Performative Prediction: When Predictions Impact the Predicted

Session Leaders: Celestine Mendler-Dünner, Tijana Zrnic

Abstract: When predictions support decisions they may influence the outcome they aim to predict. Such predictions are called *performative*; the prediction causes a change in the distribution of the target.
The motivation for studying performative prediction comes from the observation that whenever we use supervised learning in social settings, we almost never make predictions for predictions' sake, but rather to inform decision making within some broader context. Banks predict default risks to decide to whom they will allocate loans. Commuters use estimated time of arrival (ETA) prediction to choose which route to take to work. Governments predict crime rates to decide how to deploy police forces. In each of these settings, our choice of predictive model leads to changes in the way the broader system behaves and hence in the distribution over observed data. This feedback potentially invalidates what initially seemed to be accurate predictions.
Performativity is a well-studied phenomenon in policy-making that has so far been neglected in supervised learning. We will first briefly introduce a recently proposed framework for performative prediction [1]. One point to discuss is desirable properties of solution concepts in performative prediction. A first notion of optimality is called *performative stability*. Performative stability implies that the predictions are calibrated not against past outcomes, but against the future outcomes that manifest from acting on the prediction. Performatively stable models eliminate the need for retraining. Another solution concept is called *performative optimality*. This solution ensures more accurate predictions, but it does not eliminate the need for retraining.
We would also like to discuss the costs associated with both solutions; for example, existing work has pointed out that strategic adaptations to classifiers can lead to significant social burden [2].
In many applications such as recommendation systems, performativity is treated as a nuisance distribution shift, and is typically dealt with via retraining. Such retraining strategies can formally be analyzed in this new framework, and we present conditions for convergence to performatively stable points. In addition, we plan to discuss several trade-offs associated with retraining and model deployment in application domains that witness performative feedback.
Finally, the formal setup of performative prediction is quite general and captures many learning settings as a special case. We would like to discuss and understand the performative characteristics of different application domains individually.

Format: As a subfield of learning theory, performative prediction is only starting to receive attention from the community and papers in this area are largely theoretical in nature. Therefore, after spending 10-15min introducing the main concepts of performative prediction, we would like to turn to discussing the points stated above, especially focusing on the applied aspects. We will prepare questions to guide the discussion and encourage participation. Overall, we hope to raise awareness of the social implications of decision making in the machine learning community and inspire researchers to contribute to this new and interesting research direction.

[1] Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner and Moritz Hardt, *Performative Prediction*, arxiv (to appear at ICML), 2020.
[2] Smitha Milli, John Miller, Anca Dragan and Moritz Hardt, *The Social Cost of Strategic Classification*, ACM FAT*, 2019