Inductive Bias, Estimation Bias and the Bias-Variance tradeoff

Mark Craven and David Page

Computer Sciences 760

Spring 2019

Goals for the lecture

you should understand the following concepts

- generalization
- inductive bias
- estimation bias and variance
- the bias-variance tradeoff

Inductive bias

- *inductive bias* is the set of assumptions a learner uses to be able to predict y_i for a previously unseen instance $\boldsymbol{x}_i$
- two components
  - *hypothesis space bias*: determines the models that can be represented
  - *preference bias*: specifies a preference ordering within the space of models

- in order to *generalize* (i.e. make predictions for previously unseen instances) a learning algorithm must have an inductive bias

Consider the inductive bias of DT and k-NN learners

learner	hypothesis space bias	preference bias
ID3	trees with single-feature, axis-parallel splits	small trees identified by greedy search
DT learner with m-of-n splits	trees with m-of-n splits	small trees identified by greedy search
DT learner with lookahead	trees with single-feature, axis-parallel splits	small trees identified by lookahead search
k-NN	Voronoi decomposition determined by nearest neighbors	

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,* though related

Recall: 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 $D = \left\{(\boldsymbol{x}_1,\, y_1)\ldots(\boldsymbol{x}_n,\, y_n)\right\}$

indicates the dependency of model on D

- a natural measure of the accuracy 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:

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

accuracy of f as a predictor of y

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

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

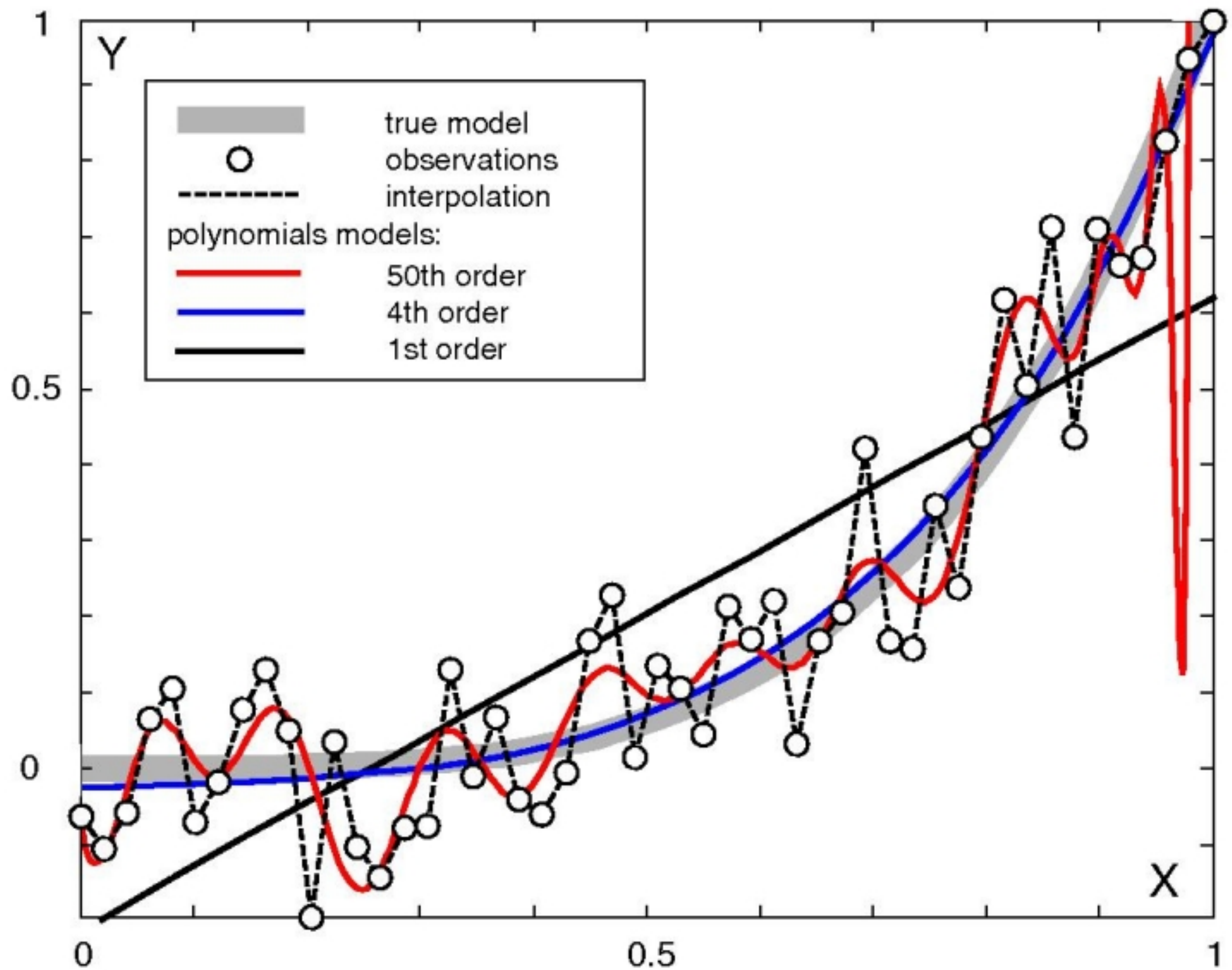
$$\left(E_D\left[f(\boldsymbol{x};\,D)\right] - E\left[y \mid \boldsymbol{x}\right]\right)^2 \qquad \text{bias}$$

