Overfitting Explained

Paul R. Cohen, David Jensen
Department of Computer Science,
University of Massachusetts
Amherst, MA 01003
cohen,jensen@cs.umass.edu

Abstract Overfitting arises when model components are evaluated against the wrong reference distribution. Most modeling algorithms iteratively find the best of several components and then test whether this component is good enough to add to the model. We show that for independently distributed random variables, the reference distribution for any one variable underestimates the reference distribution for the the highest-valued variable; thus variate values will appear significant when they are not, and model components will be added when they should not be added. We relate this problem to the well-known statistical theory of multiple comparisons or simultaneous inference.

1 Iterative Modeling Algorithms

Iterative modeling algorithms (IMAs) generate a search space $\mathcal{M}$ of models by repeatedly selecting a model $m(\cdot) \in \mathcal{M}$ and adding a component c_i from a list of components $\mathcal{C} = c_1, c_2, ..., c_n$ to $m(\cdot)$, producing $m(\cdot, c_i)$. For example, $m(\cdot)$ may be the regression equation $\hat{y} = \beta_3 c_3 + \beta_1 c_1$, and $m(\cdot, c_5)$ is $\hat{y} = \beta_3 c_3 + \beta_1 c_1 + \beta_5 c_5$. Generally, IMAs do not add every possible component to each model $m(\cdot)$—this would result in exhaustive search—but rather, they add the component that appears best according to some evaluation function $x_i = \mathcal{V}(c_i, m(\cdot), \mathcal{S})$. We call x_i the *score* of component c_i given model $m(\cdot)$ and a sample of data $\mathcal{S}$. For example, $\mathcal{V}$ might compute information gain or classification accuracy for decision tree induction algorithms, F ratios for stepwise multiple regression algorithms, and so on. We may define a general IMA algorithm as follows:

IMA: Initially, $\mathcal{M}$ contains the empty model $m()$. Now iterate:

1. Select a model $m(\cdot) \in \mathcal{M}$
2. Remove components from $\mathcal{C}$ on logical grounds if necessary, producing $\mathcal{C}'$. For example, regression models shouldn't contain multiple occurrences of the same variable; whereas decision trees can in some circumstances.
3. Find the component, $c_{max} \in \mathcal{C}'$, with the highest value $x_{max} = max(x_1, x_2, ...x_n)$, where $x_i = \mathcal{V}(c_i, m(\cdot), \mathcal{S})$
4. If $x_{max} > T_{\mathcal{V}}$, where $T_{\mathcal{V}}$ is a possibly dynamic threshold value, then add c_{max} to $m(\cdot)$.
5. Revise M by adding $m(\cdot, c_{max})$ and perhaps removing one or more models.

IMA terminates when no component can be added to any $m(\cdot) \in \mathcal{M}$ according to step 4.

A model $m(\cdot)$ *overfits* a dataset $\mathcal{S}$ when it includes one or more components c_i that have sufficient scores $x_i > T_{\mathcal{V}}$ given $\mathcal{S}$, but c_i would not have sufficient scores in general—that is, in other datasets drawn from the same population or in the population itself. Obviously, overfitting can occur if the threshold $T_{\mathcal{V}}$ is set too low. Said differently, if $T_{\mathcal{V}}$ is set in a way that underestimates the distribution of x_{max}, then overfitting will occur. In particular, if $T_{\mathcal{V}}$ is based on the distribution of scores x_i instead of the distribution of maximum scores x_{max} then overfitting is inevitable.[1] Virtually all decision tree induction algorithms, for example, base $T_{\mathcal{V}}$ on the distribution of x_i instead of x_{max}, which is why they overfit, often dramatically.

Clearly, $T_{\mathcal{V}}$ must respect the distribution of x_{max}, so we begin by examining this distribution under some simplifying assumptions. We focus on the probabilities $Pr(x_{max} \geq k)$ and $Pr(x_i \geq k)$, and on the expected values $E(x_{max})$ and $E(x_i)$. In general, the distribution of x_i underestimates the probability of x_{max}. Then we consider how $T_{\mathcal{V}}$ is set, focusing on the common view of $T_{\mathcal{V}}$ as a critical value in a reference distribution. It will then be obvious how the problem of overfitting is a version of the classical statistical problem of multiple comparisons. This equivalence suggests numerous overfitting-avoidance techniques, which have been tested empirically (see [4]).

