Contribution ID: **55** Type: **not specified**

Feature selection using distance correlation

Monday 8 May 2023 18:00 (15 minutes)

Choosing which properties of the data to use as input to multivariate decision algorithms – a.k.a. feature selection – is an important step in solving any problem with machine learning. While there is a clear trend towards training sophisticated deep networks on large numbers of relatively unprocessed inputs (so-called automated feature engineering), for many tasks in physics, sets of theoretically well-motivated and well-understood features already exist. Working with such features can bring many benefits, including greater interpretability, reduced training and run time, and enhanced stability and robustness. We develop a new feature selection method based on Distance Correlation (DisCo), and demonstrate its effectiveness on the tasks of boosted top- and W-tagging. Using our method to select features from a set of over 7,000 energy flow polynomials, we show that we can match the performance of much deeper architectures, by using only ten features and two orders-of-magnitude fewer model parameters.

Authors: SHIH, David; KASIECZKA, Gregor (Hamburg University (DE)); DAS, Ranit (Rutgers University)

Presenter: DAS, Ranit (Rutgers University)

Session Classification: Tools I