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

- bias: if on average $f(\boldsymbol{x};\,D)$ differs from $E\,[y \mid \boldsymbol{x}]$ 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

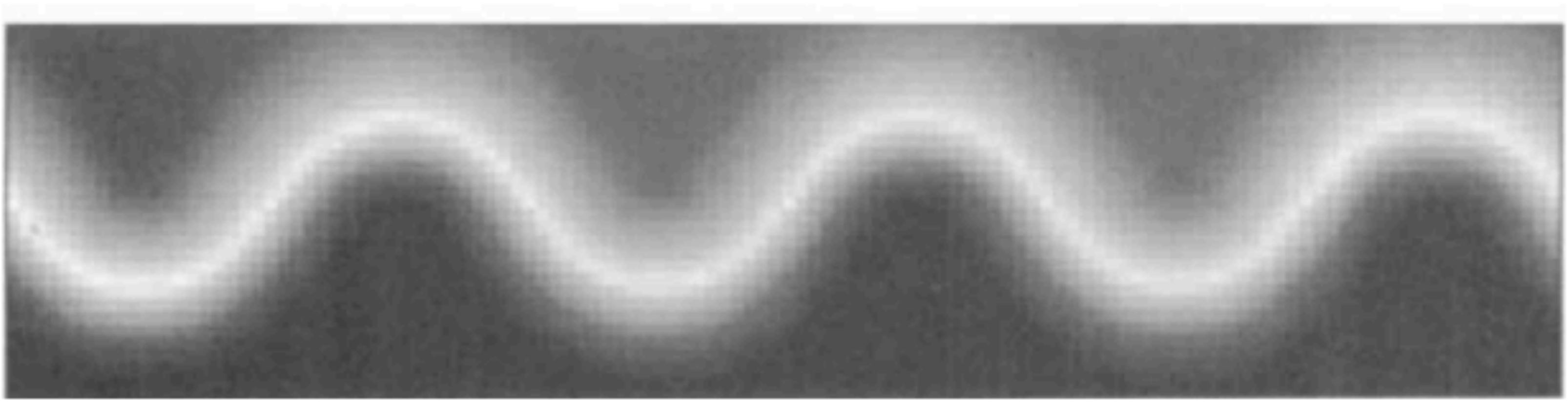
Bias/variance for polynomial interpolation

