STAT 479: Machine Learning
Lecture Notes

Sebastian Raschka
Department of Statistics
University of Wisconsin–Madison

http://stat.wisc.edu/~sraschka/teaching/stat479-fs2019/

Fall 2019

Contents

STAT 479: Machine Learning
Lecture Notes

Sebastian Raschka
Department of Statistics
University of Wisconsin–Madison

http://stat.wisc.edu/~sraschka/teaching/stat479-fs2019/

Fall 2019

8 Model Evaluation 1: Overfitting and Underfitting

8.1 Overview

- In this lecture, we discuss some of the basic terms and machine learning fundamentals that are relevant for model evaluation, namely, *bias and variance*, and *overfitting and underfitting*.

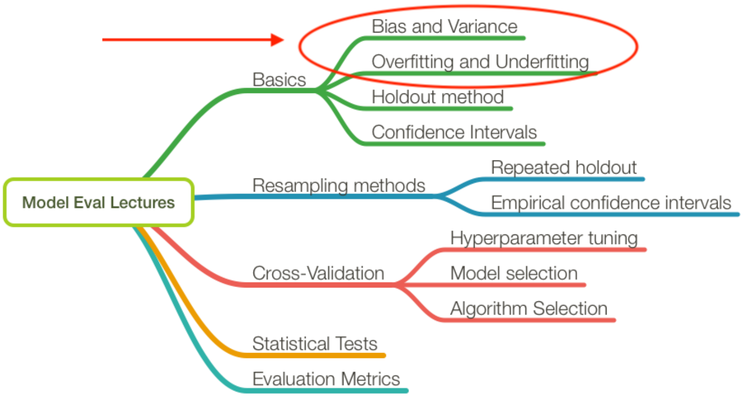


Figure 1: Overview of topics being covered in this lecture in the context of topics related to model evaluation that we will cover at a later point in time.

8.2 Overfitting and Underfitting

- The overall goal in machine learning is to obtain a model/hypothesis that generalizes well to new, unseen data.

- In other words, we want a model that *generalizes* well to unseen data, which we can measure, for example, by using an independent test set – while it sounds like this should be very straightforward, there are some pitfalls which we will discuss in the next lecture.
- Some of the evaluation metrics we can use to measure the performance on the test set are the prediction accuracy and misclassification error in the context of classification models – we say that a good model has a "high generalization accuracy" or "low generalization error" (or, simply "good generalization performance").
- The assumptions we generally make are the following:
  - i.i.d. assumption: inputs are independent, and training and test examples are identically distributed (drawn from the same probability distribution).
  - For some random model that has not been fit to the training set, we expect both the training and test error to be equal.
  - The training error or accuracy provides an (optimistically) biased estimate of the generalization performance.

Now, *overfitting* and *underfitting* are two terms that we can use to diagnose a machine learning model based on the training and test set performance. I.e., a model that suffers from underfitting does *not* perform well on the test *and* training set. In contrast, a model that overfits (e.g., from fitting the noise in the training dataset) can be usually recognized by a high training set accuracy, but low test set accuracy. Intuitively, as a rule of thumb, the larger the hypothesis space a model has access to, the higher the risk of overfitting.

A more technical term for the size of the hypothesis space is the so-called *capacity*. There are different measures for specific models and datasets that can be used to calculate the capacity of the model such as the VC dimension[1][2] (however, topics from learning theory such as VC dimension are beyond the scope of this course).

