Mathematics, Statistics, and Computer Science **@ UIC**

Computer Science Theory Seminar

Interpretability in machine learning

Gyorgy Turan (UIC)

Abstract: Interpretability is the requirement that a model obtained by machine learning, beyond having predictive power, should be comprehensible for the user, or it should be possible to reason about its properties. For example, decision trees appear to be more interpretable than neural networks. We give a brief introduction to this topic, and discuss experimental results on interpretability aspects of word embeddings in natural language processing, and a theoretical approach to interpretability for Bayesian network classifiers using ordered binary decision diagrams.

Joint work with Vanda Balogh, Gabor Berend, Karine Chubarian and Dimitris Diochnos.

Wednesday, October 30 at 4:15 PM in 1325 SEO