- 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



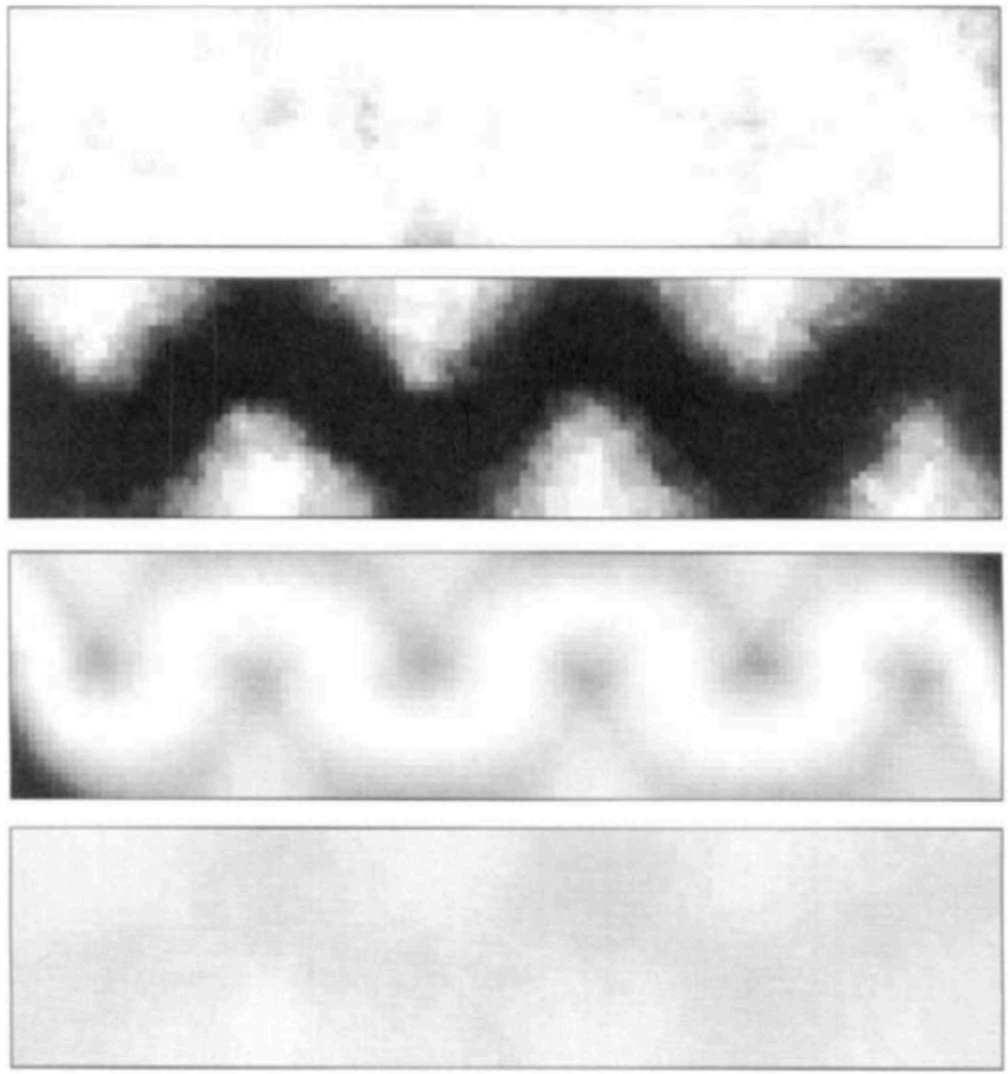
Bias/variance trade-off for nearest-neighbor regression