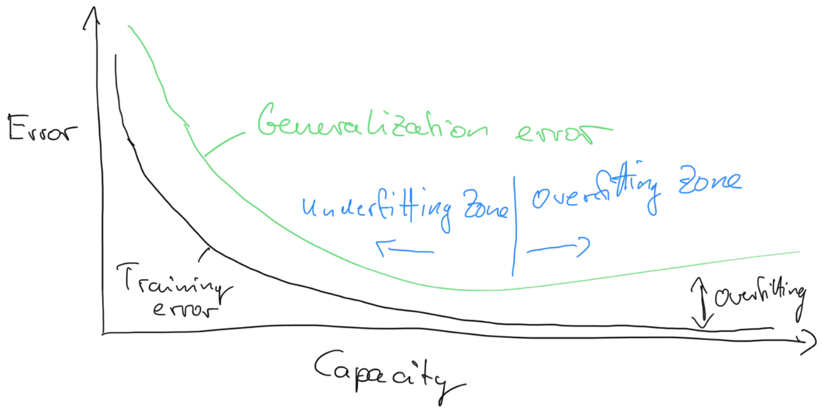


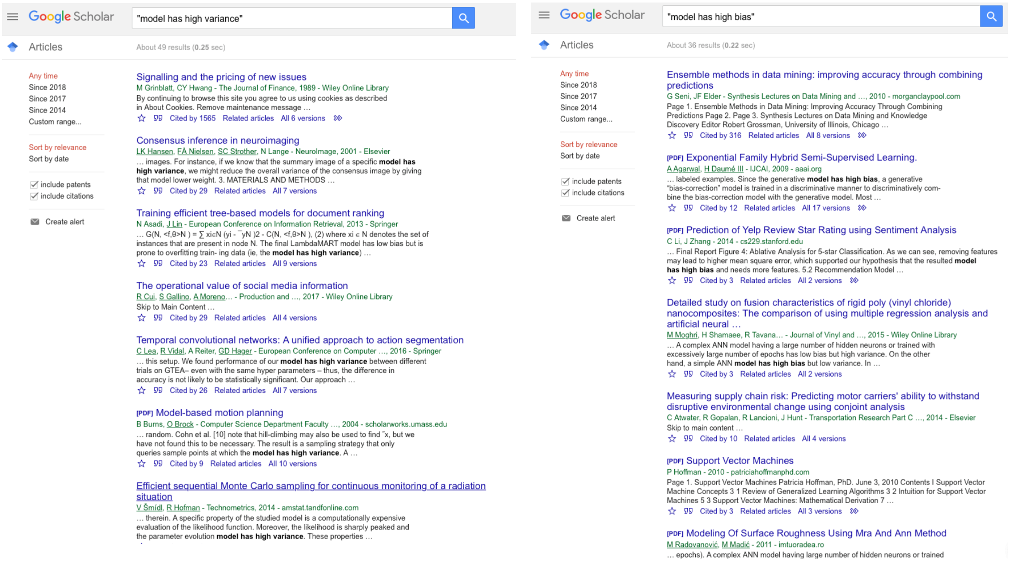
Figure 2: Illustration of overfitting and underfitting in relation to the training and test error.

[1] VC dimension stands for Vapnik-Chervonenkis dimension.

[2] Vladimir N Vapnik and A Ya Chervonenkis. "On the uniform convergence of relative frequencies of events to their probabilities". In: *Measures of complexity*. Springer, 2015, pp. 11–30.

8.3 Bias and Variance

Often, researchers use the terms *bias* and *variance* or "bias-variance tradeoff" to describe the performance of a model – i.e., you may stumble upon talks, books, or articles where people say that a model has a high variance or high bias. So, what does that mean? In general, we might say that "high variance" is proportional to overfitting, and "high bias" is proportional to underfitting. However, in this lecture, we are going to define these terms more precisely.



Signalling and the pricing of new issues
M Grinblatt, CY Hwang - The Journal of Finance, 1989 - Wiley Online Library
By continuing to browse this site you agree to us using cookies as described in About Cookies. Remove maintenance message …
☆ 99 Cited by 1565 Related articles All 6 versions

Consensus inference in neuroimaging
LK Hansen, FÅ Nielsen, SC Strother, N Lange - NeuroImage, 2001 - Elsevier
… images. For instance, if we know that the summary image of a specific **model has high variance**, we might reduce the overall variance of the consensus image by giving that model lower weight. 3. MATERIALS AND METHODS …
☆ 99 Cited by 29 Related articles All 7 versions

Training efficient tree-based models for document ranking
N Asadi, J Lin - European Conference on Information Retrieval, 2013 - Springer
… G(N, <f,θ>N) = ∑ xi∊N (yi - ¯yN)2 - C(N, <f,θ>N), (2) where xi ∊ N denotes the set of instances that are present in node N. The final LambdaMART model has low bias but is prone to overfitting train- ing data (ie, the **model has high variance**) …
☆ 99 Cited by 23 Related articles All 9 versions

