Data Mining and Mathematical Programming
October 10-13, 2006

Subspace clustering under constraints - a new kind of classifiers

Pando G. Georgiev
pgeorgie@ececs.uc.edu
ECECS Department
University of Cincinnati
ML 0030
Cincinnati, OH 45221-0030
USA

Abstract

In many practical problems the data (given by the columns of a data matrix) lie (up to some precision) on a union of m affine subspaces (called *skeleton* of the data). Finding the best possible skeleton when $m = 1$ can be performed easily by Principal Component Analysis. When $m > 1$, however, this is a hard global optimization problem. Its solution reveals an internal structure of the data set, namely, under mild assumptions, the data matrix can be decomposed uniquely (up to some unimportant ambiguities) as multiplication of a *mixing* matrix and a *source* matrix, which, more importantly, is *sparse*. We develop the idea that a binary classification tasks can be performed by finding *constrained skeletons* of the training points in the both classes. Each constrained skeleton is defined as a union of (affine) subspaces which is closest to one of the training points set (in sense of sum of the square distances from the data points to this union) and is farthest to the other. This leads to two constraint optimization problems, which are difficult global optimization problems if the number of the subspaces in each skeleton is bigger than one, but easily solvable in the case of one subspace in each skeleton. We present "kernelization" of this idea, leading to the notion of *kernel constraint skeletons*, when we work in Reproducing Kernel Hilbert Spaces. Some examples for binary classification tasks are presented with simple solutions based on constrained skeletons consisting of one subspace, which cannot be classified correctly by the classical Support Vector Machines (SVM). We emphasize that the classical SVM can be considered as a special

case of constrained clustering, when the subspaces are hyperplanes and the measure of closedness is given by the usual distance between sets.