CS446 Introduction to Machine Learning (Spring 2015)
University of Illinois at Urbana-Champaign
http://courses.engr.illinois.edu/cs446

LECTURE 5: BIAS/VARIANCE

Prof. Julia Hockenmaier
juliahmr@illinois.edu

Announcements

4-credit hour section students:

You need to find a project partner!

Talk to your friends, use Piazza.

Everybody:

Start working on Homework 1!

Last lecture's key concepts

Linear classifiers: $f(\mathbf{x}) = \mathbf{wx}$

- Decision boundary: $f(\mathbf{x}) = 0$

Loss functions:

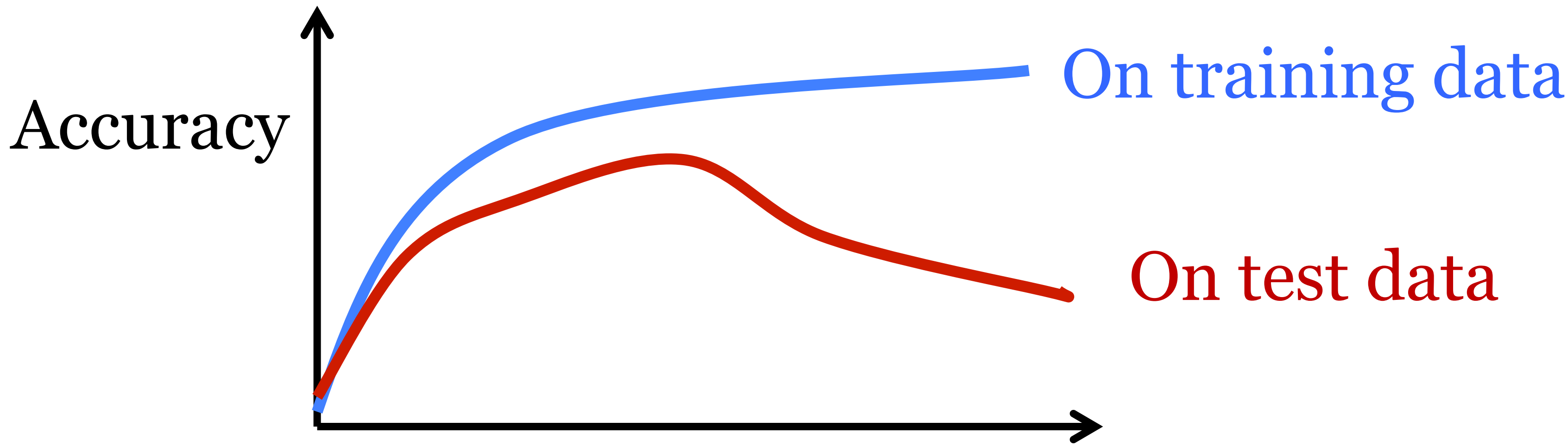
- 0-1 loss
- Square loss

(Batch) gradient descent

Stochastic gradient descent

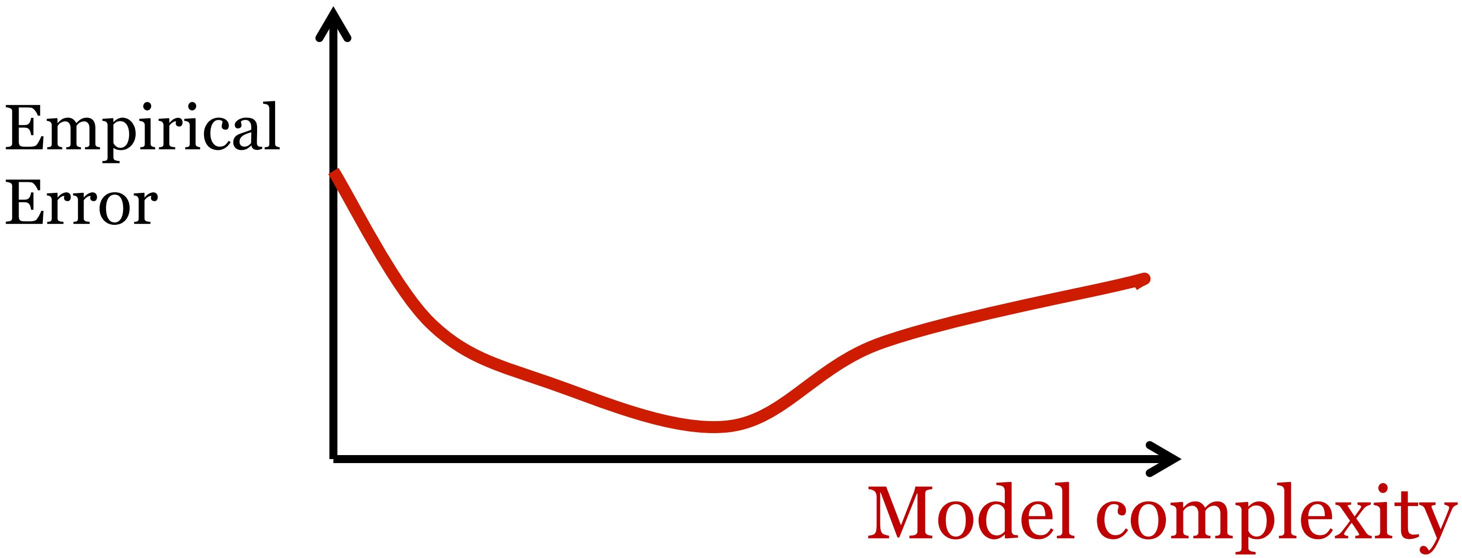
More on overfitting (informally)

