CS135
Introduction to Machine Learning

Lecture 5: Understanding the Bias and Variance Tradeoff

Bias and Variance

For prediction models, prediction errors can be decomposed into two main subcomponents we care about: error due to "bias" and error due to "variance". There is a tradeoff between a model's ability to minimize bias and variance. ***Understanding these two types of error can help us diagnose model results and avoid the mistake of over- or under-fitting.***

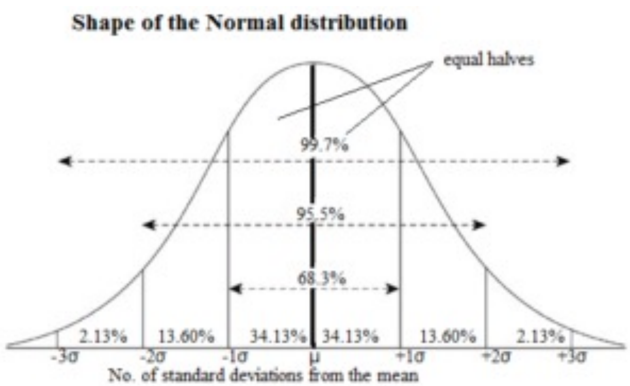
We assume variable to predict Y is related to covariates X *as follows*

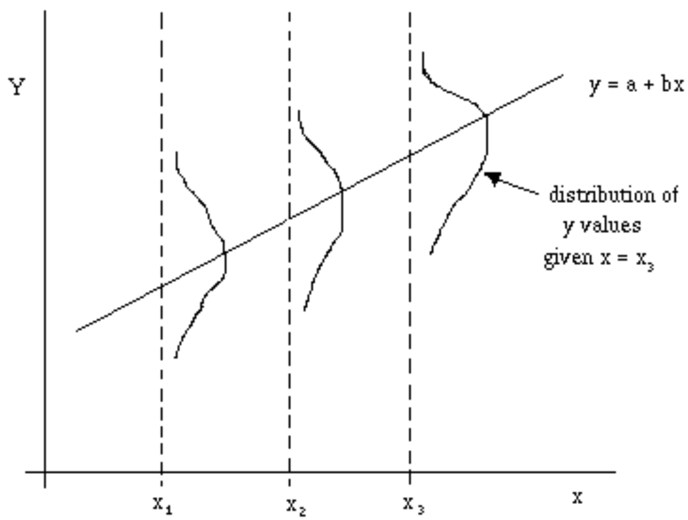
$$Y = f(X) + \varepsilon$$

where ε is normally distributed with a mean of zero, i.e., $\varepsilon \sim N(0, \sigma_\varepsilon)$.

$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is mean or expectation, σ is standard deviation, and σ^2 is the variance.





Bias and Variance

- Estimate $f(x)$ as $\hat{f}(x)$ via linear regression the expected squared error at point x is

$$Err(x) = [Y - \hat{f}(x)]^2$$

- This error can be decomposed using **bias** and **variance** components.

$$Err(x) = (E[\hat{f}(x)] - f(x))^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_\varepsilon^2$$
$$Err(x) = Bias^2 + Variance + Irreproducible\ Error$$

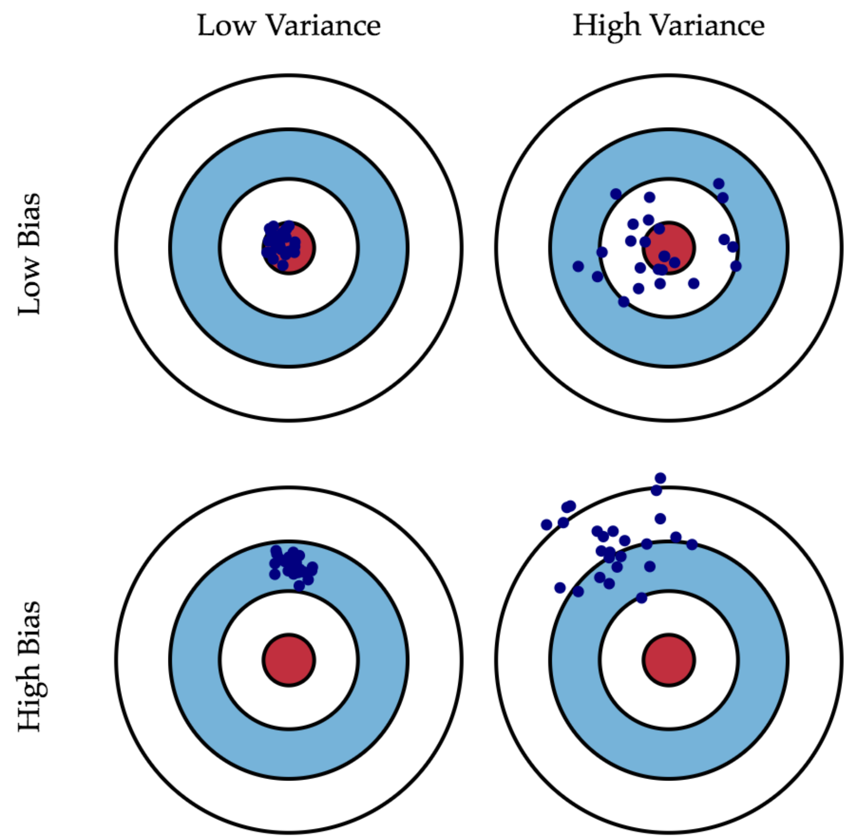
where irreproducible error cannot, hypothetically, be reduced by any model.

