Learning Theory Part 3: Bias-Variance Tradeoff

Yingyu Liang
Computer Sciences 760
Fall 2017

http://pages.cs.wisc.edu/~yliang/cs760/

Some of the slides in these lectures have been adapted/borrowed from materials developed by Mark Craven, David Page, Jude Shavlik, Tom Mitchell, Nina Balcan, Matt Gormley, Elad Hazan, Tom Dietterich, and Pedro Domingos.

Goals for the lecture

you should understand the following concepts

- estimation bias and variance
- the bias-variance decomposition

Estimation bias and variance

- How will predictive accuracy (error) change as we vary k in k-NN?

- Or as we vary the complexity of our decision trees?

- the bias/variance decomposition of error can lend some insight into these questions

note that this is a different sense of bias
than in the term *inductive bias*

Background: Expected values

- the *expected value* of a random variable that takes on numerical values is defined as:

$$E[X] = \sum_x x\, P(x)$$

this is the same thing as the *mean*

- we can also talk about the expected value of a function of a random variable

$$E[g(X)] = \sum_x g(x) P(x)$$

Defining bias and variance

- consider the task of learning a regression model $f(\boldsymbol{x}; D)$ given a training set $D = \left\{ (x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)}) \right\}$

indicates the dependency of model on D

- a natural measure of the error of f is

$$E\left[\left(y - f(\boldsymbol{x}; D) \right)^2 \mid \boldsymbol{x}, D \right]$$

where the expectation is taken with respect to the real-world distribution of instances

Defining bias and variance

- this can be rewritten as:

$$E\left[\left(y - f(\boldsymbol{x}; D)\right)^2 \mid \boldsymbol{x}, D\right] = E\left[\left(y - E[y \mid \boldsymbol{x}]\right)^2 \mid \boldsymbol{x}, D\right]$$

$$+ \left(f(\boldsymbol{x}; D) - E[y \mid \boldsymbol{x}]\right)^2$$

error of f as a predictor of y

noise: variance of y given $\boldsymbol{x}$; doesn't depend on D or f

Defining bias and variance

- now consider the expectation (over different data sets D) for the second term

$$E_D\left[\left(f(\boldsymbol{x};D) - E[y\,|\,\boldsymbol{x}]\right)^2\right] =$$

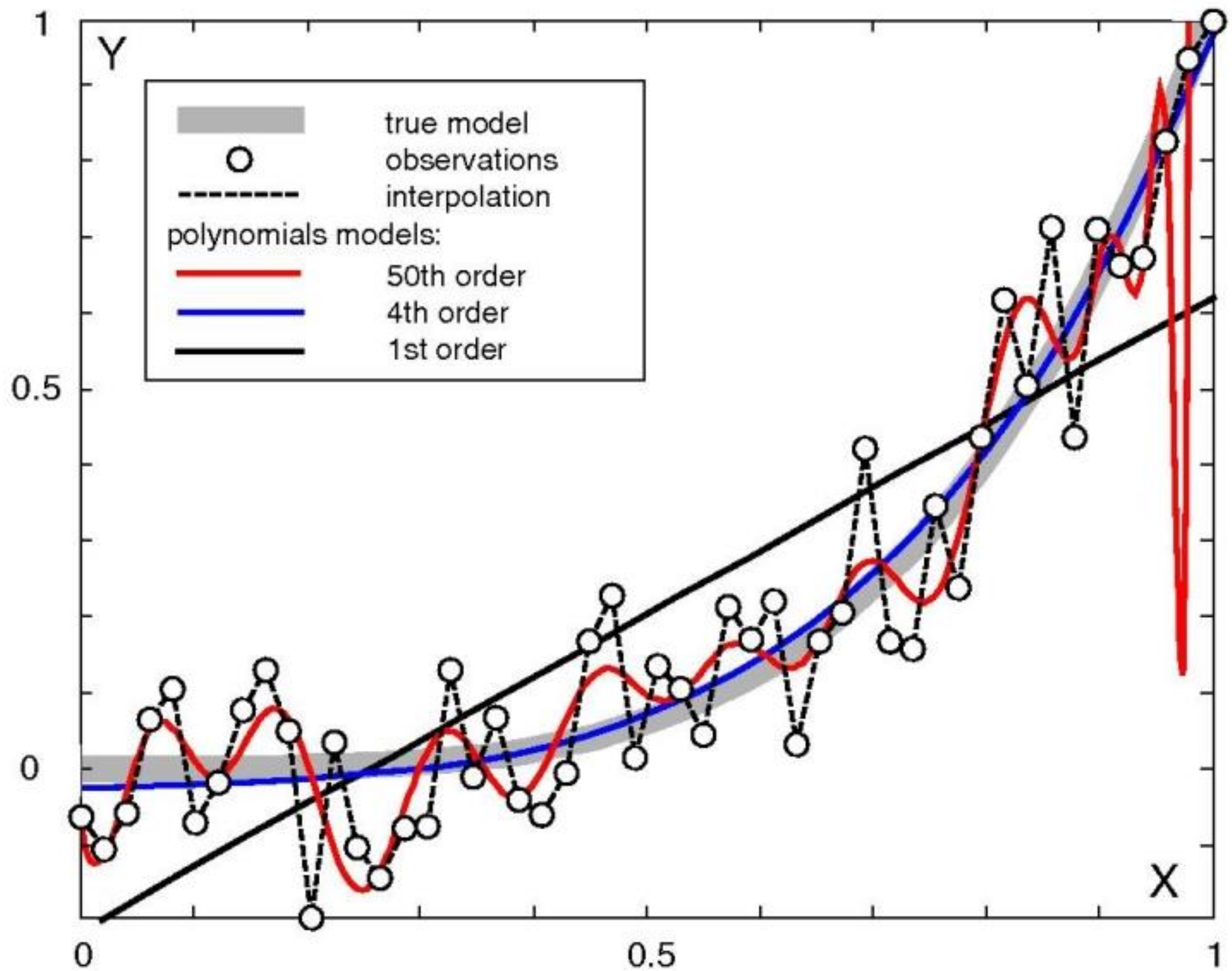
$$\left(E_D\left[f(\boldsymbol{x};D)\right] - E\left[y\,|\,\boldsymbol{x}\right]\right)^2 \qquad \text{bias}$$

