Validation of Average Error Rate Over Classifiers

Eric Bax*

July 1, 1997

Abstract

We examine methods to estimate the average and variance of test error rates over a set of classifiers. We begin with the process of drawing a classifier at random for each example. Given validation data, the average test error rate can be estimated as if validating a single classifier. Given the test example inputs, the variance can be computed exactly. Next, we consider the process of drawing a classifier at random and using it on all examples. Once again, the expected test error rate can be validated as if validating a single classifier. However, the variance must be estimated by validating all classifers, which yields loose or uncertain bounds.

Key words machine learning, Vapnik-Chervonenkis, validation.

1 Introduction

The average and variance of the test error rate over a set of classifiers are indicators of the potential performance ability of the classifiers as an ensemble, in which their decisions are combined through some form of fusion (Bishop, 1995; Jacobs et. al., 1991; Jordan and Jacobs, 1994; Perrone and Cooper, 1993; Wolpert, 1992). Estimates of these indicators can be used to select a single set of trained classifiers, from a collection of sets, to develop as an ensemble.

Estimates of the average test error rate also have applications to error estimation by inference, a method in which uniform error estimates over a large set of classifiers are produced with high confidence by exploiting similarities among the classifiers. First, uniform error estimates are derived for a small "core" set of classifiers, using standard worst-case assumptions about the underlying distributions and rates of agreement among the core classifiers. Then, for each classifier in the large set, the error is estimated by inference from the error estimates over the core classifiers. The inference is based on the fact that the difference in error rates between the classifier from the large set and each

*Computer Science Department, California Institute of Technology 256-80, Pasadena, California, 91125 (eric@cs.caltech.edu).

core classifier can be no greater than the rate of disagreement between the two classifiers.

In the simplest scheme, the error rate of the classifier from the large set is estimated by inference from the most similar core classifier. In (Bax, Cataltepe, and Sill, 1997), this scheme is used to estimate the test error rate of the classifier chosen by early stopping. A more complex method employs linear programming to estimate the error rate by inference from all core classifiers. In (Bax, 1997), linear programming is used to estimate the test error rates of voting committees and other ensembles.

If there are few core classifiers, their error rates can be uniformly estimated with high confidence. If each classifier in the large set has a high rate of agreement with one or more core classifiers, then the error estimates are accurate. Hence, we can achieve good estimates by partitioning the large set of classifiers into subsets with high rates of agreement and defining a core classifier corresponding to each subset by the following process – given an input, choose a classifier at random from the subset, and apply it. The error rate of this core classifier is the average error rate over the classifiers in the subset. One result of this paper is the development of estimates of the error rate for this type of core classifier.

2 Review of VC-Style Error Estimates

2.1 Framework

Our machine learning framework has the following structure. There is an unknown boolean-valued target function and an unknown distribution over its input space. For example, the input distribution could be typical data about credit card applicants, and the target function could be 1 if the applicant defaults within 5 years of being issued a credit card and zero otherwise.

We have a sequence of trained classifiers $g_1, \ldots, g_M$. We have d validation examples which were not used to train the classifiers. We may also have d' test inputs (but not the corresponding outputs). We assume that the validation and test inputs were drawn independently at random according to the underlying input distribution. We also assume that the validation outputs were determined by the target function.

The error rate of a classifier over a data set is the rate of disagreement over the inputs between the classifier and the target function. The ultimate goal is to produce a classifier with low error rate on the test data. In this paper, we focus on using validation data and test inputs to estimate the average and variance of test error rates over classifiers $g_1, \ldots, g_M$.

2.2 Single-Classifier Estimate

The first step to develop estimates of the average test error rate over several classifiers is to develop an estimate for the test error rate of a single classifier. We follow the development found in (Vapnik, 1982), in which a result due to Hoeffding (Hoeffding, 1963) is used to bound the confidence of estimating the test error rate by the validation error rate.

Let $g_m \in \{g_1, \ldots, g_M\}$ be a classifier chosen without reference to validation error rate. Let ν_m be the validation error rate of g_m, and let ν'_m be the test error rate. Let π_m be the expected error rate over the entire input distribution, i.e., the average test error rate over all randomly drawn test sets. We use the following bound by Hoeffding – if random variables $X_1, \ldots, X_n$ are independent and bounded by $a_i \leq X_i \leq b_i$, then, for $\epsilon > 0$,

$$\Pr\{|\overline{X} - \mu| \geq \epsilon\} \leq 2e^{-2n^2\epsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2} \tag{1}$$

where $\overline{X}$ is the sample mean of $X_1, \ldots, X_n$.

