Overfitting

In mathematical modeling, **overfitting** is "the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit to additional data or predict future observations reliably".[1] An **overfitted model** is a mathematical model that contains more parameters than can be justified by the data.[2] The essence of overfitting is to have unknowingly extracted some of the residual variation (i.e., the noise) as if that variation represented underlying model structure.[3]:45

Underfitting occurs when a mathematical model cannot adequately capture the underlying structure of the data. An **under-fitted model** is a model where some parameters or terms that would appear in a correctly specified model are missing.[2] Under-fitting would occur, for example, when fitting a linear model to non-linear data. Such a model will tend to have poor predictive performance.

The possibility of over-fitting exists because the criterion used for selecting the model is not the same as the criterion used to judge the suitability of a model. For example, a model might be selected by maximizing its performance on some set of training data, and yet its suitability might be determined by its ability to perform well on unseen data; then over-fitting occurs when a model begins to "memorize" training data rather than "learning" to generalize from a trend.

As an extreme example, if the number of parameters is the same as or greater than the number of observations, then a model can perfectly predict the training data simply by memorizing the data in its entirety. (For an illustration, see Figure 2.) Such a model, though, will typically fail severely when making predictions.

The potential for overfitting depends not only on the number of parameters and data but also the conformability of the model structure with the data shape, and the magnitude of model error compared to the expected level of noise or error in the data. Even when the fitted model does not have an excessive number of parameters, it is to be expected that the fitted relationship will appear to perform less well on a new data set than on the data set used for fitting (a phenomenon sometimes known as *shrinkage*).[2] In

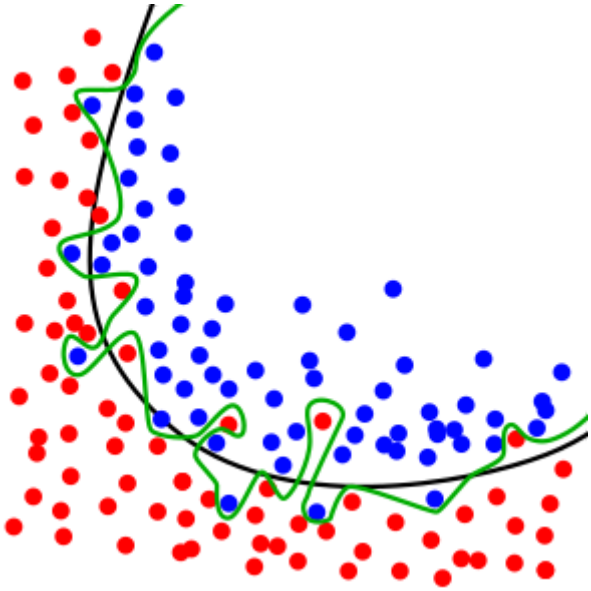

Figure 1. The green line represents an overfitted model and the black line represents a regularized model. While the green line best follows the training data, it is too dependent on that data and it is likely to have a higher error rate on new unseen data, compared to the black line.

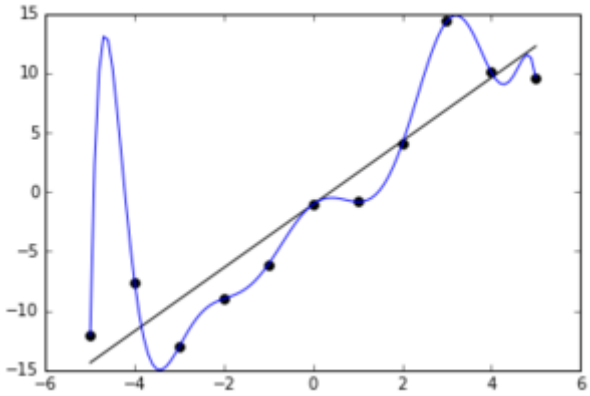

Figure 2. Noisy (roughly linear) data is fitted to a linear function and a polynomial function. Although the polynomial function is a perfect fit, the linear function can be expected to generalize better: if the two functions were used to extrapolate beyond the fitted data, the linear function should make better predictions.

particular, the value of the coefficient of determination will shrink relative to the original data.