Overfitting



A classifier overfits the training data
when its accuracy on the training data goes up
but its accuracy on unseen data goes down

Overfitting

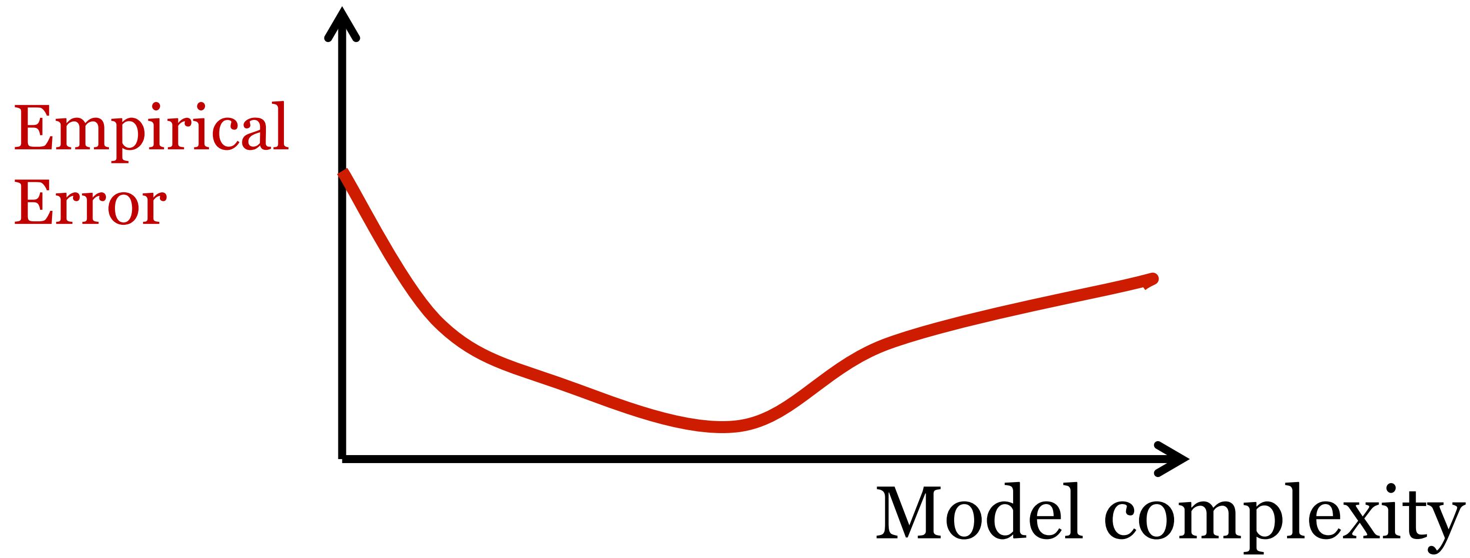


Model complexity (informally):

How many parameters do we have to learn?

Decision trees: complexity = #nodes

Overfitting



Empirical error (= on a given data set):
What percentage of items in this data set are misclassified by the classifier f ?

