Predicting Task Completion from Rich but Scarce Data

José P. González-Brenes and Jack Mostow
{joseg, mostow}@cs.cmu.edu
Project LISTEN, School of Computer Science, Carnegie Mellon University

We present a data-driven model for predicting task completion in Project LISTEN's Reading Tutor, which takes turns picking stories and listens to the child read aloud [1]. However, children do not always finish stories, and we would like to understand why, or at least detect when they are about to stop. So our EDM challenge is to learn a model to predict task completion – a widely used metric of dialogue systems' performance. Such a model could help detect imminent disengagement in time to address it, and identify factors that influence task completion, including tutor behaviors, thereby providing useful guidance to make tutors engage students longer and more effectively.

The richness of multimodal tutorial interaction over time makes the space of possible features to describe it large relative to the amount of data. When the number of features is large compared to the amount of data, classifier learners tend to overfit the data, so we need a method that learns robust models from few training examples with many features.

Consider the supervised learning problem with training data $S = \{(x^{(i)}, y^{(i)})\}$, $i = 1 \ldots n$, where each data point is a p-dimensional vector $x^{(i)}$, and $y^{(i)}$ is its label. The number of features p may exceed the number of data points ($p >> n$). A binary logistic regression model has the following form, where the vector θ contains the p parameters of the model:

$$p(y = 1 \mid x; \theta) = \frac{1}{1 + \exp(-\theta^T x)} \qquad (1)$$

ℓ_1-regularized logistic regression [2] finds the vector θ^* that maximizes this expression:

$$\theta^* = \arg\max_{\theta} \boxed{\sum_{i}^{n} \log p(y^{(i)} \mid x^{(i)}; \theta)} - \boxed{\lambda \| \theta \|_1} (2)$$

Here the term in the first box represents how well the model fits the training data according to Equation (1), and the second term penalizes the model by the sum of its parameters' absolute values ($\|\theta\|_1$). By discouraging non-zero parameters – which select the features actually used – this penalty can prevent overfitting. The hyper-parameter λ controls the trade-off between bias and variance, and can be set by internal cross-validation using a held out set of training data. For $\lambda = 0$, Equation (2) reduces to conventional logistic regression. As λ increases, the model's complexity is penalized more strongly, reducing the number of features it uses.

Our data points to test this method are 2112 story readings by 161 children, lasting four or more sentences. We want to distinguish completed readings from unfinished readings. We truncate each positive example to match the number of sentences to the one of a negative example, so as to sample potential stopping points, not just the end of the story. Negative examples are the entire unfinished readings, which can end anywhere.

We use both static and dynamic features. Static features, e.g. student grade (K-6) and story length, remain static over a story reading. Dynamic features, e.g. number of sentences read, words read per minute, or clicks logged, change throughout a reading, so we compute separate values for 1, 2, 3, and 4 sentences from the beginning and end of the reading. To avoid cheating, we exclude features of the last sentence read, e.g. whether the child clicked to exit. Altogether we have 17,163 raw, squared, and threshold features.

Figure 1 shows classification accuracy for balanced subsets of different sizes, with the same number of positive and negative examples drawn randomly from the 2112 readings. We used 10-fold cross-validation splitting data randomly. We also tried splitting across students, but this doesn't affect accuracy on the full set, yet it is noisy for small subsets due to students with sparse data. The error bars represent the 90% confidence interval. As Figure 1 shows, the method achieves 60% accuracy by training on only 500 examples, increasing to 70% with 600 examples, and asymptoting at 78% above 1500 examples. The three most predictive features are derived from the percentage of the story completed so far, consistent with the intuition that children are likelier to finish shorter stories.

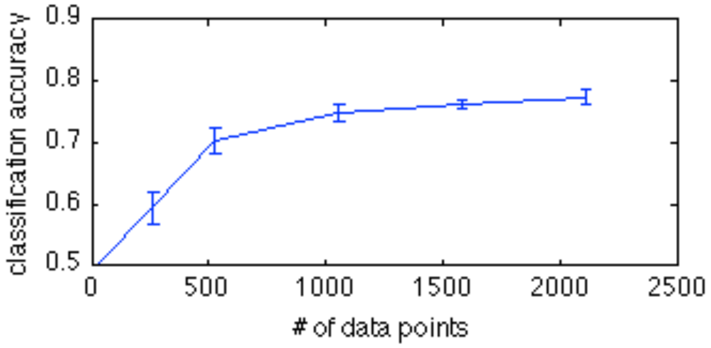

Figure 1: Classification Accuracy on Data Sets of Different Sizes

This paper has presented a novel model to predict students' task completion in a multimodal tutor, using a method that can train models from data with many dimensions but few examples. The method, used successfully elsewhere, should interest the EDM community because of its potential to cope with the curse of dimensionality inflicted by the richness of tutorial interaction.

Acknowledgements: This work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080628 to Carnegie Mellon University. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or U.S. Department of Education.

References

[1] Aist, G. and J. Mostow. Improving story choice in a reading tutor that listens. *Proceedings of the Fifth International Conference on Intelligent Tutoring Systems (ITS'2000)*, 645. 2000. Montreal, Canada.

[2] Ng, A.Y. Feature selection, L1 vs. L2 regularization, and rotational invariance. *International Conference on Machine Learning*, 78. 2004. New York, NY.