2 The Distribution of the Maximum Score

Recall that a *score* is an evaluation of a component c_i that IMA is considering adding to a model $m(\cdot)$: $x_i = \mathcal{V}(c_i, m(\cdot), \mathcal{S})$. Suppose IMA is considering n components $c_1, c_2, \ldots, c_n$ with scores $x_1, x_2, \ldots, x_n$. Each score is the value of a random variable. The distribution of the maximum score will depend on the distributions of the random variables, and, in general, the variables are not identically and independently distributed (i. i. d.). The following results are for i. i. d variables, and for independent but not necessarily identically distributed variables. We have not extended the results to non-independent variables. However, empirically we have shown that the errors introduced by non-independence are small relative to the errors incurred by not using the reference distribution for the maximum (see Figure 1 and [4]).

For simplicity and concreteness, assume x_1 and x_2 are random variables drawn from a uniform distribution of integers $(0 \ldots 6)$. The distribution of $max(x_1, x_2)$ is shown in table 1. Each entry in the table represents a joint event with the resulting maximum score; for example, $(x_1 = 3 \ \wedge \ x_2 = 4)$ has the result, $max(x_1, x_2) = 4$. Because x_1 and x_2 are i. i. d. and uniform, every joint event has the same probability, $1/49$, but the probability of a given maximum score is generally higher; for example, $Pr(max(x_1, x_2) = 4) = 9/49$. In fact, the probability $Pr(max(x_1, x_2) = k)$ increases with k; for example, $Pr(max(x_1, x_2) = 6) = 13/49$.

For i. i. d. random variables $x_1, x_2, \ldots, x_n$, it is easy to specify the relationship between cumulative probabilities of individual scores and cumulative probabilities of maximum scores:

$$\text{If } Pr(x_i < k) = q, \text{ then } Pr(max(x_1, x_2, \ldots, x_n) < k) = q^n. \tag{1}$$

[1] In fact, overfitting can occur even when the appropriate reference distribution is used, but its probability can be controlled and made arbitrarily small.

	0	1	2	3	4	5	6
0	0	1	2	3	4	5	6
1	1	1	2	3	4	5	6
2	2	2	2	3	4	5	6
3	3	3	3	3	4	5	6
4	4	4	4	4	4	5	6
5	5	5	5	5	5	5	6
6	6	6	6	6	6	6	6

Table 1: The joint distribution of the maximum of two random variables, each of which takes integer values (0...6).

For example, in table 1, $Pr(x_1 < 4) = 4/7$ (and $Pr(x_2 < 4)$ is identical, because x_1 and x_2 are i. i. d.) but $Pr(max(x_1, x_2) < 4) = (4/7)^2 = 16/49$. It is also useful to look at the upper tail of the distribution of the maximum:

$$\text{If } Pr(x_i \geq k) = p, \text{ then } Pr(max(x_1, x_2, \ldots, x_n) \geq k) = 1 - (1-p)^n \tag{2}$$

These expressions and the distribution in table 1 make clear that the distribution of any random variable x_i from i. i. d. variables $x_1, x_2, \ldots, x_n$ underestimates the distribution of the maximum $x_{max} = max(x_1, x_2, \ldots, x_n)$. $Pr(x_i \geq k)$ underestimates $Pr(max(x_1, x_2, \ldots, x_n) \geq k)$ for all values k if the distributions are continuous. Said differently, the distribution of x_{max} has a heavier upper tail than the distribution of x_i.

This disparity increases with the number of random variables, $x_1, x_2, \ldots, x_n$. Imagine three variables distributed in the same way as the two in table 1. Then,

$$\begin{aligned} Pr(x_i \geq 4) &= 3/7 = .43 \\ Pr(max(x_1, x_2, x_3) \geq 4) &= 1 - (1 - 3/7)^3 = .81. \end{aligned}$$

The distribution of x_i underestimates $Pr(max(x_1, x_2, x_3) \geq 4)$ by almost one half its value.

The expected value x_i, $E(x_i)$, generally underestimates the expected value of the maximum. This is easily demonstrated for two random variables x_1 and x_2 which are statistically independent but not necessarily identically distributed; the extension to more independent variables is obvious. The expected values of x_1 and x_2 are

$$E(x_1) = \sum_{i=1}^{n} x_{1_i} Pr(x_{1_i}), \; E(x_2) = \sum_{j=1}^{n} x_{2_j} Pr(x_{2_j}).$$

Likewise, the expected value of $max(x_1, x_2)$ is

$$E(max(x_1, x_2)) = \sum_{i=1}^{n} \sum_{j=1}^{n} max(x_{1_i}, x_{2_j}) Pr(x_{1_i}) Pr(x_{2_j}) \tag{3}$$

$$= \sum_{i=1}^{n} Pr(x_{1_i}) \sum_{j=1}^{n} max(x_{1_i}, x_{2_j}) Pr(x_{2_j}). \tag{4}$$