The i.i.d. assumption

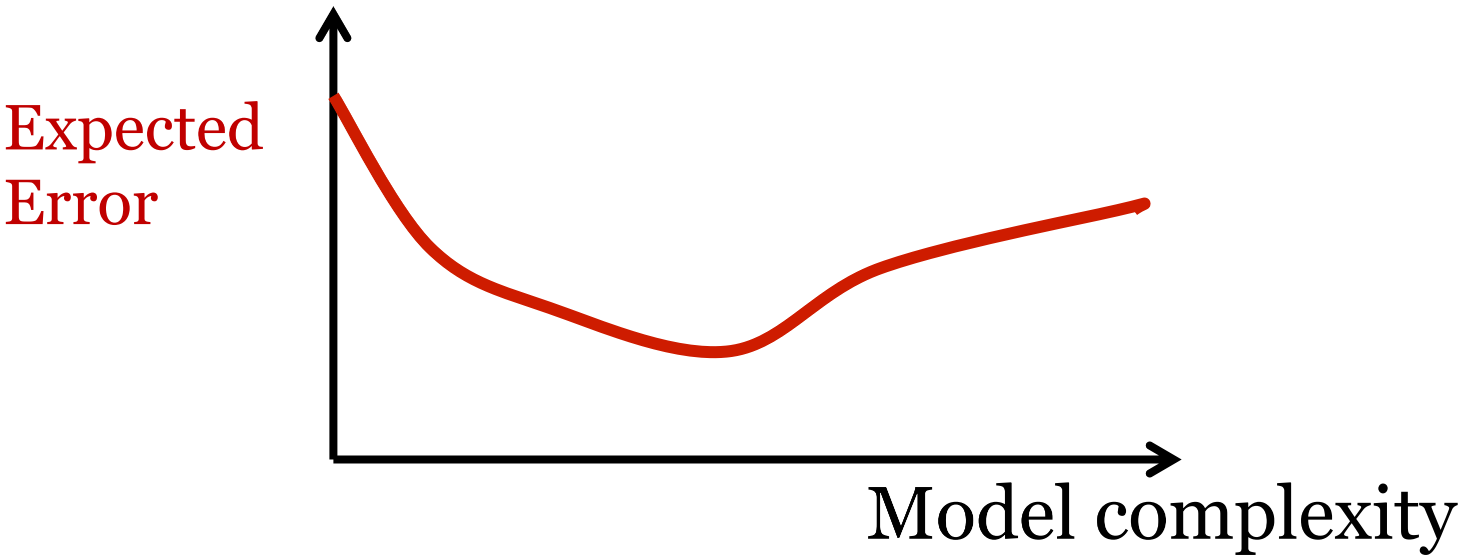
Training and test items are independently and identically distributed (i.i.d.):

- There is a distribution $P(\mathbf{X}, Y)$ from which the data $\mathcal{D} = \{(\mathbf{x}, y)\}$ is generated.

  Sometimes it's useful to rewrite $P(\mathbf{X}, Y)$ as $P(\mathbf{X})P(Y|\mathbf{X})$
  Usually $P(\mathbf{X}, Y)$ is unknown to us (we just know it exists)

- Training and test data are samples drawn from the *same* $P(\mathbf{X}, Y)$: they are identically distributed
- Each $(\mathbf{x}, y)$ is drawn independently from $P(\mathbf{X}, Y)$

