Lecture 18

Local Methods

Sasha Rakhlin

Nov 07, 2018

Today: analysis of "local" procedures such as k-Nearest-Neighbors or local smoothing. Different bias-variance decomposition (we do not fix a class $\mathcal{F}$).

Analysis will rely on local similarity (e.g. Lipschitz-ness) of regression function f^*. Idea: to predict y at a given x, look up in the dataset those Y_i for which X_i is "close" to x.

Bias-Variance

It's time to revisit the bias-variance picture. Recall that our goal was to ensure that

$$\mathbb{E}\mathbf{L}(\widehat{f}_n) - \mathbf{L}(f^*)$$

decreases with data size n, where f^* gives smallest possible $\mathbf{L}$.

For "simple problems" (that is, strong assumptions on P), one can ensure this without the bias-variance decomposition. Examples: Perceptron, linear regression in $d < n$ regime, etc.

However, for more interesting problems, we cannot get this difference to be small in "one shot" because variance (fluctuation of the stochastic part) is too large. Instead, it is more beneficial to introduce a *biased procedure* in the hope to reduce variance.

Our approach so far was to split this term into an estimation-approximation error with respect to some class $\mathcal{F}$:

$$\mathbb{E}\mathbf{L}(\widehat{f}_n) - \mathbf{L}(f_{\mathcal{F}}) + \mathbf{L}(f_{\mathcal{F}}) - \mathbf{L}(f^*)$$

Bias-Variance

In this lecture, we study a different bias-variance decomposition, typically used in nonparametric statistics. We will only work with *square loss*.

Rather than fixing $\mathcal{F}$ that controls the estimation error, we fix an algorithm (procedure/estimator) $\widehat{f}_n$ that has some *tunable parameter*.

By definition $\mathbb{E}[Y|X = x] = f^*(x)$. Then we write

$$\begin{aligned}
\mathbb{E}\mathbf{L}(\widehat{f}_n) - \mathbf{L}(f^*) &= \mathbb{E}(\widehat{f}_n(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2 \\
&= \mathbb{E}(\widehat{f}_n(X) - f^*(X) + f^*(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2 \\
&= \mathbb{E}(\widehat{f}_n(X) - f^*(X))^2
\end{aligned}$$

because the cross term vanishes (check!)

Bias-Variance

Before proceeding, let us discuss the last expression.

$$\mathbb{E}(\widehat{f}_n(X) - f^*(X))^2 = \mathbb{E}_{\mathscr{S}} \int_x (\widehat{f}_n(x) - f^*(x))^2 P(dx)$$
$$= \int_x \mathbb{E}_{\mathscr{S}} (\widehat{f}_n(x) - f^*(x))^2 P(dx)$$

We will often analyze $\mathbb{E}_{\mathscr{S}}(\widehat{f}_n(x) - f^*(x))^2$ for fixed x and then integrate.

The integral is a measure of distance between two functions:

$$\|f - g\|^2_{L_2(P)} \triangleq \int_x (f(x) - g(x))^2 P(dx).$$

Bias-Variance

Let us drop $L_2(P)$ from notation for brevity. The bias-variance decomposition can be written as

$$\begin{aligned}\mathbb{E}\left\|\widehat{f}_n - f^*\right\|^2 &= \mathbb{E}\left\|\widehat{f}_n - \mathbb{E}_{Y_{1:n}}\left[\widehat{f}_n\right] + \mathbb{E}_{Y_{1:n}}\left[\widehat{f}_n\right] - f^*\right\|^2 \\ &= \mathbb{E}\left\|\widehat{f}_n - \mathbb{E}_{Y_{1:n}}\left[\widehat{f}_n\right]\right\|^2 + \mathbb{E}\left\|\mathbb{E}_{Y_{1:n}}\left[\widehat{f}_n\right] - f^*\right\|^2,\end{aligned}$$

because the cross term is zero in expectation.

The first term is variance, the second is squared bias. One "typically" increases with the parameter, the other decreases.

Parameter is chosen either (a) theoretically or (b) by cross-validation (this is the usual case in practice).

In the rest of the lecture, we will discuss several local methods and describe (in a hand-wavy manner) the behavior of bias and variance.

For more details, consult

- "Distribution-Free Theory of Nonparametric Regression," Györfi et al
- "Introduction to Nonparametric Estimation," Tsybakov

Outline

As before, we are given $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. from P. To make a prediction of Y at a given x, we sort points according to distance $\|X_i - x\|$. Let

$$(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$$

be the sorted list (remember this depends on x).

The k-NN estimate is defined as

$$\widehat{f}_n(x) = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}.$$

If support of X is bounded and $d \geq 3$, then one can estimate

$$\mathbb{E} \left\| X - X_{(1)} \right\|^2 \lesssim n^{-2/d}.$$

That is, we expect the closest neighbor of a random point X to be no further than $n^{-1/d}$ away from one of n randomly drawn points.

Variance: Given x,

$$\widehat{f}_n(x) - \mathbb{E}_{Y_{1:n}}[\widehat{f}_n(x)] = \frac{1}{k}\sum_{i=1}^{k}(Y_{(i)} - f^*(X_{(i)}))$$

which is on the order of $1/\sqrt{k}$. Then variance is of the order $\frac{1}{k}$.

