

Research Statement

John Langford

3rd January 2006

1 Basic Approach

I am interested in understanding the process of generalization – the ability to make predictions based on past experience. I am a learning theorist – I develop theoretical tools for machine learning. However, I strongly care about whether and where learning theory is useful in practice. In particular, most methods that I developed have been tested and are being used by practitioners.

One reason learning theories fail to be useful is that they make many unverifiable assumptions. This observation motivates a basic drive of my work. Theories relying upon fewer assumptions are more likely to be true in practice and (hence) robustly applicable to a large number of real-world learning problems. I will organize my research from fewest to most unverifiable assumptions.

1.1 No unverifiable assumptions

It turns out that we *can* talk about learning without making unverifiable assumptions.

Learning Reductions

I have proposed a new style of analysis based on reductions. Reductions for learning are systems that transform real-world examples into examples for core binary learning algorithms, apply these algorithms, and then transform their predictions into predictions for the real-world problem. The analysis of this transformation is assumption-free: We can bound the loss rate of the real-world predictor in terms of the error rate of the binary classifier. This relationship holds no matter what process produces the data.

There are many benefits to the reduction approach.

1. Since any classifier learner can be paired with a reduction, developing one reduction produces many distinct learning algorithms.
2. Similarly, the code and thought of others can be reused. Reductions give new life to old algorithms by making them applicable in new domains.
3. The ease of reductions between one learning problem and another is a guide to the relative hardness of learning problems. Reductions give us a real hope of organizing the ontology of learning problems.

One specific example is class probability estimation using the Probing reduction [10]. Class probability estimation comes up in situations where learning algorithms are used for giving advice such as predicting disease presence for a doctor. The probing reduction creates a set of binary problems indexed by a parameter $0 \leq t \leq 1$. For problem t , the rate of samples from class 0 is proportional to t while the rate of samples from class 1 is proportional to $1 - t$. After learning a classifier for each binary problem, we have a spectrum of predictors, one for each t . On any individual test example, the reduction outputs a probability estimate which is the fraction of the classifiers predicting class 1. For this simple elegant algorithm, we can prove a basic mathematical property: for all binary problems, the binary regret bounds the squared error in the probability estimate. This algorithm was also tested and worked better than the best alternatives we could find.

I also developed reductions to binary classification for the following problems:

1. importance weighted classification (e.g. spam filtering) [17]
2. multiclass classification (e.g. character recognition) [8][DB95]
3. regression (e.g. function estimation) [10]
4. cost sensitive classification (when each prediction has a different associated cost) [6,8]
5. reinforcement learning, given side information (an agent wandering in a world) [4]

Every one of these reductions have been tested empirically and shown to work well, often better than the best available alternative.

Online Learning

Online Learning is an older form of assumption free analysis. In this setting a learning algorithm has a set of "experts" each of which makes predictions. The learning algorithm uses these experts to make its own predictions, then learns what the correct prediction should have been. A typical analysis compares the performance of the learning algorithm with the performance of the best expert. The optimal master algorithm can be stated as a game-theoretic strategy. Jacob Abernethy and I recently found a computationally tractable solution to computing this game-theoretic strategy [2].

1.2 IID only assumption

A standard assumption is that samples come independently and identically distributed according to some process. It is sometimes satisfied in practice, and even when it is not, the intuitions from the theory often carry over intact.

Prediction Bounds

My thesis was on prediction bounds which attempt to bound the (future) error rate of a learned classifier using statistical techniques [5,11,12,18,20,26,27,29,31,32]. The approach I took derived exceptionally tight prediction bounds. There are several outcomes of this approach.

1. New learning algorithms. Any bound is (implicitly) an algorithm: minimize the bound. That algorithm was never previously practical, but it is becoming practical now. A good example is the "set covering machine" [LMS05].
2. Better testing methods (see [5,11]). In addition to wanting good predictions, many practitioners also need to bound (or predict) their future performance. Standard approaches suffer badly in corner cases because they make false assumptions such as assuming the error rate is drawn from a Gaussian. Using prediction bounds, we can create robust testing methods.

Active Learning

In active learning there is a different source of information: the learning algorithm can ask for the correct prediction on unlabeled data. The idea here is to capture the observation that unlabeled data is very common so our goal should be minimizing the number of labels we request. A basic observation about active learning is that it can be *exponentially* better than standard batch learning. In forthcoming work, we show that this exponential improvement can be gained when data is IID. The exponential improvement holds up to twice the minimum error rate of a set of classifiers.

1.3 Many assumptions

In many natural settings, we have a sensor (such as a camera) which is taking a picture of a scene. The set of degrees of freedom in the scene is much smaller than the set of degrees of freedom in the sensor. The set of sensor observations can be thought of as belonging to a low dimensional manifold embedded in a high dimensional space. If we assume that these points are not too far apart and the embedding is well separated (doesn't come near to crossing itself), we can prove that the Isomap algorithm [28] will recover the structure. The Isomap algorithm was optimized for this, and that (theoretically motivated) optimization reduced the number of samples required by an order of magnitude.

Both the analysis of the Isomap and the Isomap algorithm have been "runaway hits". Scholar.google.com claims over 500 citations for the algorithm.

2 Other Projects

I have worked on many other projects as well, varying from establishing the possibility of steganography to practical methods for improving belief state tracking. I believe each of these is of significant importance.

1. Cover trees [1] are a datastructure supporting nearest neighbor queries. They are the first such datastructure to marry strong theoretical guarantees with fast empirical performance.
2. Steganography is like encryption, but you hide the existence of a message as well as its content. We proved that steganography is possible subject to reasonable assumptions [24]. As a followup to this work, we showed that two people can engage in a computation in such a way that neither person can even tell if the other is cooperating [9].
3. Captchas [15] change around the problem: How can we create problems which humans can solve but computers cannot? The result of the Captcha project is being used by Yahoo extensively.
4. Probabilistic graphplan [34] is an algorithm for planning in probabilistic domains. This algorithm was one of the first probabilistic planners.
5. Particle filters are a method for keeping track of a Bayesian posterior over time. I worked on a method for making sure that low approximation error occurs in important states [25].
6. Graphical games add structure to game theory by positing a graph on the interactions of the participants. Correlated equilibria are notions of equilibria which allow arbitrary dependencies between the strategies of players in a game. We showed how to efficiently solve for correlated equilibria in graphical games [16].

The Future On March 24-25, 2006 we are holding a workshop on Atomic Learning at TTI-Chicago. The basic question the workshop addresses is: What are all the ways to tear apart big learning problems into little learning problems, solve the little learning problems, and put the set back together to solve the big problem? This is one of the directions I am fundamentally interested in for the future.

Most references

All citations are in my CV except for the following references.

References

- [DB95] Tom Dietterich and Ghulum Bakiri, "Solving Multiclass Learning Problems via Error-Correcting Output Codes", *Journal of Artificial Intelligence Research*, 2:263-286, 1995.
- [LMS05] Francois Laviolette, Mario Marchand and Mohak Shah. A PAC-Bayes Approach to the Set Covering Machine. Proceedings of the 2005 conference on Neural Information Processing Systems (NIPS'05, Vancouver).