Metric Learning for Nearest Neighbor Classification

Akarin Phaibulpanich*

Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University, Bangkok, Thailand akarin@cbs.chula.ac.th

Kerby Shedden

Department of Statistics, University of Michigan, Ann Arbor, USA kshedden@umich.edu

We develop methods for constructing an A-weighted metric $(x-y)'A(x-y)$ that improves the performance of K-nearest neighbor (KNN) classifiers. KNN is known to be highly flexible, but can be somewhat inefficient and unstable. By incorporating a parametrically optimized metric into KNN, global dimension reduction is carried out efficiently, leaving the most difficult nonlinear features of the problem to be solved on a low dimensional projected feature space. Optimization over A is done by formulating a probability model that captures KNN's essential property – using only a local neighborhood of training cases to predict the class of a test case. The expected correct vote margin can be calculated under the probability model and optimized over A using gradient methods to yield a metric that is adapted to a particular problem. This framework incorporates variable selection as well as variate selection, in which certain linear combinations of the variables are deemed either informative or completely uninformative. The estimated A matrix can be used for both classification and data analysis, as it contains information about which features are informative (either in a linear or nonlinear sense), or completely uninformative about class membership.

Key words: K-Nearest Neighbor Classification, A-weighted metric learning, global dimension reduction