CMSE 890-002: Mathematics of Deep Learning, MSU, Spring 2020

Lecture 07: Bias-Variance Trade-off

January 27, 2019

Lecturer: Matthew Hirn

4.2 Bias variance tradeoff

In the previous section we considered theoretical scenarios in which we had perfect knowledge of the data generating distribution $p_{X,Y}(x, y)$ and we were able to select any model $y = f(x)$ to fit the data. In practice, we only have a finite amount of data given by our training set $T = \{(x_i, y_i)\}_{i=1}^N$, and we can only select models from the model class $\mathcal{F}$ that we specify. This leads to two sources of additional error, beyond errors that may be impossible to avoid even with perfect knowledge (e.g., if the labels are noisy, as in the model given by (4)). These two additional sources of errors are:

1. *Bias:* Since we cannot pick any model that we wish, only models from the model class $\mathcal{F}$, there is the possibility that $\mathcal{F}$ does not contain a model that fully captures the relationship between the labels $y \in \mathcal{Y}$ and the data points $x \in \mathcal{X}$. In this case, even the best model from $\mathcal{F}$ will not optimally model this relationship, and the resulting error is called a bias error. For example, if we use linear regression to fit data in which there is a nonlinear relationship between y and x, then our machine learned model will be biased since it will not be able capture the nonlinearity in the data. Furthermore, it may be that $\mathcal{F}$ contains a good model, but we are incapable of selecting it with the training set T that we are given; this will also lead to bias error.

2. *Variance:* Since we do not have complete knowledge of $p_{X,Y}(x, y)$ and in fact only have a finite training set T sampled according to $p_{X,Y}(x, y)$, we must estimate the generalization error and select the best model using only T. However, different (theoretical) draws of the training set T will lead to different models being selected, some of which may generalize to new points better than others. This randomness imparted into the model selection process can either be relatively small or drastic. The variance error encodes this source of error.

Figure 14 illustrates the difference between model classes $\mathcal{F}$ with low bias and high variance, versus those with high bias and low variance. Indeed, these two sources of error are often in tension, i.e., reducing one increases the other. This phenomenon is referred to as the *bias-variance trade-off*. One way to think about this tradeoff is as follows. Simple models, such as linear regression, may not fit the training data perfectly (or even well), and hence have a high bias, but their predictions are robust to small perturbations in the test points, and thus have a low variance. On the other hand, complex models such as the 1-nearest

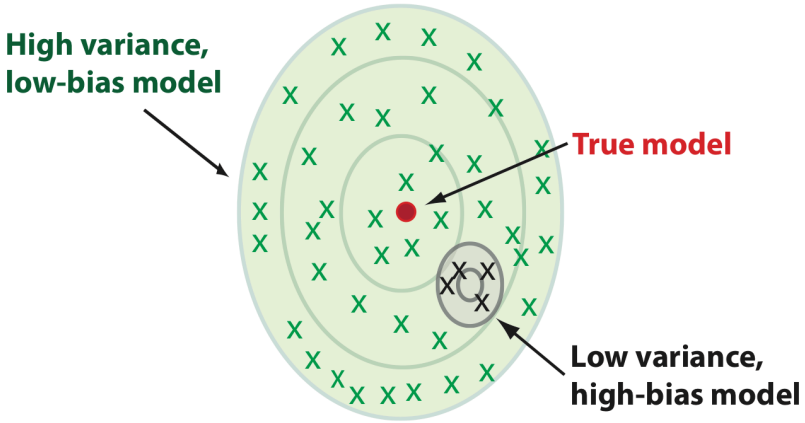

Figure 14: *Illustration of high variance, low bias model classes and low variance, high bias model classes. Each "X" denotes a model obtained through a particular draw of the training set T. In the high variance, low bias regime, indicated by the green X's, the average of all the models is close to the true model, marked with the red dot. However, there is a large variance between models as one varies the training set, meaning that any one model may be very far from the true model. In the low variance, high bias regime, indicated with the black X's, the models are not centered around the true model and thus the average model obtained from them will not be a good approximation of the true model. On the other hand, different draws of the training set lead to nearly the same model, as indicated by their tight arrangement.*

neighbor classifier may be able to fit the training data extremely well (low bias), but their high complexity means they may fit spurious patterns in the data (such as noise) and their output may change drastically with small changes on the evaluated data point, so much so that their predictive power is limited (high variance). Models, such as the $k = 15$ nearest neighbor classifier (see Figure 11), that balance reducing bias with increasing variance are the goal in predictive machine learning.

