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

Tutorial Outline

Matteo Riondato
Dept. of Computer Science
Brown University
Providence, RI 02912
matteo@cs.brown.edu

Eli Upfal
Dept. of Computer Science
Brown University
Providence, RI 02912
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 functions, of their domain (the dataset), and of the sampling process. In this tutorial, we survey the use of Rademacher Averages and the VC-dimension in 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 to use these concepts in 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.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; G.3 [**Probability and Statistics**]: [Probabilistic algorithms (including Monte Carlo)]

Keywords

Betweenness Centrality; Frequent Itemsets; Graph Mining; Pattern Mining; Randomized Algorithms; Tutorial

1. INTRODUCTION

Random sampling is a natural technique to speed up the execution of algorithms on very large datasets [6]. 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 for the deviation of the quantity of interest in the sample from its exact value in the dataset, e.g., the Chernoff-Hoeffding bounds [7]. In many 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 [4, 13], and Frequent Pattern Mining requires the computation of a potentially exponential number of quantities). In these cases, *uniform* (i.e., *simultaneous*) bounds to the deviations of all quantities are needed in order to have good approximations of all the values of interest. Classical techniques like the Union bound [11] are insufficient because excessively loose due to their worst-case assumptions that do not hold in many data mining problems, e.g., the assumption that the quantities of interest are independent from each other. *Rademacher Averages* [3] and the *Vapnik-Chervonenkis dimension* [18] have been developed by the statistical learning theory community to study the (rate of) convergence of the empirical risk of a learned function to its expectation. One of the goals of this tutorial is to show that these techniques are very flexible and powerful and their field of use is much broader. In particular, they overcome the weakness of the Union bound: they obtain much stricter uniform deviation bounds by taking into account the nature of the problem (i.e., of the quantities of interest) 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 [1, 5, 9, 13–16].

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.
ACM 978-1-4503-3664-2/15/08.
DOI: http://dx.doi.org/10.1145/2783258.2789984 .

2. OUTLINE

The tutorial is structured as follows.

Introduction. 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 start with an example involving Frequent Itemset Mining through sampling [14, 15], showing the limitations of the Union bound in solving this problem. By generalizing our settings to general data mining problem, we then introduce the key problem of learning, known as the *Glivenko-Cantelli problem for classes of functions* [18], which clarifies the strong connection with the area of statistical learning theory.

First Part: Theoretical Foundations. The theoretical foundations are presented in the first part, where we introduce the Rademacher Averages [3, 10] and the VC-dimension [19], showing how they allow to answer the questions posed in the introduction, and how they are related to each other. In particular, we show the key theorems that allow to compute an upper bound to the sample size sufficient to obtain a high-quality approximation of the quantities of interest, uniformly. The bounds depend linearly on the Rademacher averages and on the VC-dimension. We then focus on computing, estimating, and bounding the Rademacher averages and the VC-dimension, 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 (e.g., axis-aligned rectangles, half-spaces, and sinusoidal functions) to much more complex instances that are presented in research papers (e.g., graph neighborhood functions, neural networks [2], and shortest paths [1]).

Second Part: Applications. 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 eases the application of the techniques and the analysis of the algorithms. We then show a number of examples of application of this technique for different graph and data analysis problems, including network connectivity [9], shortest paths algorithms [1], betweenness centrality computation [13], and frequent pattern mining [14–16], and set covering [5].

Third part: Advanced Material. 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: PAC-Bayesian bounds [3, 17], which show a connection between the typical frequentist approach followed in statistical learning theory to the Bayesian probabilistic approach and may be useful for data mining algorithms on uncertain or probabilistic data, and a selection of the extensions of VC-dimension to real-valued or non-binary functions, including pseudodimension [12], Natarajan dimension, and fat-shattering dimension [8].

3. WEBSITE

We set up a mini-website (`http://bigdata.cs.brown.edu/vctutorial`) with links to the slides that we use for the presentation, and a bibliography of theoretical and application-oriented works about VC-dimension and Rademacher Averages.

4. ACKNOWLEDGMENTS

This work was supported by NSF grant IIS-1247581 and NIH grant R01-CA180776.

5. 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] M. Anthony and P. L. Bartlett. *Neural Network Learning - Theoretical Foundations*. Cambridge University Press, 1999.

[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] U. Brandes. A faster algorithm for betweenness centrality. *J. Math. Sociol.*, 25(2):163–177, 2001.

[5] H. Brönnimann and M. Goodrich. Almost optimal set covers in finite vc-dimension. *Discrete & Computational Geometry*, 14(1):463–479, 1995.

[6] G. Cormode and N. Duffield. Sampling for big data: A tutorial. ACM KDD'14, 2014.

[7] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. American Statistical Assoc.*, 58(301):13–30, 1963.

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

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

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

[11] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

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

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

[14] 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.

[15] M. Riondato and E. Upfal. Mining frequent itemsets through progressive sampling with rademacher averages. ACM KDD'15.

[16] M. Riondato and F. Vandin. Finding the true frequent itemsets. SIAM SDM'14.

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

[18] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1999.

[19] 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.