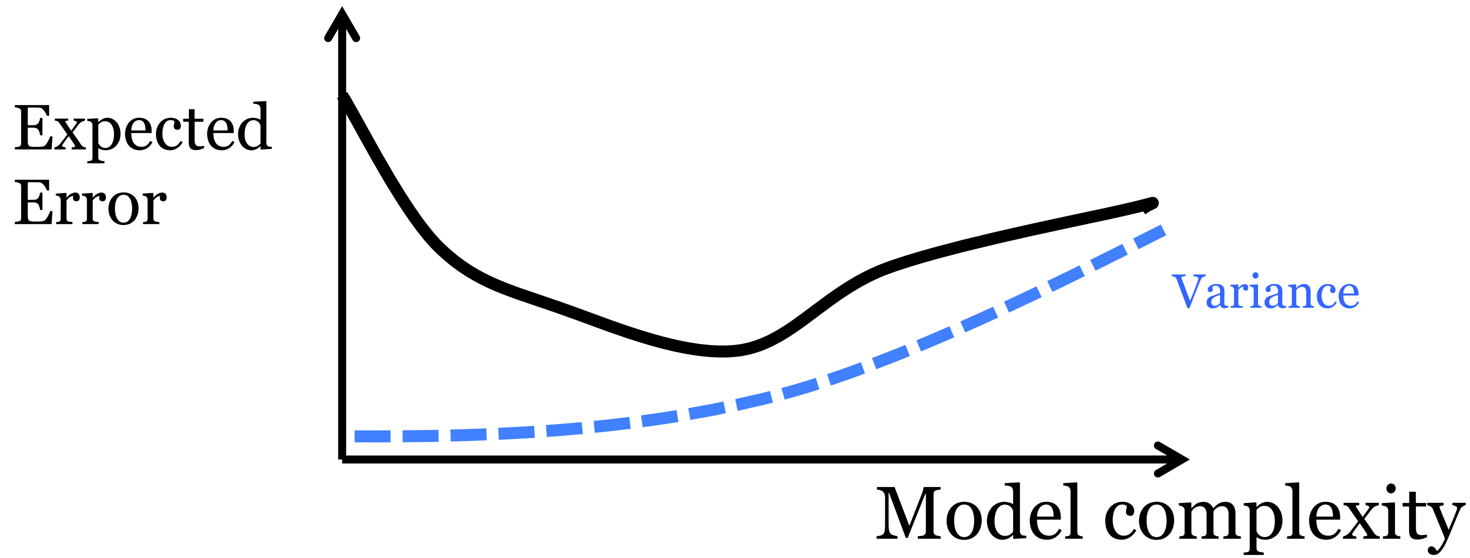
Overfitting



Expected error:

What percentage of items drawn from $P(\mathbf{x},y)$ do we expect to be misclassified by f?

Variance of a learner (informally)

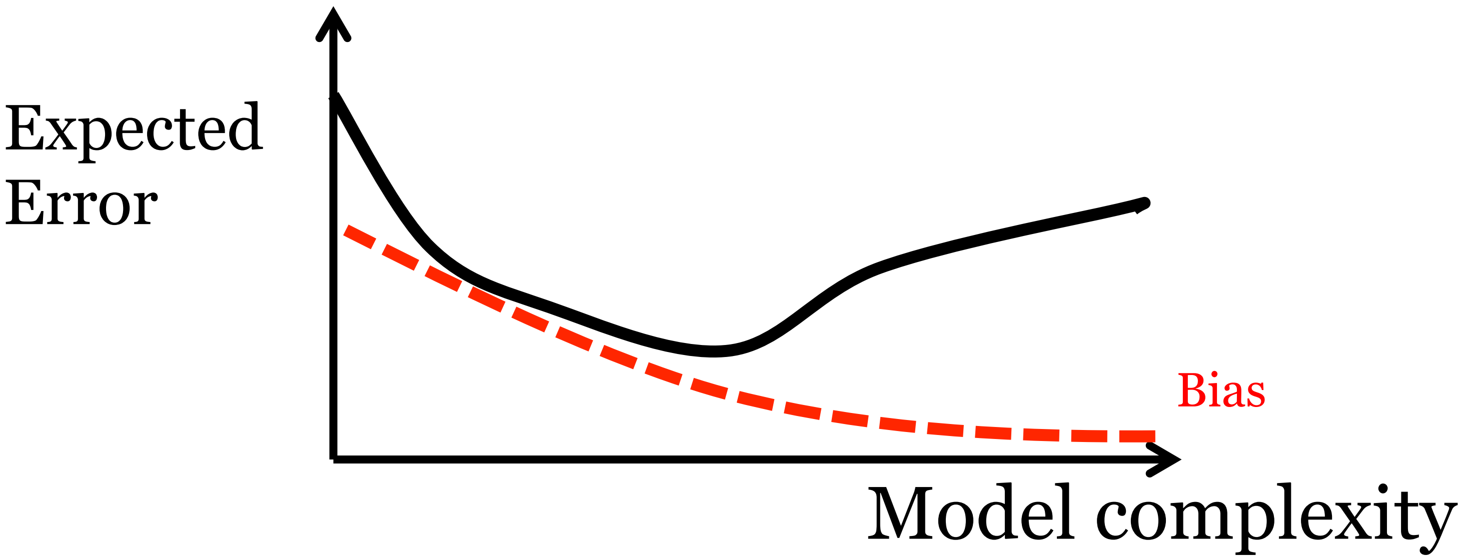


How susceptible is the learner to minor changes in the training data?