Provided true model and infinite data bias and variance could be reduced to zero

However, real-world there exists a bias-variance tradeoff

Tufts

Bias and Variance (Bull's Eye Chart)



Trade off

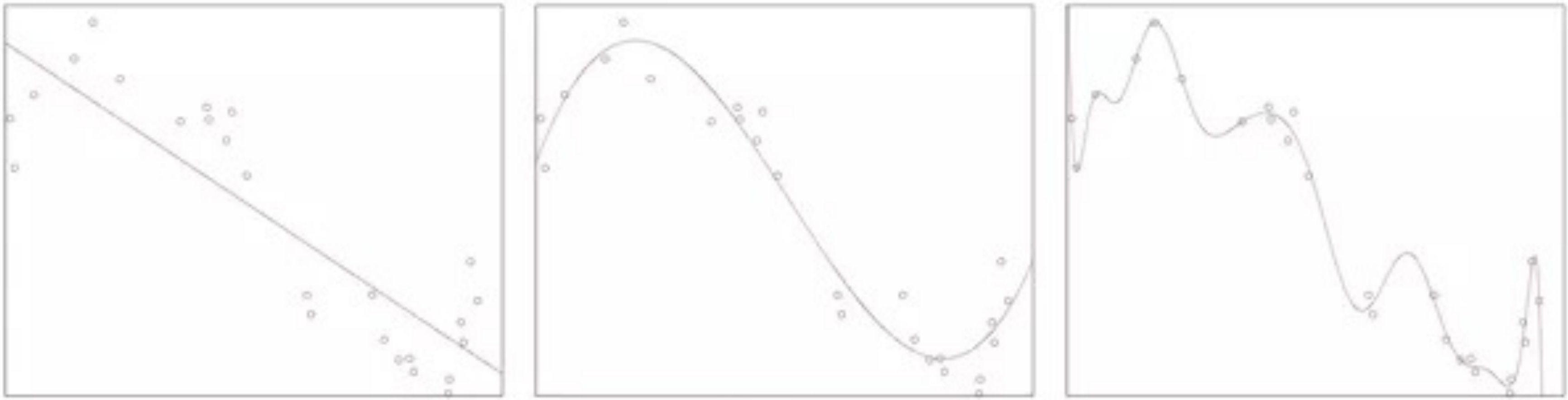
- Overfitting == low bias, high variance\
- Underfitting == high bias, low variance
- Noise is dominating!

Bias Variance Decomposition

$$Err(x) = Bias^2 + Variance + Irreproducible\ Error$$

Tufts

Bias and Variance (Overfit vs Underfit)



underfit

(degree = 1)

Ideal fit

(degree = 3)

overfit

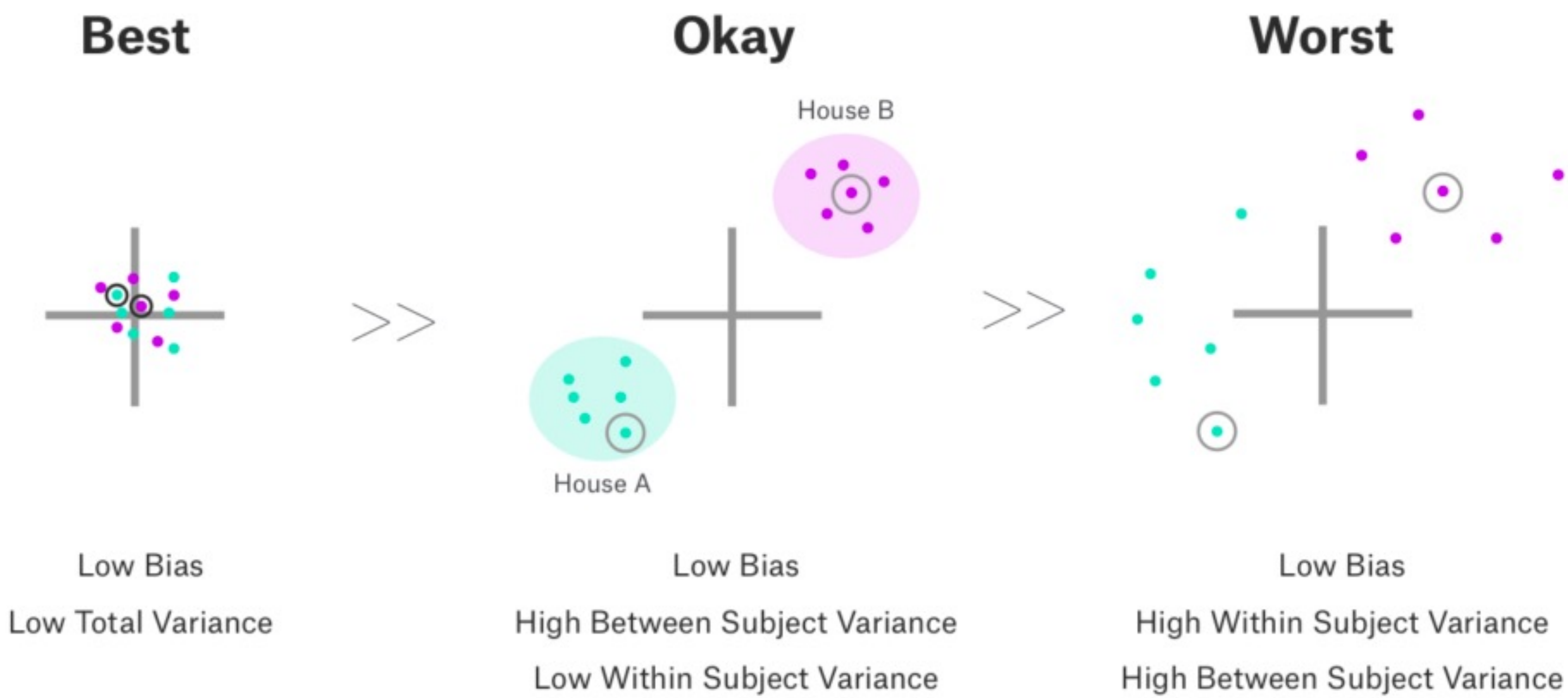
(degree = 20)

