VC-Dimension and Rademacher Averages: From Statistical Learning Theory to Sampling Algorithms

– Proposal for a KDD tutorial –

Matteo Riondato Eli Upfal
Department of Computer Science – Brown University
{matteo,eli}@cs.brown.edu

Abstract

Rademacher Averages and the *Vapnik-Chervonenkis dimension* are fundamental concepts from statistical learning theory. They allow to study simultaneous deviation bounds of empirical averages from their expectations for classes of functions, by considering properties of the problem, of the dataset, and of the sampling process. In this tutorial, we survey the use of Rademacher Averages and the VC-dimension for developing sampling-based algorithms for graph analysis and pattern mining. We start from their theoretical foundations at the core of machine learning, then show a generic recipe for formulating data mining problems in a way that allows using these concepts in the analysis of efficient randomized algorithms for those problems. Finally, we show examples of the application of the recipe to graph problems (connectivity, shortest paths, betweenness centrality) and pattern mining. Our goal is to expose the usefulness of these techniques for the data mining researcher, and to encourage research in the area.

Introduction

Random sampling is a natural technique to speed up the execution of algorithms on very large datasets. The results obtained by analyzing only a random sample of the dataset are an approximation of the exact solution. When only a single value must be computed, the trade-off between the size of the sample and the accuracy of the approximation can be studied through probabilistic bounds (e.g., the Chernoff-Hoeffding bounds) for the deviation of the quantity of interest in the sample from its exact value in the dataset. In many classical data mining problems, the number of quantities of interest can be extremely large (e.g., betweenness centrality requires to compute one quantity for each node in a graph). In these cases, *uniform* (i.e., *simultaneous*) bounds to the deviations of all quantities are needed. Classical techniques like the Union bound, are insufficient because excessively loose due to their worst-case assumptions that do not hold in many data mining problems. *Rademacher Averages* and the *Vapnik-Chervonenkis dimension* have been developed to overcome this issue: they obtain much stricter uniform deviation bounds by taking into account the nature of the problem and properties of the dataset and of the sampling process. They have been used with success in the analysis of sampling algorithms for data and graph analysis problems on very large datasets.

Target Audience and Prerequisites

The target audience of our tutorial are data mining and machine learning researchers with an interest in modern statistics and in effective use of data through the application of powerful theoretical results. Researchers whose focus is in pattern mining and graph mining should be particularly interested, as should the part of the audience with an interest in the theory of machine learning. In designing our tutorial we effectively strived in strengthen the connection between these fields: we start from theoretical concepts and results that arise in the classical learning context of classification and show how they can be used to develop practical efficient algorithms for important data mining problems. For the attending machine learning researchers, it would be a chance to be exposed to recent theoretical material in their field that has only recently started to be organized in books [10, 17], and see how it can be applied outside machine learning. At the same time, the part of the audience more interested in data mining can learn about techniques from the neighbor field of machine learning that can be used to develop algorithms for their favorite knowledge discovery problems. For these reasons, we believe that our tutorial can be of great interest to most KDD attendees.

We do not require or assume that any specific existing knowledge from the audience. The tutorial is designed for an audience of computer scientists who have a general idea of the problems and challenges in data mining and a basic understanding of probability. We believe that any advanced undergraduate student in computer science would be able to productively follow our tutorial, and we will actively engage with the audience to adapt our pace and our presentation style to ensure that every attendee can benefit from our tutorial.

Tutorial Content and Outline

We plan for a three-hours tutorial. We start with a short introduction about the use of random sampling in data mining, discussing its advantages and the challenges for the algorithm designer. The goal is to lay forward the key questions that will be answered in the rest of the tutorial. In particular, we introduce the key problem of learning, known as the Glivenko-Cantelli problem for classes of functions, and then show the limitations of the Union bound in solving this problem and ask how to overcome them in the settings arising from using sampling for data mining problems.

The theoretical foundations are presented in the first part, where we introduce the Rademacher Average and the VC-dimension, showing how they allow to answer the questions posed in the introduction, and how they are related to each other. We then focus on computing, estimating, and bounding these quantities, which is a key step in the process of using them to develop algorithms. We show a number of basic examples of classes of functions with finite and infinite VC-dimension and discuss different techniques for developing analytical bounds and empirical estimations. The examples will range from toy examples to understand the concepts (e.g., axis-aligned rectangles, half-spaces, and sinusoidal functions) to much more complex examples that are presented in research papers (e.g., graph neighborhood functions, neural networks, and shortest paths).