$$+\,E_D\left[\left(f(\boldsymbol{x};D) - E_D\left[f(\boldsymbol{x};D)\right]\right)^2\right] \qquad \text{variance}$$

- bias: if on average $f(\boldsymbol{x};D)$ differs from $E[y\,|\,\boldsymbol{x}]$ then $f(\boldsymbol{x};D)$ is a biased estimator of $E[y\,|\,\boldsymbol{x}]$
- variance: $f(\boldsymbol{x};D)$ may be sensitive to D and vary a lot from its expected value

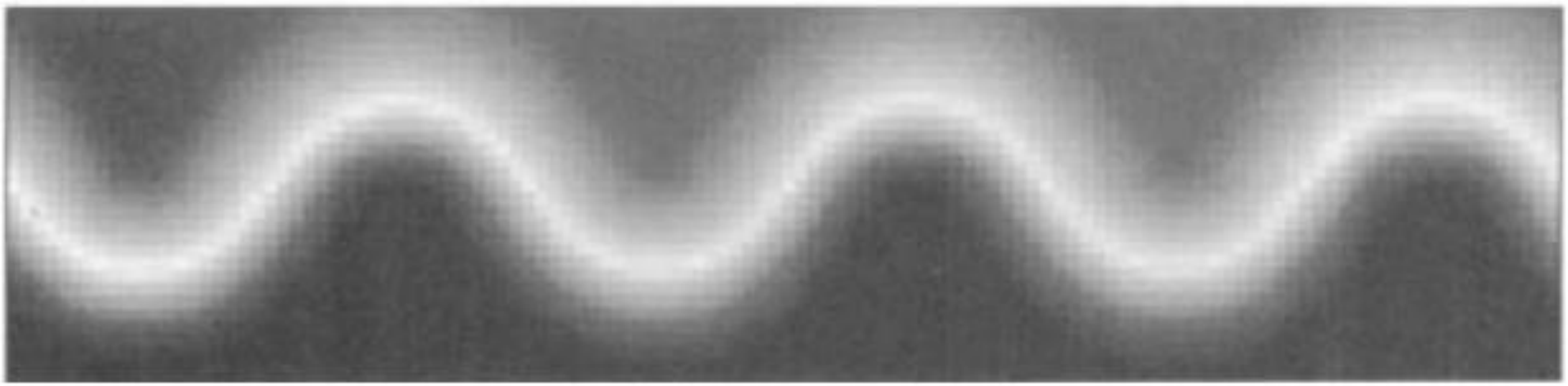
Bias/variance for polynomial interpolation

- the 1st order polynomial has high bias, low variance
- 50th order polynomial has low bias, high variance
- 4th order polynomial represents a good trade-off

