RC24756 (W0903-020) March 5, 2009
Computer Science

IBM Research Report

Learning Curves in Machine Learning

Claudia Perlich
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Learning Curves in Machine Learning

Claudia Perlich, IBM T.J. Watson Research Center

Synonyms

Experience curve, Improvement curve, Error curve, Training curve

Definition

A learning curve shows a measure of predictive performance on a given domain as a function of some measure of varying amounts of learning effort. The most common form of learning curves in the general field of machine learning shows predictive accuracy on the test examples as a function of the number of training examples as in Figure 1.

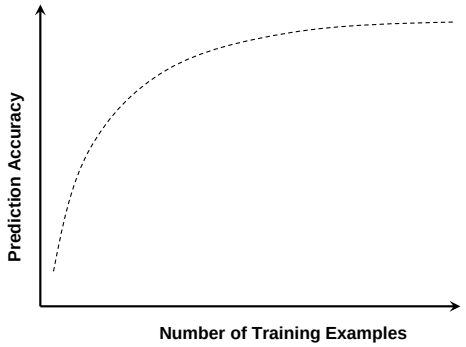

Figure 1: Stylized learning curve showing the model accuracy on test examples as function of the number of training examples.

Background

Learning curves were initially introduced in educational and behavioral/cognitive psychology. The first person to describe the learning curve was Hermann Ebbinghaus in 1885(6). He found that the time required to memorize a nonsense syllable increased sharply as the number of syllables increased. In 1936,

Theodore Paul Wright (7) described the effect of learning on labor productivity in the aircraft industry and proposed a mathematical model of the learning curve. Over time, the term has acquired related interpretation in many different fields including the above definition in machine learning and statistics.

Use of Learning Curves in Machine Learning

In the area of machine learning, the term "learning curve" is used in two different contexts which determined mostly the variable on the x-axis of the curve:

- The Artificial Neural Network literature has used the term to show the diverging behavior of in and out-of-sample performance as a function of the *number of training iterations* for a given number of training examples. Figure 2 shows this stylized effect.
- General Machine Learning uses learning curves to show the predictive generalization performance as a function of the *number of training examples*. Both graphs in Figure 3 are examples of such learning curves.

Artificial Neural Networks

The origins of artificial neural networks are heavily inspired by the social sciences and the goal of recreating the learning behavior of the brain. The original model of the 'perceptron' mirrored closely the biological foundations of neural sciences. It is likely that the notion of learning curves was to some extent carried over from the social sciences of human learning into the field of artificial neural networks. It shows the model error as a function of the training time measured in terms of the number of iterations. One iteration denotes in the context of neural network learning one single pass over the training data and the corresponding update of the network parameters (also called weights). The algorithm uses gradient descent minimizing the model error on the training data.

The learning curve in Figure 2 shows the stylized effect of the relative training and generalization error on a test set as a function of the number of iterations. After initial decrease of both types of error, the generalization error reaches a minimum and starts to increase again while the training error continues to decrease.

This effect of increasing generalization error is closely related to the more general machine learning issue of overfitting and variance error for models with high expressive power (or capacity). One of the initial solutions to this problem for neural networks was early stopping - some form of early regularization technique that picked the model at the minimum of the error curve on a validation subset of the data that was not used for training.

General Machine Learning

In the more general machine learning setting and statistics (3), learning curves represent the generalization performance of the model as a function of the size

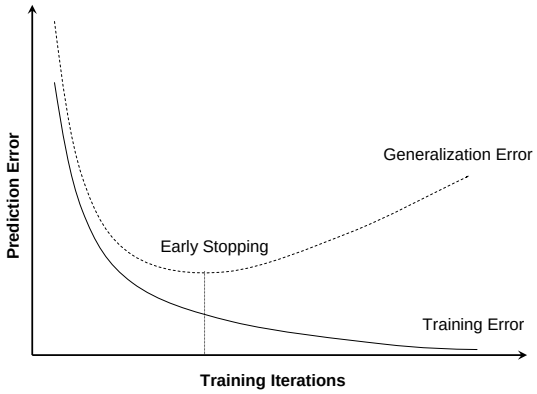

Figure 2: Learning curve for an artificial neural network.

of the training set.

Figure 3 was taken from (4) and shows two typical learning curves for two different modeling algorithms (decision tree and logistic regression) on a fairly large domain. For smaller training–set sizes the curves are steep, but the increase in accuracy lessens for larger training–set sizes. Often for very large training–set sizes the standard representation in the upper graph obscures small, but non–trivial, gains. Therefore to visualize the curves it is often useful to use a log scale on the horizontal axis and start the graph at the accuracy of the smallest training–set size (rather than at zero). In addition, one can include error bars that capture the estimated variance of the error over multiple experiments and provides some impression of the relevance of the differences between two learning curves as shown in the graphs.

