Disentangling Influence: Using Disentangled Representations to Audit Model Predictions

Charles Marx HC ’20

Motivated by the need to audit complex and black box models, there has been extensive research on quantifying how data features influence model predictions. Feature influence can be direct (a direct influence on model outcomes) and indirect (model outcomes are influenced via proxy features). Feature influence can also be expressed in aggregate over the training or test data or locally with respect to a single point. Current research has typically focused on one of each of these dimensions. In this talk, I will discuss recent work on disentangled influence audits, a procedure to audit the indirect influence of features. Specifically, we will find that disentangled representations provide a mechanism to identify proxy features in a dataset, while allowing an explicit computation of feature influence on either individual outcomes or aggregate-level outcomes. We will see through theory and experiments that disentangled influence audits can both detect proxy features and show, for each individual or in aggregate, which of these proxy features affects the classifier being audited the most.

Date: Tuesday February 4, 2020
Time: 7:00 p.m.
Place: Park 328