Statistical Pattern Recognition in High Dimensions

Carey E. Priebe
Johns Hopkins University
cep@jhu.edu

Workshop on Opportunities at the Statistics-Operations Research Interface
ASA, INFORMS, NSF
January 11-13, 2004
Santa Fe, NM USA
Abstract

Given observations in high-dimensional space, we want to find a low-dimensional manifold which captures the information relevant to statistical pattern recognition (classification, clustering, etc) for these data. One approach: write a probability model which straddles "practically relevant" and "mathematically tractable"; define an objective function whose arg opt (over manifolds) will act as a useful surrogate for "manifold with the most relevant information"; find a good approximation for the arg opt. This procedure must be accomplished in real-time in a dynamic environment to produce an "adaptive sensor" adapting its low-dimensional view based on the pattern recognition exploitation function (rather than some far-afield surrogate such as signal-to-noise).
• sensing —

\[ s : \mathcal{R} \rightarrow \mathbb{R}^p, \ p \gg 1 \]

• dimension reduction —

\[ r : \mathbb{R}^p \rightarrow \mathbb{R}^q, \ q \ll p \]

• classification/clustering —

\[ g : \mathbb{R}^q \rightarrow [0, 1]^C \]
The figure illustrates the set \( \mathcal{M}(d) \) and the function \( L(\hat{g}_n(d)) \) as \( d \to \infty \). The optimal solution is indicated by the point \( g_*(d) \), and the limit of the function as \( d \to \infty \) is 0. The optimal value of \( d \) is denoted by \( d_{opt} \).
“The wealth of your practical experience with sane and interesting problems will give to mathematics a new direction and a new impetus.”

— Leopold Kronecker to Hermann von Helmholtz