If we assign $n = d$, and define

$$X_i = \begin{cases} 1 & \text{if } g_m \text{ is correct on validation example } i \\ 0 & \text{if } g_m \text{ is incorrect on validation example } i \end{cases} \tag{2}$$

then $\mu = \pi_m$, $b_i = 1$, $a_i = 0$, and $\overline{X}$ is the validation error rate ν_m.

Set $\epsilon = \frac{\epsilon}{2}$ and use the bound.

$$\Pr\{|\nu_m - \pi_m| \geq \frac{\epsilon}{2}\} \leq 2e^{-2d^2(\frac{\epsilon}{2})^2 / \sum_{i=1}^{d}(1-0)} = 2e^{-\frac{1}{2}\epsilon^2 d} \tag{3}$$

A similar bound applies to the test error rate ν'_m.

$$\Pr\{|\nu'_m - \pi_m| \geq \frac{\epsilon}{2}\} \leq 2e^{-\frac{1}{2}\epsilon^2 d'} \tag{4}$$

Consider the probability that the validation error rate is not a good estimate of the test error rate.

$$\Pr\{|\nu'_m - \nu_m| \geq \epsilon\} \tag{5}$$

This event implies either $|\nu'_m - \pi_m| \geq \frac{\epsilon}{2}$ or $|\nu_m - \pi_m| \geq \frac{\epsilon}{2}$ or both. So

$$\Pr\{|\nu'_m - \nu_m| \geq \epsilon\} \leq \Pr\{|\nu'_m - \pi_m| \geq \frac{\epsilon}{2} \text{ or } |\nu_m - \pi_m| \geq \frac{\epsilon}{2}\} \tag{6}$$

Bound the probability of the union event by the sum of event probabilities.

$$\Pr\{|\nu'_m - \nu_m| \geq \epsilon\} \leq 2e^{-\frac{1}{2}\epsilon^2 d'} + 2e^{-\frac{1}{2}\epsilon^2 d} \leq 4e^{-\frac{1}{2}\epsilon^2 D} \tag{7}$$

where $D = \min(d, d')$. To obtain a bound on the confidence of the estimate, take the complement of the LHS, subtract the RHS from 1, and reverse the inequality.

$$\Pr\{|\nu'_m - \nu_m| < \epsilon\} \geq 1 - 4e^{-\frac{1}{2}\epsilon^2 D} \tag{8}$$

Hence, with confidence at least $1 - 4e^{-\frac{1}{2}\epsilon^2 D}$, the test error rate is within ϵ of the validation error rate.

2.3 Uniform Estimate

For comparison with the single-classifier estimate and the estimates of averages that we derive later, we now derive a bound for the confidence that the validation error rates are uniformly good estimates of corresponding test error rates over $g_1, \ldots, g_M$. We follow the derivation for VC analysis found in (Vapnik, 1982).

Consider the probability of failure for at least one single-classifier estimate.

$$\Pr\{|\nu_1' - \nu_1| \geq \epsilon \text{ or } \ldots \text{ or } |\nu_M' - \nu_M| \geq \epsilon\} \tag{9}$$

Bound the probability of the union event by the sum of event probabilities.

$$\leq \Pr\{|\nu_1' - \nu_1| \geq \epsilon\} + \ldots + \Pr\{|\nu_M' - \nu_M| \geq \epsilon\} \tag{10}$$

Use the bound (7) for each probability. The result is

$$\Pr\{|\nu_1' - \nu_1| \geq \epsilon \text{ or } \ldots \text{ or } |\nu_M' - \nu_M| \geq \epsilon\} \leq 4Me^{-\frac{1}{2}\epsilon^2 D} \tag{11}$$

To obtain the bound, take the complement of the LHS, subtract the RHS from 1, and reverse the inequality.

$$\Pr\{|\nu_1' - \nu_1| < \epsilon \text{ and } \ldots \text{ and } |\nu_M' - \nu_M| < \epsilon\} \geq 1 - 4Me^{-\frac{1}{2}\epsilon^2 D} \tag{12}$$

