A Bayesian expected cost reduction approach to active learning

Richard Fredlund

April 2011

Submitted by Richard Fredlund, to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Computer Science, April 2011.

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

(signature) ..........................................................................................................................
Declaration

Chapter 3 is partially based on the previously published work:


Richard Fredlund
To my son Daniel Fredlund who is never far from my mind.
Acknowledgements

I would especially like to thank my supervisors Richard Everson and Jonathan Fieldsend who always managed to exceed my expectations. I would also like to thank the various friends and family who were helpful and supportive along the way.
Abstract

There has been growing recent interest in the field of active learning for binary classification. This thesis develops a Bayesian approach to active learning which aims to minimise the objective function on which the learner is evaluated, namely the expected misclassification cost. We call this approach the expected cost reduction approach to active learning. In this form of active learning queries are selected by performing a ‘lookahead’ to evaluate the associated expected misclassification cost.

Firstly, we introduce the concept of a query density to explicitly model how new data is sampled. An expected cost reduction framework for active learning is then developed which allows the learner to sample data according to arbitrary query densities. The model makes no assumption of independence between queries, instead updating model parameters on the basis of both which observations were made and how they were sampled. This approach is demonstrated on the probabilistic high-low game which is a non-separable extension of the high-low game presented by Seung et al. [1992]. The results indicate that the Bayes expected cost reduction approach performs significantly better than passive learning even when there is considerable overlap between the class distributions, covering 30% of input space. For the probabilistic high-low game however narrow queries appear to consistently outperform wide queries. We therefore conclude the first part of the thesis by investigating whether or not this is always the case, demonstrating examples where sampling broadly is favourable to a single input query.

Secondly, we explore the Bayesian expected cost reduction approach to active learning within the pool-based setting. This is where learning is limited to a finite pool of unlabelled observations from which the learner may select observations to be queried for class-labels. Our implementation of this approach uses Gaussian process classification with the expec-
tation propagation approximation to make the necessary inferences. The implementation is demonstrated on six benchmark data sets and again demonstrates superior performance to passive learning.
# Contents

1. Introduction 10  
   1.1. Expected cost reduction ........................................ 12  
   1.2. Outline of thesis ................................................. 14  

2. Background and Related Work 17  
   2.1. Supervised Learning .............................................. 17  
       2.1.1. Maximum likelihood ......................................... 21  
       2.1.2. Bayesian approach .......................................... 23  
       2.1.3. Monte Carlo integration ................................... 28  
   2.2. Active Learning .................................................. 31  
       2.2.1. Theoretic bounds on Active Learning ....................... 36  
       2.2.2. QBC approaches ............................................. 38  
       2.2.3. Uncertainty based approaches ............................... 42  
       2.2.4. Expected cost reduction approaches ........................ 43  
       2.2.5. Other approaches ........................................... 44  
       2.2.6. Performance measures for active learning ................. 45  
       2.2.7. Stopping criteria ........................................... 47  
   2.3. Discussion ....................................................... 48  

3. A Bayesian Framework for Active Learning 51  
   3.1. Introduction .................................................... 51  
   3.2. A framework for Bayesian active learning .................... 51  
       3.2.1. Sample oracle and query density ............................ 52  
       3.2.2. Parameter update ........................................... 53  
       3.2.3. Expected misclassification cost ............................ 55
3.2.4.  Algorithm overview ........................................... 57
3.3.  Probabilistic high-low game ................................. 59
3.4.  Results for the high-low game .............................. 63
3.5.  Examples where wide queries are better ..................... 81
   3.5.1.  Example 1: Three Gaussians ........................... 81
   3.5.2.  Example 2: Pie example .................................. 85
3.6.  Discussion ...................................................... 86

4.  Pool-based Active Learning .................................. 88
   4.1.  Introduction .................................................. 88
   4.2.  Point queries ................................................ 89
   4.3.  Pool-based active learning ................................ 92
   4.4.  Bayesian cost reduction approach to pool-based active learning ................................. 93
   4.5.  Gaussian process active learning ......................... 95
      4.5.1.  GP regression ............................................. 97
      4.5.2.  GP classification ....................................... 98
   4.6.  Implementation and results ................................ 101
      4.6.1.  Benchmark data-sets .................................. 107
      4.6.2.  Synthetic data-sets ..................................... 110
      4.6.3.  Non-synthetic data-sets ................................. 117
   4.7.  Performance measures for benchmark data-sets ............. 122
   4.8.  Comparison with other algorithms ....................... 125
      4.8.1.  Hyperparameters ........................................ 125
      4.8.2.  Data-sets .............................................. 126
      4.8.3.  Results ............................................... 127
   4.9.  Summary ..................................................... 133

5.  Discussion ...................................................... 134
   5.1.  Query density framework .................................... 134
   5.2.  Pool based setting .......................................... 136
   5.3.  Closing remarks ............................................. 137
A. Misclassification costs for the Pie Example
   A.0.1. Single point query ........................................ 138
   A.0.2. Random query ........................................... 140

B. Parameter posterior $p(\alpha, \beta \mid x_1, \phi)$ for a uniform prior in the probabilistic high-low game 142