The operational value of social media information
R Cui, S Gallino, A Moreno… - Production and …, 2017 - Wiley Online Library
Skip to Main Content …
☆ 99 Cited by 29 Related articles All 4 versions

Temporal convolutional networks: A unified approach to action segmentation
C Lea, R Vidal, A Reiter, GD Hager - European Conference on Computer …, 2016 - Springer
… this setup. We found performance of our **model has high variance** between different trials on GTEA– even with the same hyper parameters – thus, the difference in accuracy is not likely to be statistically significant. Our approach …
☆ 99 Cited by 26 Related articles All 7 versions

[PDF] Model-based motion planning
B Burns, O Brock - Computer Science Department Faculty …, 2004 - scholarworks.umass.edu
… random. Cohn et al. [10] note that hill-climbing may also be used to find ˜x, but we have not found this to be necessary. The result is a sampling strategy that only queries sample points at which the **model has high variance**. A …
☆ 99 Cited by 9 Related articles All 10 versions

Efficient sequential Monte Carlo sampling for continuous monitoring of a radiation situation
V Šmídl, R Hofman - Technometrics, 2014 - amstat.tandfonline.com
… therein. A specific property of the studied model is a computationally expensive evaluation of the likelihood function. Moreover, the likelihood is sharply peaked and the parameter evolution **model has high variance**. These properties …

Ensemble methods in data mining: improving accuracy through combining predictions
G Seni, JF Elder - Synthesis Lectures on Data Mining and …, 2010 - morganclaypool.com
Page 1. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions Page 2. Page 3. Synthesis Lectures on Data Mining and Knowledge Discovery Editor Robert Grossman, University of Illinois, Chicago …
☆ 99 Cited by 316 Related articles All 8 versions

[PDF] Exponential Family Hybrid Semi-Supervised Learning.
A Agarwal, H Daumé III - IJCAI, 2009 - aaai.org
… labeled examples. Since the generative **model has high bias**, a generative "bias-correction" model is trained in a discriminative manner to discriminatively combine the bias-correction model with the generative model. Most …
☆ 99 Cited by 12 Related articles All 17 versions

[PDF] Prediction of Yelp Review Star Rating using Sentiment Analysis
C Li, J Zhang - 2014 - cs229.stanford.edu
… Final Report Figure 4: Ablative Analysis for 5-star Classification. As we can see, removing features may lead to higher mean square error, which supported our hypothesis that the resulted **model has high bias** and needs more features. 5.2 Recommendation Model …
☆ 99 Cited by 3 Related articles All 2 versions

Detailed study on fusion characteristics of rigid poly (vinyl chloride) nanocomposites: The comparison of using multiple regression analysis and artificial neural …
M Moghri, H Shamaee, R Tavana… - Journal of Vinyl and …, 2015 - Wiley Online Library
… A complex ANN model having a large number of hidden neurons or trained with excessively large number of epochs has low bias but high variance. On the other hand, a simple ANN **model has high bias** but low variance. In …
☆ 99 Cited by 3 Related articles All 2 versions

Measuring supply chain risk: Predicting motor carriers' ability to withstand disruptive environmental change using conjoint analysis
C Atwater, R Gopalan, R Lancioni, J Hunt - Transportation Research Part C …, 2014 - Elsevier
Skip to main content …
☆ 99 Cited by 10 Related articles All 4 versions

[PDF] Support Vector Machines
P Hoffman - 2010 - patriciahoffmanphd.com
Page 1. Support Vector Machines Patricia Hoffman, PhD. June 3, 2010 Contents I Support Vector Machine Concepts 3 1 Review of Generalized Learning Algorithms 3 2 Intuition for Support Vector Machines 5 3 Support Vector Machines: Mathematical Derivation 7 …
☆ 99 Cited by 3 Related articles All 3 versions

[PDF] Modeling Of Surface Roughness Using Mra And Ann Method
M Radovanović, M Madić - 2011 - imtuoradea.ro
… epochs). A complex ANN model having large number of hidden neurons or trained

Figure 3: Results from searching the terms "model has high variance" and "model has high bias" on GoogleScholar

Note that the so-called Bias-Variance decomposition we are talking about in this lecture was initially formulated for regression losses (i.e., mean squared error[3]); however, we are also going to look into formulations for the 0-1 loss that we use to measure the misclassification error (or accuracy).

