Expected Error Analysis for Model Selection*

Tobias Scheffer
Technische Universität Berlin
FR 5-8, Franklinstr. 28/29
10587 Berlin, Germany
scheffer@cs.tu-berlin.de

Thorsten Joachims
Universität Dortmund
LS VIII, Baroper Str. 301
44221 Dortmund, Germany
joachims@ls8.cs.uni-dortmund.de

Abstract

In order to select a good hypothesis language (or *model*) from a collection of possible models, one has to assess the generalization performance of the hypothesis which is returned by a learner that is bound to use that model. This paper deals with a new and very efficient way of assessing this generalization performance. We present a new analysis which characterizes the expected generalization error of the hypothesis with least training error in terms of the distribution of error rates of the hypotheses in the model. This distribution can be estimated very efficiently from the data which immediately leads to an efficient model selection algorithm. The analysis predicts learning curves with a very high precision and thus contributes to a better understanding of why and when over-fitting occurs. We present empirical studies (controlled experiments on Boolean decision trees and a large-scale text categorization problem) which show that the model selection algorithm leads to error rates which are often as low as those obtained by 10-fold cross validation (sometimes even lower). However, the algorithm is much more efficient (because the learner does not have to be invoked at all) and thus solves model selection problems with as many as thousand relevant attributes and 12,000 examples.

Proceedings of the Sixteenth International Conference on Machine Learning, 1999

1 Introduction

Consider the following situation: In order to solve a learning problem we need to choose a hypothesis language from a set of possible languages (or models) – decision trees of variable depth perhaps, or a neural network with a variable number of hidden neurons. Our goal is to select the model which gives us the highest generalization performance. If we choose too simple a model (a tree of depth one say) even the best hypothesis in that model is likely to incur both a high error on the training data and a high true error. On the other hand, if we choose too rich a model (*e.g.,* a neural network with a hundred hidden units) our hypothesis might be poor due to over-fitting. The task of choosing a model such that the error of the hypothesis returned by a learner (which uses this model) is low is referred to as *model selection*. Approaches to model selection fall into the categories hold-out testing, complexity penalization, and Bayesian learning. See [17] or [7] for a more detailed overview.

Hold-out testing algorithms stratify the potential hypothesis language H into subsets ("models") $H_1, H_2, \ldots \subset H$. Starting with the smallest model, the learning algorithm returns one hypothesis from each model. The hold-out set (which has not been used for learning) is then used to obtain an estimate of the expected generalization error; the model with the lowest estimate is selected and the learner is invoked for this model with the whole sample. In order to minimize the variance of the estimate the idea of n-fold cross validation, *e.g.,* [10, 19], and bootstrapping [5] is to average many error measurements which are generated on re-sampled data sets. For many applications, n-fold cross validation works quite well, but it does not scale well for very large-scale problems as the learner has to be invoked about at least once for each model (*e.g.,* [8]).

Complexity penalization based methods minimize a demerit criterion which consists of the empirical error of the hypothesis and an additional term that penalizes the complexity of the model which the hypothesis came from. Effectively, complexity penalization algorithms try to reconstruct the learning curve (*i.e.,* the dependency between model index and generalization error) from only the empirical error rate and some measure of the model complexity. Unfortunately, the learning curve can exhibit various shapes [18] and therefore no single complexity penalization scheme can perform well for all model selection problems. For each complexity penalization algorithm one can construct a pair of model selection problems such that the algorithm performs well for one problem and fails (*i.e.,* incurs an additional error λ that does not vanish when the sample size grows) for the other one [7]. Therefore, all practical learners which conduct complexity penalization (*e.g.,* the Support Vector Machine [20], decision tree pruning algorithms [9], neural weight decay algorithms [3]) possess a regularization parameter that can be adapted for the given problem (*e.g.,* by cross validation).

Bayesian learners [1] focus on the posterior probability of a function f having generated the sample S: $P(f|S)$. Under certain ideal conditions, one can, under high computational effort, derive the Bayes hypothesis from $P(f|S)$ which is guaranteed to have the least generalization error. The Bayes hypothesis is a weighted majority (where $P(f|S)$ defines the weights) over all hypotheses. Usually, the posterior is used for less expensive heuristics such as MAP (the *maximum a posteriori* hypothesis maximizes the chance of having generated the data) or MDL [11] (the MDL hypotheses minimizes the description length required for the data by compressing it to a hypothesis and the exceptions to the hypothesis in the data). Using Bayes' rule, the posterior $P(f|S)$ works out to $P(S|f)P(f)/P(S)$. $P(S)$ is constant for a problem (sometimes, $P(S)$ can be derived from $P(S|f)$ and $P(f)$), and $P(S|f)$ can usually be determined. However, the prior distribution $P(f)$ is assumed to be known in advance – which is indeed a very strong assumption. The general belief is that Bayesian learners are fairly robust against some degree of misalignment between the actual and the assumed $P(f)$ (this setting is referred to as robust Bayesian analysis [2]). The No-Free-Lunch Theorems [23] explain that Bayesian learners perform better than randomly guessing only if the actual and the assumed prior are "not completely unaligned".

