Learning in High Dimensions

Rich Caruana
Nikos Karampatziakis
Ainur Yessenalina
Department of Computer Science
Cornell University, Ithaca, NY 14853

caruana@cs.cornell.edu
www.cs.cornell.edu/~caruana

We extend an empirical evaluation of supervised learning methods we performed several years ago to examine learning from high dimensional data. New experiments with data of varying dimension confirm that the results of our previous study hold until the data has about 1000 dimensions. But, as the dimensionality increases beyond this point, things change in surprising ways. As dimensionality increases from about 5,000 to 500,000 dimensions, the performance of boosting and kernelized SVMs begins to taper off, while other learning methods such as Random Forests and Neural Nets remain strong and ultimately take top honors. This pattern is observed for a variety of losses including squared loss, 0/1 loss, and AUC.

Perhaps the biggest surprise is that although linear methods such as logistic regression and linear SVMs are much more competitive in high dimensions than they were in low dimensions (where their overall performance was weak), they still do not compete with non-linear methods such as random forests and neural nets when learning from data with very high dimension. Even problems of very high dimension benefit from non-linear models.

Another effect that we had not anticipated is that calibration appears to be more important when learning from high-D data than when learning from low-D data (we had expected the opposite). Even learning methods that are well calibrated in low-D such as neural nets and bagged trees benefited from calibration when learning from high-dimension data. Effects that we had anticipated such as the decrease in performance of memory-based methods like KNN with increasing dimensionality are clearly evident in our results. Our experiments allow us to roughly quantify how much learning methods such as MBL fall behind as dimensionality increases.