- Why are we doing this? The Decomposition of the loss into bias and variance help us understand learning algorithms, concepts are correlated to underfitting and overfitting.
- Thinking back of the ensemble lecture, the bias-variance decomposition and tradeoff help explain why ensemble methods might perform better than single models (i.e., why bagging reduces the variance, and why a boosting model has a lower bias than individual weak learners like decision tree stumps).

[3] In particular, in statistics, we evaluate the goodness of an estimator in relation to the true parameter or function

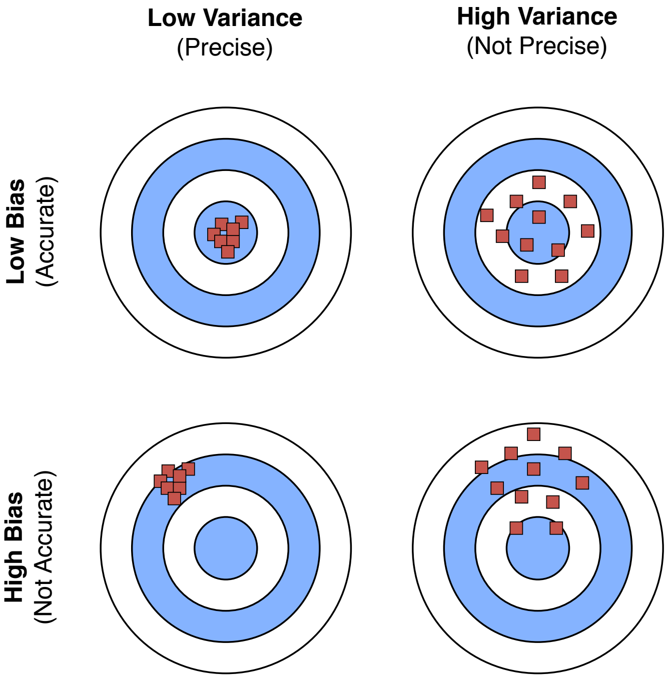


Figure 4: Bias-variance intuition.

To use the more formal terms for bias and variance, assume we have a point estimator $\hat{\theta}$ of some parameter or function θ. Then, the bias is commonly defined as the difference between the expected value of the estimator and the parameter that we want to estimate:

$$\text{Bias} = E[\hat{\theta}] - \theta. \tag{1}$$

If the bias is larger than zero, we also say that the estimator is positively biased, if the bias is smaller than zero, the estimator is negatively biased, and if the bias is exactly zero, the estimator is unbiased. Similarly, we define the variance as the difference between the expected value of the squared estimator minus the squared expectation of the estimator:

$$\text{Var}(\hat{\theta}) = E\big[\hat{\theta}^2\big] - \bigg(E\big[\hat{\theta}\big]\bigg)^2. \tag{2}$$

Note that in the context of this lecture, it will be more convenient to write the variance in its alternative form:

$$\text{Var}(\hat{\theta}) = E[(E[\hat{\theta}] - \hat{\theta})^2]. \tag{3}$$

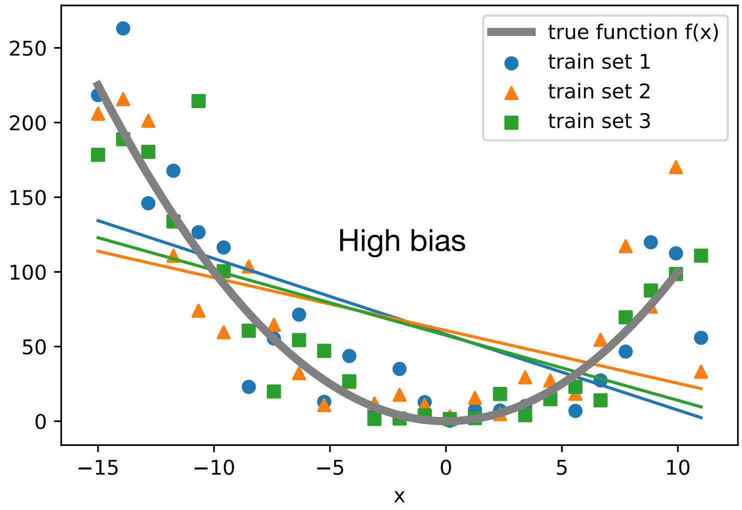


Figure 5: Suppose that there is an unknown target function or "true function" $f(x)$ which we want to approximate. Now, suppose we have different training sets drawn from an unknown distribution defined as "true function + noise." This plot shows different linear regression models, each fit to a different training set. None of these hypotheses approximate the true function well, except at two points (around x=-10 and x=6). In general, we can say that the bias is large because the difference between the true value and the predicted value, on average (here, average means "expectation of the training sets" not "expectation over examples in the training set"), is large (for most points).

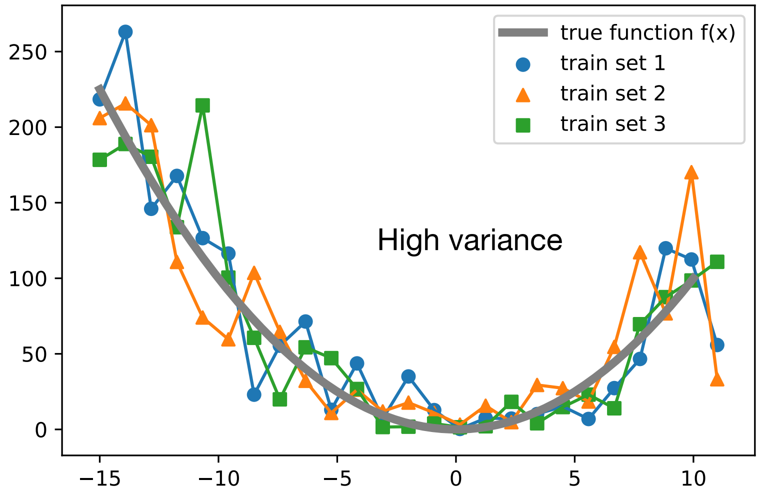


Figure 6: Suppose that there is an unknown target function or "true function" which we want to approximate. Now, suppose we have different training sets drawn from an unknown distribution defined as "true function + noise." This plot shows different unpruned decision tree models, each fit to a different training set. Note that these hypotheses fit the training data very closely. However, if we would consider the expectation over training sets, the average hypothesis would fit the true function perfectly (given that the noise is unbiased and has an expected value of 0). As we can see, the variance is very high, since on average, a prediction differs a lot from the expectation value of the prediction.

8.4 Bias-Variance Decomposition of the Squared Loss

We can decompose a loss function such as the squared loss into three terms, a variance, bias, and a noise term (and the same is true for the decomposition of the 0-1 loss later). However, for simplicity, we will ignore the noise term in this lecture (some of the literature referenced in later sections include the noise term if you are eager to learn more about variance-bias decomposition.).

Before we introduce the bias-variance decomposition of the 0-1 loss for classification, let us start with the decomposition of the squared loss as an easy warm-up exercise to get familiar with the overall concept.

The previous section already listed the common formal definitions of bias and variance, however, let us define them again for convenience:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta, \quad \text{Var}(\hat{\theta}) = E[(E[\hat{\theta}] - \hat{\theta})^2]. \tag{4}$$

Recall that in the context of these machine learning lecture (notes), we defined

- the true or target function as $y = f(x)$,
- the predicted target value as $\hat{y} = \hat{f}(x) = h(x)$,
- and the squared loss as $S = (y - \hat{y})^2$. (I use S here because it will be easier to tell it apart from the E, which we use for the *expectation* in this lecture.)

Note that unless noted otherwise, the expectation is over training sets!

