Maxim Kachalov, Google Prediction API on Adult dataset write-up

Final Results

Total Percent Correct/# Instances	Percent Correct ">=50K"/# Instances	Percent Correct "<50K"/#Instances
85.04% / 16273	42.12% / 3843	93.42% / 12430

     Benefiting from the fact that the Adult dataset is cleaned, curated, and widely studied, there were available resources for examining the quality of the predictions. As compared to the most recent publication (released over 10 years ago) that utilized this data and recorded a best-performance of ~91% [1], the final results are sub-optimal. However, the accuracy is on-par with the results obtained in the original paper (released in 1996) which achieved a best performance of ~85% [2]. When quantifying the accuracy, it was anticipated that one of the classes should be significantly smaller as the nominal median wage in 1995 just in the United States was about $34,000. It was also anticipated that this should skew the results as simply achieving a high accuracy rate on the larger class would have a strong effect on the overall accuracy. As can be seen in the results, both these hypotheses were correct.

     Although the specifications of what is happening under the hood in Google's Prediction API could not be definitively found, some have suggested that it uses a Naive Bayes Classifier, which is the same algorithm used in the original paper. This makes sense given the highly similar accuracies achieved by both approaches. What was interesting about this is the fact that Naive Bayes is often said to be robust to the issue of a small number of positive examples in a large dataset; the results show that this was not the case here. Additionally, though the most recent paper that could be obtained is over 10 years old, the Google predictions, comparatively, are only 5% less than the best-achieved accuracy using Boosted Stumps and on-par with the other algorithms examined in the paper. Reiterating that these results are not optimal in the general case, in the context of its use as a black box for machine learning, the Google API predictions can be deemed acceptable.

1. Caruana, Rich, and Alexandru Niculescu-Mizil. "An Empirical Evaluation of Supervised Learning for ROC Area." *ROCAI*. 2004.

2. Kohavi, Ron. "Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid." *KDD*. 1996.