For any value x_{1_i}, $max(x_{1_i}, x_{2_j}) \geq x_{2_j}$. Consequently,

$$\sum_{j=1}^{n} max(x_{1_i}, x_{2_j}) Pr(x_{2_j}) \geq E(x_2) \tag{5}$$

Thus, expression 4 becomes an inequality:

$$\begin{aligned} E(max(x_1, x_2)) &\geq \sum_{i=1}^{n} Pr(x_{1_i}) E(x_2) \\ &\geq E(x_2) \sum_{i=1}^{n} Pr(x_{1_i}) \\ &\geq E(x_2) \end{aligned}$$

We can prove $E(max(x_1, x_2)) \geq E(x_1)$ in the same way. In sum,

$$E(max(x_1, x_2)) \geq max(E(x_1), E(x_2)) \tag{6}$$

In fact, $max(E(x_1), E(x_2))$ nearly always underestimates $E(max(x_1, x_2))$; more dramatically as the number of random variables increases.

These properties of the distribution of x_{max} depend on $x_1, x_2, \ldots, x_n$ being independently (if not identically) distributed. In the general case, where $x_1, x_2, \ldots, x_n$ are dependent, the probability $Pr(max(x_{1_i}, x_{2_j}, \ldots, x_{n_m}) \geq k)$ is not so easy to estimate (but see [6]). It is not simply a product of probabilities, as in expressions 2 and 4, because $Pr(a, b) \neq Pr(a)Pr(b)$ when a and b are dependent. In empirical studies of overfitting (e.g., Figure 1), we see that the errors introduced by assuming independent variables to derive a reference distribution for x_{max} are small relative to the errors introduced by relying on the reference distribution for x_i instead of the x_{max} distribution.

3 Underestimation and Overfitting

Underestimating the maximum of n random variables can lead to overfitting. Recall that IMA adds component c_i to model $m(\cdot)$ when c_i is the best component (step 3) and c_i's score, x_i, exceeds the threshold $T_\mathcal{V}$ (step 4). There are many ways to set $T_\mathcal{V}$, but however one does it, $T_\mathcal{V}$ ought to reflect the number of components being considered, the variances of the distributions of the components, the size of sample $\mathcal{S}$, and the number of components already in model $m(\cdot)$. These factors suggest treating $T_\mathcal{V}$ as a critical value in a reference distribution; said differently, $x_i \geq T_\mathcal{V}$ can be tested with the machinery of statistical hypothesis testing. In fact, this is how many IMA algorithms decide whether to add components. We will briefly review the logic of statistical hypothesis testing.

Suppose we want to test whether a component, c_1, contributes enough to model $m(\cdot)$ to warrant generating a new model $m(\cdot, c_1)$. The usual approach is to derive a reference distribution F_1, for the scores, x_{1_i}, under the *null hypothesis*, H_0, that c_1 contributes *nothing* to $m(\cdot)$. Then, given a particular score $x_1 = k$, one calculates the probability $p = Pr(x_1 \geq k)$, and if it is very low, one rejects H_0 and concludes that c_1 probably does contribute something to $m(\cdot)$. The probability p bounds one's confidence in this conclusion. Typically, one selects a high quantile of F_1, say, the 95th quantile, $F_1(95)$. If $x_1 > F_1(95)$,

then one rejects the hypothesis that c_1 contributes nothing to $m(\cdot)$, with a probability of error $p \leq .05$. $F_1(95)$ is called the .05 critical value for the reference distribution F_1.

The hypothesis testing strategy can be misapplied in incremental modeling algorithms, with overfitting as the consequence. Here is the *incorrect* implementation of hypothesis testing in IMA:

Incorrect Hypothesis Testing in IMA: For a given model $m(\cdot)$, and components $C' = c_1, c_2, \ldots, c_n$ with scores $x_1, x_2, \ldots, x_n$,

1. Find the best component c_i for which $x_i = x_{max} = max(x_1, x_2, \ldots, x_n)$.
2. Formulate the null hypothesis that c_i contributes nothing to $m(\cdot)$ and derive the reference distribution F_i under this hypothesis.
3. Set $T_\mathcal{V} = F_i(95)$ (or some other confidence level). If $x_i \geq T_\mathcal{V}$ reject the null hypothesis and add c_i to $m(\cdot)$.

In this procedure, the null hypothesis, and thus the reference distribution, are incorrect. The correct null hypothesis is, "The *best* of n components adds nothing to the model," and the correct reference distribution is the distribution of F_{max} under this null hypothesis. It is easy to see how one might erroneously use F_i to test x_i when x_i is the maximum score, but F_i underestimates F_{max}—as we demonstrated earlier for i. i. d. variables, and have shown to be generally true even for non-independent variables—so x_i might easily exceed $F_i(95)$ but fall short of $F_{max}(95)$.