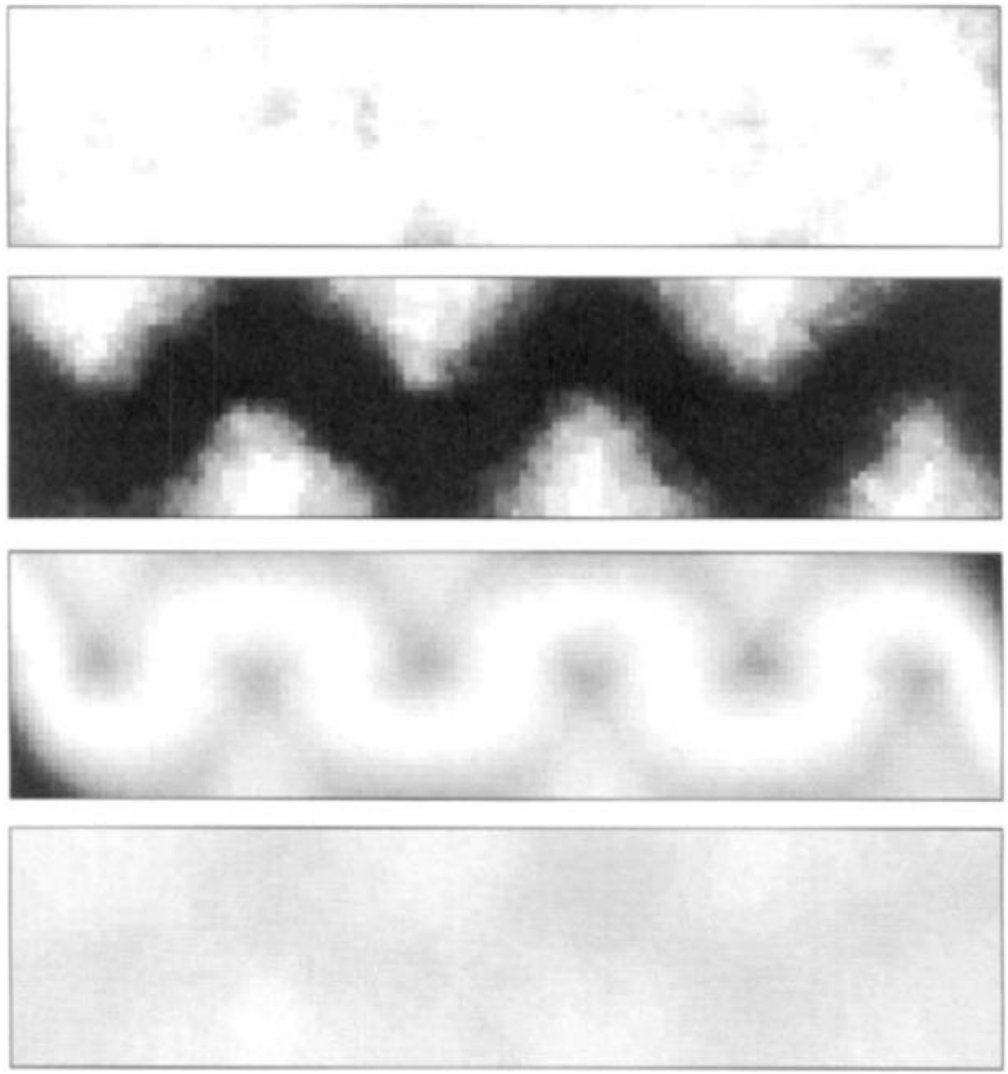


Bias/variance trade-off for nearest-neighbor regression

- consider using k-NN regression to learn a model of this surface in a 2-dimensional feature space



