Informative labels in Semi-Supervised Learning

Aude Sportisse*[1,2], Charles Bouveyron[1,2], and Pierre-Alexandre Mattei[1,2]

[1]3iA Côte d'Azur – 3iA Côte d'Azur – France
[2]Modèles et algorithmes pour l'intelligence artificielle – Inria Sophia Antipolis - Méditerranée, Scalable and Pervasive softwARe and Knowledge Systems, Université Nice Sophia Antipolis (... - 2019), Laboratoire Jean Alexandre Dieudonné – France

Abstract

In semi-supervised learning, we have access to features but the outcome variable is missing for a part of the data. In real life, although the amount of data available is often huge, labeling the data is costly and time-consuming. It is particularly true for image data sets: images are available in large quantities on image banks but they are most of the time unlabeled. It is therefore necessary to ask experts to label them. In this context, people are more inclined to label images of some classes which are easy to recognize. The unlabeled data are thus informative missing values, because the unavailability of the labels depends on their values themselves. Typically, the goal of semi-supervised learning is to learn predictive models using all the data (labeled and unlabeled ones). However, classical methods lead to biased estimates if the missing values are informative. We aim at designing new semi-supervised algorithms that handle informative missing labels.

*Speaker