It is now clear how this procedure causes overfitting: In general a reference distribution F_i will underestimate F_{max}, so any value $T_\mathcal{V}$ based on F_i will be too low. Thus, components will be added because their scores seem statistically unlikely (e.g., $x_i \geq F_i(95)$) when, according to the correct reference distribution, they are not unlikely at all (i.e, $x_i < F_{max}(95)$).

Equation 2 provides an estimate of the probability of overfitting for any given model $m(\cdot) \in \mathcal{M}$. For example, if any one of ten components could be added to a model, and the components' scores are i. i. d., and we use a 0.10 critical value for F_i instead of for F_{max} as $T_\mathcal{V}$, then the probability of overfitting is

$$1 - (1 - 0.10)^{10} = .6513.$$

Keep in mind that this result characterizes the probability of incorrectly adding a *single* component to a model. After adding one component, most modeling algorithms then consider adding another, and another, and each of these decisions also has an elevated probability of being incorrect. One can easily build models in which *most* of the components shouldn't be there. Decision tree induction algorithms, for instance, are exquisitely prone to overfitting [4, 7].

Figure 1 illustrates how non-independence of the scores $x_1, \ldots, x_n$ affects the probability of incorrectly rejecting the null hypothesis and thus accepting a model component incorrectly. In each trial, ten binary attribute with equal class probability and 50 instances were compared to a randomly-generated binary classification variable. The scores for these attributes, $x_1, \ldots, x_{10}$, measure strength of association between the attribute and the binary classification variable. These scores are expected to be small because, as noted, the

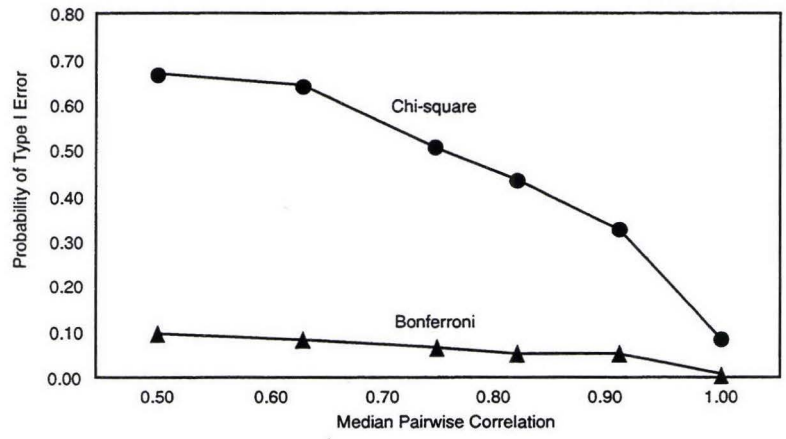

Figure 1: The joint effects of underestimating F_{max} and non-independent scores.

classification variable is random. The horizontal axis of Figure 1 is the median pairwise correlation between the attributes. The leftmost value, 0.50, means the attributes are i. i. d. and higher values reflect increasing dependence among the attributes and thus their scores. The chi-square scores were compared with conventional reference distributions for chi-square tests with critical value $F(90)$. That is, the probability of incorrectly rejecting the null hypothesis is 0.10. When the attributes are i. i. d., the probability of spuriously accepting one into a model on the basis of chi-square scores is roughly 0.65—no 0.10 as it should be. As the attribute scores become more dependent, the probability of overfitting drops. Intuitively this is because as attribute scores become more highly correlated, the number of independent opportunities to reject the null hypothesis is effectively reduced. In fact, we may think of highly correlated attributes as equivalent to having fewer attributes; in the extreme case of perfect correlation, all the attributes behave identically, so either they all reject the null hypothesis or none does.

To avoid overfitting, simply replace F_i with F_{max} in the procedure, above. To do this, one must estimate F_{max}, which is easy to do by randomization, bootstrapping or some other Monte Carlo procedure [3, 8]. That is, once we have an estimate of F_{max} we can select a critical value T_ν to give us any desired probability of incorrectly rejecting the null hypothesis and accepting a spurious model component.

Alternatively, one might adjust the critical value in the F_i reference distribution to ensure that the probability of falsely rejecting the null hypothesis on the basis of x_{max} is, say, 0.10 as desired. This approach is reminiscent of the Bonferroni adjustment, and it works quite well [1, 2, 4], although the adjustment tends to be conservative, especially when the variables are not i. i. d. The line marked "Bonferroni" in Figure 1 is Bonferroni-adjusted chi-square scores, and when the attributes are i. i. d., the adjustment gives us exactly the probability of overfitting that we stipulated, 0.10, but as the attribute scores become more correlated, the Bonferroni adjustment becomes overly stringent. While it prevents overfitting, it also prevents us adding *any* model components.