degree *n* →

Increasing Model Complexity

Tufts

Bias and Variance (Graphical View)



Bias and Variance (Toy Example)

Voting Republican	Voting Democrat	Non-Respondent	Total
13	16	21	50

- Probability voting republican

Tufts

Bias and Variance (Toy Example)

Voting Republican	Voting Democrat	Non-Respondent	Total
13	16	21	50

- Probability voting republican

$$\frac{13}{13+16} = 44.8\%$$

- Press release list democrats as winning by margin of 10%; however contrary is true. How could this be?

Bias and Variance (Toy Example)

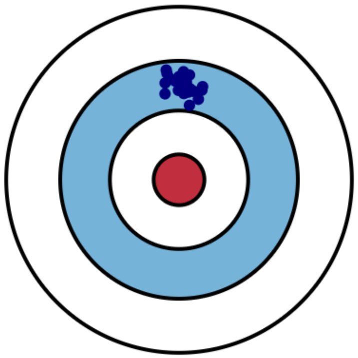
Voting Republican	Voting Democrat	Non-Respondent	Total
13	16	21	50

- Probability voting republican

$$\frac{13}{13+16} = 44.8\%$$

- Press release list democrats as winning by margin of 10%; however contrary is true. How could this be?

Bias introduced by only using phonebook + not following up non respondents



Tufts

Bias and Variance (Toy Example)

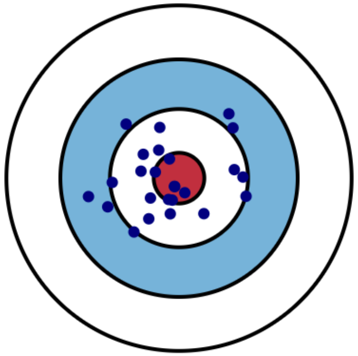
Voting Republican	Voting Democrat	Non-Respondent	Total
13	16	21	50

