lewis Weights Preconditioning and Stochastic Gradient De

Case study: Logistic regression

Logistic regression: Algorithm

Training data. A sequence of pairs $(x_i, y_i) \in \mathbb{R}^p \times \{\pm 1\}$, $i = 1, \ldots, n$.

Goal. Find a vector $\hat{w}_n \in \mathbb{R}^p$, such that

$$\operatorname{sign}(\langle x, \hat{w}_n \rangle) \approx y,$$

where $(x, y) \in \mathbb{R}^p \times \{\pm 1\}$ denotes the possibly unseen *test data*.

Logistic regression. Set $\hat{w}_n$ as a minimizer of the function

$$f_n(w) := \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-y_i \langle x_i, w \rangle}\right).$$

Statistical perspective (1/3)

Model. Given $x \in \mathbb{R}^p$, consider the distribution

$$\mathsf{P}\left(y = 1\right) = 1 - \mathsf{P}\left(y = -1\right) = \frac{1}{1 + \mathrm{e}^{-\langle x, w^\natural \rangle}},$$

for some unknown parameter $w^\natural \in \mathbb{R}^p$. Each y_i is a random variable following the model with x_i, and $y_1, \ldots, y_n$ are statistically independent.

Logistic regression $=$ maximum-likelihood (ML) estimation.
Then, $\hat{w}_n$ is also a maximizer of the *likelihood function*

$$\Phi(w) := \prod_{i=1}^{n} \left(\frac{1}{1 + \mathrm{e}^{-y_i \langle x_i, w \rangle}} \right).$$

Statistical perspective (2/3)

Suppose the statistical model is true.

Statistical consistency. As the data size n goes to infinity, the ML estimator converges in probability to the unknown parameter, i.e.,

$$\lim_{n\to\infty} \mathsf{P}\left(\|\hat{w}_n - w^\natural\|_2 > \varepsilon\right) = 0, \quad \forall \varepsilon > 0.$$

Asymptotic normality. In distribution,

$$\sqrt{n}\left(\hat{w}_n - w^\natural\right) \to \mathcal{N}(0, F^{-1}) \quad \text{as} \quad n \to \infty,$$

where $F \in \mathbb{R}^{p\times p}$ denotes the *Fisher information matrix*.

A. W. van der Vaart. *Asymptotic Statistics*. 1998.

L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. 1986.

Statistical perspective (3/3)

Why should we expect the data follows the (weird) probability distribution?

What if the random variables $y_1, \ldots, y_n$ are not statistically independent?

“All models are wrong, but some are useful.” — G. Box

Critique

"The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems." — L. Breiman

L. Breiman. Statistical modeling: The two cultures. 2001.

Statistical learning perspective (1/5)

Model. Assume that $(x_1, y_1), \ldots, (x_n, y_n)$ are independent and identically distributed (i.i.d.) random variables following an *unknown probability distribution.*

Goal. Find some $w^\star \in \mathbb{R}^p$ that minimizes the function

$$R(w) \coloneqq \mathsf{P}\left(y \left\langle x, w \right\rangle < 0\right) = \mathsf{E}\, \mathbb{1}_{\{\, y\langle x,w\rangle \,\}<0}.$$

Empirical risk minimization. (ERM) Approximate $w^\star$ by a minimizer of the *empirical risk* function

$$\hat{R}_n(w) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\, y_i\langle x_i,w\rangle<0 \,\}}.$$

Statistical learning perspective (2/5)

Theorem. Computing an empirical risk minimizer is NP-hard.

Convex surrogate. Consider the logistic loss function

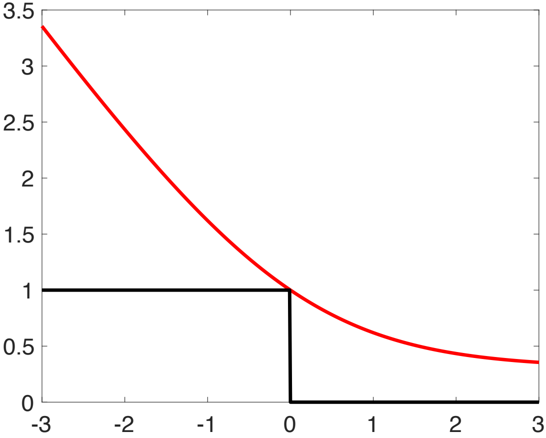
$$\rho(w; y, x) \coloneqq \log\left(1 + \mathrm{e}^{-y\langle x,w\rangle}\right)$$

as a convex approximation of the function $\mathbb{1}_{\{\, y\langle x,w\rangle<0\,\}}$. Then a minimizer of the associated empirical risk function

$$\tilde{R}_n(w) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \rho(w; y_i, x_i)$$

can be efficiently computed by an appropriate convex optimization algorithm. Then, we get logistic regression.

