Double data piling leads to perfect classification in high dimensions

Sungkyu Jung[1]; Woonyoung Chang[1]; Jeongyoun Ahn[2]

[1] Seoul National University
[2] University of Georgia

Abstract:

Data piling refers to the phenomenon that training data vectors from each class project to a single point for classification. While this interesting phenomenon has been a key to understanding many distinctive properties of high-dimensional discrimination, the theoretical underpinning of data piling is far from properly established. In this work, high-dimensional asymptotics of data piling is investigated under a spiked covariance model, which reveals its close connection to the well-known ridged linear classifier. In particular, by projecting the ridge discriminant vector onto the subspace spanned by the leading principal component directions and the maximal data piling vector, we show that a negatively ridged discriminant vector can asymptotically achieve data piling of independent test data, essentially yielding a perfect classification. The second data piling direction is obtained purely from training data and shown to have a maximal property. In this setting, asymptotic perfect classification occurs only along the second data piling direction. The findings are extended to a multi-category classification setting, in which double data piling also occurs.

Keywords:

Maximal data piling ; high-dimension, low-sample-size; ridged classifier; linear classification