- Probability voting republican

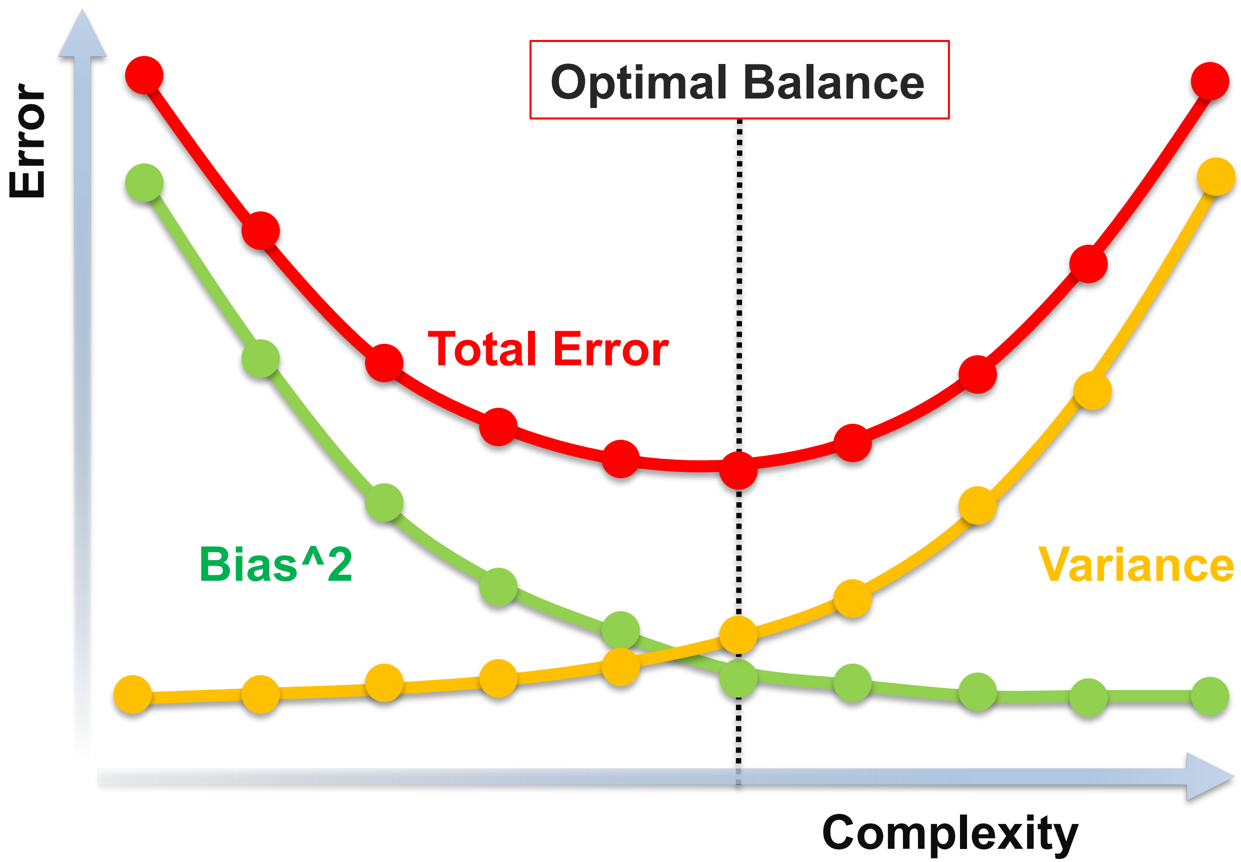
$$\frac{13}{13+16} = 44.8\%$$

- Press release list democrats as winning by margin of 10%; however contrary is true. How could this be?