(i.e. to different samples from $P(\mathbf{X}, Y)$)

Variance increases with model complexity

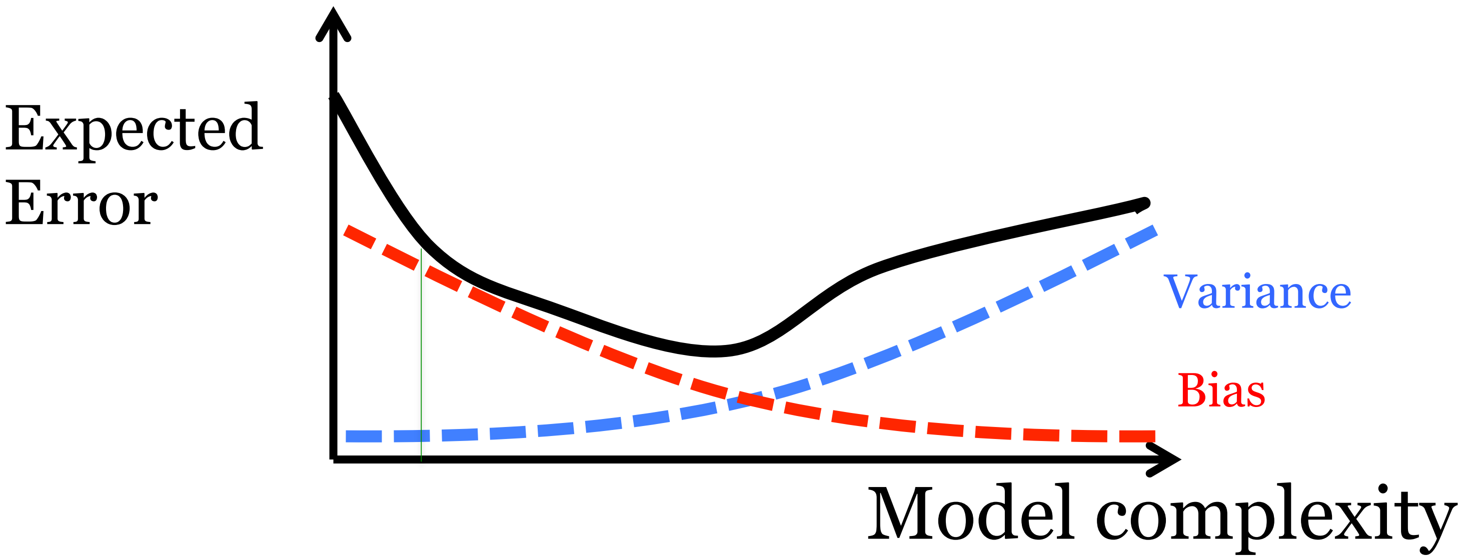
Bias of a learner (informally)



How likely is the learner to identify the target hypothesis?

