Contribution ID: **43** Type: **not specified**

Learning Invariant Representations using Mutual Information Regularization

Tuesday 16 April 2019 09:45 (20 minutes)

Invariance of learned representations of neural networks against certain sensitive attributes of the input data is a desirable trait in many modern-day applications of machine learning, such as precision measurements in experimental high-energy physics and enforcing algorithmic fairness in the social and financial domain. We present a method for enforcing this invariance through regularization of the mutual information between the target variable and the classifier output. Applications of the proposed technique to rare decay searches in experimental high-energy physics are presented, and demonstrate improvement in statistical significance over conventionally trained neural networks and classical machine learning techniques.

Preferred contribution length

20 minutes

Author: Mr TAN, Justin (University of Melbourne)

Presenter: Mr TAN, Justin (University of Melbourne)

Session Classification: Submitted contributions