Variance introduced by small sample size

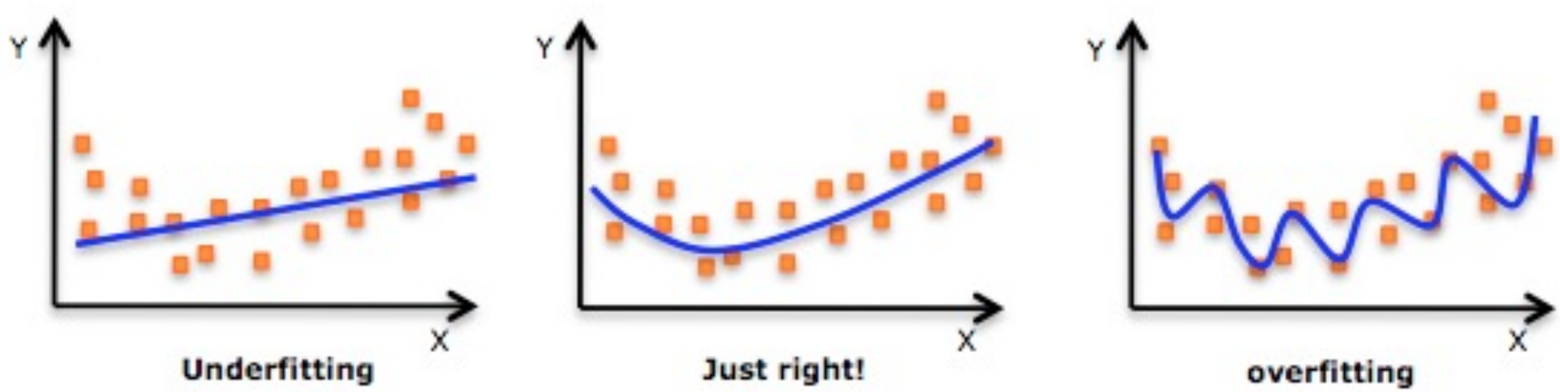


Bias and Variance on Total Error



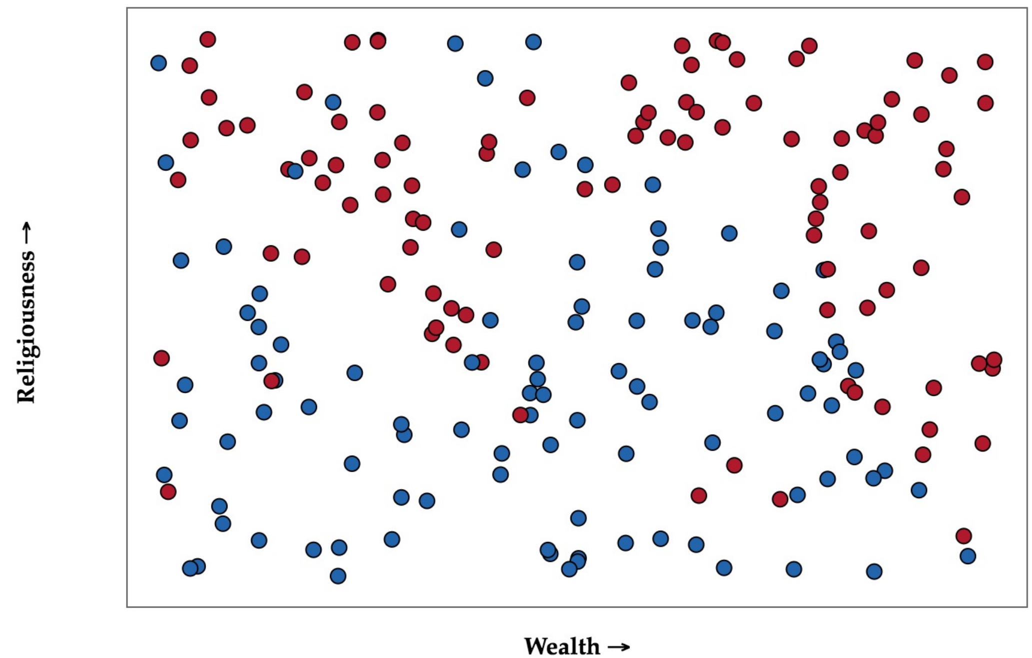
Bias and Variance (Over- and Under- Fitting)

- At the core, Bias-Variance Tradeoff corresponds to over- and under-fitting
  - Bias reduces and variance increases with increasing model complexity



An Applied Example: Voter Party Registration

- Assume we have a training data of voters tagged with 3 properties
  1. voter party registration
  2. voter wealth
  3. quantitative measure of voter religiousness.
- We want to predict voter registration provided wealth and religiousness features



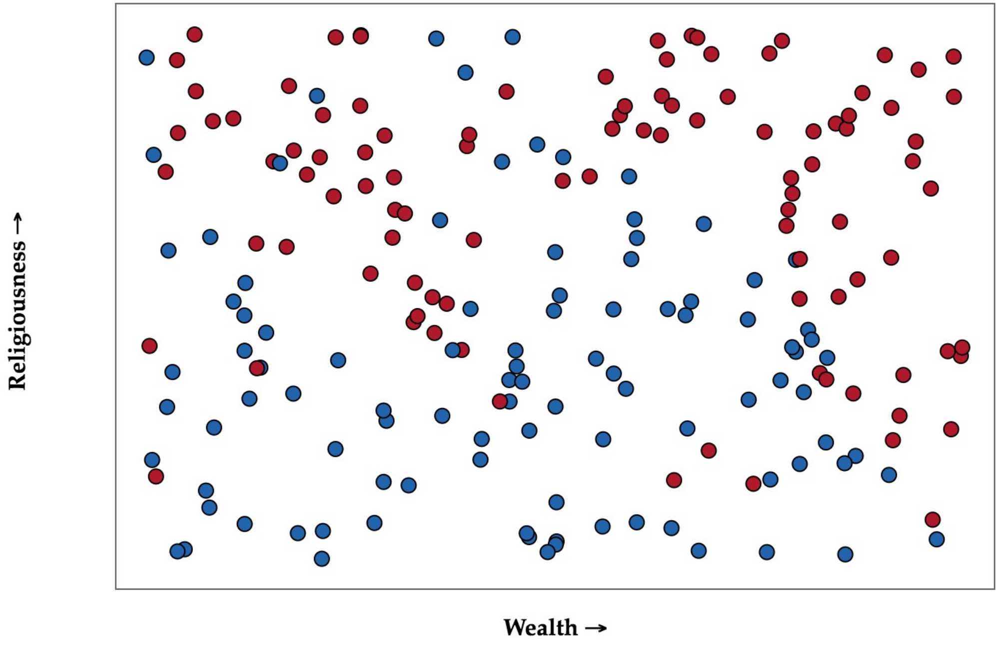
Red and blue circles are Republican and Democrat, respectively

Tufts

An Applied Example: Voter Party Registration

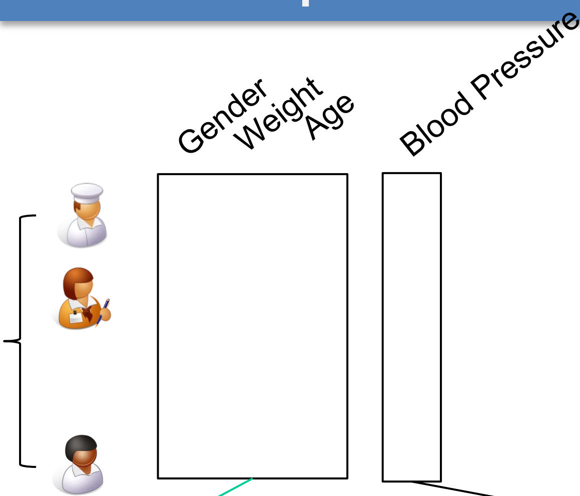
The K-Nearest Neighbor Algorithm

- Many ways to model this task
  - For binary outcome like our, logistic regressions are often used (next topic)
- Considering the non-linearity in relationships in our variables
  - A more flexible, data adaptive approach might be desired (i.e., KNN)