The second part focuses on showing how to use Rademacher averages and VC-dimension to develop sampling-based algorithms for data and graph mining problems. We start by presenting a generic recipe for developing such algorithms, which makes it easier to applying the techniques and perform the analysis. We then show a number of examples of application of this technique for different graph and data analysis problems, including network connectivity, shortest paths

algorithms, betweenness centrality computation, and frequent pattern mining, and set covering. In this part of the tutorial, we will partially discuss research done by one or both the tutorial proposers. This is quite unavoidable, as we were among the firsts to explore the use of these techniques in data mining. Naturally, a number of works by other researchers will also be discussed here and in the other parts of the tutorial.

In the third part, we will focus on more advanced material, to encourage the audience to further explore the field of statistical learning theory, and to stimulate discussion and research on using the results from this field to develop data mining algorithms. Specifically, we will discuss: *1.* PAC-Bayesian bounds, which shows a connection between the typical frequentist approach followed in statistical learning theory to the Bayesian probabilistic approach, *2.* the connection between VC-dimension and the Minimum Description Length principle from information theory, and *3.* a selection of the extensions of VC-dimension to real-valued or non-binary functions, including pseudodimension, Natarajan dimension, and fat-shattering dimension.

The following is a preliminary outline of the tutorial.

1. Introduction and Theoretical Foundations (1 hour)
   - 1.1 Sampling and Data Mining: a happy marriage?
   - 1.2 The Glivenko-Cantelli problem: uniform convergence of classes of functions [19]
   - 1.3 Beyond the Union Bound: Rademacher averages [8] and the Vapnik-Chervonenkis dimension [20]. The sampling theorems [2]
   - 1.4 Computing the VC-dimension is hard [11]. Estimating the VC-dimension [18, 21]. Bounding the VC-dimension: axis-aligned rectangles, sine functions, paths and stars in a graph [9]
2. Applications to Graph and Pattern Mining (1 hour and 30 minutes)
   - 2.1 A generic recipe for sampling-based data mining algorithms
   - 2.2 Applications to Graph Mining: identifying network disruption [7], shortest path algorithms [1], betweenness centrality [14]
   - 2.3 Applications to Pattern Mining: static and progressive sampling algorithms for frequent itemsets [15, 16]
   - 2.4 Other applications: set covering [4]
3. Recent developments and advanced topics (30 minutes)
   - 3.1 PAC-Bayesian bounds [3, 17]
   - 3.2 Statistical Risk Minimization and the connection with the Minimum Description Length principle [19]
   - 3.3 Beyond binary functions: pseudodimension [12], fat-shattering dimension [6], and related measures [10]

Time Plan for Tutorial Materials Preparation

We plan to create a mini-website for the tutorial at `http://bigdata.cs.brown.edu/vctutorial/`. The website will contain the abstract of the tutorial, a more detailed outline with short a description of each item of the outline, a full list of references complete with links to electronic editions, and naturally the slides used in the tutorial. A preliminary version of the website will be available 15

days after the tutorial is accepted. We plan to work on it and enrich the contents continuously. A preliminary version of the slides will be available 30 days before the conference, or in any case by any deadline given to us by the conference organizers, and the final version will be available 15 days before the conference.

Tutors

This tutorial is developed by Matteo Riondato and Eli Upfal.

Matteo Riondato is a postdoctoral research associate at Brown University, USA, supervised by Prof. Eli Upfal. He received his Ph.D. from Brown in May 2014, with a dissertation on sampling-based randomized algorithms for data analytics, which received the Best Student Poster Award at SIAM SDM 2014. He presented a nectar talk about modern sampling algorithms at ECML PKDD 2014. His research focuses on exploiting theoretical results for practical algorithms in pattern and graph mining.