Bias: a bit more complicated. For a given x,

$$\mathbb{E}_{Y_{1:n}}[\widehat{f}_n(x)] - f^*(x) = \frac{1}{k}\sum_{i=1}^{k}(f^*(X_{(i)}) - f^*(x)).$$

Suppose f^* is 1-Lipschitz. Then the square of above is

$$\left(\frac{1}{k}\sum_{i=1}^{k}(f^*(X_{(i)}) - f^*(x))\right)^2 \leq \frac{1}{k}\sum_{i=1}^{k}\left\|X_{(i)} - x\right\|^2$$

So, the bias is governed by how close the closest k random points are to x.

Claim: enough to know the upper bound on the closest point to x among n points.

Argument: for simplicity assume that $J = n/k$ is an integer. Divide the original (unsorted) dataset into k blocks, n/k size each. Let X^i be the closest point to x in ith block. Then the collection $X^1, \ldots, X^J$, a k-subset which is no closer than the set of k nearest neighbors. That is,

$$\frac{1}{k} \sum_{i=1}^{k} \left\| X_{(i)} - x \right\|^2 \leq \frac{1}{k} \sum_{i=1}^{k} \left\| X^i - x \right\|^2$$

Taking expectation (with respect to dataset), the bias term is at most

$$\mathbb{E} \left\{ \frac{1}{k} \sum_{i=1}^{k} \left\| X^i - x \right\|^2 \right\} = \mathbb{E} \left\| X^1 - x \right\|^2$$

which is expected squared distance from x to the closest point in a random set of n/k points. When we take expectation over X, this is at most

$$(n/k)^{-2/d}$$

Putting everything together, the bias-variance decomposition yields

$$\frac{1}{k} + \left(\frac{k}{n}\right)^{2/d}$$

Optimal choice is $k \sim n^{\frac{2}{2+d}}$ and the overall rate of estimation at a given point x is

$$n^{-\frac{2}{2+d}}.$$

Since the result holds for any x, the integrated risk is also

$$\mathbb{E}\left\|\widehat{f}_n - f^*\right\|^2 \lesssim n^{-\frac{2}{2+d}}.$$

Summary

- We sketched the proof that k-Nearest-Neighbors has sample complexity guarantees for prediction or estimation problems with square loss if k is chosen appropriately.
- Analysis is very different from "empirical process" approach for ERM.
- Truly nonparametric!
- No assumptions on underlying density (in $d \geq 3$) beyond compact support. Additional assumptions needed for $d \leq 3$.

Outline

Fix a kernel $K : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$. Assume K is zero outside unit Euclidean ball at origin (not true for e^{-x^2}, but close enough).

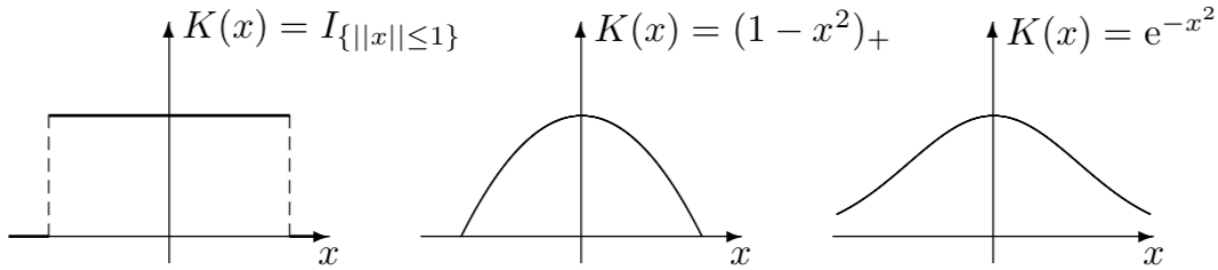

Figure 5.1. Examples for univariate kernels.

(figure from Györfi et al)

Let $K_h(x) = K(x/h)$, and so $K_h(x - x')$ is zero if $\|x - x'\| \geq h$.

h is "bandwidth" – tunable parameter.

Assume $K(x) > cI\{\|x\| \leq 1\}$ for some $c > 0$. This is important for the "averaging effect" to kick in.

Nadaraya-Watson estimator:

$$\widehat{f}_n(x) = \sum_{i=1}^{n} Y_i W_i(x)$$

with

$$W_i(x) = \frac{K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)}$$

(Note: $\sum_i W_i = 1$).

Unlike the k-NN example, bias is easier to estimate.