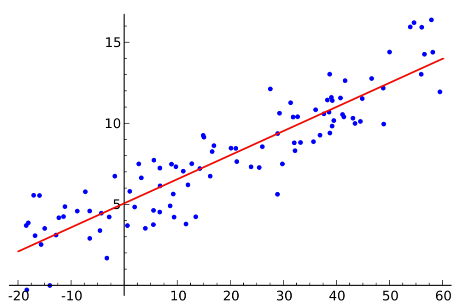
Tufts

Least Squares Estimator (LSE)

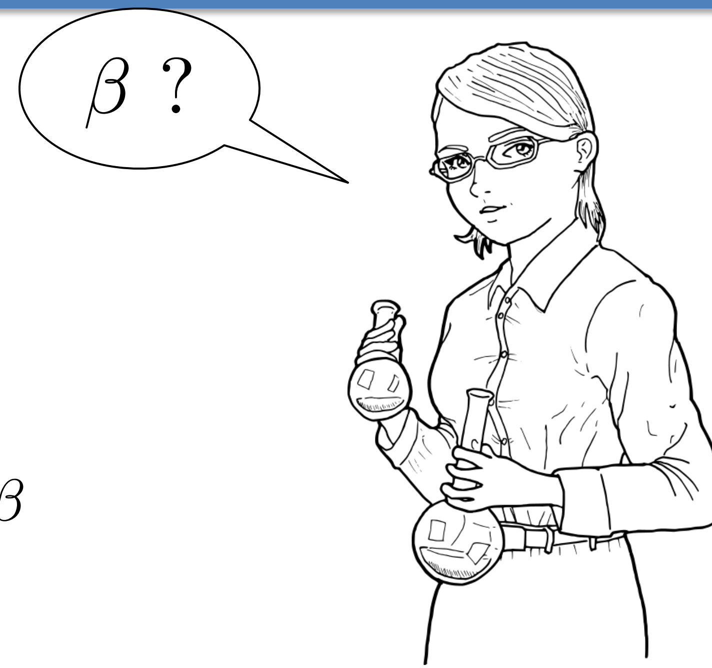


$X \in \mathbb{R}^{n \times d}$ $y \in \mathbb{R}^n$

$$y_i \approx f(x_i) = \beta^\top x_i = \sum_{k=1}^{d} \beta_k x_{ik}$$



Why LSE?

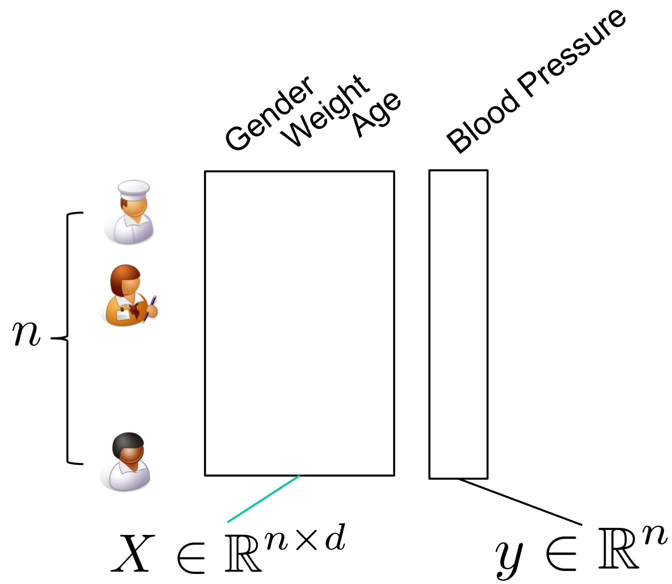


Estimate of β

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - \langle \beta, x_i \rangle)^2$$
$$= \arg\min_{\beta \in \mathbb{R}^d} \|X\beta - y\|_2^2$$
$$= (X^T X)^{-1} X^T y$$

Tufts

Reason: If Noise is Gaussian, LSE is an MLE!



$$y_i = \beta^\top x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

ε_i i.i.d., $\mathbb{E}[\varepsilon_i] = 0, \mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$

- Suppose, in addition, that

$$\varepsilon_i \sim N(0, \sigma^2)$$

Then, the negative log-likelihood of the labels is:

$$-\log\left(P\left(y|\beta, X\right)\right) = -\log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-(y_i - \beta^\top x_i)^2/2\sigma^2}\right)$$