To lessen the chance or amount of overfitting, several techniques are available (e.g., model comparison, cross-validation, regularization, early stopping, pruning, Bayesian priors, or dropout). The basis of some techniques is either (1) to explicitly penalize overly complex models or (2) to test the model's ability to generalize by evaluating its performance on a set of data not used for training, which is assumed to approximate the typical unseen data that a model will encounter.

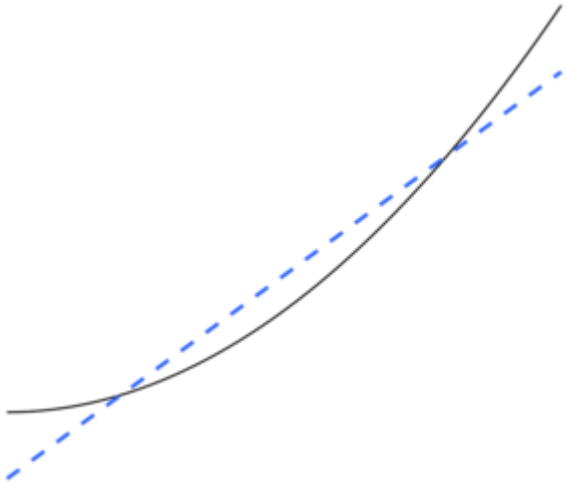

Figure 3. The blue dashed line represents an underfitted model. A straight line can never fit a parabola. This model is too simple.

Contents

Statistical inference

In statistics, an inference is drawn from a statistical model, which has been selected via some procedure. Burnham & Anderson, in their much-cited text on model selection, argue that to avoid overfitting, we should adhere to the "Principle of Parsimony".[3] The authors also state the following.[3]:32–33

> Overfitted models … are often free of bias in the parameter estimators, but have estimated (and actual) sampling variances that are needlessly large (the precision of the estimators is poor, relative to what could have been accomplished with a more parsimonious model). False treatment effects tend to be identified, and false variables are included with overfitted models. … A best approximating model is achieved by properly balancing the errors of underfitting and overfitting.

Overfitting is more likely to be a serious concern when there is little theory available to guide the analysis, in part because then there tend to be a large number of models to select from. The book *Model Selection and Model Averaging* (2008) puts it this way.[4]

Given a data set, you can fit thousands of models at the push of a button, but how do you choose the best? With so many candidate models, overfitting is a real danger. Is the monkey who typed Hamlet actually a good writer?

Regression

In regression analysis, overfitting occurs frequently.[5] As an extreme example, if there are *p* variables in a linear regression with *p* data points, the fitted line can go exactly through every point.[6] For logistic regression or Cox proportional hazards models, there are a variety of rules of thumb (e.g. 5–9,[7] 10[8] and 10–15[9] — the guideline of 10 observations per independent variable is known as the "one in ten rule"). In the process of regression model selection, the mean squared error of the random regression function can be split into random noise, approximation bias, and variance in the estimate of the regression function. The bias–variance tradeoff is often used to overcome overfit models.

With a large set of explanatory variables that actually have no relation to the dependent variable being predicted, some variables will in general be falsely found to be statistically significant and the researcher may thus retain them in the model, thereby overfitting the model. This is known as Freedman's paradox.

Machine learning

Usually a learning algorithm is trained using some set of "training data": exemplary situations for which the desired output is known. The goal is that the algorithm will also perform well on predicting the output when fed "validation data" that was not encountered during its training.

