Center for Artificial Intelligence in Society presents...

Dr. Yevgeniy Vorobeychik

The Art and Science of Adversarial Machine Learning

The success of machine learning has led to numerous attempts to apply it in adversarial settings like spam and malware detection. The core challenge in this class of applications is that adversaries are not static data generators, but make a deliberate effort to either evade the classifiers deployed to detect them, or degrade the quality of the data used to train the classifiers. I will discuss our recent research into the problem of adversarial classifier evasion, specifically the theoretical foundations of black-box attacks on classifiers, and several of our efforts in designing evasion-robust classifiers on binary feature spaces, including a principled, theoretically-grounded, retraining method.

Second, I will discuss scientific foundations of classifier evasion modeling. A dominant paradigm in the machine learning community is to model evasion in "feature space" through direct manipulation of classifier features. In contrast, the cyber security community developed several "problem space" attacks, where actual instances (e.g., malware) are modified, and features are then extracted from the evasive instances. I will show, through a case study of PDF malware detection, that feature-space models are a very poor proxy for problem space attacks. Then I will demonstrate a simple "fix" to identify a small set of features which are invariant (conserved) with respect to evasion attacks, and constrain these features to remain unchanged in feature-space models. Lastly, I will show that such conserved features exist and *cannot* be inferred using standard regularization techniques, but can be automatically identified for a given problem-space evasion model.

Monday, July 24th, 11 a.m.—12 p.m.

Tutor Hall (RTH) 217

Yevgeniy Vorobeychik is an Assistant Professor of Computer Science, Computer Engineering, and Biomedical Informatics at Vanderbilt University. He received Ph.D. (2008) and M.S.E. (2004) degrees in Computer Science and Engineering from the University of Michigan, and a B.S. degree in Computer Engineering from Northwestern University. His work focuses on adversarial reasoning in AI, computational game theory, security and privacy, network science, and agent-based modeling. He received an NSF CAREER award in 2017, was an invited early career spotlight speaker at IJCAI 2016.

Please RSVP to Hailey at hwinetro@usc.edu by July 20.