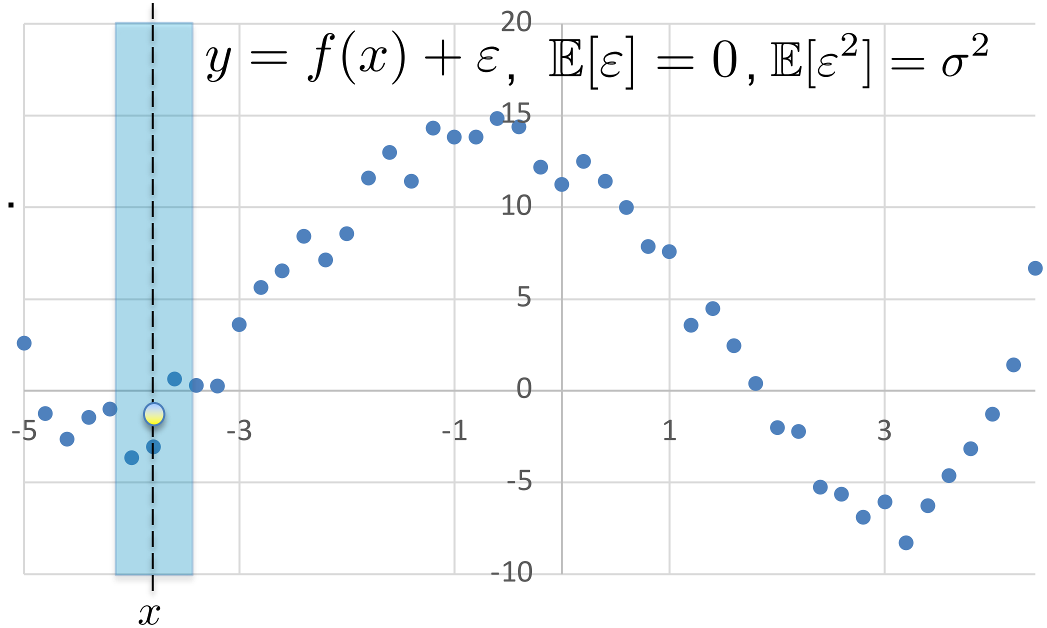
$$= \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^\top x_i)^2 + C$$

Bias vs. Variance Trade-off

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

ε_i i.i.d., $\mathbb{E}[\varepsilon_i] = 0\,, \mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty\,.$

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$



Expected Prediction Error (EPE):

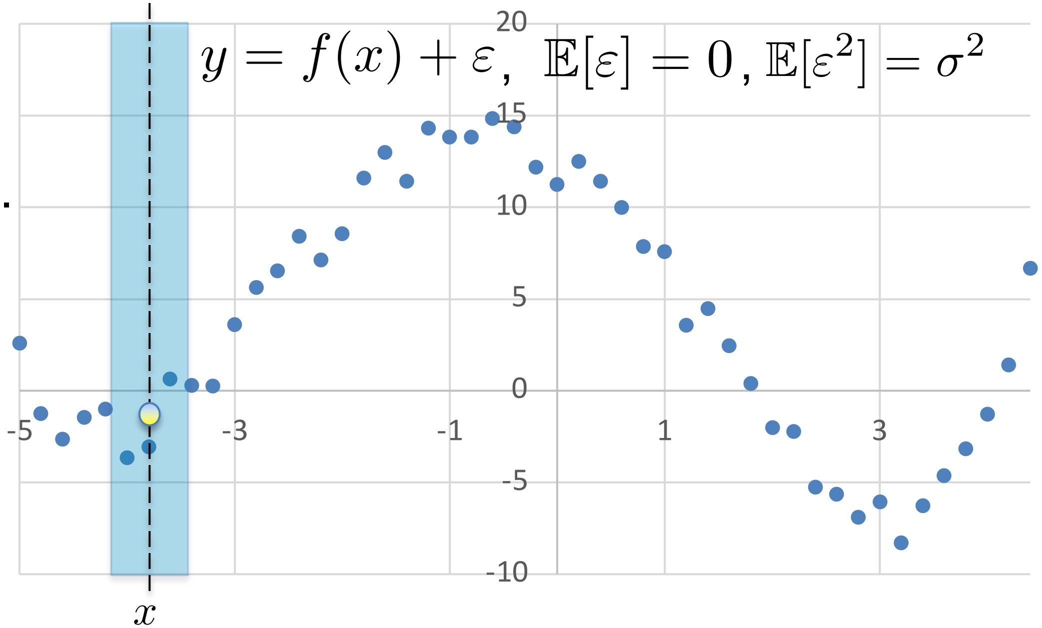
$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathbb{E}\left[(y - \mathbb{E}[y])^2\right] + \left(\mathbb{E}[y] - \mathbb{E}[\hat{f}(x)]\right)^2 + \mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right]$$

Bias vs. Variance Trade-off

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

ε_i i.i.d., $\mathbb{E}[\varepsilon_i] = 0\,, \mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty\,.$

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$



Expected Prediction Error (EPE):

$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \underbrace{\mathbb{E}\left[(y - \mathbb{E}[y])^2\right]}_{\text{inherent noise}} + \left(\mathbb{E}[y] - \mathbb{E}[\hat{f}(x)]\right)^2 + \mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right]$$