Note that the confidence of uniform estimation over $g_1, \ldots, g_M$, i.e., $1-4Me^{-\frac{1}{2}\epsilon^2 D}$, is much lower than the confidence for a single classifier, i.e., $1 - 4e^{-\frac{1}{2}\epsilon^2 D}$, when the number of classifiers M is large or when the size of the data set D is small. In the following sections, we show that the average error rate over classifiers $g_1, \ldots, g_M$ can be estimated with the same high degree of confidence that is achieved in the estimation of error rate for a single classifier.

3 Draw a Classifier for Each Example

Consider the following random process. For each example in a data set (with inputs drawn i.i.d. from an underlying distribution), draw a classifier from $\{g_1, \ldots, g_M\}$ uniformly at random, and apply the classifier to the example input.

3.1 Validation of Average Test Error Rate

Let random variable ν_r be the error rate achieved by the process on the validation data, and let ν_r' be the error rate on the test data. We will show

$$\Pr\{|\nu_r' - \nu_r| < \epsilon\} \geq 1 - 4e^{-\frac{1}{2}\epsilon^2 D} \tag{13}$$

By (8), the result holds for a single nonrandom classifier with validation error rate ν_r and test error rate ν_r'. Move the random selection of a classifier to the input distribution by adding an input variable that takes on values $1, \ldots, M$ with equal probabilities, independent of the distribution over the old variables. Define a new classifier that examines the new variable and applies the corresponding classifier to the old variables. This single nonrandom classifier has error rate distributions identical to the distributions of ν_r and ν'_r.

3.2 Calculation of Test Error Rate Variance

Given the test example inputs, we can compute the variance of the test error rates over applications of the random classifier. Let w_j be a random variable that has value 1 if a randomly selected classifier misclassifies test example j. Then the test error rate is the random variable

$$\sum_{j=1}^{d'} \frac{w_j}{d'} \tag{14}$$

The variables $\frac{w_j}{d'}$ are independent since the choice of classifier for each example is independent. Since the test error rate is the sum of independent variables, the variance is the sum of variances over the individual variables.

Let p_j be the fraction of classifiers that return 1 for example j. If the correct label is 0, then

$$w_j = \begin{cases} 1 & \text{with probability } p_j \\ 0 & \text{with probability } 1 - p_j \end{cases} \tag{15}$$

Since w_j is a Bernoulli variable with "success" probability p_j

$$\mathrm{Var}(w_j) = p_j(1 - p_j) \tag{16}$$

If the correct label is 1, then

$$w_j = \begin{cases} 1 & \text{with probability } 1 - p_j \\ 0 & \text{with probability } p_j \end{cases} \tag{17}$$

So

$$\mathrm{Var}(w_j) = (1 - p_j)p_j \tag{18}$$

The variance is the same in both cases.

Summing over individual variables, we compute the variance of the test error rate.

$$\sum_{j=1}^{d'} \mathrm{Var}(\frac{w_j}{d'}) = (\frac{1}{d'})^2 \sum_{j=1}^{d'} p_j(1 - p_j) \tag{19}$$

where p_j is the fraction of classifiers that return 1 for test example j.

4 Draw a Single Classifier for All Examples

Now consider the following random process. Draw a classifier uniformly at random from $\{g_1, \ldots, g_M\}$, and apply it to all examples in the validation and test sets.

4.1 Validation of Average Test Error Rate

We will show that the average error rate over classifiers can be estimated as if validating a single classifier, i.e.

$$\Pr\{|\frac{1}{M}\sum_{m=1}^{M}\nu'_m - \frac{1}{M}\sum_{m=1}^{M}\nu_m| \geq \epsilon\} \leq 4e^{-\frac{1}{2}\epsilon^2 D} \tag{20}$$

Since expectations commute, the average over classifiers of the average error over examples equals the average over examples of the average error over classifiers. This is important because the average error over classifiers is i.i.d. from example to example.