4 Underestimation and Multiple Comparisons

The Bonferroni adjustment is popular for problems involving multiple comparisons, or simultaneous inference. There is a direct mapping from the problem of estimating the distribution of the maximum to the problem of multiple or simultaneous comparisons.

Suppose $\mathcal{C} = c_1, c_2, \ldots, c_n$ with scores $x_1, x_2, \ldots, x_n$, and assume these scores are independently and identically distributed (i. i. d.) random variables. Consider two null hypotheses:

Simultaneous : Every component c_i contributes nothing to model $m(\cdot)$.

Max : The best component, c_{max}, contributes nothing to model $m(\cdot)$.

Suppose one tests each of the **simultaneous** null hypotheses against a reference distribution, F_i (which is the same for all scores because they are i. i. d.) For example, one tests c_1 by comparing x_1 to F_i, then one tests c_2 by comparing x_2 to F_i, and so on. Alternatively, one might test the **max** null hypothesis by comparing x_{max} to F_i. Assuming i. i. d. scores, these testing strategies produce identical *type I errors.*

A type I error involves rejecting the null hypothesis when it is true. In the simultaneous case, above, α_c denotes the probability that a test of a single component will erroneously reject the null hypothesis, and α_e denotes the probability that at least one test of n components will erroneously reject the null hypothesis. Think of α_c as a bias on a coin: if $\alpha_c = .05$, then with probability .95, a toss will land tails, and no error will occur. Clearly, if one tosses the coin twice, the probability of landing tails twice, and avoiding a type I error, is $.95^2$. Clearly, if one performs n independent statistical tests, each with α_c probability of a type I error, then the probability of at least one type I error in all n tests is

$$\alpha_e = 1 - (1 - \alpha_c)^n \tag{7}$$

α_e is called the *experimentwise* type I error.

Now suppose we have $x_1, x_2, \ldots, x_n$ i. i. d. random variables, and we set k so that $Pr(x_i \geq k) = \alpha_c$. What is the probability that the maximum of the variables exceeds k? From expression 2 we see

$$Pr(max(x_1, x_2, \ldots, x_n) \geq k) = 1 - (1 - \alpha_c)^n. \tag{8}$$

That is, the probability of a type I error committed by comparing $max(x_1, x_2, \ldots, x_n)$ to a reference distribution for F_i is identical to the experimentwise probability of a type I error in n comparisons. The **simultaneous** and **max** null hypotheses, above, are identical in terms of the resulting type I error probabilities. This is not surprising, because finding the maximum of n random variables and then testing whether it exceeds a critical value requires n pairwise comparisons of random variables.

The upshot of this result is that we may apply techniques developed for problems of multiple comparisons (such as the Bonferroni adjustment) to the overfitting problem [2, 1]. All these techniques adjust $T_\mathcal{V}$ to account for the fact that we are testing not one, but the best of several, model components.

5 Acknowledgments

This research is supported by DARPA/Rome Laboratory contract F30602-93-C-0076 and by a Subcontract from Sterling Software Inc. 7335-UOM-001 (DARPA/Rome Laboratory F30602-95-C-0257). The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of the Defense Advanced Research Projects Agency, Rome Laboratory or the U.S. Government.

References

[1] Cohen, P.R. (1995). *Empirical Methods for Artificial Intelligence.* Cambridge, MA: MIT Press.

[2] J. Galambos (1978). The Asymptotic Theory of Extreme Order Statistics. New York: Wiley.

[3] Jensen, D. (1992). Induction with Randomization Testing: Decision-Oriented Analysis of Large Data Sets, Doctoral Dissertation, Sever Institute of Technology, Washington University, St. Louis, Missouri.

[4] Jensen, D. (1997). Adjusting for multiple testing in decision tree pruning. Accepted to the Sixth International Workshop on Artificial Intelligence and Statistics (poster).

[5] Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* 29(2):199–127.

[6] Kotz, S. and N.L. Johnson (Eds.) (1982-1989). *Encyclopedia of Statistical Sciences.* New York: Wiley.

[7] Miller, Rupert G. (1981). Simultaneous Statistical Inference. New York: Springer-Verlag.

[8] Oates, T. and D. Jensen (1997). The effects of training set size on decision tree complexity. Accepted to the Sixth International Workshop on Artificial Intelligence and Statistics.