Let us now derive the bias variance tradeoff in a general statistical setting using the squared loss. For this we will assume there exists a deterministic function $F : \mathbb{R}^d \to \mathbb{R}$ that determines the uncorrupted label of $x \in \mathbb{R}^d$, and that the labels we observe are corrupted by white noise:

$$y_i = F(x_i) + \varepsilon_i \,.$$

The variables ε_i independently and identically distributed normal random variables with mean zero and variance σ^2, i.e.,

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \,,$$

which implies, in particular, that $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] = \sigma^2$.

Given a training set T drawn from $p_{X,Y}(x, y)$ and a parameterized model class $\mathcal{F} = \{f(\cdot; \theta) : \theta \in \mathbb{R}^n\}$, we obtain a model by minimizing the squared loss (empirical risk) over

the training set:

$$\widehat{\theta}_T = \arg\min_{\theta \in \mathbb{R}^n} R_{\text{emp}}(f(\cdot;\theta)) = \arg\min_{\theta \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i;\theta))^2 \,,$$

where the subscript T emphasizes that $\widehat{\theta}_T$ depends upon the particular training set used to solve for the model. Given this model, we define the *conditional test error*, which is expected test error given the draw of the training set T:

$$\text{err}(\mathcal{F}, T) = \mathbb{E}_{X,Y}[(Y - f(X;\widehat{\theta}_T))^2] \,.$$

Note that $\text{err}(\mathcal{F}, T)$ is often a quantity we are very interested in, as it encodes the ability of our selected model $f(\cdot;\widehat{\theta}_T)$, obtained through our specific training set T, to generalize to new points. A related quantity is the *expected test error*, which is defined as:

$$\text{Err}(\mathcal{F}) = \mathbb{E}_T[\text{err}(\mathcal{F}, T)] \,.$$

The expectation $\mathbb{E}_T$ is an expectation over all possible training sets T with N elements, drawn according to $p_{X,Y}(x, y)$. Note that the expected test error, $\text{Err}(\mathcal{F})$, measures the quality of the model class $\mathcal{F}$ and not any particular model $f(x;\widehat{\theta}_T)$ selected from it using a specific training set T.

Figure 15 shows that minimizing the empirical risk by increasing model complexity is not a good way to minimize the conditional test error or the expected test error. Indeed, the empirical risk decreases with model complexity assuming that increasing complexity allows us to obtain increasingly better approximations of F, or to even include F at some point. However, such complex models, particularly when the model is more complex than F, generalize poorly as they fit spurious patterns generated by the additive noise ε_i, which is reflected in the increase of the conditional and expected test errors at a certain point.

Let us now derive the bias variance tradeoff, which will give a quantitative interpretation for the empirical results we have observed. The result will decompose the expected test error at an arbitrary, but fixed point $x \in \mathcal{X}$. We thus define the *expected test error at* x as:

$$\begin{aligned} \text{Err}(\mathcal{F}, x) &= \mathbb{E}_T \mathbb{E}_{X,Y}[(Y - f(X;\widehat{\theta}_T))^2 \mid X = x] \\ &= \mathbb{E}_T \mathbb{E}_{Y|X}[(Y - f(X;\widehat{\theta}_T))^2 \mid X = x] \end{aligned}$$

Theorem 4.2 (Bias variance tradeoff)**.** *Let $T = \{(x_i, y_i)\} \subset \mathbb{R}^d \times \mathbb{R}$ be an arbitrary training set drawn from a joint distribution $p_{X,Y}(x, y)$ with $y_i = F(x_i) + \varepsilon_i$ for some deterministic function $F : \mathbb{R}^d \to \mathbb{R}$ and iid random variables $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Then the expected test error at $x \in \mathcal{X}$ can be decomposed as:*

$$\text{Err}(\mathcal{F}, x) = \sigma^2 + (F(x) - \mathbb{E}_T[f(x;\widehat{\theta}_T)])^2 + \mathbb{E}_T[(f(x;\widehat{\theta}_T) - \mathbb{E}_T[f(x;\widehat{\theta}_T)])^2] \,.$$

Theorem 4.2 decomposes the test error at an arbitrary point $x \in \mathcal{X}$ into three components:

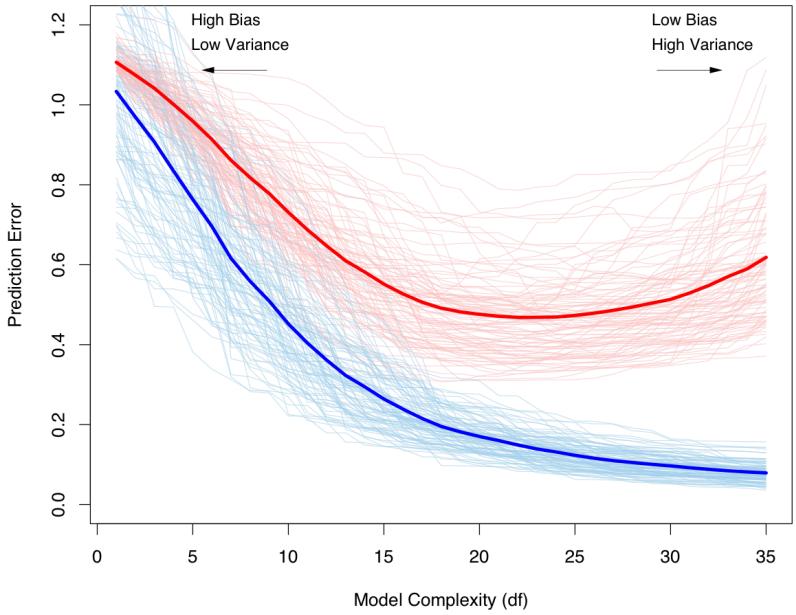

Figure 15: *Behavior of test error and training error as the model class complexity is increased. The plot depicts* 100 *different training sets drawn from the same distribution. The light blue curves show the empirical risk* $R_{emp}(f(\cdot;\widehat{\theta}_T))$ *for each of the* 100 *training sets,* T*. The light red curves show the corresponding conditional test error,* $\text{err}_T(\mathcal{F}, f(\cdot;\widehat{\theta}_T))$ *for each of the training set. The dark blue curve is the expected empirical risk over all draws of the training set,* $\mathbb{E}_T[R_{emp}(f(\cdot;\widehat{\theta}_T))]$*, estimated by averaging the light blue curves. The dark red curve is the expected test error* $\text{Err}(\mathcal{F})$*, estimated by averaging the light red curves.*

- The irreducible noise error: σ^2
- The bias induced by model class: $F(x) - \mathbb{E}_T[f(x;\widehat{\theta}_T)]$
- The variance of the selected model from the model class: $\mathbb{E}_T[(f(x;\widehat{\theta}_T) - \mathbb{E}_T[f(x;\widehat{\theta}_T)])^2]$

The irreducible noise error is an error that is incurred regardless of the model class used to fit the training data; it is a fundamental limitation of any model learned from training data generated by $p_{X,Y}(x,y)$. The bias measures the capacity of the model class to contain a model that fits the underlying function $F(x)$ that generates the labels. The bias error will decrease (or at least stay flat) as the complexity of the model class increases. The variance measures the stability of the selected model, $f(x;\widehat{\theta}_T)$, over different draws of the training set T. If the variance is low, then the model class is stable and different draws of the training set (theoretical or not) will not influence the learned model too much. On the other hand, if the variance is high, then the model is unstable and a new training set may result in a very different model. Since the underlying, noiseless label $F(x)$ is deterministic, this may result

in poor predictions because it is likely to get "unlucky" with a high variance model class. Figure 14 illustrates these ideas. Let us now prove the theorem.

Proof. We write $Y(x) = F(x) + \varepsilon$ to denote the random variable Y conditioned on $X = x$ for some $x \in \mathcal{X}$. Our first goal is to extract the irreducible noise error through the following calculation:

