Progress Report

MDST ASSISTments 2017 Team

2017-12-18

The competition organizers requested a brief description of our methods. We here explain how we arrived at our submission of 2017-10-22.

Executive Summary

Our best-performing entry is one of our baseline models. This model fits a logistic curve not by MAP estimation but instead based on an SVM's decision function; this more gently penalizes outliers in accordance with the competition's scoring rules. We addressed the main challenge of feature engineering by using features such as "the median concentration over time" and "the slope of correctness over time" and nothing more sophisticated. Since the overall model's decision function involves both unsupervised (PCA) and supervised learning, we were able to learn from the unlabeled students; however, for the submission in question, we did not. In following the data before our intuitions, we have arrived at a model with the potential to give insight into factors predictive of future careers. For instance, one of our preliminary findings is that scores on the Massachusetts Comprehensive Assessment System were highly predictive of the STEM label. We look forward to further interrogating our model for insights.

Technical Summary

Featurization

We began with variable-length time-series of fixed-length "click vectors", each of which had a timestamp, several categorical entries, several integer-valued entries, and several real-valued entries. We embedded each click vector in $\mathbb{R}^D$ through one-hot encoding, then summarized each time-series through element-wise aggregate statistics such as mean, standard deviation, median, interquartile range, and least-squares slope over time. Some features were independent of time, so we included them directly; of these, we found the MCAS scores and school-ID to be essential: performance significantly deteriorated when we ablated them. In our only compelling use of domain knowledge, we also included an "MCAS is negative" binary flag feature.

We then pre-process our data as follows: normalize each element to 0 mean and unit variance. We then perform PCA to return a whitened, lower-dimensional dataset. Because PCA is unsupervised, we can perform it on the entirety of the ASSISTments dataset, not just on the minority of labeled student records. However, to do so without care would lead us to overestimate our confidence in our performance estimates, so we have not yet done this.

Model and Hyperparameter Selection

We used a linear, L^2-regularized Support Vector Machine (SVM) with regularizer $\lambda_{\text{svm}} = C^{-1} \in [10^{0.0}, 10^{2.5}]$ and PCA dimension $d \in [10^{0.5}, 10^{1.5}] \cap \mathbb{Z}$. The SVM decision function we converted to a "probability score" by training L^2-regularized logistic regression to predict binary labels from the decision function values ($\lambda_{\text{logreg}} = 10^{0.0}$).

We sampled each hyperparameter independently on a logarithmic scale, and we tested each configuration thus obtained on $L = 16$ random $80-20$ train-val splits. We did this for $M = 16$ configurations, then picked the best configuration according to a variance-penalized average validation score. This winning configuration we trained on the whole data.

If we instead maintain a global train-test split and at the end train the winning configuration on only the global train set, then we obtain an estimate for out-of-sample performance. Average estimates from $N = 16$ random $80 - 20$ train-test splits per algorithm, we compared linear methods, random forests, and neural networks. We found L^2-regularized SVM to perform well.