This paper's scope. Suppose that we have to solve some learning problem for which two possible models are available. Model H_1 contains just one single hypothesis h_1 while model H_2 contains two hypotheses, h_{21} and h_{22}. All three hypotheses incur (unknown) generalization errors. When we draw a sample, each hypothesis exhibits an empirical error which is an *unbiased* estimate of its true error. Unbiased means that the expected empirical error of each hypothesis is just its generalization error. The relation between true and empirical error is known: the empirical error is governed by the binomial distribution with the true error as its mean value. Now suppose that we minimize the empirical error in H_2; let h_2^* be the hypothesis in H_2 with the least empirical error (H_1 contains only one hypothesis, so minimizing the empirical error in H_1 is trivial). When both hypotheses in H_2 incur equal empirical errors, we draw one at random. Unfortunately, the empirical error of h_2^* is not an unbiased estimate of its true error. Why is that? With a chance of almost $\frac{1}{2}$, the empirical error of each hypothesis is an optimistic estimate of its true error, and with a chance of almost $\frac{1}{2}$ it is a pessimistic estimate. An optimistically assessed hypothesis has a greater chance of being selected as h_2^* than a pessimistically assessed one. In return, the empirical error of h_2^* is, on average, optimistically biased. So, what should we do when the empirical error of h_2^* is less than the empirical error of h_1? The empirical error of h_1 is unbiased while the empirical error of h_2^* (which is lower) is known to be optimistic. Here, the question is how strong this bias is. In this paper, we present an answer to this question. We will see that the expected error of the hypothesis which minimizes the empirical error in model H_2 depends on the distribution of error values in that model. The distribution of error values of all hypotheses in H_2 contains (at most) two occurring error values; however, the true error rates are unknown. But we can estimate the distribution of true error rates by recording the distribution of empirical error rates in H_2. The distribution of empirical error values in H_2 is, in this case, very simple because only (at most) two distinct values occur which can be observed. When these two error values are known, this suffices to determine an estimate of the expected generalization error of h_2^*. The most interesting aspect of our solution is that no learning has to be conducted in order to determine the error of the hypothesis that would be the result of learning.

2 Preliminaries

In this paper, we focus on classification learning from labeled examples where the target criterion is the expected zero-one loss.

Instances. We assume that there is a set of *instances* X and a finite set of *class labels* Y. A classification problem is defined by an unknown distribution $D_{XY} = D_{Y|X} D_X$ over labeled instances $(X \times Y)$, which we want to approximate as closely as possible. Sometimes, when this is more convenient, we speak of target functions. Note that target distributions can "emulate" target functions ($D_{Y|X}(y|x) = 1$ iff $y = f(x)$, 0 otherwise).

Hypotheses and Error. A *hypothesis* $h : X \rightarrow Y$ is a mapping from instances to class labels. The *true (or generalization) error of a hypothesis*, with respect to the (unknown) distribution D_{XY} is $E_D(h) = \int_{(x,y)\in X\times Y} \ell(h(x), y) dD_{XY}(x, y)$, where ℓ is the zero-one loss function. Let the *sample* S be a sequence of m labeled instances drawn *independently and identically distributed* according to D_{XY}. We then define the *empirical or observed error* as $E_S(h) = \frac{1}{m}\sum_{(x,y)\in S} \ell(h(x), y)$.

Hypothesis Language and Model. We assume the existence of a given hypothesis language H which may be infinite and may even have an infinite VC-dimension. A *stratification* of the hypothesis language is a finite sequence of models $\langle H_1, \ldots, H_m\rangle$, $H_i \subseteq H$. We do not assume the models to be properly nested, but we require the models H_i to be *finite* subsets of H.

Learner. A *learner* takes as input a sample S and a model H_i and determines the set $H_i^*(S) = \{h \in H_i : E_S(h) = \min_{h'\in H_i}(E_S(h'))\}$ of hypotheses with least empirical error. There is at least one such hypothesis. If there is more than one, we assume that the learner draws one at random under uniform distribution.

Notations. We generally write probability distributions and densities in the form $P_{\{x\}}(f(x) = y)$ where the subscript x indicates that x is a random variable. The distribution of x should become clear in the given context. $P_{\{x\}}(f(x) = y)$ refers to the chance of drawing an x such that $f(x) = y$. Similarly, we write $\mathbf{E}_{\{x\}}(f(x))$ for the expectation of $f(x)$ over all x (again, the distribution of x becomes clear in the context). We write the binomial distribution as $B[p, n](x)$, denoting the probability of making x mistakes on n trials when the chance of a mistake is p.

3 Expected Error Analysis

Let us look at a model H_i and a target distribution D_{XY}. The target D_{XY} defines an error $E_D(h)$ for each hypothesis $h \in H_i$. These error values define a distribution of error values in H_i, which we write as $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ and which is the "prior" in our analysis. (Here, "prior" means prior to observing the sample and minimizing the empirical error.) $P_{\{h\}}(E_D(h) = e_D|H_i, D_{XY})$ is the chance of drawing a hypothesis h from H_i (when drawing under uniform distribution) which incurs an error of e_D.

Suppose that H_i contains two hypotheses (as an easy example). The prior $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ tells us which error values occur in H_i. There are either two values with a chance of $\frac{1}{2}$ or one value with a chance of 1 (if both hypotheses have equal errors). Let us invent names h_1 and h_2 for the hypotheses and let $E_D(h_1)$ and $E_D(h_2)$ be the two occurring true error values. When we will draw a sample S, the hypotheses will show empirical error values of $E_S(h_1)$ and $E_S(h_2)$, respectively. The empirical errors on the sample S of size m are distributed according to $B[E_D(h_1), m]$ and $B[E_D(h_2), m]$, where B is the binomial distribution. (Each example is classified correctly or wrongly, the chance of a wrong answer being $E_D(h_1)$ and $E_D(h_2)$, respectively. This results in a binomial distribution.) Let us now select the hypothesis with the smaller empirical error, call it h_L. The chance that h_L has a particular error value e_D is now no longer $P_{\{h\}}(E_D(h) = e_D|H_i, D_{XY})$, because h_L is not a randomly drawn hypothesis. It is, instead, the hypothesis which minimizes the empirical error. The expected true error of h_L is likely to be greater than its empirical error (which is optimistically biased) but less than the error of a randomly drawn hypothesis. But precisely what is the expected error of h_L? Our analysis characterizes the expected error of h_L (the distribution of error values of h_L is our "posterior") in terms of the prior $P_{\{h\}}(E_D(h)|H_i, D_{XY})$. And, as we will see, the prior can be estimated from a sample S by recording $P_{\{h\}}(E_S(h)|H_i, S)$, the empirical counterpart of $P_{\{h\}}(E_D(h)|H_i, D_{XY})$. This is the principle idea of the expected error analysis which we will discuss in the following.

