Data Visualization
STAT 890

Assignment 4 (**Only for STAT 890**)
Fall 2006
Department of Statistics and Actuarial Science
University of Waterloo

Due: Tuesday December 5. Hand in to Joan Hatton at MC 6028
Instructor: Ali Ghodsi
MS 6081G x37316, aghodsib@uwaterloo.ca

Policy on Lateness: Slightly late assignments (up to 24 hs after due date) are accepted with 10% penalty. No assignment are accepted after 24 hs after the due date.

Learning a Metric from Class-Equivalence Side Information

We have seen in many cases that it may be possible to obtain a small amount of information regarding the similarity of points in a particular data set. We learned how such side information can be used to learn a new metric.

In this approach, given a set of t points, $\{x_i\}_{i=1}^{t} \in R^n$, we identify two kinds of class-related side information. The first is a set of pairs of similar or class-equivalent points (they belong to the same class)

$$S: \quad (x_i, x_j) \in \mathcal{S} \quad \text{if } x_i \text{ and } x_j \text{ are similar}$$

and the second is a set of dissimilar or class-inequivalent pairs (they belong to different classes)

$$O: \quad (x_i, x_j) \in \mathcal{O} \quad \text{if } x_i \text{ and } x_j \text{ are dissimilar}$$

We then wish to learn a matrix A that induces a distance metric $D^{(A)}$ over the points

$$D^{(A)}(x_i, x_j) = \|x_i - x_j\|_A = \sqrt{(x_i - x_j)^T A (x_i - x_j)}$$

where $A \succeq 0$.

We define the following loss function, which, when minimized attempts to minimize the squared induced distance between similar points and maximize the squared induced distance between dissimilar points

$$L(A) = \sum_{(x_i,x_j)\in\mathcal{S}} \|x_i - x_j\|_A^2 - \sum_{(x_i,x_j)\in\mathcal{O}} \|x_i - x_j\|_A^2$$

The optimization problem then becomes

$$\begin{aligned} & \min_A L(A) \\ \text{s.t.} \quad & A \succeq 0 \\ & \mathrm{Tr}(A) = 1 \end{aligned}$$

Prove or disprove that matrix A obtained by this technique is always rank 1. (i.e., the data are always projected onto a line).

Note: This approach follows [1]. In that work Xing *et al.* use side information identifying pairs of points as "similar". They then construct a metric that minimizes the distance between all such pairs of points. At the same time, they attempt to ensure that all "dissimilar" points are separated by some minimal distance. By default, they consider all points not explicitly identified as similar to be dissimilar. They proof their algorithm results in A always being rank 1. You may want to read their paper (provided on the course web page) and get some idea.

References

[1] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, pages 505–512, 2003.