Eli Upfal is a professor of computer science at Brown University, where he was also the department chair from 2002 to 2007. Prior to joining Brown in 1998, he was a researcher and project manager at the IBM Almaden Research Center in California, and a professor of Applied Mathematics and Computer Science at the Weizmann Institute of Science in Israel. Upfal's research focuses on the design and analysis of algorithms. In particular he is interested in randomized algorithms, probabilistic analysis of algorithms, and computational statistics, with applications ranging from combinatorial and stochastic optimization to routing and communication networks, computational biology, and computational finance. Upfal is a fellow of the IEEE and the Association for Computing Machinery (ACM). He received the IBM Outstanding Innovation Award, and the IBM Research Division Award. His work at Brown has been funded in part by the National Science Foundation (NSF), the Defense Advanced Research Projects Agency (DARPA), the Office of Naval Research (ONR), and the National Institute of Health (NIH). He is co-author of a popular textbook "Probability and Computing: Randomized Algorithms and Probabilistic Analysis" (with M. Mitzenmacher, Cambridge University Press 2005).

Contact Information

Matteo Riondato and Eli Upfal
{matteo,eli}@cs.brown.edu
Box 1910, 115 Waterman Street, Providence, RI 02912, USA
Tel.: +14018631000

Previous Venues and Similar Tutorials

The tutorial was never presented before. Matteo Riondato gave a nectar talk with limited overlap with the tutorial content at ECML PKDD 2014 [13]. The tutorial we propose is much richer in content and more structured in its organization.

A tutorial on sampling, but *not covering any of the material in our proposed content*, was presented at ACM KDD 2014 [5]. It can be seen as complementary to ours, introducing different techniques for different applications (mostly in streaming settings). We believe that the fact that

a tutorial on sampling was presented last year at KDD reinforces the fact that machine learning and data mining researchers should be interested in sampling-related techniques. Moreover VC-dimension can be used for much more than sampling (for example, for computing set coverings). We believe that data mining research can greatly benefit from the rich theory, techniques, and tools developed in the context of statistical learning. This tutorial aims to spread the knowledge of these concepts to the wider data mining community. This goal sets this tutorial apart from the KDD 2014 sampling tutorial.

References

[1] I. Abraham, D. Delling, A. Fiat, A. V. Goldberg, and R. F. Werneck. VC-dimension and shortest path algorithms. ICALP'11, 2011.

[2] N. Alon and J. H. Spencer. *The Probabilistic Method*. Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Hoboken, NJ, USA, third edition, 2008.

[3] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification : A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[4] H. Brönnimann and M.T. Goodrich. Almost Optimal Set Covers in Finite VC-Dimension SCG'94, 1994.

[5] G. Cormode and N. Duffield. Sampling for Big Data: A tutorial. KDD'14, 2014.

[6] M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. FOCS'90, 1990.

[7] J. M. Kleinberg, M. Sandler, and A. Slivkins. Network failure detection and graph connectivity. *SIAM J. Comput.*, 38(4):1330–1346, 2008.

[8] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inf. Theory*, 47(5):1902–1914, July 2001.

[9] E. Kranakis, D. Krizanc, B. Ruf, J. Urrutia, and G. Woeginger. The VC-dimension of set systems defined by graphs. *Discrete Applied Mathematics*, 77(3):237–257, 1997.

[10] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. The MIT Press, 2012.

[11] C. H. Papadimitriou and M. Yannakakis. On limited nondeterminism and the complexity of the V-C dimension. *J. Comput. Syst. Sci.*, 53(2):161–170, 1996.

[12] D. Pollard. *Convergence of stochastic processes*. Springer, 1984.

[13] M. Riondato. Sampling-based data mining algorithms: Modern techniques and case studies. ECML PKDD'14, 2014.

[14] M. Riondato and E. M. Kornaropoulos. Fast approximation of betweenness centrality through sampling. WSDM'14, 2014.

[15] M. Riondato and E. Upfal. Mining frequent itemsets through progressive sampling with rademacher averages. Extended Version. URL `http://cs.brown.edu/%7Ematteo/papers/progrsamplfi-ext.pdf`.

[16] M. Riondato and E. Upfal. Efficient discovery of association rules and frequent itemsets through sampling with tight performance guarantees. *ACM Trans. Knowl. Disc. from Data*, 8(4):20, 2014.

[17] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[18] X. Shao, V. Cherkassky, and W. Li. Measuring the VC-dimension using optimized experimental design. *Neural Computation*, 12(8):1969–1986, 2000.

[19] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for engineering and information science. Springer-Verlag, New York, NY, USA, 1999.

[20] V. N. Vapnik and A. J. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[21] V. N. Vapnik, E. Levin, and Y. Le Cun. Measuring the VC-dimension of a learning machine. *Neural Computation*, 6:851–876, 1994.