Since the prior tells us how many hypotheses incur which error values, we can assign names h_1 through $h_{|H_i|}$ to the occurring error values $E_D(h_1)$ through $E_D(h_{|H_i|})$. If, for example, $P_{\{h_i\}}(E_D(h) = 0|H_i, D_{XY})$ equals $\frac{2}{|H_i|}$, we know that there are exactly two hypotheses with true error of zero. Although we do not know *which* hypotheses these are, we can name them, for instance, h_1 and h_2. Now we can claim that $E_D(h_1) = E_D(h_2) = 0$. But note that assigning hypothesis names to the known error values is only a notational "trick" that will make life easier for us during the main proof; we do not actually know (and our derivation does not exploit) *which* individual hypothesis incurs a particular error value. Now the following happens. On a sample S of size m, each hypothesis incurs an empirical error which is binomially distributed: $P_S(E_S(h) = e_S|E_D(h), m) = B[E_D(h), m](e_S)$. The learner first determines the set $H_i^*(S)$ of hypotheses with least empirical error. Then the learner returns a hypothesis h_L from $H_i^*(S)$ (the set of hypotheses with least empirical error), breaking ties by drawing randomly under uniform distribution. Assume that the sample size m is fixed and given *a priori*. By contrast, the sample S itself is a random variable, distributed according to $(D_{XY})^m$. This implies that $H_i^*(S)$ is a random variable (as it depends on S) and so is h_L; h_L is drawn randomly from $H_i^*(S)$. This leads us to the posterior distribution $P_{\{S,h_L\}}(E_D(h_L) = e_D|H_i, D_{XY}, m, h_L \in H_i^*(S))$ which is the chance of drawing a sample S (of fixed size m) and, consequently, a hypothesis h_L from $H_i^*(S)$, such that the true error of h_L is e_D. The principle difference between the prior and posterior distribution is that the prior gives the distribution of error values of hypotheses which are drawn uniformly from H_i, whereas the posterior gives the distribution of error values for hypotheses which have been generated by an error minimization process. Knowing the posterior $P_{\{S,h_L\}}(E_D(h_L) = e_D|H_i, D_{XY}, m, h_L \in H_i^*(S))$ we can easily derive the expectation $\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, D_{XY}, m, h_L \in H_i^*(S))$. It is the expected true error of h_L (the hypothesis returned by the learner) using model H_i and a sample of size m. This expectation is a natural (and often used [21, 19]) measure of the quality of a model and will serve as the model selection criterion in our study.

We quantify the expected true error of h_L, $\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, D_{XY}, m, h_L \in H_i^*(S))$ in Theorem 1. The crucial part of the proof is how to determine the minimum error e_S and the number of hypotheses $|H_i^*(S)|$ which achieve this error. The idea is that we can determine the chance that $H_i^*(S)$ is a particular subset H^* by factorizing the least error e_S and calculating the chances that each hypothesis in H^* has an empirical error of e_S and each hypothesis outside incurs a strictly greater error. This, however, imposes another difficulty. We know that the empirical error of a hypothesis is distributed binomially, given the true error. But here we have to determine the chance of *several hypotheses* incurring a certain empirical error e_S. Unfortunately, this probability cannot be determined unless we assume the empirical errors of distinct hypotheses to be independent (in reality, the empirical errors are somewhat dependent because they are measured on the same sample). The empirical error rates depend on the corresponding true error rates and are not identically distributed. We do not make any assumptions on the true error rates. This independence assumption is often made implicitly; for instance, the calculation of p-values which is required to compare n-fold cross validation results (the p-value gives the chance that one learner does better than another learner for some problem, given the cross validation results) is based on the assumptions that the hold-out errors are independent estimates.

Assumption 1 (Independence Assumption) *The empirical errors $E_S(h)$ of hypotheses $h \in H$ are independent given the corresponding true errors $E_D(h)$. $P(E_S(h_1), \ldots, E_S(h_{|H_i|})|H_i,\ D_{XY},\ m,\ E_D(h_1), \ldots, E_D(h_{|H_i|})) = \prod_{j=1}^{|H_i|} P(E_S(h_j)|H_i, D_{XY}, m, E_D(h_j))$.*

Theorem 1 (Expected error of an empirical error minimizing hypothesis) *For a distribution D_{XY} of labeled instances and a finite model H_i let $E_D(h)$ be the true error of each hypothesis $h \in H$. Let m be the fixed sample size. Let h_L be a hypothesis which is drawn uniformly from the set $H_i^*(S)$ of hypotheses in H_i with least empirical error with respect to a sample S, drawn according to $(D_{XY})^m$. Then, under the independence assumption, the expected error of h_L is*

$$\begin{aligned}
&\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, D_{XY}, m, h_L \in H_i^*(S)) \\
&\quad = \sum_{j=1}^{|H_i|} E_D(h_j) \sum_{e_S} B[E_D(h_j), m](e_S) \sum_{m=1}^{|H_i|} \frac{1}{n} \qquad (1)\\
&\quad \sum_{\substack{H^* \subseteq H_i \setminus \{h_j\} \\ |H^*| = n-1}} \prod_{h^* \in H^*} P_{\{S\}}(E_S(h^*) = e_S|E_D(h^*), m) \\
&\quad \prod_{h \in H \setminus \{h_j\} \setminus H^*} P_{\{S\}}(E_S(h) > e_S|E_D(h), m)
\end{aligned}$$

The proof is given in Appendix A. Equation 1 can, in principle, be evaluated, given the distribution of true error values $P_{\{h\}}(E_D(h)|H_i, D_{XY})$, $|H_i|$, and the sample size m. Later in this Section, we will see how the prior $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ can be estimated efficiently when only H_i and a sample are known. We then have a means of estimating the expected generalization error of a hypothesis which minimizes the empirical error without actually having to invoke a learner. A straightforward implementation of Theorem 1 would run in time exponential in $|H_i|$ which is clearly unacceptable. Therefore, we make the technical assumption that removing one single hypothesis from H_i does not considerably alter the distribution of error values.

