AMS280B-01: Seminars in Statistics for Oct-8th (Monday 4pm at BE 156)

Speaker: Sindhu Ghanta, Research Scientist at Parelles

Title: Machine Learning in Production

Abstract: Bringing the research advances in Machine Learning (ML) to production is necessary for businesses to gain value from ML. A key challenge of production ML is the monitoring and management of real-time prediction quality. This is complicated by the variability of live production data, the absence of real-time labels and the non-determinism posed by ML techniques themselves where real-time predictions are impacted by the datasets used much earlier during training. We present an approach to ML Health -- the real time assessment of ML prediction quality -- and a solution to monitor and manage ML Health within a realistic full production ML lifecycle. Empirical results on these datasets show that a combination of these techniques can be utilized to assess the quality of predictions in the absence of labels. We also describe an end-to-end system to deploy ML pipelines and monitor their health in production. Our system handles production realities such as scale, heterogeneity and distributed runtimes. Via the combination, we present what we believe is the first solution to production ML Health explored at both an empirical and complete system implementation level.