Statistics Seminar Department of Mathematics and Statistics

| DATE: | Thursday, April 24, 2025 |
|-----------|---------------------------------------|
| TIME: | 1:15pm - 2:15pm |
| LOCATION: | WH 100E |
| SPEAKER: | Bruce Phillips, Binghamton University |
| TITLE: | Data Twinning |

Abstract

In this work, we develop a method named Twinning for partitioning a dataset into statistically similar twin sets. Twinning is based on SPlit, a recently proposed model-independent method for optimally splitting a dataset into training and testing sets. Twinning is orders of magnitude faster than the SPlit algorithm, which makes it applicable to Big Data problems such as data compression. Twinning can also be used for generating multiple splits of a given dataset to aid divide-and-conquer procedures and k-fold cross validation.

Reference: A. Vakayil, V. R. Joseph, Data Twinning, Stat. Anal. Data Min.: ASA Data Sci. J.. 15 (2022), 598–610. https://doi.org/10.1002/sam.11574

From:

 ${\it https://www2.math.binghamton.edu/- \textbf{Department of Mathematics and Statistics, Binghamton University}$

Permanent link:

https://www2.math.binghamton.edu/p/seminars/stat/apr242025

Last update: 2025/04/22 00:49