Assumption 2 *We assume that* $P(|H^*| = n | h_i \in H^*, E_S(h_i) = e_{min}, H_i, m, D_{XY}) = P(|H^*| = n | h_j \in H^*, E_S(h_j) = e_{min}, H_i, m, D_{XY})$.

Assumption 2 means that the chance of the set of hypotheses with least empirical error being of size m when it is known that a hypothesis h_i with empirical error e_{min} belongs to this set is not dependent on *which* hypothesis with empirical error e_{min} is known to be in this set. This is always true when H_i is "large". This assumption is reasonable in all practical cases as $|H_i|$ grows doubly exponential for Boolean functions and at least singly exponential for languages such as monomials.

Theorem 2 (Efficient evaluation of Theorem 1) *For a distribution D_{XY} of labeled instances and a finite model H_i, let $E_D(h)$ be the true error of each hypothesis $h \in H$. Under assumptions* 1 *and* 2*, the expected error of the hypothesis h_L returned by the learner given an* i.i.d. *sample S drawn according to D_{XY} is*

$$\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, m, D_{XY}, h_L \in H_i^*(S)) = \frac{\int e_D P_{\{h\}}(E_D(h) = e_D|H_i, D_{XY})}{\int P_{\{h\}}(E_D(h) = e_D|H_i, D_{XY})} \frac{dP_{\{S\}}(h_{eD} \in H_i^*(S)|H_i, m, E_D(h_{e_D}))}{dP_{\{S\}}(h_{e_D} \in H_i^*(S)|H_i, m, E_D(h_{e_D}))} \quad (2)$$

where

$$P_{\{S\}}(h_{eD} \in H_i^*(S)|H_i, m, E_D(h_{e_D})) = \sum_{e_S} B[e_D, m](e_S) \prod_{e'_D} \left(\sum_{e \geq e_S} B[e'_D, m](e) \right)^{f(e_D, e'_D)} \quad (3)$$

$$f(e_D, e'_D) = \begin{cases} |H_i| P_{\{h\}}(E_D(h) = e'_D | H_i, D_{XY}) & \text{iff } e_D \neq e'_D \\ |H_i| P_{\{h\}}(E_D(h) = e'_D | H_i, D_{XY}) - 1 & \text{iff } e_D = e'_D \end{cases}$$

and h_{eD} is an arbitrary hypothesis with true error $E_D(h_{eD}) = e_D$.

The proof can be found in Appendix B. Note that the only input to Theorem 2 are the error prior, $|H_i|$, and m. Theorem 2 (which effectively replaces Theorem 1) solves the primary complexity problem by removing the product over all subsets of H_i from Equation 2. A careful implementation of the formula (see the full paper [17] for details) now runs in $O(m^2)$.

Estimating $P_{\{h\}}(E_D(h)|H_i, D_{XY})$. As $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ depends on D_{XY}, it cannot be determined exactly. All information on D_{XY} which we can access is contained in S. We can use S to measure $P_{\{h\}}(E_S(h)|H_i, S)$, which will serve as an estimate of $P_{\{h\}}(E_D(h)|H_i, D_{XY})$. As m grows, $P_{\{h\}}(E_S(h)|H_i, S)$ converges towards $P_{\{h\}}(E_D(h)|H_i, D_{XY})$; so it is a consistent estimate and for a reasonably large S we can hope to obtain a good estimate of $P_{\{h\}}(E_D(h)|H_i, D_{XY})$. In fact, the experiments presented in the following sections show that even samples of size 50 allow for fairly accurate estimates. Note that this is a one-dimensional distribution only; the dimensionality does not increase when H_i grows. $P_{\{h\}}(E_S(h)|H_i, S)$ can be estimated by drawing a small number of hypotheses uniformly from H_i and recording the empirical error values.

Algorithm. We now have an efficient model selection algorithm: For each model (a) record $P_{\{h\}}(E_S(h)|H_i, S)$ by drawing a small number of hypotheses and use it as an estimate of $P_{\{h\}}(E_D(h)|H_i, D_{XY})$, (b) use Theorem 2 to determine the expected generalization error. (c) Select the model with the lowest estimated expected generalization error. Copies of the implementation can be obtained from the authors upon request.

4 What is Over-Fitting?

A learning curve is a function which maps a model index i to the error of the hypothesis which is generated by a given learner on model H_i for a given learning problem. Often, the error increases for large models which is generally referred to as over-fitting. This is frequently considered to be due to the high hypothesis complexity in models with high indeces. However, over-fitting does not necessarily occur in all settings. Boosting algorithms have often exhibited a complementary behavior [14]; experiments by Schaffer [13] support the view that whether over-fitting occurs at all is a property of the problem.

Using Chernoff bounds, one can guarantee that (with high probability) the difference between true and empirical error of *no* hypothesis in some model H_i exceeds a certain threshold. This immediately leads to worst-case error bounds. This is the way that PAC and VC theory argue. The empirical error is binomially distributed, so even a poor hypothesis has a small chance (depending on the sample size) of exhibiting a low empirical error. When H_i grows, the chance of *some* hypothesis in H_i having a large difference between true and empirical error grows steeply. Therefore, given two hypotheses with equal empirical error which come from distinct models, PAC theory gives better guarantees for the one which comes from the smaller model. But just because *there is* a hypothesis with a large difference between true and empirical error does not mean that the expected error of the returned hypothesis is (absolutely) high. In fact, from our expected error analysis we can derive that, when the prior distribution of error values in the model remains constant, the expected error of the returned hypothesis *converges from above* as H_i grows.

Let us look at the expected error of the returned hypothesis h_L when the model size, $|H_i|$, approaches infinity and $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ stays constant.