bias for 1-NN

variance for 1-NN

bias for 10-NN

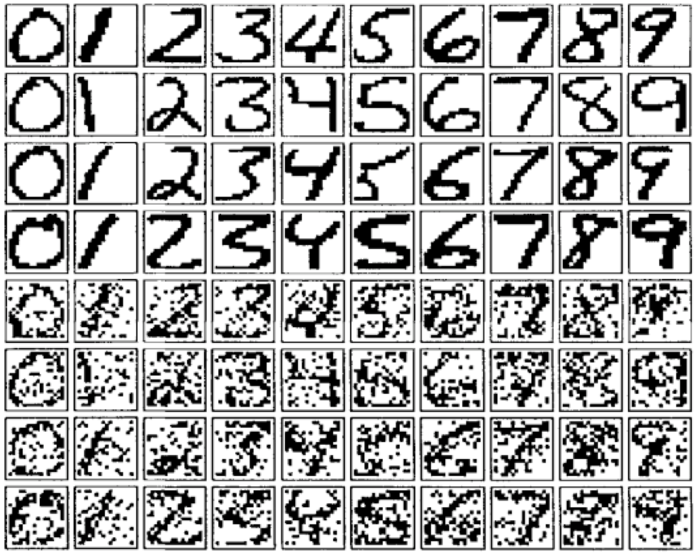
variance for 10-NN

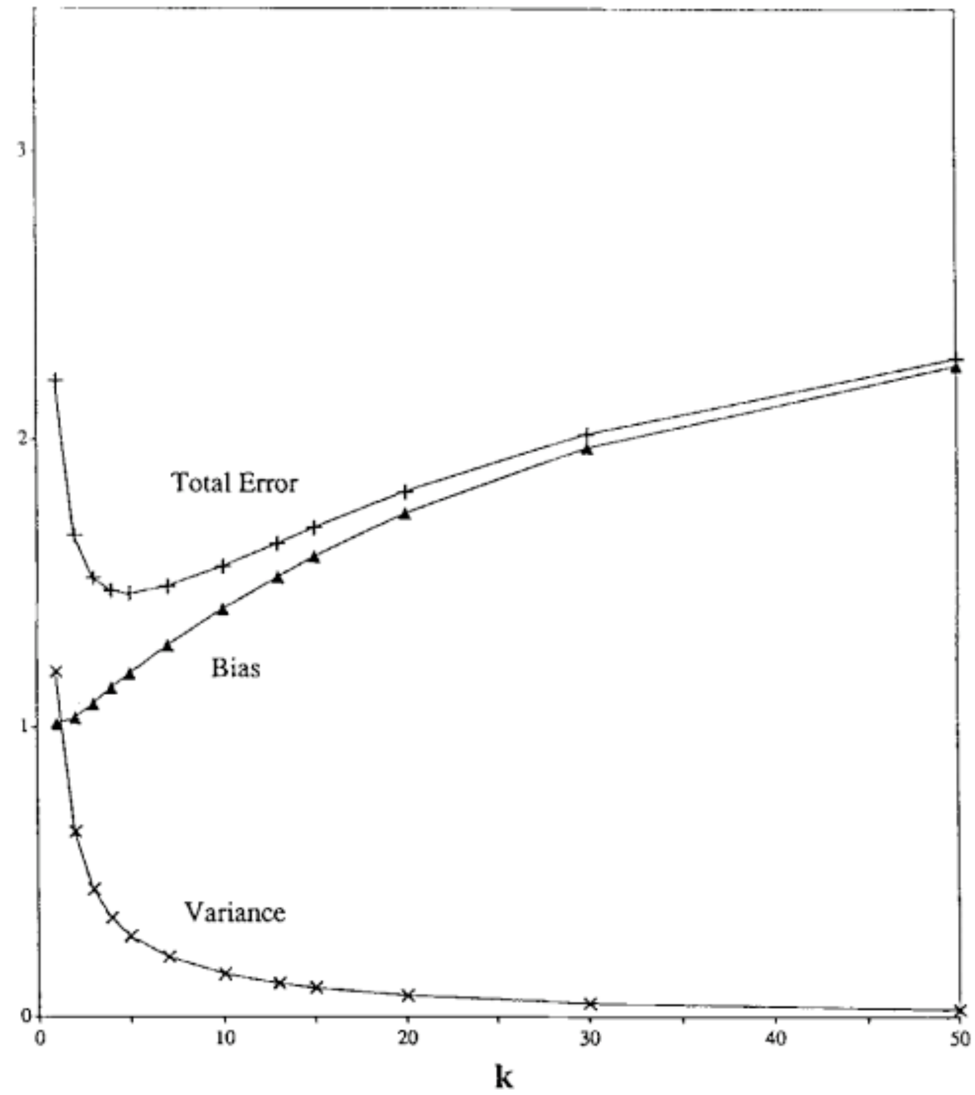


darker pixels correspond to higher values

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