Chapter 1
Machine Learning Techniques—Reductions Between Prediction Quality Metrics

Alina Beygelzimer, John Langford, and Bianca Zadrozny

Abstract Machine learning involves optimizing a loss function on unlabeled data points given examples of labeled data points, where the loss function measures the performance of a learning algorithm. We give an overview of techniques, called reductions, for converting a problem of minimizing one loss function into a problem of minimizing another, simpler loss function. This tutorial discusses how to create robust reductions that perform well in practice. The reductions discussed here can be used to solve any supervised learning problem with a standard binary classification or regression algorithm available in any machine learning toolkit. We also discuss common design flaws in folklore reductions.

1.1 Introduction

Machine learning is about learning to make predictions from examples of desired behavior or past observations. Learning methods have found numerous applications in performance modeling and evaluation (see, for example, [33, 22, 37, 41, 43, 39]). One natural example of a machine learning application is fault diagnosis: based on various observations about a system, we may want to predict whether the system is in its normal state or in one of several fault states. Machine learning techniques are preferred in situations where engineering approaches like hand-crafted models simply can not cope with the complexity of the problem. In the fault diagnosis prob-

Alina Beygelzimer
IBM Thomas J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532 e-mail: beygel@us.ibm.com

John Langford
Yahoo! Research, New York, NY e-mail: jl@yahoo-inc.com

Bianca Zadrozny
Fluminense Federal University, Brazil e-mail: bianca@ic.uff.br

lem, it is reasonably easy to collect examples of resolved faults, but writing robust diagnosis rules is very difficult.

A basic difficulty in applying machine learning in practice is that we often need to solve problems that don't quite match the problems solved by standard machine learning algorithms. In fault diagnosis, for example, the cost of misclassifying a faulty state as a normal state is sometimes much higher than the cost of misclassifying a normal state as a faulty state. Thus binary classification algorithms, which don't take misclassification costs into account, do not perform well on this problem.

Reductions are techniques that transform practical problems into well-studied machine learning problems. These can then be solved using any existing base learning algorithm whose solution can, in turn, be used to solve the original problem. Reductions have several desirable properties.

- They yield highly automated learning algorithms. Reductions convert *any* learner for the base problem into a learning algorithm for the new problem. Any future progress on the base problem immediately translates to the new problem.
- Reductions are modular and composable. A single reduction applied to N base learners gives N new learning algorithms for the new problem. Simple reductions can be composed to solve more complicated problems.
- The theory of learning has focused mostly on binary classification and regression. Reductions transfer existing learning theory to the new problem.
- Reductions help us organize and understand the relationship between different learning problems.

An alternative to reductions is designing new learning algorithms or modifying existing ones for each new problem. While this approach is quite attractive to learning algorithm designers, it is undesirable in some situations. For example, some algorithms cannot be easily modified to handle different learning problems, as evidenced, for example, by inconsistent proposals for extending Support Vector Machines to multiclass classification (see [30]). More generally, we can expect that people encountering new learning problems may not have the expertise or time for such adaption (or simply don't have access to the source code of the algorithm), implying that a reduction approach may be more desirable.

A critical question when comparing the two approaches is performance. Our experience is that both approaches can be made to work well. There is fairly strong empirical evidence that reductions analysis produces learning algorithms that perform well in practice (see, for example, [18, 10, 47, 13, 38]). This tutorial shows how reductions can be easily used by nonexperts.

1.2 Basic Definitions

Data points, called *examples*, are typically described by their values on some set of *features*. In fault diagnosis, for example, each event can be represented as a binary vector describing which observations have been made (ping latency from one node