Statistical learning perspective (3/5)



$$f(x) \coloneqq \log\left(1 + \mathrm{e}^{-x}\right) + (1 - \log 2), \qquad g(x) \coloneqq \mathbb{1}_{\{\, x<0 \,\}}$$

Statistical learning perspective (4/5)

Probably approximately correct (PAC) guarantee. Set $\tilde{w}_n$ as a minimizer of $\tilde{R}_n$ on the set

$$\mathcal{W} := \{ \; \|w\|_1 \leq \nu \mid w \in \mathbb{R}^p \; \}$$

for some $\nu > 0$. Then, with probability at least $1 - \delta$, $\delta \in]0, 1[$, it holds that

$$R(\tilde{w}_n) - \inf_{w \in \mathbb{R}^p} R(w) = O\left(\left[\frac{\nu^2 \log(1/\delta)}{n} \right]^{1/4} \right) + O\left(\inf_{w \in \mathcal{W}} R(w) - \inf_{w \in \mathbb{R}^p} R(w) \right).$$

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. 2004.

S. Boucheron *et al.* Theory of classification: A survey of some recent advances. 2005.

Learning theoretic perspective (5/5)

~~Why should we expect the data follows the (weird) probability distribution?~~

What if the random variables $y_1, \ldots, y_n$ are not statistically independent?

Induction vs. transduction.
Induction: Recover the probability distribution from the data, and then find a classifier.
Transduction: Directly find a classifier from the data.

V. Vapnik. *Estimation of Dependences Based on Empirical Data*. 1982.

Attempts to relax the i.i.d. assumption

One may assume the following more general models for the data.

- α-, $\beta-$, and other mixing processes
- Ergodic stationary process
- ...

Question. What if the data does not satisfy the assumption?

H. Hang and I. Steinwart. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. 2017.

L. Györfi *et al.* Empirical log-optimal portfolio selections: A survey. 2012.

Another perspective (1/2)

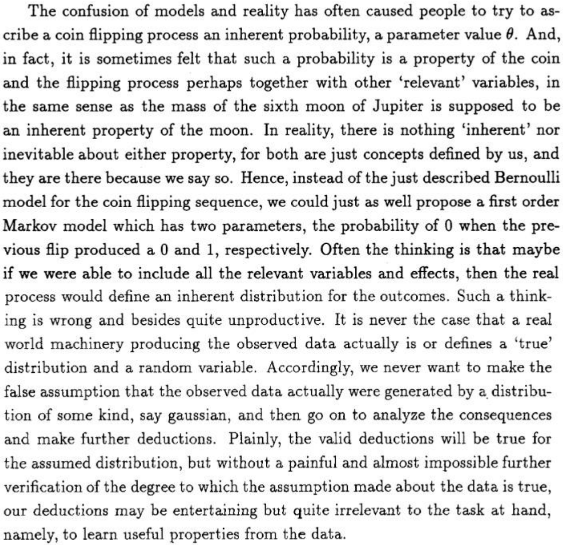
"It will appear that a measurable uncertainty, or 'risk' proper, as we shall use the term, is so far different from an unmeasurable one that it is not in effect an uncertainty at all." — F. Knight

"Data are just numbers. Except in special situations, assuming that the data are random is something we do by reflex but it is often not justifiable" — L. Wasserman

F. Knight. *Risk, Uncertainty, and Profit.* 1921.

L. Wasserman. Statistics without probability (individual sequences). 2012.

Another perspective (2/2)



