An Iterative Algorithm for Extending Learners to a Semi-supervised Setting

Mark Culp and George Michailidis

In this paper, we present an iterative self-training algorithm, whose objective is to extend learners from a supervised setting into a semi-supervised setting. The algorithm is based on using the predicted values for observations where the response is missing (unlabeled data) and then incorporates the predictions appropriately at subsequent stages. Convergence properties of the algorithm are investigated for particular learners, such as linear/logistic regression and linear smoothers with particular emphasis on kernel smoothers. Further, implementation issues of the algorithm with other learners such as generalized additive models, tree partitioning methods, partial least squares, etc. are also addressed. The connection between the proposed algorithm and graph-based semi-supervised learning methods is also discussed. The algorithm is illustrated on a number of real data sets using a varying degree of labeled responses.

Key Words: Semi-supervised learning; linear smoothers; convergence; iterative algorithm

Supplementary Materials

1. Computer Code

jcgscode.zip This file is directory containing the following R scripts. Each script is stand alone.

ksmoothex.R This R script computes the kernel smoother example (refer to Figure 2).

gamex.R This R script computes the the gam example (refer to Figure 5).
	treeex.R This R script computes the tree example (refer to Figure 6).

Mark Culp is an assistant professor in the Department of Statistics at West Virginia University and George Michailidis is an associate professor in the Department of Statistics at the University of Michigan.

© 2008 American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America

This document describes supplementary material to an article published in the Journal of Computational and Graphical Statistics.
loggamex.R This R script computes the logistic gam example (refer to Figure 7).

airex.R This R script computes the semi-supervised gam with loess for the air quality data.

**Remark:** The code is designed to give the results in the paper for specific R package implementations (i.e. `spa, gam, rpart`). In addition, warning messages are expected when running certain procedures on small labeled partitions.