Statistical learning on measures: an application to persistence diagrams

Olympio Hacquard*[1,2]

[1]Laboratoire mathématique d'Orsay – Université Paris-Sud - Université Paris-Saclay – France
[2]Understanding the Shape of Data – Inria Saclay - Ile de France – France

Abstract

We consider a binary supervised learning classification problem where instead of having data in a finite-dimensional Euclidean space, we observe measures supported in a compact space $\mathbf{X}$. Given a set of functions on $\mathbf{X}$, we build corresponding classifiers in the space of measures. We provide upper and lower bounds on the Rademacher complexity of this new class of classifiers that can be expressed simply in terms of corresponding quantities for the class of functions on $\mathbf{X}$. If the measures are uniform over a finite set, this classification task boils down to a multi-instance learning problem. However, our approach allows more flexibility and diversity in the input data we can deal with. While such a framework has many possible applications, this work puts a strong emphasis on classifying data via topological descriptors called persistence diagrams. These objects are discrete measures on $\mathbf{R}^2$, where the coordinates of each point correspond to the range of scales at which a topological feature exists. We will present several classifiers on measures and show how they can heuristically and theoretically enable a good classification performance in various settings in the case of persistence diagrams.

Keywords: Statistical learning, measures, topological data analysis, persistence diagrams

*Speaker