DVR 0065528

ESI SENIOR RESEARCH FELLOW LECTURE COURSE
Winter Term 2019/20

The Erwin Schrödinger International Institute for Mathematics and Physics (ESI) of the University of Vienna offers the following Lecture Course held by a Senior Research Fellow in residence during the Winter Term 2019/20:

Geometric Aspects of Statistical Learning Theory
Shahar Mendelson (Australian National U, Canberra)

Lecture Course (250108 V)): January 7 - 22, 2020
Time: Tuesday and Wednesday, 13:15 - 14:45
Start: Tuesday, January 7, 2020
End: Wednesday January 22, 2020

Venue: ESI, Schrödinger Lecture Hall

Statistical learning theory plays a central role in modern data science, and the question we focus on in this course has been the key question in the area since the late 60s.

To describe the problem, let F be a class of functions defined on a probability space (Ω, μ), and consider a random variable Y. The goal is to find some function that is almost as close to Y as the best approximation to Y in F. The twist is the limited information at one's disposal: while the class F is given, Y is not known, nor is the underlying measure μ. Instead, if X is distributed according to μ and (X,Y) is the joint distribution of X and Y, one is given N independent sample points $\{(X_1,Y_1),...,(X_N,Y_N)\}$, selected according to (X,Y). Using the sample and the identity of F, one has to generate a (data-dependent) function $\hat{f}$ that approximates Y. The success of the choice is determined by the accuracy-confidence tradeoff: one has to ensure that with probability $1-\delta$ with respect to $(X_i,Y_i)_{i=1}^N$,

$$\mathbb{E}\left((\hat{f}(X)-Y)^2|(X_i,Y_i)_{i=1}^N\right) \leq \inf_{f\in F} \mathbb{E}(f(X)-Y)^2+\varepsilon. \tag{1}$$

Naturally, high accuracy and high confidence are conflicting requirements: the higher the wanted accuracy, the more difficult it is to ensure that (**??**) holds with high confidence.

Question: Given a class F, a distribution (X,Y), and a sample size N, what is the optimal tradeoff between the wanted accuracy ε and the confidence $1-\delta$? And, what is the right choice of $\hat{f}$ that attains the optimal tradeoff?

The aim of this course: is to show that this question has a strong geometric flavour and to highlight some of the ideas in empirical processes theory and in asymptotic geometric analysis that have led to its solution—under minimal assumptions on the class F and on (X,Y).

The plan of the lecture course:

(1) Why is learning possible? The definition of a learning problem; what can we hope for; the quadratic and multiplier processes; complexity measures of classes of functions (2 hours).

(2) The small-ball method and (some of) its applications (4 hours).

(3) Median-of-means tournaments and the solution for convex classes (2 hours).

(4) Complexity measures of classes revisited: chaining methods for Bernoulli and gaussian processes; combinatorial dimension and metric entropy (4 hours).

Prerequisites: To get the most out of the course, one is likely to require the knowledge of (graduate level) probability/measure theory and functional analysis, as well as some mathematical maturity. Most of the material I will cover can be found in the course's lecture notes, and because of the nature of the course, some of the details will not be presented during the lectures.

Course website: https://www.esi.ac.at/events/e32/
Link to the course directory

Christoph Dellago
Director