TOPICS IN STATISTICAL LEARNING THEORY

Peter L. Bartlett & Sanjoy Dasgupta
(Stéphane Boucheron & Nicolas Vayatis, editors)

Panoramas et Synthèses

Numéro 57

SOCIÉTÉ MATHÉMATIQUE DE FRANCE

Diffusion

Maison de la SMF
Case 916 - Luminy
13288 Marseille Cedex 9
France
`christian.smf@cirm-math.fr`

AMS
P.O. Box 6248
Providence RI 02940
USA
`www.ams.org`

Tarifs

Vente au numéro : 38 € ($ 57)

Des conditions spéciales sont accordées aux membres de la SMF.

Secrétariat

Panoramas et Synthèses
Société Mathématique de France
Institut Henri Poincaré, 11, rue Pierre et Marie Curie
75231 Paris Cedex 05, France
Tél : (33) 01 44 27 67 99 • Fax : (33) 01 40 46 90 96
`panoramas@smf.emath.fr` • `http://smf.emath.fr/`

ISSN 1272-3835

ISBN 978-2-85629-964-7

Directeur de la publication : Fabien Durand

PANORAMAS ET SYNTHÈSES 57

TOPICS IN STATISTICAL LEARNING THEORY

Peter L. Bartlett & Sanjoy Dasgupta

(Stéphane Boucheron & Nicolas Vayatis, editors)

Société mathématique de France

Nicolas Vayatis
Centre Borelli, ENS Paris-Saclay

Stéphane Boucheron
LPSM Université Paris Cité

Sanjoy Dasgupta
Department of Computer Science, University of California, San Diego, La Jolla CA 92093

Peter L. Bartlett
Departments of Computer Science and Statistics, Berkeley AI Research Lab, University of California Berkeley

***Classification mathématique par sujets.* (2010) —** 68T10, 68Q32, 68T09 (62R07), 6860E15, 62G08, 62H30

***Mots-clés et phrases.* —** Apprentissage statistique, processus empiriques, classification automatique, apprentissage en ligne, régression non paramétrique.

***Keywords and phrases.* —** Computational learning theory, statistical learning theory, empirical processes, clustering, active learning, online learning, non parametric regression.

TOPICS IN STATISTICAL LEARNING THEORY

Peter L. Bartlett & Sanjoy Dasgupta
(Stéphane Boucheron & Nicolas Vayatis, editors)

***Abstract.* —** This volume is the outcome of a series of three lectures on statistical learning theory given at Institut Henri Poincaré in 2011 under the auspices of the Société Mathématique de France. The introductory chapter provides an overview of the history of Statistical Learning Theory, its roots, its mathematical tools and the questions that make it. The chapter "Algorithms for minimally supervised learning" by Sanjoy Dasgupta describes the progress of theoretical computer science on the issues of unsupervised learning (clustering) and active learning. Surprisingly, much of this progress is due to the confrontation of concentration of measure theory, complexity theory and established practices in numerical statistics.

The chapter "Online prediction" by Peter Bartlett focuses on online learning. It is a confrontation between statistics, game theory and optimization.

***Résumé* (Questions de théorie de l'apprentissage statistique). —** Ce volume est le résultat d'une série de trois cours sur la théorie statistique de l'apprentissage données à l'Institut Henri Poincaré en 2011 sous l'égide de la Société Mathématique de France.

Le chapitre de présentation propose un survol de l'histoire de la théorie statistique de l'apprentissage, de ses racines, de ses outils mathématiques et des questions qui la constituent.

Le chapitre «Algorithms for minimally supervised learning» par Sanjoy Dasgupta décrit les progrès de l'informatique théorique sur les questions d'apprentissage non supervisé (clustering) et sur l'apprentissage dit actif. De façon assez surprenante, ces progrès sont dus en grande partie à la confrontation de la théorie de la concentration de la mesure, de la théorie de la complexité et des pratiques établies en statistique numérique.

Le chapitre «Online prediction» par Peter Bartlett porte sur l'apprentissage en ligne. Il s'agit d'une confrontation entre statistique, théorie des jeux et optimisation.

CONTENTS

RÉSUMÉS DES ARTICLES

L'article *Minimally supervised learning* de Sanjoy Dasgupta étudie les problèmes d'apprentissage non supervisé et d'apprentissage actif. Dans l'apprentissage non supervisé, l'algorithme d'apprentissage manipule un échantillon de points non étiquetés issus d'une loi inconnue et tente de les regrouper en classes de sorte que les points proches soient affectés à la même classe et les points éloignés à des classes distinctes. Cette tâche a fait l'objet d'une grande attention en statistique mathématique car elle recoupe l'analyse des modèles de variables latentes. Les modèles à variables latentes sont notoirement difficiles à traiter, tant du point de vue statistique que du point de vue informatique, et sont abordés à l'aide de méthodes itératives à motivation heuristique comme EM. Sanjoy Dasgupta fournit des preuves convaincantes que EM devrait fonctionner bien et rapidement lorsqu'il est confronté à des données tirées de certains mélanges de distributions bien concentrées. Ses arguments combinent des outils issus des processus empiriques et de la théorie de la concentration de la mesure avec une analyse algorithmique.

Nous examinons les modèles de prédiction basés sur la théorie des jeux, dans lesquels le processus générant les données est modélisé comme un adversaire avec lequel la méthode de prédiction est en compétition. Nous présentons une formulation qui englobe une grande variété de problèmes de décision, et nous nous concentrons sur la relation entre la prédiction dans ce cadre de la théorie des jeux et la prédiction dans le cadre probabiliste plus standard. En particulier, nous présentons une vision des stratégies de prédiction standard comme des méthodes de décision bayésiennes, et nous montrons comment le regret des stratégies optimales dépend de mesures de complexité qui sont étroitement liées à celles qui apparaissent dans les cadres probabilistes.