Bias: for a given x,

$$\mathbb{E}_{Y_{1:n}}[\widehat{f}_n(x)] = \mathbb{E}_{Y_{1:n}}\left[\sum_{i=1}^n Y_i W_i(x)\right] = \sum_{i=1}^n f^*(X_i) W_i(x)$$

and so

$$\mathbb{E}_{Y_{1:n}}[\widehat{f}_n(x)] - f^*(x) = \sum_{i=1}^n (f^*(X_i) - f^*(x)) W_i(x)$$

Suppose f^* is 1-Lipschitz. Since K_h is zero outside the h-radius ball,

$$|\mathbb{E}_{Y_{1:n}}[\widehat{f}_n(x)] - f^*(x)|^2 \leq h^2.$$

Variance: we have

$$\widehat{f}_n(x) - \mathbb{E}_{Y_{1:n}}[\widehat{f}_n(x)] = \sum_{i=1}^{n}(Y_i - f^*(X_i))W_i(x)$$

Expectation of square of this difference is at most

$$\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - f^*(X_i))^2 W_i(x)^2\right]$$

since cross terms are zero (fix X's, take expectation with respect to the Y's).

We are left analyzing

$$n\mathbb{E}\left[\frac{K_h(x - X_1)^2}{(\sum_{i=1}^{n} K_h(x - X_i))^2}\right]$$

Under some assumptions on density of X, the denominator is at least $(nh^d)^2$ with high prob, whereas $\mathbb{E}K_h(x - X_1)^2 = O(h^d)$ assuming $\int K^2 < \infty$. This gives an overall variance of $O(1/(nh^d))$. *Many* details skipped here (e.g. problems at the boundary, assumptions, etc)

Overall, bias and variance with $h \sim n^{-\frac{1}{2+d}}$ yield

$$h^2 + \frac{1}{nh^d} = n^{-\frac{2}{2+d}}$$

Summary

- Analyzed smoothing methods with kernels. As with nearest neighbors, slow (nonparametric) rates in large d.
- Same bias-variance decomposition approach as k-NN.

Outline

Let us revisit the following question: can a learning method be successful if it interpolates the data?

Consider the Nadaraya-Watson estimator. Take a kernel that approaches a large value τ at 0, e.g.

$$K(x) = \max\{1/\|x\|^{\alpha}, \tau\}$$

Note that large τ means $\widehat{f}_n(X_i) \approx Y_i$ since the weight $W_i(X_i)$ is large. In fact, if $\tau = \infty$, we get *interpolation* $\widehat{f}_n(X_i) = Y_i$ of all training data. Yet, the sketched proof still goes through. Hence, "memorizing the data" (governed by parameter τ) is completely decoupled from the bias-variance trade-off (as given by parameter h).

Contrast with conventional wisdom: fitting data too well means overfitting.

NB: Of course, we could always redefine any $\widehat{f}_n$ to be equal to Y_i on X_i, but our example shows more explicitly how memorization is governed by a parameter that is independent of bias-variance.

Bias-Variance and Overfitting

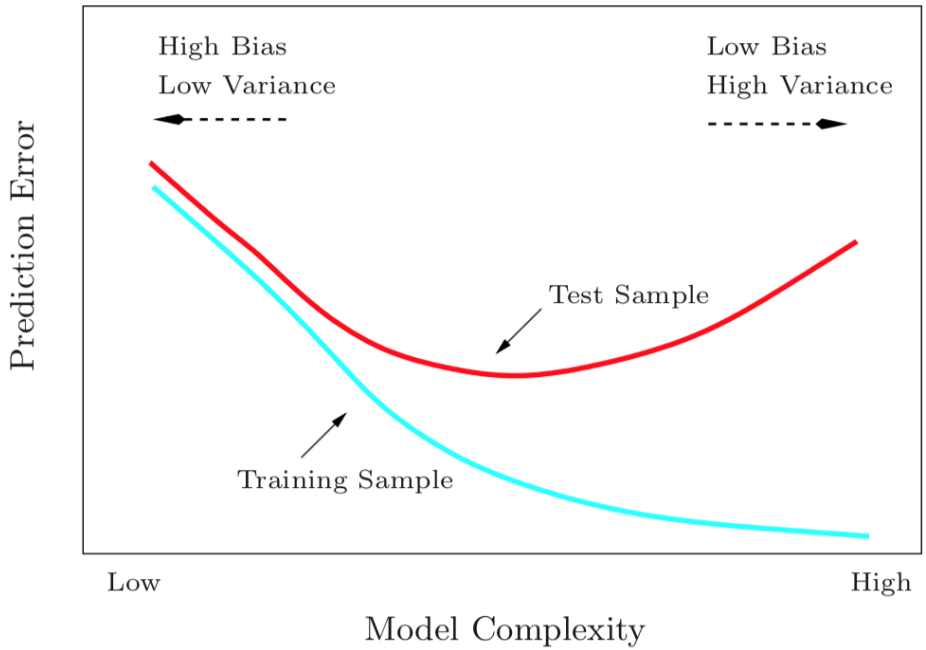

FIGURE 2.11. *Test and training error as a function of model complexity.*

Figure 2.11 shows the typical behavior of the test and training error, as model complexity is varied. The training error tends to decrease whenever we increase the model complexity, that is, whenever we fit the data harder. However with too much fitting, the model adapts itself too closely to the training data, and will not generalize well (i.e., have large test error). In

"Elements of Statistical Learning," Hastie, Tibshirani, Friedman

What is *overfitting*?

- Fitting data too well?
- Bias too low, variance too high?

Key takeaway: we should not conflate these two.