Appeared in poster session of LEARNING 2005. Snowbird, Utah. April 5-8, 2005.

Active Learning for Multi-Class Logistic Regression

Andrew I. Schein and Lyle H. Ungar[1]
Department of Computer and Information Science
Levine Hall
3330 Walnut Street
Philadelphia, PA 19104-6389
{ais,ungar}@cis.upenn.edu

Which active learning methods can we expect to yield good performance in learning logistic regression classifiers? Addressing this question is a natural first step in providing robust solutions for active learning across a wide variety of exponential models including maximum entropy, generalized linear, loglinear, and Markov random field models. We extend previous work on active learning using explicit objective functions [1, 2] by developing a framework for implementing a wide class of loss functions for active learning of logistic regression, including variance (A-optimality) and log loss reduction. We then run comparisons against different variations of the most widely used heuristic schemes, query by committee and uncertainty sampling, to discover which methods work best for different classes of problems and why.

Our approach to loss functions for active learning borrows from the field of optimal experimental design in statistics. We exploit several properties of nonlinear regression models that allow us to compute the variance of a prediction with respect to the model's input distribution. The strategy of minimizing prediction variance is referred to as A-optimality. A Taylor series around many loss functions conveniently factorizes into alternative weightings of this variance computation. We investigate squared and log loss within this framework, and compare our results against a recently proposed heuristic approximation of log loss [3].

Our empirical evaluations are the largest effort to date to evaluate explicit objective function methods in active learning. We employ seven data sets in the evaluation from domains such as image recognition and document classification. The data sets vary in number of categories from 2 to 26 and have as many as 6,191 predictors. Our work establishes the benefits of these often cited (but rarely used) strategies, and counters the claim that experimental design methods are too computationally complex to run on interesting data sets.

The same data were used to evaluate several heuristic methods. Uncertainty sampling was tested using two different measures of uncertainty: Shannon entropy and margin size. Margin-based uncertainty sampling was found to be superior; however, both methods perform worse than random sampling at times. We show that these failures to match random sampling can be caused by predictor space regions of varying noise or model mismatch. The query by committee algorithm was uniformly dominated by the margin-based uncertainty sampling. A-optimality and the margin-based uncertainty sampling were the only methods that almost always outperformed random sampling.

References

[1] David A. Cohn. Neural network exploration using optimal experimental design. *Neural Networks*, 9(6):1071–1083, 1996.

[2] David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):589–603, 1992.

[3] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. 18th International Conf. on Machine Learning*, pages 441–448. Morgan Kaufmann, San Francisco, CA, 2001.

[1] presenting author.

Topic: learning algorithms
Preference: oral/poster