Bias is high when the model is (too) simple

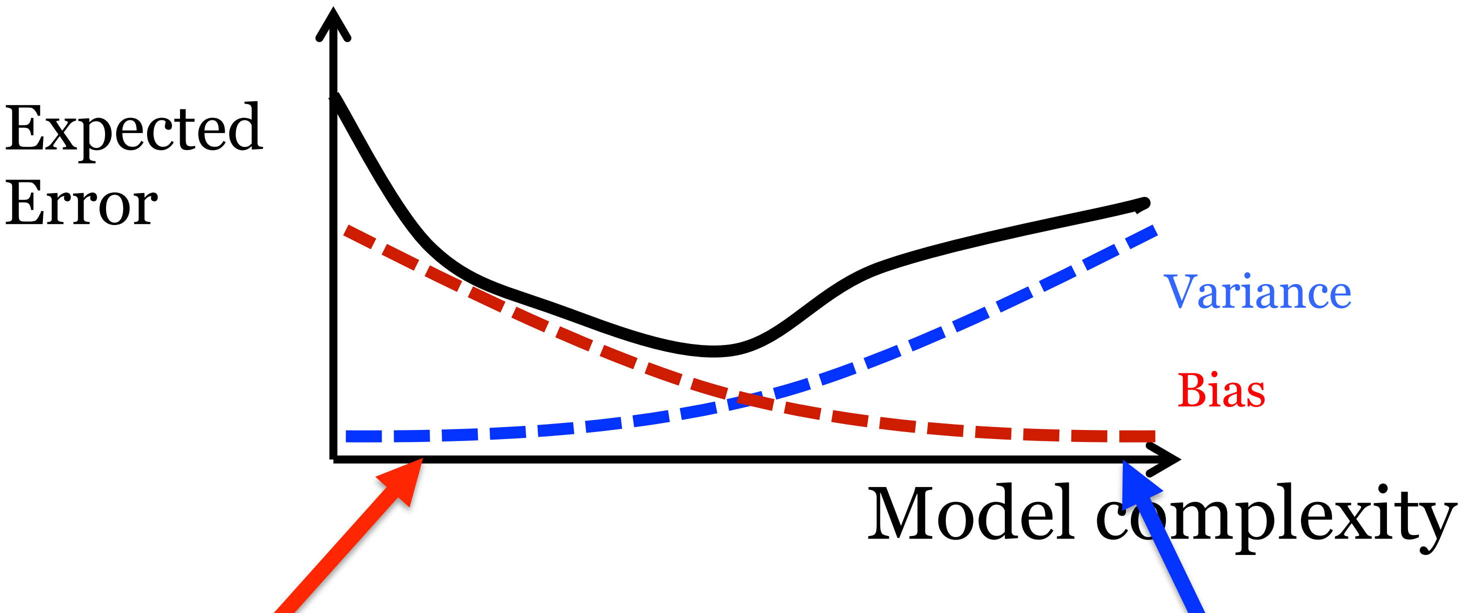
Impact of bias and variance



Expected error of a learner ≈ bias² + variance

(+noise)