In more detail, let e_{mj} be the error of classifier g_m on validation example j. Hence, $e_{mj} = 1$ if g_m misclassifies example j, and $e_{mj} = 0$ if g_m correctly classifies the example. Define e'_{mj} similarly for test examples. Note

$$\frac{1}{M}\sum_{m=1}^{M}\nu_m = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{d}\sum_{j=1}^{d}e_{mj} = \frac{1}{d}\sum_{j=1}^{d}(\frac{1}{M}\sum_{m=1}^{M}e_{mj}) \tag{21}$$

The random variables $\frac{1}{M}\sum_{m=1}^{M}e_{mj}$ are i.i.d. since the example inputs are drawn i.i.d. from the input distribution.

By the Hoeffding bound (7), the bound for a single classifier,

$$\Pr\{|\nu'_m - \nu_m| \geq \epsilon\} = \Pr\{|\frac{1}{d'}\sum_{j=1}^{d'}e'_{mj} - \frac{1}{d}\sum_{j=1}^{d}e_{mj}| \geq \epsilon\} \leq 4e^{-\frac{1}{2}\epsilon^2 D} \tag{22}$$

holds because the errors on independently drawn examples are random variables in $[0, 1]$. Since the average error over classifiers is also a random variable in $[0, 1]$,

$$\Pr\{|\frac{1}{d'}\sum_{j=1}^{d'}(\frac{1}{M}\sum_{m=1}^{M}e'_{mj}) - \frac{1}{d}\sum_{j=1}^{d}(\frac{1}{M}\sum_{m=1}^{M}e_{mj})| \geq \epsilon\} \leq 4e^{-\frac{1}{2}\epsilon^2 D} \tag{23}$$

4.2 Bound on Average Test Error Rate

Given the test example inputs, we can bound the average test error rate. For each test example, the average error over classifiers is either the fraction of classifiers that return 1 or the fraction that return 0. Let p_j be the fraction of classifiers that return 1 for example j. Then the average test error rate is in the range

$$[\frac{1}{d'}\sum_{j=1}^{d'}\min(p_j, 1-p_j), \frac{1}{d'}\sum_{j=1}^{d'}\max(p_j, 1-p_j)] \tag{24}$$

The range is centered about $\frac{1}{2}$ since $\min(p_j, 1-p_j) = 1 - \max(p_j, 1-p_j)$.

4.3 Validation of Test Error Rate Variance

Now consider using validation data to estimate the variance of test error rates over classifiers when we draw one classifier at random for the entire data set. Since the test error rates are generally not independent among the classifiers, we cannot partition the estimate by examples as we did to estimate the average. Instead, we validate all classifiers and use the obtained bounds to bound the accuracy of estimating test error variance by validation error variance.

Choose some bound tightness $\epsilon \in (0, 1)$. Recall from (12)

$$\Pr\{\max_m |\nu'_m - \nu_m| < \epsilon\} \geq 1 - 4Me^{-\frac{1}{2}\epsilon^2 D} \tag{25}$$

Let δ_m be the residual of estimating the test error rate by the validation error rate, i.e. $\delta_m = \nu'_m - \nu_m$. With confidence $1 - 4Me^{-\frac{1}{2}\epsilon^2 D}$, $|\delta_m| < \epsilon$ for all classifiers g_m.

To bound the error due to estimating the test error variance over classifiers by the validation error variance over classifiers, first expand the test error variance using a well-known identity from probability theory (Feller, 1968, p.128).

$$\mathrm{Var}(\nu') = \mathrm{E}(\nu'^2) - (\mathrm{E}(\nu'))^2 \tag{26}$$

Replace test error rates by validation error rates and residuals.

$$\mathrm{Var}(\nu') = \mathrm{E}((\nu + \delta)^2) - (\mathrm{E}(\nu + \delta))^2 \tag{27}$$

Expectation is a linear operator.

$$\mathrm{Var}(\nu') = \mathrm{E}((\nu + \delta)^2) - (\mathrm{E}(\nu) + \mathrm{E}(\delta))^2 \tag{28}$$

Expand.

$$\mathrm{Var}(\nu') = \mathrm{E}(\nu^2) + \mathrm{E}(2\nu\delta) + \mathrm{E}(\delta^2) - \mathrm{E}(\nu)^2 - 2\mathrm{E}(\nu)\mathrm{E}(\delta) - \mathrm{E}(\delta)^2 \tag{29}$$

Collect variances.