Bias/variance trade-off for nearest-neighbor regression



bias for 1-NN

variance for 1-NN

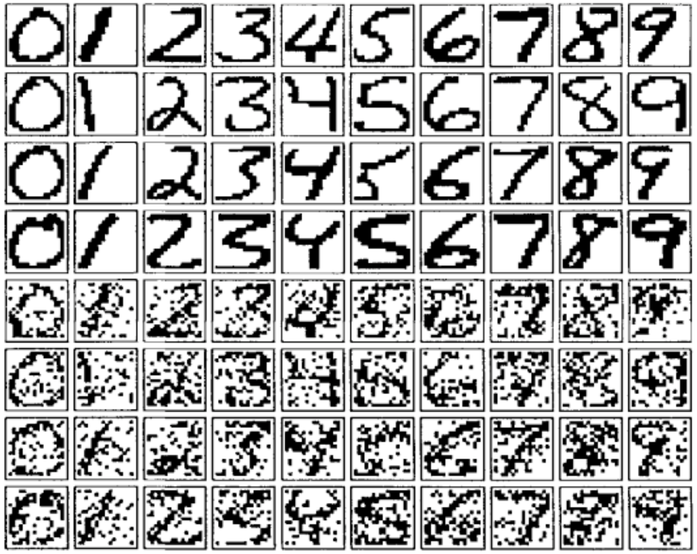
bias for 10-NN

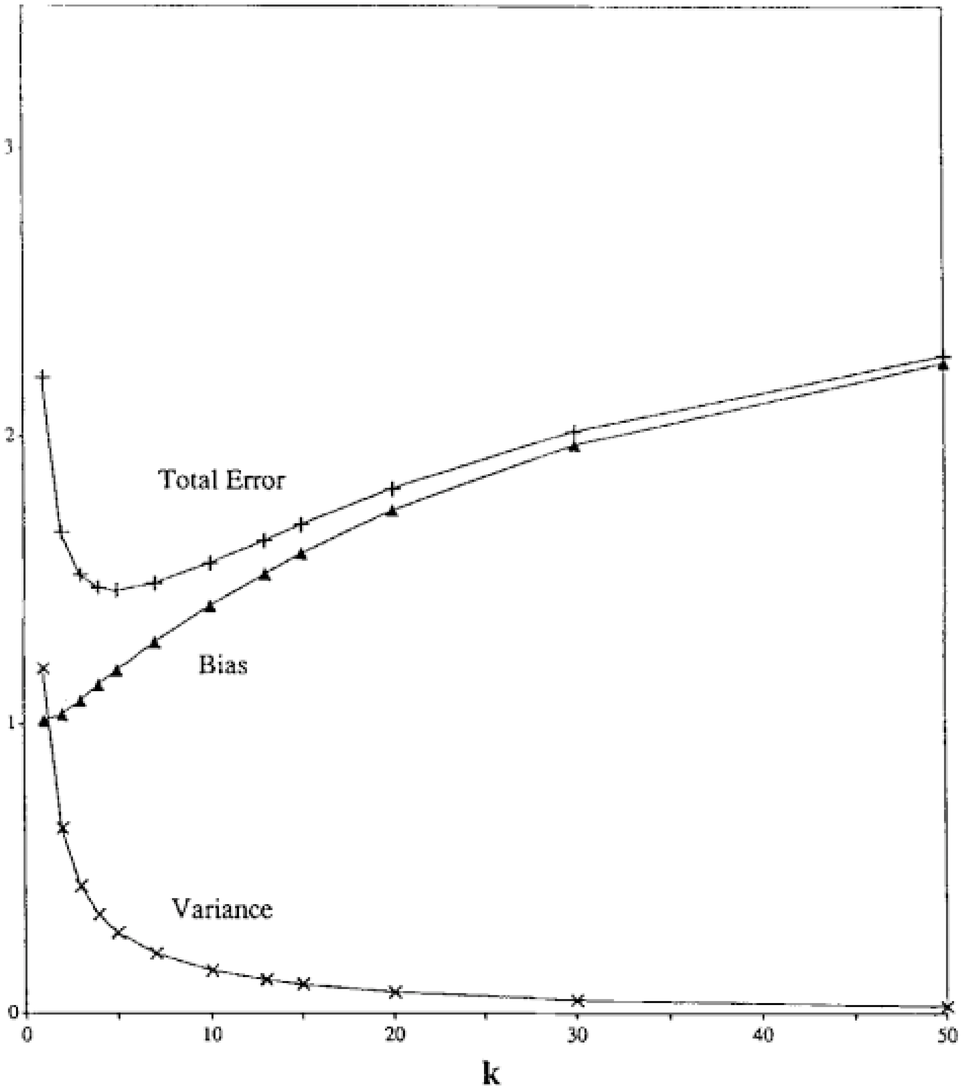
variance for 10-NN

darker pixels correspond to higher values

Bias/variance trade-off

- consider k-NN applied to digit recognition





Bias/variance discussion

- predictive error has two controllable components
  - expressive/flexible learners reduce *bias*, but increase *variance*

- for many learners we can trade-off these two components (e.g. via our selection of k in k-NN)

- the optimal point in this trade-off depends on the particular problem domain and training set size

- this is not necessarily a strict trade-off; e.g. with ensembles we can often reduce bias and/or variance without increasing the other term

Bias/variance discussion

the bias/variance analysis

- helps explain why simple learners can outperform more complex ones
- helps understand and avoid overfitting