GR 6103, (Spring 2021) Applied Statistics III
Nonparametric Theory in Machine Learning

Schedule

Time: Mondays 3:10-5:30pm, **Location:** Zoom:
https://columbiauniversity.zoom.us/j/97983713376?pwd=V05ENGNBaVJOdm9kU3NvdmtETzlSQT09

Instructor: Samory Kpotufe. *email:* skk2175@columbia.edu
Office hours: Mondays 5:45 pm to 6:30 pm, Zoom (same as above).

Description

First, despite the "applied" title, this course is on **Statistical Machine Learning Theory** (emphasis on *theory*, while *applied* is simply what the department calls non-traditional aspects of Statistics). The plan is to cover new theoretical insights on the performance of *nonparametric methods* in ML, while keeping in mind practical realities of modern ML, e.g., computational constraints, costs of data acquisition, etc. *Aim:* at the end of the class you should be able to parse most theoretical papers on the subject.

We will start with traditional nonparametric methods, arising early in Statistics proper, and dive into the rich philosophy and the key mathematical concepts underlying nonparametrics (e.g. minimax upper and lower-bounds, adaptivity, smoothness classes, etc). We will then spend much time on the more common, and often better performing, *kernel machines* arising in ML (but with roots in other traditional areas of applied math).

• *Traditional Nonparametric Methods.* These are the usual suspects, namely, kernel density estimation, k-NN classification and regression, tree-based prediction, nonlinear regression via basis expansion, etc. Many of these approaches appear in ML either directly, or as part of more sophisticated learning pipelines (e.g. the use of k-NN in prediction layers of neural networks, e.g., for transfer tasks).

• *Kernel Machines.* Common examples are SVMs, Kernel Ridge, Gaussian Processes, Kernel PCA, Maximum Mean Discrepancy, etc. They remain one of the most successful and better understood methods in Machine Learning. In fact, there has been recent attempts to cast less understood methods like Neural Nets as forms of kernel machines, in the hope for better insights. In parallel, much recent research effort has gone into scaling up kernel methods to meet the various computational challenges of real-world ML. All of this has called for refined understanding of their statistical properties, rooted in mathematical tools from functional analysis and (linear) operator theory.

Basic background

While we'll try to have self-contained discussions, familiarity with the following will be helpful.

- Basic probability concepts, e.g., measurability, integration, characterictic functions, $\mathcal{L}_p(\mu)$ spaces, ...
- Basic Linear Algebra, e.g., vector spaces, Spectral and Singular Value theorems, ...
- Basic Real Analysis, e.g. completeness, compactness, forms of continuity, ...
- Basic Statistical concepts, e.g., ℓ_p Risks, Regularization, basic concentration inequalities such as Chernoff, Bernstein, ...

Useful Reading: I'll be giving out recommendations on papers and books as class progresses. Some authors of books on the subject (to get a sense): Lászlo Györfi, Alexandre Tsybakov, Ingo Steinwart, ...

Grading: This is yet to be ironed out, but the idea is to mostly base evaluation on class participation and engagement, along with projects, group homeworks, and or paper presentations.