$$\mathrm{Var}(\nu') = [\mathrm{E}(\nu^2) - \mathrm{E}(\nu)^2] + 2\mathrm{E}(\nu\delta) - 2\mathrm{E}(\nu)\mathrm{E}(\delta) + [\mathrm{E}(\delta^2) - \mathrm{E}(\delta)^2] \tag{30}$$

$$\mathrm{Var}(\nu') = \mathrm{Var}(\nu) + 2\mathrm{E}(\nu\delta) - 2\mathrm{E}(\nu)\mathrm{E}(\delta) + \mathrm{Var}(\delta) \tag{31}$$

Subtract Var(ν) from both sides and take absolute values.

$$|\mathrm{Var}(\nu') - \mathrm{Var}(\nu)| = |2[\mathrm{E}(\nu\delta) - \mathrm{E}(\nu)\mathrm{E}(\delta)] + \mathrm{Var}(\delta)| \tag{32}$$

Assume $|\delta_m| < \epsilon$ for all classifiers g_m. For $\nu \in [0,1]$ and $\delta \in [-\epsilon, \epsilon]$, the RHS is maximized when $(\nu, \delta) = (0, -\epsilon)$ with probability $\frac{1}{2}$, and $(\nu, \delta) = (1, \epsilon)$ with probability $\frac{1}{2}$. In this case,

$$|2[\mathrm{E}(\nu\delta) - \mathrm{E}(\nu)\mathrm{E}(\delta)] + \mathrm{Var}(\delta)| = |2[\frac{1}{2}\epsilon - 0] + \epsilon^2| = \epsilon + \epsilon^2 \tag{33}$$

Since $|\delta_m| < \epsilon$ for all classifiers with confidence $1 - 4Me^{-\frac{1}{2}\epsilon^2 D}$,

$$\Pr\{|\mathrm{Var}(\nu') - \mathrm{Var}(\nu)| < \epsilon + \epsilon^2\} \geq 1 - 4Me^{-\frac{1}{2}\epsilon^2 D} \tag{34}$$

Because we use uniform estimation of residuals over all classifiers to estimate test error rate variance, the confidence is much weaker than for the estimate of average test error rate.

5 Conclusion

We have shown that the average test error rate over a set of classifiers can be estimated as if validating a single classifier. We have also shown that the variance of test error for the process of drawing a classifier at random for each example can be computed exactly from the test inputs. However, for the process of drawing a single classifier to use on all data, our estimate of the variance requires us to validate all classifiers.

It may not be possible to obtain a better estimate for variance in the general case. However, there are better approaches for specific cases. The difference in error rates between classifiers is bounded by the rate of disagreement. Hence, if the classifiers agree on many test examples, then the variance can be bounded using bounds on error rate differences.

6 Acknowledgements

Thanks to Zehra Cataltepe and Joseph Sill for their instructive conversations and helpful pointers. Thanks to Dr. Yaser Abu-Mostafa for teaching – the results in this paper were inspired by his class on learning theory. Thanks to Dr. Joel Franklin for advice and guidance. Also, thanks to an anonymous referee for invaluable advice on the presentation of these results.

7 References

Bax, E., Z. Cataltepe, and J. Sill. (1997). Alternative error bounds for the classifier chosen by early stopping. CalTech-CS-TR-97-08.

Bax, E. (1997). Validation of voting committees. CalTech-CS-TR-97-13.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition* pp. 364-371. Oxford University Press.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* p. 128. John Wiley & Sons.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Am. Stat. Assoc. J.* pp. 13-30.

Jacobs, R. A., M. I. Jordan, S. J. Nowlan, and G E Hinton. (1991). Adaptive mixtures of local experts. *Neural Computation* 3 (1) pp. 79-87.

Jordan, M. I. and R. A. Jacobs. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6 (2) pp. 181-214.

Perrone, M. P., and L. N. Cooper. (1993). When networks disagree: ensemble methods for hybrid neural networks. in R. J. Mammone (Ed.). *Artificial Neural Networks for Speech and Vision* pp. 126-142. Chapman & Hall, London.

Vapnik, V. N. (1982). *Estimation of Dependences Based on Empirical Data* p.31. Springer-Verlag New York. Inc.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks.* 5 (2) pp. 241-259.