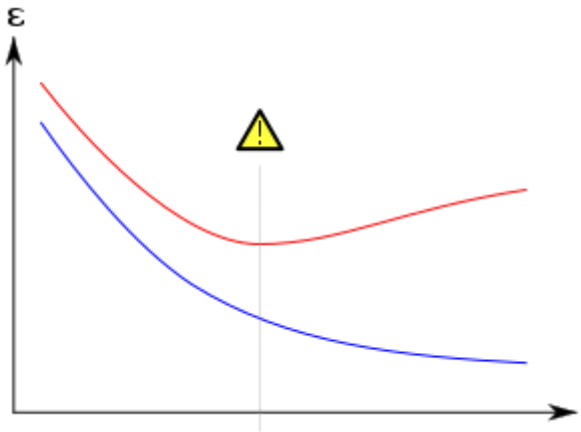

Figure 4. Overfitting/overtraining in supervised learning (e.g., neural network). Training error is shown in blue, validation error in red, both as a function of the number of training cycles. If the validation error increases (positive slope) while the training error steadily decreases (negative slope) then a situation of overfitting may have occurred. The best predictive and fitted model would be where the validation error has its global minimum.

Overfitting is the use of models or procedures that violate Occam's razor, for example by including more adjustable parameters than are ultimately optimal, or by using a more complicated approach than is ultimately optimal. For an example where there are too many adjustable parameters, consider a dataset where training data for y can be adequately predicted by a linear function of two independent variables. Such a function requires only three parameters (the intercept and two slopes). Replacing this simple function with a new, more complex quadratic function, or with a new, more complex linear function on more than two independent variables, carries a risk: Occam's razor implies that any given complex function is *a priori* less probable than any given simple function. If the new, more complicated function is selected instead of the simple function, and if there was not a large enough gain in training-data fit to offset the complexity increase, then the new complex function "overfits" the data, and the complex overfitted function will likely perform worse than the simpler function on validation data outside the training dataset, even though the complex function performed as well, or perhaps even better, on the training dataset.[10]

When comparing different types of models, complexity cannot be measured solely by counting how many parameters exist in each model; the expressivity of each parameter must be considered as well. For example, it is nontrivial to directly compare the complexity of a neural net (which can track curvilinear relationships) with m parameters to a regression model with n parameters.[10]

Overfitting is especially likely in cases where learning was performed too long or where training examples are rare, causing the learner to adjust to very specific random features of the training data that have no causal relation to the target function. In this process of overfitting, the performance on the training examples still increases while the performance on unseen data becomes worse.

As a simple example, consider a database of retail purchases that includes the item bought, the purchaser, and the date and time of purchase. It's easy to construct a model that will fit the training set perfectly by using the date and time of purchase to predict the other attributes, but this model will not generalize at all to new data, because those past times will never occur again.

Generally, a learning algorithm is said to overfit relative to a simpler one if it is more accurate in fitting known data (hindsight) but less accurate in predicting new data (foresight). One can intuitively understand overfitting from the fact that information from all past experience can be divided into two groups: information that is relevant for the future, and irrelevant information ("noise"). Everything else being equal, the more difficult a criterion is to predict (i.e., the higher its uncertainty), the more noise exists in past information that needs to be ignored. The problem is determining which part to ignore. A learning algorithm that can reduce the risk of fitting noise is called "robust."

Consequences

The most obvious consequence of overfitting is poor performance on the validation dataset. Other negative consequences include:[10]

- A function that is overfitted is likely to request more information about each item in the validation dataset than does the optimal function; gathering this additional unneeded data can be expensive or error-prone, especially if each individual piece of information must be gathered by human observation and manual data-entry.
- A more complex, overfitted function is likely to be less portable than a simple one. At one extreme, a one-variable linear regression is so portable that, if necessary, it could even be done by hand. At the other extreme are models that can be reproduced only by exactly duplicating the original modeler's entire setup, making reuse or scientific reproduction difficult.

Remedy

The optimal function usually needs verification on bigger or completely new datasets. There are, however, methods like minimum spanning tree or life-time of correlation that applies the dependence between correlation coefficients and time-series (window width). Whenever the window width is big enough, the correlation coefficients are stable and don't depend on the window width size anymore. Therefore, a correlation matrix can be created by calculating a coefficient of correlation between investigated variables. This matrix can be represented topologically as a complex network where direct and indirect influences between variables are visualized. Dropout regularisation can also improve robustness and therefore reduce over-fitting by probabilistically removing inputs to a layer.

