

1124-68-299

Mireille Boutin* (mboutin@purdue.edu), School of ECE, 465 Northwestern Avenue, West Lafayette, IN 47907. *The Unexpected Geometry of "Real" High-Dimensional Data*. Preliminary report.

We describe a surprising property of high-dimensional data sets, which we first observed while studying large sets of images and subsequently found in many other “real” data sets as well. Our observations suggest that real high-dimensional data sets have a tremendous amount of structure, so much so that a mere projection onto a random 1D subspace of the data space is likely to uncover some of that structure. This has important implications in the area of automatic recognition and database indexing. For example, we have developed an ultra-fast method based on random 1D projections for clustering a high-dimensional data set. We have also developed methods for automatic recognition of patterns; the random heuristic behind these methods is very simple, yet they can lead to very accurate results. Building on these simple heuristics, we proposed pattern recognition benchmarks that can be used to determine if an automatic classifier with a pre-determined error rate can be obtained from a given training dataset without the need to use any sophisticated method such as support vector machines or deep neural networks.

This work is in collaboration with my graduate student Sangchun Han (now at Google) and Tarun Yellamraju. (Received September 12, 2016)