Seminar – Tuesday, June 13, in 111 Rhodes Annex **2:00 – 3:00 p.m.**

Holcombe Department of Electrical and Computer Engineering Seminar Series

A Maximum Entropy Framework for Semisupervised and Active Learning with Unknown and Label-Scarce Classes

Dr. George Kesidis

Professor, School of Electrical Engineering & Computer Science
Pennsylvania State University

Abstract

We investigate semisupervised learning (SL) and pool-based active learning (AL) of a classifier for domains with label-scarce (LS) and unknown categories, *i.e.* defined categories for which there are initially no labeled examples. This scenario manifests *e.g.* when a category is rare, or expensive to label. There are several learning issues when there are unknown categories: i) it is *a priori* unknown which subset of (possibly many) measured features are needed to discriminate unknown from common classes; ii) label scarcity suggests overtraining is a concern. Our classifier exploits the inductive bias that an unknown class consists of the subset of the unlabeled pool's samples that are *atypical* (relative to the common classes) with respect to certain key (albeit *a priori* unknown) features and feature interactions. Accordingly, we treat negative log-p-values on raw features as *non-negatively weighted* derived feature inputs to our class posterior, with zero weights identifying irrelevant features. Via a hierarchical class posterior, our model accommodates multiple common classes, multiple LS classes, and unknown classes. For learning, we propose a novel semisupervised objective customized for the LS/unknown category scenarios. While several works minimize class decision uncertainty on unlabeled samples, we instead *preserve* this uncertainty (maximum entropy) to avoid overtraining. Our experiments on a variety of UCI ML domains show: 1) use of p-value features coupled with weight constraints leads to sparse solutions and gives significant improvement over use of raw features; 2) for label-scarce SL and AL, unlabeled samples are helpful, and should be used to *preserve* decision uncertainty (maximum entropy), rather than to minimize it, especially during the early stages of AL. Our AL system, leveraging a novel sample selection scheme, discovers unknown classes and discriminates label-scarce classes from common ones, with sparing use of oracle labeling. Work in collaboration with D.J. Miller and Zhicong Qiu.

Biography of Speaker

George Kesidis received his MS (in 1990) and PhD (in 1992) from UC Berkeley in EECS. Following eight years as a professor of ECE at the University of Waterloo, he has been a professor of CSE and EE (now EECS) at the Pennsylvania State University since 2000. His research interests include many aspects of networking, cyber security, machine learning, and, more recently, cloud computing. His work has been supported by over a dozen NSF research grants, DARPA and AFOSR grants, and several Cisco Systems URP gifts, the latter supporting applied work on cyber security. His web site is http://www.cse.psu.edu/~gik2