Theorem 3 *When* $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ *is constant, the expected error of a randomly drawn empirical minimizer* h_L *converges as* $|H_i|$ *grows.*

$$\lim_{|H_i|\to\infty} \mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, m, D_{XY}, h_L \in H_i^*(S)) \qquad (4)$$

$$= \frac{\int_{e_D} e_D \times (1-e_D)^m dP_{\{h\}}(E_D(h) = e_D|H_i, D_{XY})}{\int_{e_D} (1-e_D)^m dP_{\{h\}}(E_D(h) = e_D|H_i, D_{XY})}$$

The proof can be found in [17]. Note that Theorem 3 is subject to the assumption that $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ remains fixed while H_i grows. Let us now study how $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ actually behaves when we want to learn Boolean functions under uniform distribution of the Boolean instances. In this case, the prior can be determined analytically. When the target function uses attributes x_1 through x_n and the model H_i consists of hypotheses over attributes x_1 through x_i, then the prior is a certain binomial distribution depending on i and n (function and hypothesis must agree on all possible $2^{\max\{i,n\}}$ instances which can be distinguished by target function or hypothesis). By plugging the exact prior into Theorem 1 we can determine the learning curve analytically. Figure 1 shows some prior distributions of error values $P_{\{h,D_{XY}\}}(E_D(h)|H_i)$ (D_{XY} is now a random variable as we draw Boolean functions at random) for various models when the target function is over three attributes. Note that models H_1 through H_3 have identical error priors; the increase in the model size leads to a decrease in error. When irrelevant attributes are added (models H_4 through H_6) the tails of the distribution become skinnier – intuitively, the concentration of really good (and really bad) hypotheses decreases and more hypotheses incur an error of close to $\frac{1}{2}$. This causes the learning curve to rise; Figures 2 and 3 show some predicted learning curves (for various sample sizes) and compare them to learning curves measured in a simulation (the predicted curves are the ones labeled "predicted learning curve, exact prior"). The "simulation" curve shows the error measured in an experiment, averaged over 200 randomly drawn target functions. The deviation originates from three sources: The simulation curve is only measured in an experiment and hence subject to some inaccuracy, the independence assumption 1 on the empirical error causes a modest bias, and the simplification 2 on which the implementation is based incurs a very small error. However, the learning curve is still predicted very accurately. While PAC theory predicts learning curves very poorly for non worst-case problems, our analysis incorporates a joint property of the hypothesis language and the problem – namely $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ – and is therefore able to explain learning curves with high accuracy.

5 Experiments

In our experiments, we study the accuracy and variance of the predicted error (in comparison to 10-fold cross validation).

Learning Boolean decision trees. In this set of experiments, we used randomly drawn Boolean functions as targets. Figures 2 and 3 compare some predicted learning curves (where the prior has been estimated from the sample) to the average of 200 learning curves measured empirically. As we expected, the quality of

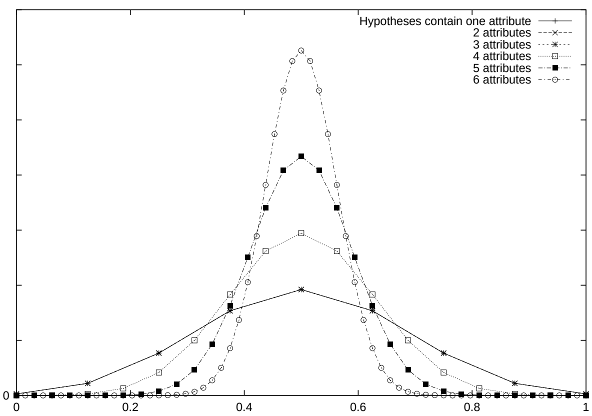
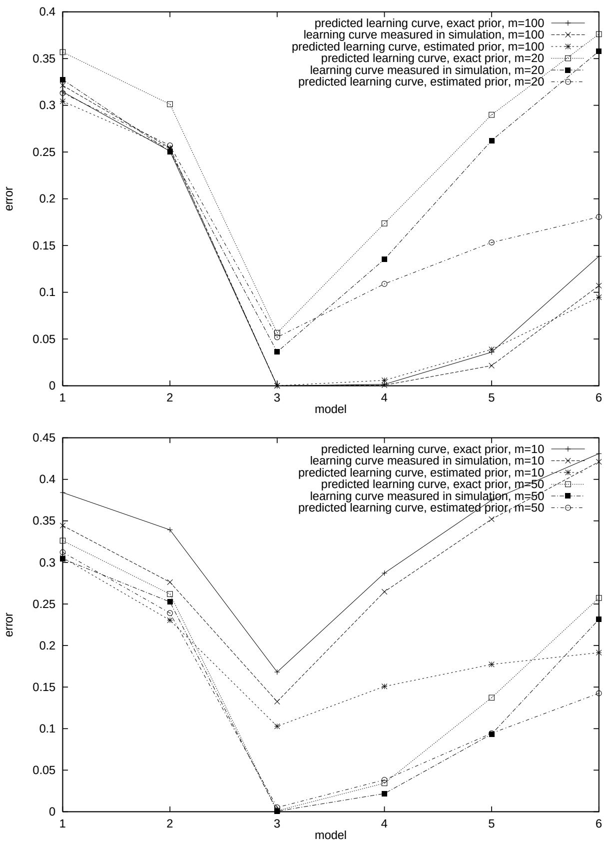


Figure 1 (horizontral axis: error rate; vertical axis: frequency of hypotheses in H_i which incur that error rate): Various shapes of $P_{\{h,D_{XY}\}}(E_D(h)|H_i, D_{XY})$

curves for models which consist of i Boolean attributes when the target function is a randomly drawn function over attributes x_1, x_2, x_3.



Figures 2 and 3: Averaged learning curves for 200 randomly drawn Boolean target functions. The target functions have 3 attributes (X-axis: number of attributes i in the model).

