Active Learning for Multi-Class Logistic Regression

Andrew I. Schein and Lyle H. Ungar

The University of Pennsylvania
Department of Computer and Information Science
3330 Walnut Street
Philadelphia, PA 19104-6389 USA
{ais, ungar}@cis.upenn.edu

Goals

- Discover best practices for active learning with logistic regression.
- Develop loss function methods for logistic regression:
  - Squared Loss
  - Log Loss
- Evaluate heuristic methods:
  - uncertainty sampling
  - query by bagging
  - classifier certainty
- Explain method underperformance, when it occurs.

Logistic Regression: A Method for Class Probability Estimation

- Binary

$$f(\mathbf{x}_n) = \frac{1}{1 + \exp(-\mathbf{x}_n \cdot \mathbf{w})} \tag{1}$$

- Multiple Classes (> 2)

$$f(\mathbf{x}_n, y) = \frac{\exp(\mathbf{x}_n \cdot \mathbf{w}_y)}{\sum_{y'} \exp(\mathbf{x}_n \cdot \mathbf{w}_{y'})} \tag{2}$$

$f(\mathbf{x}, y)$ is an inexact model for $t(\mathbf{x}, y)$ introducing a bias.

$\mathbf{x}_n$ is a vector of predictors for observation n.
$\mathbf{w}_y$ is a vector of weights indexed by class.

Pool-based Active Learning for Classification

We focus on pool-based settings:

Observations x_n are given without their corresponding class labels y_n.

Our goal: Sequentially pick and label x_n to best improve classifier.

Example:

We have a pool of 2000 documents with no topic label.

Which documents do we label to build the best topic classifier?

A Variance Reduction Approach Based On A-Optimality

Using a Taylor expansion:

$$\sum_{n \in \text{Pool}} \sum_{c} E_T[\hat{f}(x_n, c|T) - E_T f(x_n, c|T)]^2 \quad \simeq \quad \text{tr}\left\{AF^{-1}\right\}$$

A encodes the distribution of predictors in the pool.

F^{-1} encodes the variance of the model parameter estimates.

The expectation is over training sets of fixed size.

The objective is to make $\hat{f}$ close to $E_T f$ in squared loss.

We also explore a variant for log loss.

Methods Evaluated

- Baseline
  random instance selection
  bagging (interesting since it is used in QBB methods)

- Loss Function Approaches
  variance reduction (A-optimality)
  log loss reduction

- Heuristics
  CC Classifier Certainty
  QBB-MN Query by Bagging – KL divergence measure
  QBB-AM Query by Bagging – ensemble margin
  entropy-based uncertainty Sampling
  margin-based uncertainty Sampling

Data Sets and Evaluation

Data Set	Classes	Obs	Pred	Maj	Data Type
Art	20	20,000	5	3635	artificial
ArtNoisy	20	20,000	5	3047	artificial
ArtConf	20	20,000	5	3161	artificial
Comp2a	2	1989	6191	997	document
Comp2b	2	2000	8617	1000	document
LetterDB	26	20,000	16	813	char. image
NewsGroups	20	18,808	16,400	997	document
OptDigits	10	5620	64	1611	char. image
TIMIT	20	10,080	12	1239	voice
WebKB	4	4199	7543	1641	document

- Split data into equal size pool and test set for 10-fold x-validation.
- Start training at 20 observations, stop at 300, run significance tests.

Results - Accuracies

Data Set	random	bagging	variance	log loss
Art	0.809	**0.792**	**0.862**	**0.867**
ArtNoisy	0.565	**0.557**	**0.579**	**0.579**
ArtConf	0.837	0.830	0.842	0.840
Comp2a	0.821	**0.794**	0.805	0.821
Comp2b	0.799	0.793	0.807	0.796
LetterDB	0.609	**0.593**	**0.644**	**0.646**
NewsGroups	0.483	**0.422**	–	–
OptDigits	0.927	0.931	**0.937**	**0.944**
TIMIT	0.413	**0.397**	0.405	0.423
WebKB	0.830	**0.803**	–	–

	CC	QBB-MN	QBB-AM	entropy	margin
Art	**0.821**	**0.848**	**0.861**	**0.832**	**0.867**
ArtNoisy	0.567	**0.577**	0.571	**0.536**	**0.572**
ArtConf	0.845	0.843	0.835	**0.723**	**0.749**
Comp2a	**0.788**	0.814	0.818	0.826	0.818
Comp2b	0.796	0.804	0.808	0.805	0.800
LetterDB	**0.625**	**0.599**	**0.637**	**0.548**	**0.633**
NewsGroups	–	**0.464**	**0.444**	**0.356**	**0.438**
OptDigits	**0.942**	**0.941**	**0.949**	**0.951**	**0.952**
TIMIT	**0.395**	0.408	**0.438**	**0.327**	**0.440**
WebKB	–	**0.844**	**0.860**	**0.855**	**0.860**

Results - equivalent random samples required

Data Set	random	bagging	variance	log loss
Art	300	220	>600	>600
ArtNoisy	300	240	450	450
ArtConf	120	100	130	120
Comp2a	150	110	130	210
Comp2b	150	130	170	140
LetterDB	300	250	380	380
NewsGroups	300	230	–	–
OptDigits	300	310	350	430
TIMIT	300	240	290	310
WebKB	300	220	–	–

	CC	QBB-MN	QBB-AM	entropy	margin
Art	330	480	>600	370	>600
ArtNoisy	310	420	350	160	350
ArtConf	140	140	110	50	50
Comp2a	90	150	150	190	150
Comp2b	140	160	170	160	150
LetterDB	340	250	360	180	360
NewsGroups	–	290	280	170	260
OptDigits	400	400	>600	>600	>600
TIMIT	230	290	420	90	380
WebKB	–	360	570	460	530

Analysis of Heuristic Method Performance

- Entropy Sampling fails on the noisier data sets
  - Noise is defined as training set independent squared error.
  - Estimated using very large training sets.
- Margin Sampling fails on data with hierarchical category structure
  - NewsGroups data set contains 5 computer and 5 politics topics.
  - ArtConf is an artificial data set constructed with this property.
- Difficult to explain failures of QBB-MN and Classifier Certainty.

NewsGroup Hierarchy of Topics

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey
talk.religion.misc alt.atheism soc.religion.christian	misc.forsale
sci.crypt sci.electronics sci.med sci.space	talk.politics.misc talk.politics.guns talk.politics.mideast

Evaluation Conclusions

- Loss function approaches are robust
  - but relatively slow
  - variance and log loss perform comparably, log loss possibly better
- Heuristic methods often fail to beat random selection
  - Noise level and type predicts some failures
- Best heuristic: margin-based uncertainty sampling
  - Very fast
  - Failures are predictable– opportunity to improve the method.
- Effect of bagging: mostly negative