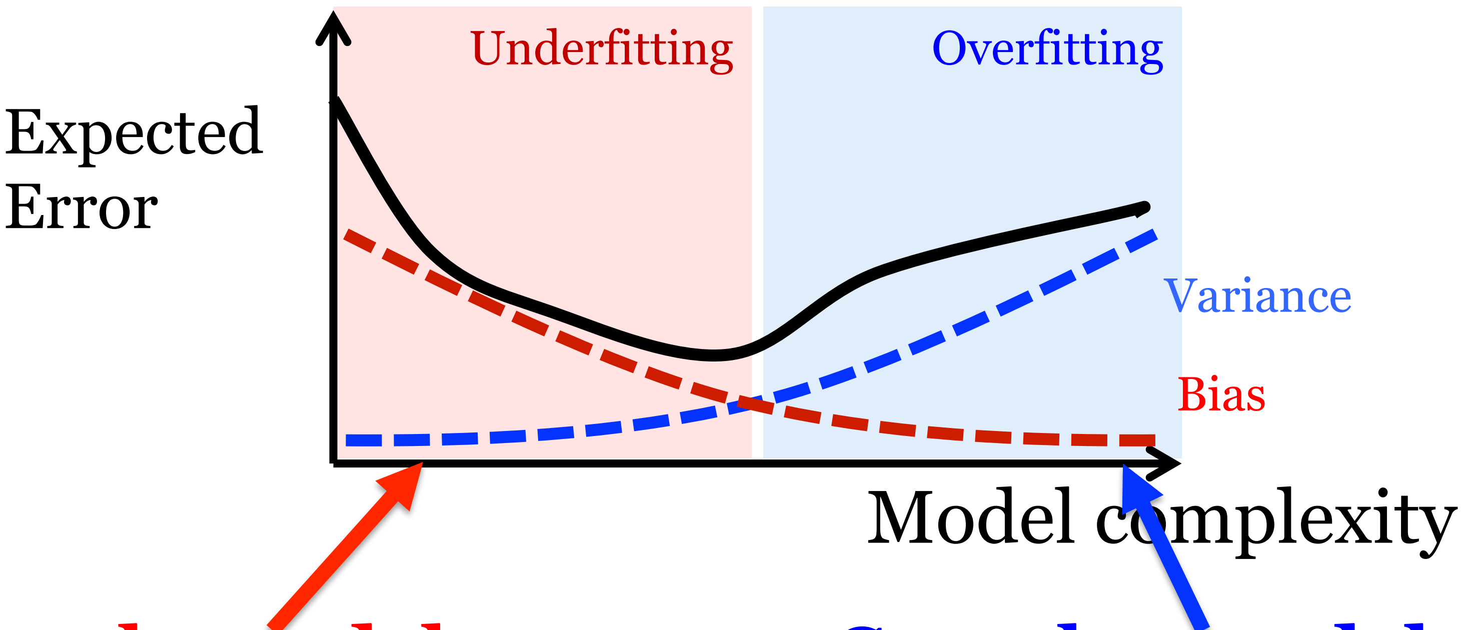
Model complexity



Simple models:
High bias and low variance

Complex models:
High variance and low bias

Underfitting and Overfitting

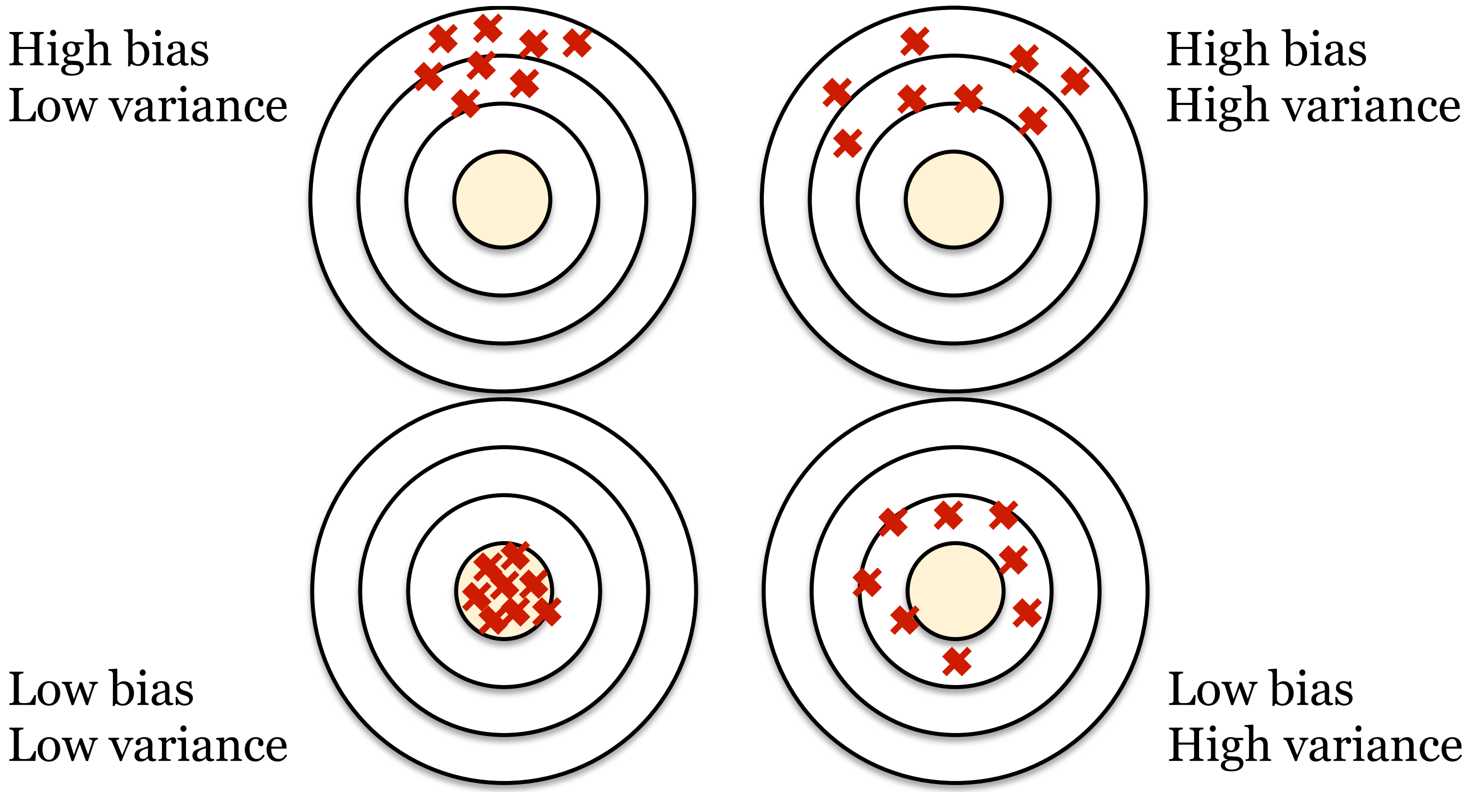


Simple models:
High bias and low variance

Complex models:
High variance and low bias

Bias-variance tradeoffs

Dartboard = hypothesis space
Bullseye = target function
Darts = learned models



K Nearest Neighbor classifier

kNN classifier

Classify an item **x** by:

– finding the *k* training examples that are closest to **x** (**= x**'s *k* nearest neighbors)

– assign the majority class of these *k* training examples to **x**

How does the bias and variance of kNN vary with k (for k = 1… N)?

