Bias-Variance analysis

Mark Craven and David Page
Computer Sciences 760
Spring 2018

www.biostat.wisc.edu/~craven/cs760/

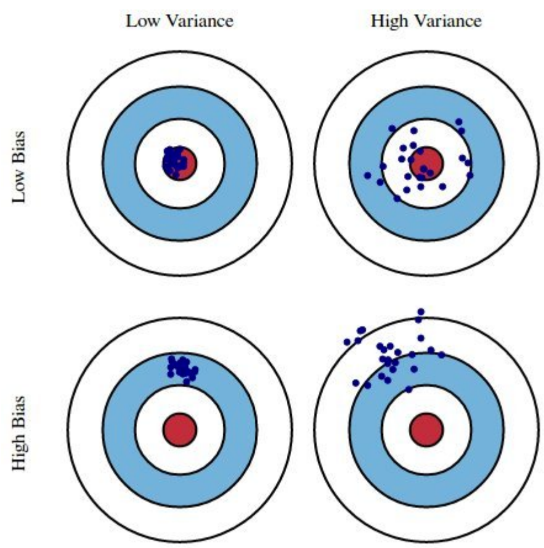
Some of the slides in these lectures have been adapted/borrowed from materials developed by Tom Dietterich, Pedro Domingos, Tom Mitchell, David Page, and Jude Shavlik

Goals for the lecture

you should understand the following concepts

- estimation bias and variance
- the bias-variance decomposition

Bias and variance in one figure



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?
- Why are regularization approaches like L_2 penalties and dropout often effective?
- 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*

Defining bias and variance for regression tasks

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

indicates the dependency of model on D

- a natural measure of the error of f is

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

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

Defining bias and variance

- this can be rewritten as:

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

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

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

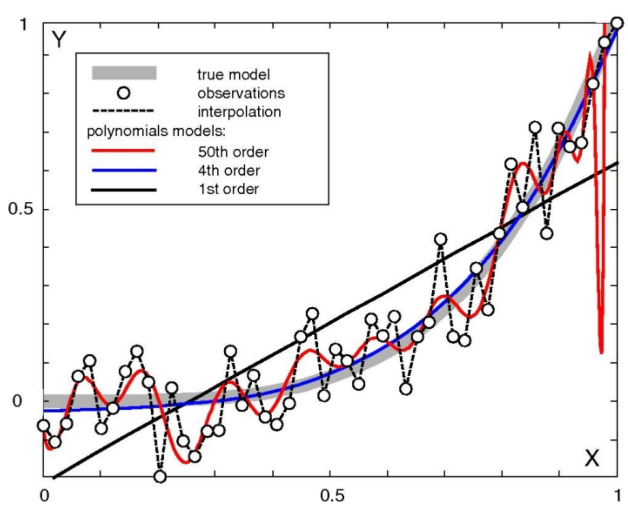
$$\left(E_D[f(\boldsymbol{x};\, D)] - E[y \mid \boldsymbol{x}]\right)^2 \qquad \text{bias}$$

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

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

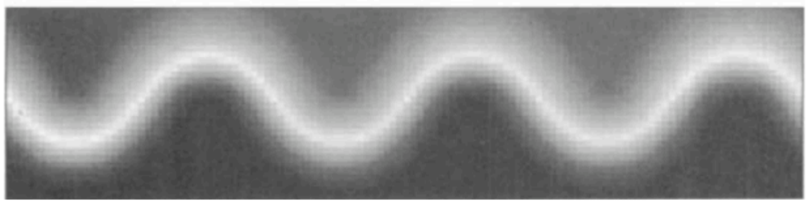
Bias/variance for polynomial interpolation

- 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

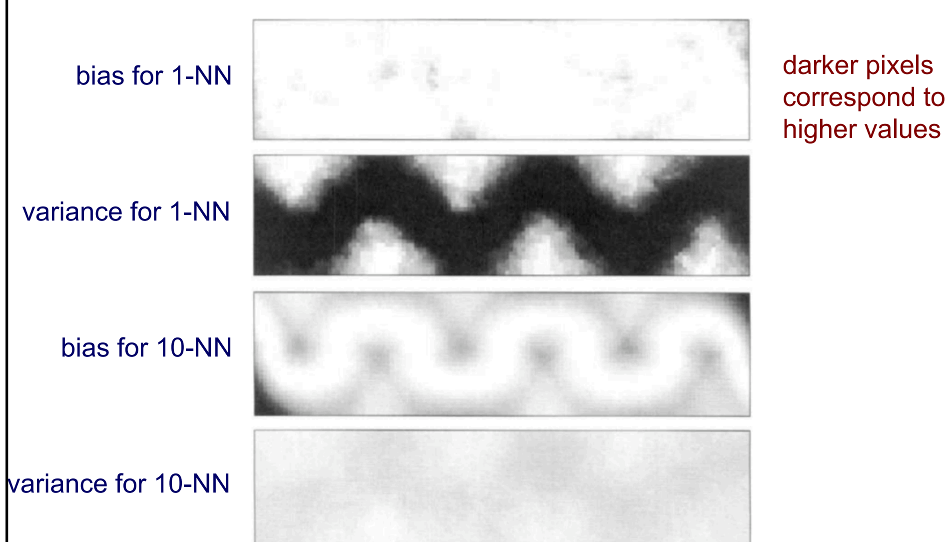


Bias/variance trade-off for nearest-neighbor regression

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

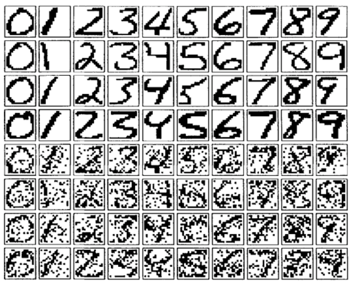


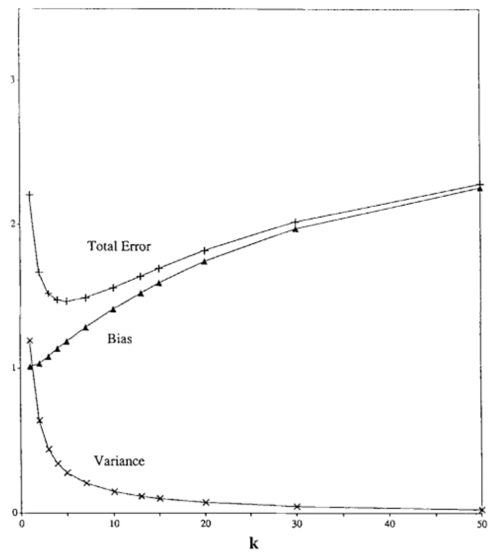
Bias/variance trade-off for nearest-neighbor regression



Bias/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