University of California, Irvine
Statistics Seminar

Generalized Data Thinning Using Sufficient Statistics

Jacob Bien
Associate Professor, Data Science & Operations
USC

4 p.m., Thursday, May 25, 2023
6011 Donald Bren Hall

Sample splitting is one of the most tried-and-true tools in the data scientist toolbox. It breaks a data set into two independent parts, allowing one to perform valid inference after an exploratory analysis or after training a model. A recent paper (Neufeld, et al. 2023) provided a remarkable alternative to sample splitting, which the authors showed to be attractive in situations where sample splitting is not possible. Their method, called convolution-closed data thinning, proceeds very differently from sample splitting, and yet it also produces two statistically independent data sets from the original.

In this talk, we will show that sufficiency is the key underlying principle that makes their approach possible. This insight leads naturally to a new framework, which we call generalized data thinning. This generalization unifies both sample splitting and convolution-closed data thinning as different applications of the same procedure. Furthermore, we show that this generalization greatly widens the scope of distributions where thinning is possible. This work is a collaboration with Ameer Dharamshi, Anna Neufeld, Keshav Motwani, Lucy Gao, and Daniela Witten.