Bias/Variance decomposition

Bias of a learner

We distinguish between

- Inductive bias: What assumptions about the target function does the learner use?

  Absolute bias: Target function has a particular form (e.g. only consider linear decision boundaries)

  Relative bias: Prefer some hypotheses over others (e.g. consider smaller decision trees before larger ones)

- Statistical bias: Systematic error that the learner is expected to make

  (for a particular target function, over data sets of size M)

Inductive bias

Absolute biases can be…

… *appropriate*: Hypothesis space contains good approximations to target function.

… *inappropriate*: Hypothesis space does not contain good approximations to target.
High statistical bias.

Inductive bias

Relative biases can be…

… *too strong*: Poor approximations to the target function are preferred.

Statistical bias is high, Variance is low.

… *too weak*: *H* not sufficiently constrained.

Statistical bias is low, Variance is high.

Inductive and statistical bias/variance

Inductive bias		Statistical bias	Variance
Absolute	**Relative**		
appropriate	too strong	high	low
appropriate	**ok**	**low**	**low**
appropriate	too weak	low	high
inappropriate	too strong	high	low
inappropriate	ok	high	moderate
inappropriate	too weak	high	high

"Sweet spot". Often difficult to achieve in practice.

Bias/variance decomposition

The **expected error of a learner on a particular target function** decomposes into a statistical **bias** term and a **variance** term (which both depend on the learner) and a constant **noise** term (which depends on the target function).

Theoretical analysis: Useful to know about, although knowledge of P(**x**,y) or access to lots of data sets is required to actually compute these terms for a particular target and learner.

Recap (Probability/Stats Cheat Sheet)

Expectation/Mean of (discrete) random variable *X*

The weighted average of X

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

Expectation of a function of *X*, *f*(*X*):

The weighted average of *f*(*X*)

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

Variance of *X*:

$$Var(X) = E[\,(X - \mu_X)^2\,] = \sigma_X^2$$

The expected value of the squared difference between X and its mean

Standard deviation of *X*:

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{Var(X)}$$

The square root of the variance

Statistical bias

- Assume each $\mathbf{x}$ has a target value $h(\mathbf{x})$.
- Assume we sample L data sets $D_1,\ldots,D_L$ of size M, and train L learners $f_1,\ldots,f_L$, one on each D_l
- $f_{avg}(\mathbf{x})$ is the expected predicted value of $f(\mathbf{x})$

$$f_{avg}(\mathbf{x}) = \lim_{L\to\infty} \frac{1}{L}\sum_{l=1}^{L} f_l(\mathbf{x})$$

- The statistical bias of f (for sample size M) at $\mathbf{x}$ is the difference between $f_{avg}(\mathbf{x})$ and $h(\mathbf{x})$

$$\mathrm{Bias}(f, \mathrm{M}, \mathbf{x}) = f_{avg}(\mathbf{x}) - h(\mathbf{x})$$

Variance of a learner

The variance of a learner $f(\mathbf{x})$ is the expected value (over all data sets of size M) of the squared difference between $f(\mathbf{x})$ and $f_{avg}(\mathbf{x})$

$$\mathrm{Var}(f, M, \mathbf{x}) = E[\,(f_l(\mathbf{x}) - f_{avg}(\mathbf{x})\,)^2\,]$$

CF.: The variance of a random variable X is the expected value of the squared difference between X and its mean:

$$\mathrm{Var}(X) = E[\,(X - \mu_X\,)^2\,]$$

Error = bias² + variance

The expected mean-squared error of f on $\mathbf{x}$ is equal to the squared bias of f on $\mathbf{x}$ plus the variance of f on $\mathbf{x}$:

$$E[\,(f_l(\mathbf{x}) - h(\mathbf{x}))^2\,] = \mathrm{Bias}(f, M, \mathbf{x})^2 + \mathrm{Var}(f, M, \mathbf{x})$$

($\mathbf{x}$ is fixed; the expectation is taken over all data sets $D_1, \ldots, D_l, \ldots, D_L$ of size M)

Today's key concepts

Overfitting

kNN classifier

Inductive bias

Statistical bias and variance