Statistical Learning

Fengnan Gao[1,2]

13.9.2017

[1]School of Data Science, Fudan University

[2]Shanghai Center for Mathematical Sciences

fngao@fudan.edu.cn

Course Overview

Overview i

- The course is primarily modeled after **Alexander Rakhlin**'s course.
- This course will focus on theoretical aspects of Statistical Learning.
- We will analyze learning with i.i.d. data using classical tools: concentration inequalities, random averages, covering numbers, and combinatorial parameters.
- The minimax approach, which we emphasize throughout the course, offers a systematic way of comparing learning problems. Beyond the theoretical analysis, we will discuss learning algorithms and, in particular, an important connection between learning and optimization.
- Our framework will give a handle on developing near-optimal and computationally efficient algorithms.

Overview ii

- We will illustrate this on the problems of matrix completion, link prediction, and other.
- Time permitting, we will make excursions into Information Theory and Game Theory, and show how our new tools seamlessly yield a number of interesting results.
- **Prerequisites**: Probability Theory and Linear Algebra
- Course Homepage: `http://www.sdspeople.fudan.edu.cn/gaofengnan/teaching/1718FSSCMS.html`
- I am teaching another course *Statistics: Principles, Methods and R (I)* for the School of Data Science.

Contents

Some logistic Issues

- The first lecture is given by Hou Yanxi.
- The second lecture on September 20 is cancelled, and will be rearranged pending further discussions.
- The lecture on October 4 will not happen because of the compulsory Mid-Autumn holiday.
- The final (and only) exam will take place on January 3, 2018.

References

- The primary lecture notes: `http://www-stat.wharton.upenn.edu/~rakhlin/courses/stat928/`
- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. New York: Springer series in statistics, 2001.
- James, Gareth, et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.

Exams and Grading

- This is a course focused on mathematical theory, not on applications.
- Time permitting, we may arrange paper reading sessions — students form groups and read papers on a certain list relevant to the course subject.
- The final grades will be a weighted average of the following evaluations:
  - 30% — Homework and participations
  - 70% — Final exam

Introduction to the Learning problem

What is Statistical Learning? i

- *Learning* can be loosely defined as the *ability to improve performance after observing data*.
- Applications of learning methods include:
  - Facial detection/recognition
  - Predication of stock markets and weather patterns
  - Speech recognition
  - Learning user's search preferences, placement of relevant ads
- A "learning" program has to adapt. The goal should be
  - To encode, for a particular application, as much as the domain-specific knowledge as necessary
  - and leave *enough flexibility* for the system to improve upon observing the data
- We will make it precise that *no single learning algorithm works universally*.

What is Statistical Learning? ii

- It is *not* our goal to make computers learn everything at once: each application requires prior domain knowledge from the expert.
- The goal of learning theory is to
  - develop general guidelines and algorithms,
  - and prove guarantees about various natural assumptions
- We will make it precise whether the "prior knowledge" of the expert should be encoded in the learning algorithm with the minimax approach.
- Looking at problems through a minimax lens!
- Three key aspects in learning theory
  1. how data are generated
  2. how the performance is measured
  3. where we place prior knowledge

What are the learning problems i

- Starting with *supervised learning*
- Suppose you are statistical consultants hired to provide advice on how to improve sales of some product
- The `Advertising` data set consists of the `sales` of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: `TV`, `radio` and `newspaper`.
- The client can control the advertising expenditures in each of three media, hoping for better sales.
- If we determine that there is an association between advertising and sales, we can instruct the client to adjust advertising budgets to increase sales.
- In the above setting,

What are the learning problems ii

  - the advertising budgets are *input variables*, typically denoted using X, with a subscript to distinguish them
    - X_1 might be the `TV` budge, X_2 `radio`, ...
    - Also called *predictors*, *independent variables*, *features*, ...
  - the `sales` is an *output variable*, often called the *response* or *dependent varible*, and typically denoted using Y
- More generally, suppose that we observe
  - a quantitative response Y, and
  - p different predictors X_1,... X_p.
- Assume the relation between Y and $X = (X_1, \ldots, X_p)$

$$Y = f(X) + \varepsilon$$

- f is some fixed but unknown function of $X_1, \ldots, X_p$, and ε is a random *error term*, independent of X and has mean zero.
- f represents the *systematic* information that X provides about Y
- Statistical learning refers to a set of approaches for estimating f

Why estimate f?—Prediction i

- Two main reasons—*prediction* and *inference*
- Suppose that a set of inputs X are readily available, but the output Y cannot be easily obtained.
- If $\hat{f}$ is our estimate for f, then we may predict Y using

$$\hat{Y} = \hat{f}(X)$$

- For example, $X_1, \ldots, X_p$ are characteristics of a patient's blood sample, easily measured in a lab, and Y is a variable encoding the patient's risk for a severe adverse reaction to a particular drug.
- With a accurate prediction of Y given X, we can avoid giving the drug to patients with high risk of an adverse reaction
- f is a black box!

Why estimate f?—Prediction ii

- The accuracy of $\hat{Y}$ depends on two quantities

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \varepsilon - \hat{f}(X)]^2 \\ &= E[f(X) - \hat{f}(X)]^2 + \text{Var}[\varepsilon] \end{aligned}$$

- $[f(X) - \hat{f}(X)]^2$ can be made smaller by having better estimate of f, but $\text{Var}[\varepsilon]$ is not reducible.
- Suppose that $f \in \mathcal{F}$, then we are interested in finding an estimator to minimize the following difference

$$E[f(X) - \hat{f}(X)]^2 - \inf_{f' \in \mathcal{F}} E[f(X) - f'(X)]^2$$

- The above display quantifies the *loss* relative to the functional class $\mathcal{F}$, which will be studies extensively in the course.

Why estimate f?—Inference

- Suppose that we are interested in understanding the way Y is affected as $X_1, X_2, ; X_p$ change, not in predicting Y
- f can no longer be treated as a black box, because we need the exact form to interpret how Y changes as a function of $X_1, \ldots X_p$
- In the inference setting, one may ask the following questions
  1. Which predictors are associated with the response?
  2. What is the relationship between the response and each predictor?
  3. Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?
- Recall the `Advertising` data set, one may be interested in asking
  - Which media contribute to sales?
  - Which media generate the biggest boost in sales? or
  - How much increase in sales is associated with a given increase in TV advertising?

How to estimate f?—Parametric Methods

- Suppose that we have *training data* $(X_i, Y_i)_{i=1}^n$ where X_i's are p-dimensional vectors and are fixed a priori. ($p << n$)
- Suppose that $f(X)$ is linear in X

$$f(X) = \sum_{i=1}^{p} \beta_i X_i$$

- The problem reduces to the *estimation* of $\beta = (\beta_i)_{i=1}^p$
- This problem is commonly known as *linear regression*
- A reasonable estimate of β is obtained via

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2$$

How to estimate f—Nonparametric Methods

- Suppose that we have *training data* $(X_i, Y_i)_{i=1}^n$ where X_i's are p-dimensional vectors and are fixed a priori.
- We do *not* assume the linear form of β, but instead imposes some *regularity* conditions on f, e.g., the smoothness of f
- It is much more difficult as we cannot reduce the problem to the estimation of a few numbers.
- We still need to find a way to estimate f and quantify how good the estimator is
- This is, again, achievable with the minimax framework
- We will discuss in details in the upcoming lectures.

Thank you!