The figure also highlights a very important issue in comparative analysis of different modeling techniques: learning curves for the same domain and different models can cross. This implies an important pitfall as pointed out by Langley(1): 'Typical empirical papers report results on training sets of fixed size, which tells one nothing about how the methods would fare given more or less data, rather than collecting learning curves ...'. A corollary on the above observation is the dangers of selecting an algorithm on a smaller subset of the ultimately available training data either in the context of a proof of concept pre-study or some form of cross-validation.

Aside from its empirical relevance there has been significant theoretical work on learning curves - notably by Cortes(5). She is addressing the question of predicting the expected generalization error from the training error of a model. Her analysis provided many additional insights about the generalization performance of different models as a function of not only training size but in addition the model capacity.

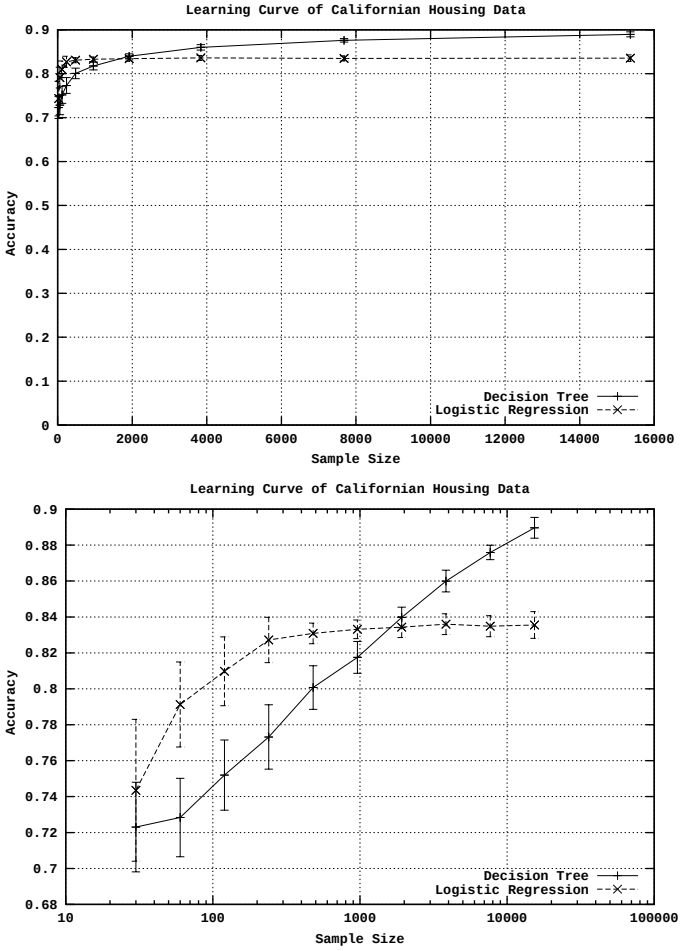

Figure 3: Typical learning curves in original and log scale.

Cross References

Computational Learning Theory, Overfitting, Artificial Neural Networks, Generalization Performance, Decision Tree, Logistic Regression

References

[1] Kibler, D. and Langley, P. (1988), "Machine Learning as an Experimental Science," *Proceedings of the Third European Working Session on Learning*, Pittman, Glasgow, 81–92.

[2] Shavlik, J.W., Mooney, R.J., and Towell, G.G. (1991), "Symbolic and Neural Learning Algorithms: An Experimental Comparison", *Machine Learning*, **6**, 111–143.

[3] Flury, B.W. and Schmid, M.J. (1994), "Error Rates in Quadratic Discrimination with Constraints on the Covariance Matrices," *Journal of Classification*, **11**, 101–120.

[4] Perlich, C., Provost, F. and Simonoff, J., (2003), "Tree Induction vs. Logistic Regression: A Learning-curve Analysis", *Journal of Machine Learning Research*, **4**, 211–255.

[5] Cortes, C., Jackel, L.D., Solla, S.A., Vapnik, V., and Denker, J.S., (1994), "Learning curves: Asymptotic values and rate of convergence", *Advances in Neural Information Processing Systems*,6: 327–334.

[6] Wozniak, R. H. (1999). "Introduction to memory: Hermann Ebbinghaus (1885/1913)". *Classics in the history of psychology.*

[7] Wright, T.P. (1936) , "Factors Affecting the Cost of Airplanes", *Journal of Aeronautical Sciences*, **3(4)**: 122-128.

Definitions of key terms used above

Artificial Neural Networks: is a computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.

Generalization performance: is a measurement of the performance (accuracy or error) of the model prediction on test examples that were in no way involved in the model estimation or selection.

Overfitting is traditionally defined as training some flexible representation so that it memorizes the data but fails to predict well in the future.