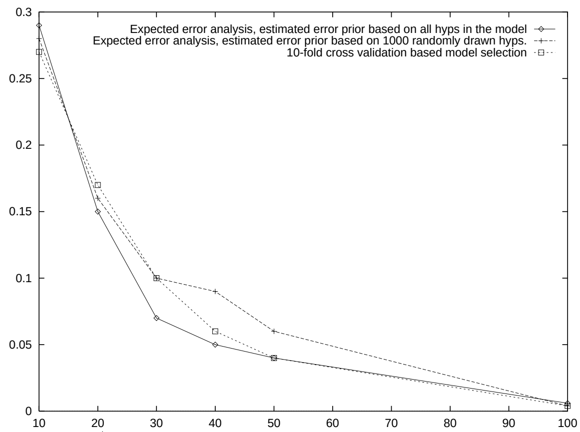


Figure 4 (horizontal axis: sample size, vertical axis: error): True error of hypothesis returned by model selection based learning. Average over 200 randomly drawn target functions with i attributes (i distributed uniformly).

the prediction depends crucially on how well the prior has been estimated. Sample sizes of 10 and 20 lead to unsatisfactory results; when the sample size is 50, the estimated curve approaches the measured curve and for $m = 100$ the predicted curve is a fair estimate of the measured curve. We now want to compare this means of error estimation to 10-fold cross validation. We use the following setting: First, the depth of the target tree is drawn uniformly between 2 and 6 attributes, the target tree is then determined at random with the chosen number of attributes. Next the sample is drawn. We then compare a 10-fold cross validation based model selection algorithm to expected error analysis based model selection. In both cases, models H_1 through H_6 are available where H_i contains all decision trees over i attributes. The cross validation based algorithm uses the averaged hold-out errors to select a model while the expected error analysis based algorithm makes a decision on grounds of the predicted generalization error for each model. After the model is selected, a learner is invoked which uses the whole sample and the true error of the resulting hypothesis is determined. The true error can be determined because the target function is known (but unknown to the learner). Figure 4 shows the results; each point is averaged over 200 randomly drawn target functions. The resulting error decreases, of course, with growing sample size. For a sample size of 30, expected error analysis does significantly better than 10-fold cross validation (p-value is .002); for $m = 40$ cross validation does better than the expected error analysis when the prior is only estimated by drawing 1000 hypotheses. The error rate achieved by expected error analysis as at least comparable to cross validation based model selection. The principle advantage of our expected error analysis is that the estimate is obtained in an extremely efficient manner since no learning has to be done.

Scaling up: text categorization. Now we want to demonstrate that the expected error analysis easily scales up to large learning problems with as many as thousand relevant attributes. Text categorization is the problem of mapping texts to semantic categories. Interesting applications of this are classification of newspaper articles and classification of web pages. Documents are represented as a vector of the word stems occurring in them (*i.e.,* the word ordering is ignored). We used the Reuters-21578 corpus of newspaper articles. It contains roughly 12,000 documents with 10,000 features and we used the ten most frequent categories. Experiments by Joachims [6] have shown that decision trees learned on this corpus are highly unbalanced. So we decided to use 2-DL (decision lists with two literals per conjunction) [12] as hypothesis language. Training was done with a very efficient greedy learner (based on the coverage approach) using an inverse indexing technique to determine empirical errors quickly. Assessing a model by means of the expected error analysis requires approximately 2 minutes on a PC while running hold-out testing for one fixed model and one category takes about 2-3 hours. As we wanted to compare the predicted error rates to cross validation error rates, we had to settle for a small number of models. Initial experiments indicated that almost all attributes are relevant. So we used 7 models where H_i contains decision lists with monomials that cover at least $1, 10, 20, 30, 50, 70$, and 100 examples, respectively. The parameter which is adapted here influences the model by only allowing monomials which are supported by a certain sample size. One could think of this number of examples which have to be covered by each monomial as a pruning threshold which is adapted. We refer the reader to the full paper for a more detailed discussion of the experimental settings. Figure 5 shows the predicted error rates and the error rates estimated by one-fold cross validation, averaged over the ten most frequent categories. The predicted values are subject to a pessimistic bias of .03 (which we already observed in the previous experiments), but the shapes of the learning curves are fairly similar. By means of the expected error analysis we can obtain a fair error estimate for a large number of models while hold-out testing (or even cross validation) is unfeasible for all but an extremely small number of models.

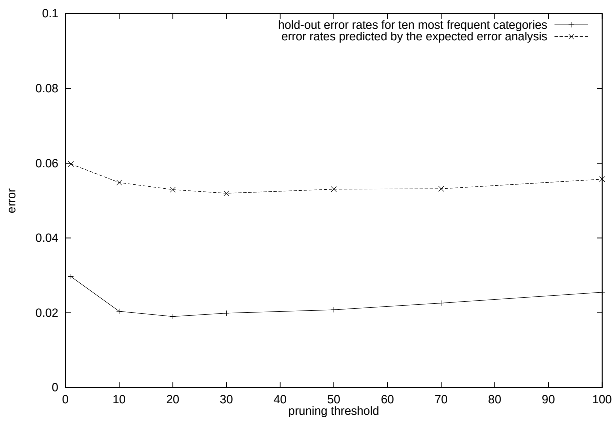


Figure 5: Learning curves (one-fold cross validation and expected error analysis for the text categorization problem), averaged over 10 categories.

6 Discussion

Many attempts have been made to find an efficient means of assessing models. The potential benefit of analyzing $P_{\{S\}}(E_D(h_L)|H_i, D_{XY}, m, L)$ has been discussed by Wolpert [22]. Scheffer and Joachims [15, 16] have found a solution which is based on two independence assumptions, one of which is rather strong. Domingos [4] presented a solution which is based on several very strong assumption (*e.g.,* it is assumed that the hypotheses which are contained in one model have equal true errors[1]). In this paper, we presented a very general solution for finite models which is only based on the relatively mild assumption that the empirical error rates of distinct hypotheses are independent given the corresponding true error rates (and no assumption is made on the true error rates). We presented experiments which showed that, when the prior distribution of error values $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ is known (as is the case when the target is a Boolean function and the instances are uniformly distributed), our analysis predicts and thus explains learning curves with a very high precision. When the error prior is estimated from the sample, the error estimate is poor when the sample is too small to estimate the error distribution of h well. In our experiments estimates become reasonably accurate for sample sizes of 50 or above. We observed a small pessimistic bias which is caused by Assumption 1, but this bias was almost independent of the model. Note that an almost constant bias does not hurt too much when one wants to know which of several hypotheses is the best one (as opposed to when one wants to know the precise error of a particular hypothesis). Therefore, for the Boolean decision tree problem, the expected error analysis based model selection algorithm performed at least as good as 10-fold cross validation (even for small samples). However, the expected error analysis is much more efficient than cross validation, because no learning has to be done. Therefore it can be applied to large scale problems (such as text categorization or knowledge discovery in databases) with many (tens of) thousands of relevant attributes and many (tens of) thousands of examples.

