University of Idaho

JON A. WELLNER LECTURE
DEPARTMENT OF MATHEMATICS and STATISTICAL SCIENCE

Tuesday, January 30, 2024
Talk begins at 2:30 p.m. in TLC 029
Refreshments at 3:30 p.m. in IRIC Atrium

Data thinning and its applications

Daniela Witten, Ph.D.
Professor of Biostatistics and Dorothy Gilford Endowed Chair of Mathematical Statistics, University of Washington

ABSTRACT

We propose data thinning, a new approach for splitting an observation from a known distributional family with unknown parameter(s) into two or more independent parts that sum to yield the original observation, and that follow the same distribution as the original observation, up to a (known) scaling of a parameter. This proposal is very general, and can be applied to a broad class of distributions within the natural exponential family, including the Gaussian, Poisson, negative binomial, Gamma, and binomial distributions, among others. Furthermore, we generalize data thinning to enable splitting an observation into two or more parts that can be combined to yield the original observation using an operation other than addition; this enables the application of data thinning far beyond the natural exponential family. Data thinning has a number of applications to model selection, evaluation, and inference. For instance, cross-validation via data thinning provides an attractive alternative to the "usual" approach of cross-validation via sample splitting, especially in unsupervised settings in which the latter is not applicable. We will present an application of data thinning to single-cell RNA-sequencing data, in a setting where sample splitting is not applicable. This is joint work with Anna Neufeld (Fred Hutch), Ameer Dharamshi (University of Washington), Lucy Gao (University of British Columbia), and Jacob Bien (University of Southern California).

For More Information:

Department of Mathematics and Statistical Science
208-885-6742
uidaho.edu/sci/mathstat/news/wellner-lecture