$$\begin{aligned}
&\mathbb{E}_{Y|X}\mathbb{E}_T[(Y - f(X;\widehat{\theta}_T))^2 \mid X = x] \\
&\quad = \mathbb{E}_{Y|X}\mathbb{E}_T[(Y(x) - f(x;\widehat{\theta}_T))^2] \\
&\quad = \mathbb{E}_{Y|X}\mathbb{E}_T[(Y(x) - F(x) + F(x) - f(x;\widehat{\theta}_T))^2] \\
&\quad = \mathbb{E}_{Y|X}\mathbb{E}_T[(Y(x) - F(x))^2] \\
&\qquad + \mathbb{E}_{Y|X}\mathbb{E}_T[(F(x) - f(x;\widehat{\theta}_T))^2] \\
&\qquad + 2\mathbb{E}_{Y|X}\mathbb{E}_T[(Y(x) - F(x))(F(x) - f(x;\widehat{\theta}_T))] \\
&\quad = \mathbb{E}_\varepsilon[\varepsilon^2] + \mathbb{E}_T[(F(x) - f(x;\widehat{\theta}_T))^2] + 2\mathbb{E}_\varepsilon\mathbb{E}_T[\varepsilon(F(x) - f(x;\widehat{\theta}_T))] \\
&\quad = \sigma^2 + \mathbb{E}_T[(F(x) - f(x;\widehat{\theta}_T))^2] + 2\mathbb{E}_\varepsilon[\varepsilon]\mathbb{E}_T[F(x) - f(x;\widehat{\theta}_T)] \\
&\quad = \sigma^2 + \mathbb{E}_T[(F(x) - f(x;\widehat{\theta}_T))^2]
\end{aligned}$$

where we used $\mathbb{E}[\varepsilon] = 0$, $\mathbb{E}[\varepsilon^2] = \sigma^2$, and the fact that the noise ε on the label of a test point x is independent of the training set T.

Now we further decompose the second term into the bias term and the variance term:

$$\begin{aligned}
&\mathbb{E}_T[(F(x) - f(x;\widehat{\theta}_T))^2] \\
&\quad = \mathbb{E}_T[(F(x) - \mathbb{E}_T[f(x;\widehat{\theta}_T)] + \mathbb{E}_T[f(x;\widehat{\theta}_T)] - f(x;\widehat{\theta}_T))^2] \\
&\quad = \mathbb{E}_T[(F(x) - \mathbb{E}_T[f(x;\widehat{\theta}_T))^2] + \mathbb{E}_T[(f(x;\widehat{\theta}_T) - \mathbb{E}_T[f(x;\widehat{\theta}_T)])^2] \\
&\qquad + \mathbb{E}_T[(F(x) - \mathbb{E}_T[f(x;\widehat{\theta}_T)])(f(x;\widehat{\theta}_T) - \mathbb{E}_T[f(x;\widehat{\theta}_T)])] \\
&\quad = (F(x) - \mathbb{E}_T[f(x;\widehat{\theta}_T)])^2 + \mathbb{E}_T[(f(x;\widehat{\theta}_T) - \mathbb{E}_T[f(x;\widehat{\theta}_T)])^2] \\
&\qquad + (F(x) - \mathbb{E}_T[f(x;\widehat{\theta}_T)])\mathbb{E}_T[f(x;\widehat{\theta}_T) - \mathbb{E}_T[f(x;\widehat{\theta}_T)]] \\
&\quad = (F(x) - \mathbb{E}_T[f(x;\widehat{\theta}_T)])^2 + \mathbb{E}_T[(f(x;\widehat{\theta}_T) - \mathbb{E}_T[f(x;\widehat{\theta}_T)])^2]
\end{aligned}$$

The proof is thus completed. □

Let us now apply Theorem 4.2 to the k-nearest neighbors algorithm and the linear model. Starting with k-nearest neighbors, we obtain:

$$\mathrm{Err}(k, x) = \sigma^2 + \left(F(x) - \frac{1}{k}\sum_{x_i \in N_k(x)} (F(x_i) + \varepsilon_i)\right)^2 + \frac{\sigma^2}{k},$$

where we have assumed for simplicity that the training set is fixed and the randomness comes from the labels only. We see that the variance, σ^2/k, decreases as the number of neighbors

k increases (recall that larger k means a less complex model). On the other hand, larger k means that $F(x)$ is estimated from a larger neighborhood around x, potentially increasing the bias, especially for irregular functions F.

For linear models $f(x;\theta) = \langle x, \theta \rangle$, we obtain the following for the expected training error:

$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{Err}(\mathcal{F}_{\mathrm{linear}}, x_i) = \sigma^2 + \frac{1}{N}\sum_{i=1}^{N} (F(x_i) - \mathbb{E}_{\varepsilon}[f(x_i; \widehat{\theta}_T)])^2 + \frac{d}{N}\sigma^2 \,.$$

Here the model complexity increases with the dimension d, but the variance term only increases linearly in d. This observation motivates our discussion in the next section.

References

[1] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.

[4] Larry Greenemeier. AI versus AI: Self-taught AlphaGo Zero vanquishes its predecessor. *Scientific American*, October 18, 2017.

[5] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to Machine Learning for physicists. arXiv:1803.08823, 2018.

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York, 2nd edition, 2009.