The confusion of models and reality has often caused people to try to ascribe a coin flipping process an inherent probability, a parameter value θ. And, in fact, it is sometimes felt that such a probability is a property of the coin and the flipping process perhaps together with other 'relevant' variables, in the same sense as the mass of the sixth moon of Jupiter is supposed to be an inherent property of the moon. In reality, there is nothing 'inherent' nor inevitable about either property, for both are just concepts defined by us, and they are there because we say so. Hence, instead of the just described Bernoulli model for the coin flipping sequence, we could just as well propose a first order Markov model which has two parameters, the probability of 0 when the previous flip produced a 0 and 1, respectively. Often the thinking is that maybe if we were able to include all the relevant variables and effects, then the real process would define an inherent distribution for the outcomes. Such a thinking is wrong and besides quite unproductive. It is never the case that a real world machinery producing the observed data actually is or defines a 'true' distribution and a random variable. Accordingly, we never want to make the false assumption that the observed data actually were generated by a distribution of some kind, say gaussian, and then go on to analyze the consequences and make further deductions. Plainly, the valid deductions will be true for the assumed distribution, but without a painful and almost impossible further verification of the degree to which the assumption made about the data is true, our deductions may be entertaining but quite irrelevant to the task at hand, namely, to learn useful properties from the data.

J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. 1998.

Formalization of PAC learning theory

Framework.

- Data: I.i.d. random variables $z_1, \dots, z_n \in \mathcal{Z}$ following an unknown probability distribution P.
- Hypothesis class: A set $\mathcal{H}$.
- Loss: A function $\rho : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$.
- Risk: The expected loss $R : h \mapsto \mathsf{E}_z\, \rho(h, z)$, where z is a random variable following P.
- Goal: Given the data, find the optimal hypothesis

$$h^\star \in \arg\min_h \{\, R(h) \mid h \in \mathcal{H} \,\}\,.$$

Online learning perspective (1/3)

An online learning protocol. Let the initial cumulative loss $L_0 = 0$. For $t = 1, 2, \ldots, T$, the following occur in order.

1. $\textsc{Learner}$ announces $h_t \in \mathcal{H}$.
2. $\textsc{Reality}$ chooses z_t and announces $f_t : h \mapsto \rho(h, z_t)$.
3. Update $\textsc{Learner}$'s loss: $L_t = L_{t-1} + f_t(h_t)$.

Regret. The regret is defined as

$$R_T := \sum_{t=1}^{T} f_t(h_t) - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} f_t(h).$$

Goal. Achieve a sublinear regret (in T).

Online learning perspective (2/3)

Characteristics.

- There is no probability in the model. The data is just a sequence of elements in $\mathscr{Z}$.
- The expected loss (i.e., the risk) is not defined. Therefore, we use the average loss as a performance index.
- There should be no theory-practice mismatch. A regret guarantee holds even when the data is generated by a strategic adversary.

Online learning perspective (3/3)

Applications.

- Data compression.
- Portfolio selection.
- Game theory.
- Prediction with expert advice.
- Complexity of a sequence
- Convex optimization.
- PAC learning.
- Routing (bandit).
- Online advertisement (bandit)..
- ...

Case study: Calibrated forecasting

Weather forecasting

How do we estimate the *chance of rain* tomorrow?

Classic approach.

1. Mathematical modeling (probabilistic model).
2. Design an algorithm for the probabilistic model.
3. Test the algorithm empirically to evaluate the theory-practice mismatch (due to model-reality mismatch).
4. Poor empirical performance → Design another algorithm, or propose another probabilistic model.

Falsifiability

"I shall require that [the] logical form [of the theory] shall be such that it can be singled out, by means of empirical tests, in a negative sense: it must be possible for an empirical scientific system to be refuted by experience." — K. Popper

K. Popper. *The logic of scientific discovery*. 1959.

Weather forecasting as an online problem

Protocol. For $t = 1, \ldots, T$, the following occur.

- $\textsc{Learner}$ announces $x_t \in [0, 1]$.
- $\textsc{Reality}$ announces $y_t \in \{ 0, 1 \}$.

Observation. There is not any probabilistic model in the protocol.