A Proof of Theorem 1

Proof. The expected error $\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, D_{XY}, m, h_L \in H_i^*(S))$ is the average over all errors $E_D(h_j)$ weighted

[1] We have been reported that Domingos eliminated this particular assumption in a new, not yet published paper.

by the chance that h_j is selected by the learner. $\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, D_{XY}, m, h_L \in H_i^*(S)) = \sum_{j=1}^{|H_i|} E_D(h_j)\ P_{\{S,h_L\}}(h_L = h_j|H_i, D_{XY}, m, h_L \in H_i^*(S))$. The chance of a hypothesis being selected which is not in $H_i^*(S)$ (the set of minimum error hypotheses) is zero (Equation 5). We now factorize the number of minimum error hypotheses n (Equation 6). The chance of h_j being chosen as h_L when h_j is a minimum error hypothesis is $\frac{1}{n}$ (Equation 7). Now we factorize the empirical error e_S of h_j.

$$\begin{aligned}
&P_{\{S,h_L\}}(h_L = h_j|H_i, D_{XY}, m, h_L \in H_i^*(S)) && (5)\\
&= P_{\{S,h_L\}}(h_L{=}h_j|H_i, D_{XY}, m, h_j{\in} H_i^*(S), \\
&\quad h_L{\in} H_i^*(S))P_{\{S,h_L\}}(h_j{\in} H_i^*(S)|H_i, D_{XY}, m) \\
&= \sum_{n=1}^{|H_i|} P_{\{S,h_L\}}(h_L = h_j|h_j \in H_i^*(S), |H_i^*(S)| = n, \\
&\quad H_i, D_{XY}, m, h_L \in H_i^*(S)) \\
&\quad P_{\{S\}}(h_j \in H_i^*(S), |H_i^*(S)| = n|H_i, D_{XY}, m) && (6)\\
&= \sum_{n=1}^{|H_i|} \frac{1}{n} P_{\{S\}}(h_j \in H_i^*(S), |H_i^*(S)| = n| \\
&\quad H_i, D_{XY}, m) && (7)\\
&= \sum_{e_S} P_{\{S\}}(E_S(h_j) = e_S|E_D(h_j), H_i, m) \\
&\quad \sum_{n=1}^{|H_i|} \frac{1}{n} P_{\{S\}}(h_j \in H_i^*(S), | \\
&\quad H_i^*(S)| = n|E_S(h_i) = e_S, H_i, D_{XY}, m) && (8)
\end{aligned}$$

The empirical error (given the true error is binomially distributed, so $P_{\{S\}}(E_S(h_j) = e_S|E_D(h_j), H_i, m)$ in Equation 8 equals $B[E_D(h_j), m](e_S)$. Now we need to determine the unknown term $P_{\{S\}}(h_j \in H_i^*(S), |H_i^*(S)| = n|E_S(h_i) = e_S, H_i, D_{XY}, m)$. In Equation 9, we factorize all possible $H_i^*(S)$ of size n. When the hypotheses in $H_i^*(S)$ incur an empirical error of e_S, all other hypotheses have to incur a strictly higher error (according to the definition of $H_i^*(S)$. In Equation 10, we exploit the independence assumption to resolve the quantifiers.

$$\begin{aligned}
&P(h_j \in H_i^*(S), |H_i^*(S)| = n|E_S(h_j) = e_S, H_i, D_{XY}, m) && (9)\\
&= \sum_{\substack{H^* \subseteq H_i \\ |H^*|=n}} P(\forall h^* {\in} H^*{:} E_S(h^*) {=} e_S, \forall h {\in} H_i \backslash H^*{:} E_S(h) {>} e_S| \\
&\quad E_S(h_j) {=} e_S, E_D(h), E_D(h^*), m) \\
&= \sum_{\substack{H^* \subseteq H_i \\ |H^*|=n}} \prod_{h^* \in H^*} P(E_S(h^*) = e_S|E_S(h_j) = e_S, E_D(h^*), m) \\
&\quad \prod_{h \in H \backslash H^*} P(E_S(h) > e_S|E_S(h_j) = e_S, E_D(h), m) && (10)\\
&= \sum_{\substack{H^* \subseteq H_i \backslash \{h_j\}, \\ |H^*|=n-1}} \prod_{h^* \in H^*} B[E_D(h^*), m](e_S) \\
&\quad \prod_{h \in H_i \backslash \{h_i\} \backslash H^*} \sum_{e > e_S} B[E_D(h^*), m](e) && (11)
\end{aligned}$$

This completes the proof. ∎

B Proof of Theorem 2

Remember that the learner draws a hypothesis from $H_i^*(S)$ under uniform distribution and that assigning error values to individual hypothesis names was a notational trick in the first place. This means, if $E_D(h_i) = E_D(h_j)$ then $P_{\{S\}}(h_L = h_i|H_i, m) = P_{\{S\}}(h_L = h_j|H_i, m)$. Let h_{e_D} be an arbitrary hypothesis with $E_D(h_{e_D}) = e_D$. Then,

$$\begin{aligned}
&\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, m, D_{XY}, h_L \in H_i^*(S)) && (12)\\
&= \int_{e_D} e_D dP_{\{h\}}(E_D(h) = e_D|H_i, D_{XY}) \\
&\quad P_{\{S,h_L\}}(h_L = h_{e_D}|H_i, D_{XY}, m, h_L \in H_i^*(S))
\end{aligned}$$

Now we have to take care of $P_{\{S,h_L\}}(h_L = h_{eD}|H_i, D_{XY}, m, h_L \in H_i^*(S))$. We insert Equation 8 into Equation 12.