Underfitting

Underfitting is the inverse of overfitting, meaning that the statistical model or machine learning algorithm is too simplistic to accurately represent the data. A sign of underfitting is that there is a high bias and low variance detected in the current model or algorithm used (the inverse of overfitting: low bias and high variance). This can be gathered from the Bias-variance tradeoff which is the method of analyzing a model or algorithm for bias error, variance error and irreducible error. With a high bias and low variance the result of the model is that it will inaccurately represent the data points and thus insufficiently be able to predict future data results (see Generalization error). Shown in Figure 5 the linear line could not represent all the given data points due to the line not resembling the curvature of the points. We would expect to see a parabola shaped line as shown in Figure 6 and Figure 1. As previously mentioned if we were to use Figure 5 for analysis we would get false predictive results contrary to the results if we analyzed Figure 6.

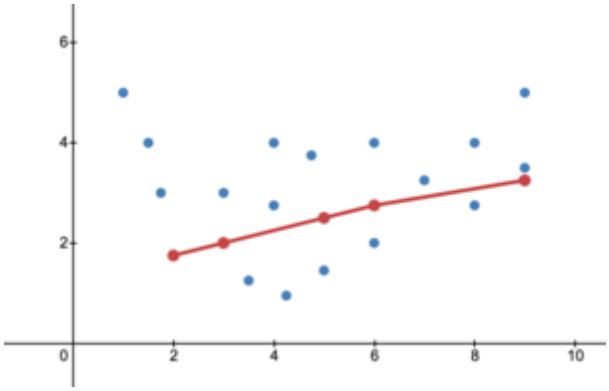

Figure 5. The red line represents an underfitted model of the data points represented in blue. We would expect to see a parabola shaped line to represent the curvature of the data points.

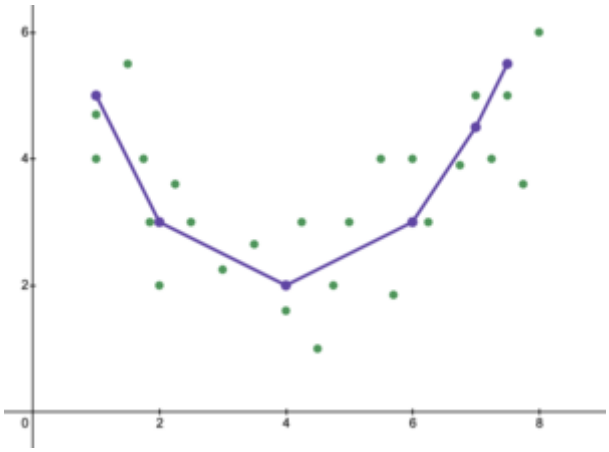

Figure 6. The blue line represents a fitted model of the data points represented in green.

Burnham & Anderson state the following.[3]:32

> … an underfitted model would ignore some important replicable (i.e., conceptually replicable in most other samples) structure in the data and thus fail to identify effects that were actually supported by the data. In this case, bias in the parameter estimators is often substantial, and the sampling variance is underestimated, both factors resulting in poor confidence interval coverage. Underfitted models tend to miss important treatment effects in experimental settings.

Resolving underfitting

Resolving underfitting can be handled in multiple ways, a possible method could be to increase the model's parameters, or to add more training data. Adding more training data could be obtained from getting new features from the current features (known as Feature engineering). Another possible method would be to move away from the current statistical model or machine learning algorithm to a different one that could better represent the data.

See also

- Bias–variance tradeoff
- Curve fitting
- Data dredging
- Feature selection
- Feature engineering
- Freedman's paradox

- Generalization error
- Goodness of fit
- Life-time of correlation
- Model selection
- Occam's razor
- Primary model
- VC dimension – larger VC dimension implies larger risk of overfitting