To get started with the squared error loss decomposition into bias and variance, let use do some algebraic manipulation, i.e., adding and subtracting the expected value of $\hat{y}$ and then expanding the expression using the quadratic formula $(a + b)^2 = a^2 + b^2 + 2ab$):

$$\begin{aligned} S = (y - \hat{y})^2 && (5)\\ (y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 & (6)\\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - \hat{y})^2 + 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y}). & (7) \end{aligned}$$

Next, we just use the expectation on both sides, and we are already done:

$$\begin{aligned} E[S] &= E[(y - \hat{y})^2] & (8)\\ E[(y - \hat{y})^2] &= (y - E[\hat{y}])^2 + E[(E[\hat{y}] - \hat{y})^2] & (9)\\ &= [\text{Bias}]^2 + \text{Variance} & (10) \end{aligned}$$

You may wonder what happened to the "$2ab$" term $(2(y - E[\hat{y}])(E[\hat{y}] - \hat{y}))$ when we used the expectation. It turns that it evaluates to zero and hence vanishes from the equation, which can be shown as follows:

$$\begin{aligned} E[2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})] &= 2E[(y - E[\hat{y}])(E[\hat{y}] - \hat{y})] & (11)\\ &= 2(y - E[\hat{y}])E[(E[\hat{y}] - \hat{y})] & (12)\\ &= 2(y - E[\hat{y}])(E[E[\hat{y}]] - E[\hat{y}]) & (13)\\ &= 2(y - E[\hat{y}])(E[\hat{y}] - E[\hat{y}]) & (14)\\ &= 0. & (15) \end{aligned}$$

So, this is the canonical decomposition of the squared error loss into bias and variance. The next section will discuss some approaches that have been made to decompose the 0-1 loss that we commonly use for classification accuracy or error.

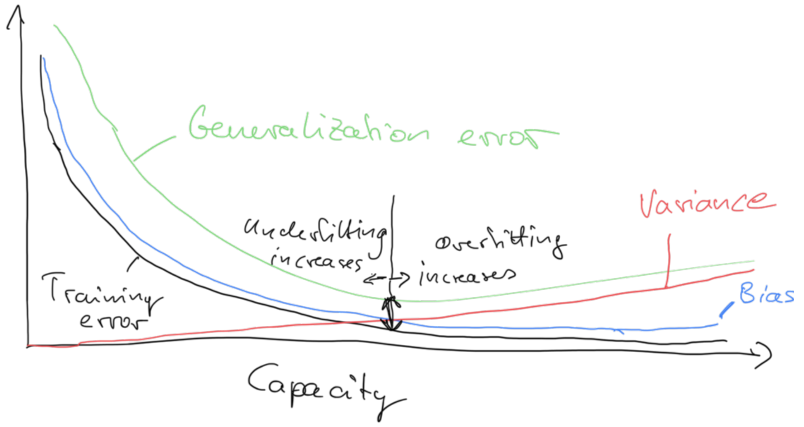


Figure 7: A sketch of variance and bias in relation to the training error and generalization error – how high variance related to overfitting, and how large bias relates to underfitting.

8.5 Bias-Variance Decomposition of the 0-1 Loss

Note that decomposing the 0-1 loss into bias and variance components is not as straightforward as for the squared error loss. To quote Pedro Domingos, a well-known machine learning researcher and professor at University of Washington: "several authors have proposed bias-variance decompositions related to zero-one loss (Kong & Dietterich, 1995; Breiman, 1996b; Kohavi & Wolpert, 1996; Tibshirani, 1996; Friedman, 1997). However, each of these decompositions has significant shortcomings."[4]. In fact, the paper this quote was taken from may offer the most intuitive and general formulation at this point. However, we will first, for simplicity, go over Kong & Dietterich formulation[5] of the 0-1 loss decomposition, which is the same as Domingos's but excluding the noise term (for simplicity).

The table below summarizes the relevant terms we used for the squared loss in relation to the 0-1 loss. Recall that the 0-1 loss, L, is 0 if a class label is predicted correctly, and it is 1, otherwise. The main prediction for the squared error loss is simply the average over the predictions $E[\hat{y}]$ (the expectation is over training sets), for the 0-1 loss Kong & Dietterich and Domingos defined it as the mode. I.e., if a model predicts the label one more than 50% of the time (considering all possible training sets), then the main prediction is 1, and 0 otherwise.