$$\begin{aligned}
&\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, m, D_{XY}, h_L \in H_i^*(S)) \\
&= \int_{e_D} e_D dP_{\{h\}}(E_D(h) = e_D|H_i, D_{XY}) \\
&\quad \sum_{e_{min}} \left(\sum_{n=1}^{|H_i|} \frac{1}{n} P_{\{S\}}(|H_i^*(S)| = n|h_{e_D} \in H_i^*(S), \right. \\
&\quad \left. E_S(h) = e_{min}, H_i, D_{XY}, m) \right) \\
&\quad P(h_{e_D} \in H_i^*(S)|E_S(h_{e_D}) = e_{min}, H_i, m) \\
&\quad B[e_D, m](e_{min}) && (13)
\end{aligned}$$

Exploiting assumption 2 we can claim that

$$\begin{aligned}
const &= \sum_{n=1}^{|H_i|} \frac{1}{n} P(|H_i^*(S)| = n|h \in H_i^*(S), \\
&\quad E_S(h) = e_{min}, H_i, D_{XY}, m) && (14)
\end{aligned}$$

is constant for all hypotheses h. $P_{\{S,h_L\}}(E_D(h_L)|H_i, m, D_{XY}, h_L \in H_i^*(S))$ should integrate to 1. This fixes the scaling factor *const* to

$$\begin{aligned}
const = 1/\Big(&\sum_{e_{min}} B[e_D, m](e_{min}) && (15)\\
&P(h_{e_D} \in H_i^*(S)|E_S(h_{e_D}) = e_{min}, H_i, m)\Big).
\end{aligned}$$

This proves Equation 2. Now we will focus on Equation 3. Equation 16 follows from the straightforward observation that h_{e_D} is in $H_i^*(S)$ iff h_{e_D} incurs an empirical error of e_S and all other h_j incur at least an error of e_S. Note that this equation exploits assumption 1. In Equation 17, we group all hypotheses with equal true error e'_D into one factor and take this factor to the number of hypotheses with that error. One hypothesis (h_{e_D}) has already been assigned an empirical error and is thus not included in the product.

$$
\begin{aligned}
&P_{\{S\}}(h_{eD} \in H_i^*(S)|H_i, m, E_D(h_{e_D})) && (16)\\
&= \sum_{e_S} B[e_D, m](e_S) \prod_{\substack{j=1\\ h_j \neq h_{e_D}}}^{|H_i|} \left(\sum_{e \geq e_S} B[E_D(h_j), m](e) \right)\\
&= \sum_{e_S} B[e_D, m](e_S) \prod_{e'_D}\\
&\quad \left(\sum_{e \geq e_S} B[e'_D, m](e) \right)^{|\{h : h \in H \setminus \{h_{e_D}\}, E_D(h) = e'_D\}|} && (17)
\end{aligned}
$$

Now Equations 13, 15, and 17 can be rewritten as Equation 3. This completes the proof. ∎

ACKNOWLEDGMENT

We would like to thank Uschi Sondhauss and Fritz Wysotzki for discussions about earlier versions of this paper. This work was partially supported by grants WY 20/1-2 of the German Research Council (DFG), the DFG collaborative research center SFB 475, and an Ernst-von-Siemens fellowship held by Tobias Scheffer.

References

[1] J. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, 1985.

[2] J. Berger. An overview of robust Bayesian analysis. Technical Report 93-53C, Department of Statistics, Purdue University, 1993.

[3] Y. Le Cun, J. Denker, and S. Solla. Optimal brain damage. In *NIPS-89*, pages 598–605, 1989.

[4] P. Domingos. A process-oriented heuristic for model selection. In *ICML-98*, pages 127–135, 1998.

[5] B. Efron. Bootstrap methods: another look at the jacknife. *Annals of Statistics*, 7(1):1–26, 1979.

[6] T. Joachims. Text categorization with support vector machines. In *Proceedings of the European Conference on Machine Learning*, 1998.

[7] M. Kearns, Y. Mansour, A. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning Journal*, 27:7–50, 1997.

[8] R. Kohavi and G. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):245–271, 1997.

[9] J. Mingers. An empirical comparison of selection measures for decision-tree induction. *Machine Learning*, 3:319–342, 1989.

[10] C. Mosier. Problems and designs of cross validation. *Educational and Psychological Measurement*, 11:5–11, 1951.

[11] J. Rissanen. Minimum-description-length principle. *Ann. Statist.*, 6:461–464, 1985.

[12] R. L. Rivest. Learning decision lists. *Machine Learning*, 2(2):229–246, 1987.

[13] C. Schaffer. Overfitting avoidance as bias. *Machine Learning*, 10:153–178, 1993.

[14] R. Schapire, Y. Freund, P. Bartlett, and W. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *ICML-97*, pages 322–330, 1997.

[15] T. Scheffer and T. Joachims. Estimating the expected error of empirical minimizers for model selection (abstract). AAAI-98, 1998.

[16] T. Scheffer and T. Joachims. Estimating the expected error of empirical minimizers for model selection. Technical Report TR 98-9, Technische Universitaet Berlin, 1998.

[17] T. Scheffer and T. Joachims. Expected error analysis for model selection. Preprint, TU Berlin, 1999. Available at http:ki.cs.tu-berlin.de/ scheffer.

[18] D. Schuurmans, L. Ungar, and D. Foster. Characterizing the generalization performance of model selection strategies. In *ICML-97*, pages 340–348, 1997.

[19] G. Toussaint. Bibliography on estimation of misclassification. *IEEE Transactions on Information Theory IT20*, 4:472–479, 1974.

[20] V. Vapnik. *Statistical Learning Theory.* Wiley, 1998.

[21] W. Wapnik and A. Tscherwonenkis. *Theorie der Zeichenerkennung.* Akademie Verlag, Berlin, 1979.

[22] D. H. Wolpert. The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework. In *The Mathematics of Generalization*, pages 117–214, 1995.

[23] David Wolpert. On the connection between in-sample testing and generalization error. *Complex Systems*, 6(1):47, 1992.