Question. How do we interpret the following sentence? *The chance of rain tomorrow is 30%.*

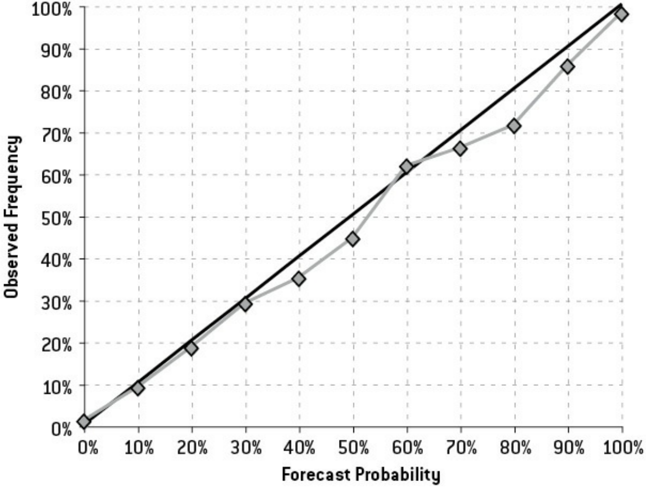
Question. When can we claim that a weather forecaster is accurate?

Calibration as a criteria

"'Suppose that, in a long (conceptually infinite) sequence of weather forecasts, we look at all those days for which the forecast probability of precipitation was, say, close to some given value ω and (assuming these form an infinite sequence) determine the long run proportion p of such days on which the forecast event (rain) in fact occurred. The plot of p against ω is termed the forecaster's empirical calibration curve. If the curve is the diagonal $p = \omega$, the forecaster may be termed (empirically) well calibrated" — A. P. Dawid

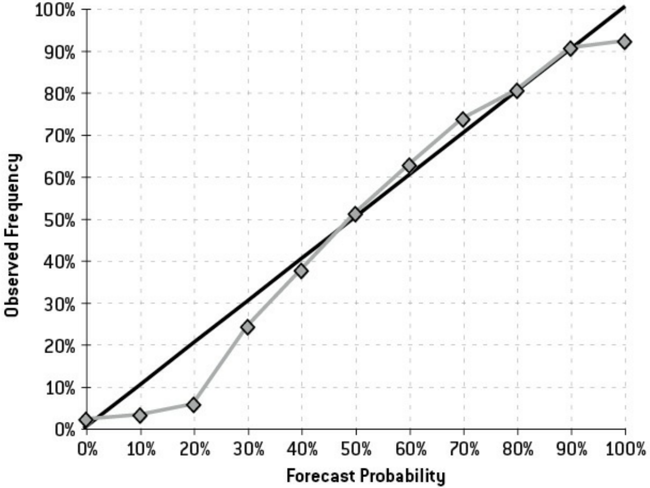
A. P. Dawid. The well-calibrated Bayesian. 1982.

Empirical calibration curve of U.S. National Weather Service



N. Silver. *The Signal and the Noise*. 2012.

Empirical calibration curve of a weather channel



N. Silver. *The Signal and the Noise*. 2012.

Is calibrated weather forecasting possible?

Observation. If $\textsc{Reality}$ may choose y_t after seeing x_t, then calibrated weather forecasting is impossible.

Observation. If $\textsc{Learner}$ adopts a deterministic algorithm, then calibrated weather forecasting is impossible.

Theorem. (informal version) There exists a randomized algorithm for $\textsc{Learner}$, such that almost surely, the weather forecasts are asymptotically calibrated as T goes to infinity.

D. Oakes. Self-calibrating priors do not exist. 1985.

D. P. Foster and R. V. Vohra. Asymptotic calibration. 1998.

Implications

Theorem. (informal version) Suppose that in a sequential game, each player plays optimally with regard to the forecasts of a calibrated forecasting algorithm for other players' actions. Then, the empirical distribution of the players' actions converge to a Nash equilibrium (or a generalized notion of it).

Theorem. (informal version) Expert testing is difficult. That is, if a test does not reject the truth, then the test can be passed ignorantly.

S. Kakade and D. P. Foster. Deterministic calibration and Nash equilibrium. 2008.
W. Olszewski. Calibration and expert testing. 2015.

Conclusions

Summary

- Check the course website for all information.
- Statistical, statistical learning, and online learning perspectives.
- Calibrated weather forecasting is doable but may not be beneficial.

Next lecture

- Crash course of learning with probability.