	Squared Loss	0-1 Loss
Single loss	$(y-\hat{y})^2$	$L(y,\hat{y})$
Expected loss	$E[(y-\hat{y})^2]$	$E[L(y,\hat{y})]$
Main prediction $E[\hat{y}]$	mean (average)	mode
Bias^2	$(y-E[\hat{y}])^2$	$L(y,E[\hat{y}])$
Variance	$E[(E[\hat{y}]-\hat{y})^2]$	$E[L(\hat{y},E[\hat{y}])]$

[4] Pedro Domingos. "A unified bias-variance decomposition". In: *Proceedings of 17th International Conference on Machine Learning.* 2000, pp. 231–238.

[5] Thomas G Dietterich and Eun Bae Kong. *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms.* Tech. rep. Technical report, Department of Computer Science, Oregon State University, 1995.

Hence, as result from using the mode to define the main prediction of the 0-1 loss, the bias is 1 if the main prediction does not agree with the true label y, and 0 otherwise:

$$Bias = \begin{cases} 1 \text{ if } y \neq E[\hat{y}], \\ 0 \text{ otherwise.} \end{cases} \tag{16}$$

The variance of the 0-1 loss is defined as the probability that the predicted label does not match the main prediction:

$$Variance = P(\hat{y} \neq E[\hat{y}]). \tag{17}$$

Next, let us take a look at what happens to the loss if the bias is 0. Given the general definition of the loss, loss = bias + variance, if the bias is 0, then we define the loss as the variance:

$$Loss = 0 + Variance = Loss = P(\hat{y} \neq y) = Variance = P(\hat{y} \neq E[\hat{y}]). \tag{18}$$

In other words, if a model has zero bias, it's loss is entirely defined by the variance, which is intuitive if we think of variance in the context of being proportional overfitting.

The more surprising scenario is if the bias is equal to 1. If the bias is equal to 1, as explained by Pedro Domingos, the increasing the variance can decrease the loss, which is an interesting observation. This can be seen by first rewriting the 0-1 loss function as

$$Loss = P(\hat{y} \neq y) = 1 - P(\hat{y} = y). \tag{19}$$

(Note that we have not done anything new, yet.) Now, if we look at the previous equation of the bias, if the bias is 1, we have $y \neq E[\hat{y}]$. If y is not equal to the main prediction and y is is equal to $\hat{y}$, then $\hat{y}$ cannot be equal to the main prediction. Using the "inverse" ("1 minus"), we can then write the loss as

$$Loss = P(\hat{y} \neq y) = 1 - P(\hat{y} = y) = 1 - P(\hat{y} \neq E[\hat{y}]). \tag{20}$$

Since the bias is 1, the loss is hence defined as "loss = bias - variance" if the bias is 1 (or "loss = 1 - variance"). This might be quite unintuitive at first, but the explanations Kong, Dietterich, and Domingos offer was that if a model has a very high bias such that it main prediction is always wrong, increasing the variance can be beneficial, since increasing the variance would push the decision boundary, which might lead to some correct predictions just by chance then. In other words, for scenarios with high bias, increasing the variance can improve (decrease) the loss!

8.6 Conclusion

In this lecture, we decomposed the squared error loss into variance and bias terms and discussed how these components relate to overfitting and underfitting. Then, we referred to a bias-variance decomposition that Kong & Dietterich defined for the 0-1 loss. Pedro Domingos later generalized this further, including the noise term, which we did not discuss in this lecture. However, interested students are encouraged to read the original paper:

- Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).[6]

[6] https://homes.cs.washington.edu/~pedrod/bvd.pdf

Now, we should be more familiar with the terms bias and variance (or, as statistics students, consider this as a refresher) and how it relates to overfitting and underfitting if we say that a model has a high variance or high bias, respectively.

In the next lecture, we will take a closer look at the holdout method for model evaluation (and estimating the generalization performance). Also, we will discuss several methods for constructing confidence intervals.