Notes

1. Definition of "overfitting (https://en.oxforddictionaries.com/definition/overfitting)" at OxfordDictionaries.com: this definition is specifically for statistics.
2. Everitt B.S., Skrondal A. (2010), *Cambridge Dictionary of Statistics*, Cambridge University Press.
3. Burnham, K. P.; Anderson, D. R. (2002), *Model Selection and Multimodel Inference* (2nd ed.), Springer-Verlag.
4. Claeskens, G.; Hjort, N.L. (2008), *Model Selection and Model Averaging*, Cambridge University Press.
5. Harrell, F. E., Jr. (2001), *Regression Modeling Strategies*, Springer.
6. Martha K. Smith (2014-06-13). "Overfitting" (http://www.ma.utexas.edu/users/mks/statmistakes/ovefitting.html). University of Texas at Austin. Retrieved 2016-07-31.
7. Vittinghoff, E.; McCulloch, C. E. (2007). "Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression" (https://doi.org/10.1093%2Faje%2Fkwk052). *American Journal of Epidemiology*. **165** (6): 710–718. doi:10.1093/aje/kwk052 (https://doi.org/10.1093%2Faje%2Fkwk052). PMID 17182981 (https://pubmed.ncbi.nlm.nih.gov/17182981).
8. Draper, Norman R.; Smith, Harry (1998). *Applied Regression Analysis* (3rd ed.). Wiley. ISBN 978-0471170822.
9. Jim Frost (2015-09-03). "The Danger of Overfitting Regression Models" (http://blog.minitab.com/blog/adventures-in-statistics/the-danger-of-overfitting-regression-models). Retrieved 2016-07-31.
10. Hawkins, Douglas M (2004). "The problem of overfitting". *Journal of Chemical Information and Modeling*. **44** (1): 1–12. doi:10.1021/ci0342472 (https://doi.org/10.1021%2Fci0342472). PMID 14741005 (https://pubmed.ncbi.nlm.nih.gov/14741005).

References

- Leinweber, D. J. (2007). "Stupid data miner tricks". *The Journal of Investing*. **16**: 15–22. doi:10.3905/joi.2007.681820 (https://doi.org/10.3905%2Fjoi.2007.681820). S2CID 108627390 (https://api.semanticscholar.org/CorpusID:108627390).
- Tetko, I. V.; Livingstone, D. J.; Luik, A. I. (1995). "Neural network studies. 1. Comparison of Overfitting and Overtraining" (http://www.vcclab.org/articles/jcics-overtraining.pdf) (PDF). *Journal of Chemical Information and Modeling*. **35** (5): 826–833. doi:10.1021/ci00027a006 (https://doi.org/10.1021%2Fci00027a006).
- *Tip 7: Minimize overfitting*. Chicco, D. (December 2017). "Ten quick tips for machine learning in computational biology" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5721660). *BioData Mining*. **10** (35): 35. doi:10.1186/s13040-017-0155-3 (https://doi.org/10.1186%2Fs13040-017-0155-3). PMC 5721660 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5721660). PMID 29234465 (https://pubmed.ncbi.nlm.nih.gov/29234465).

Further reading

- Christian, Brian; Griffiths, Tom (April 2017), "Chapter 7: Overfitting", *Algorithms To Live By: The computer science of human decisions*, William Collins, pp. 149–168, ISBN 978-0-00-754799-9

External links

- Overfitting: when accuracy measure goes wrong (http://blog.lokad.com/journal/2009/4/22/overfitting-when-accuracy-measure-goes-wrong.html) – introductory video tutorial
- The Problem of Overfitting Data (http://www3.cs.stonybrook.edu/~skiena/jaialai/excerpts/node16.html) – Stony Brook University
- What is "overfitting," exactly? (https://statmodeling.stat.columbia.edu/2017/07/15/what-is-overfitting-exactly/) – Andrew Gelman blog
- CSE546: Linear Regression Bias / Variance Tradeoff (http://courses.cs.washington.edu/courses/cse546/12wi/slides/cse546wi12LinearRegression.pdf) – University of Washington
- Underfitting and Overfitting in machine learning and how to deal with it !!! (https://towardsdatascience.com/underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6fe4a8a49dbf) – Towards Data Science
- What is Underfitting (https://www.ibm.com/cloud/learn/underfitting) – IBM
- ML | Underfitting and Overfitting (https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/) – Geeks for Geeks article - Dewang Nautiyal