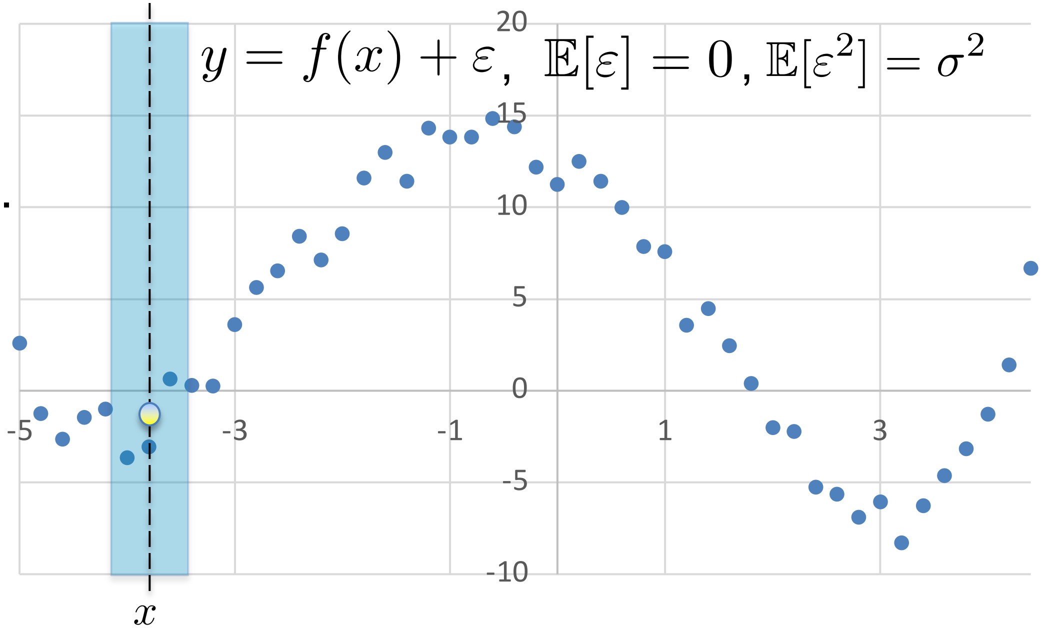
Tufts

Bias vs. Variance Trade-off

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

ε_i i.i.d., $\mathbb{E}[\varepsilon_i] = 0\,, \mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty\,.$

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$



Expected Prediction Error (EPE):

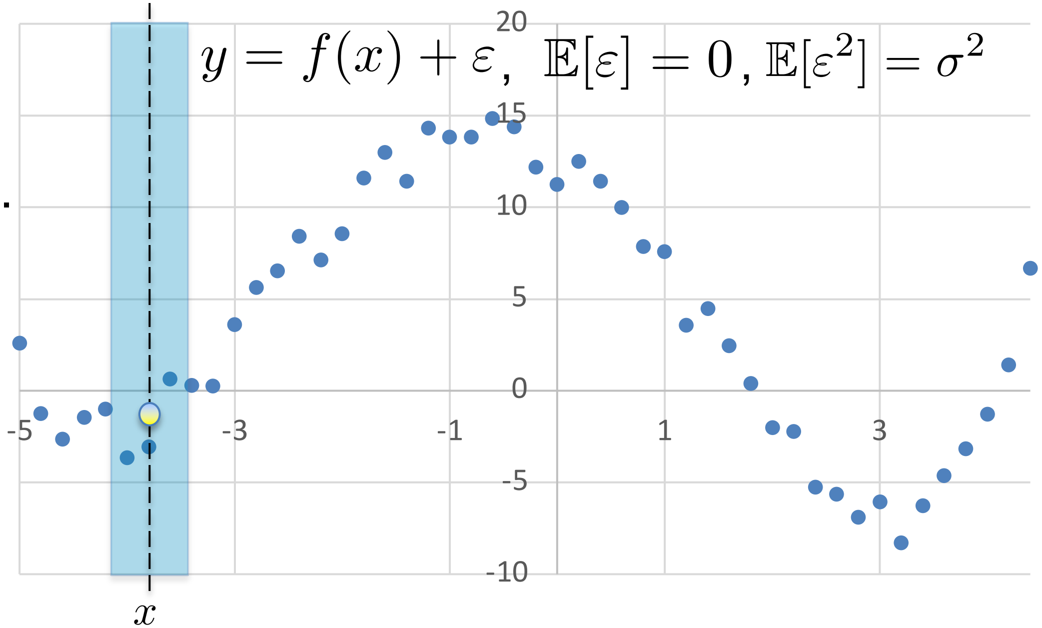
$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathbb{E}\left[(y - \mathbb{E}[y])^2\right] + \underbrace{\left(\mathbb{E}[y] - \mathbb{E}[\hat{f}(x)]\right)}_{\text{estimator } \textbf{bias}}^2 + \mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right]$$

Bias vs. Variance Trade-off

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

ε_i i.i.d., $\mathbb{E}[\varepsilon_i] = 0\,, \mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty\,.$

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$



Expected Prediction Error (EPE):

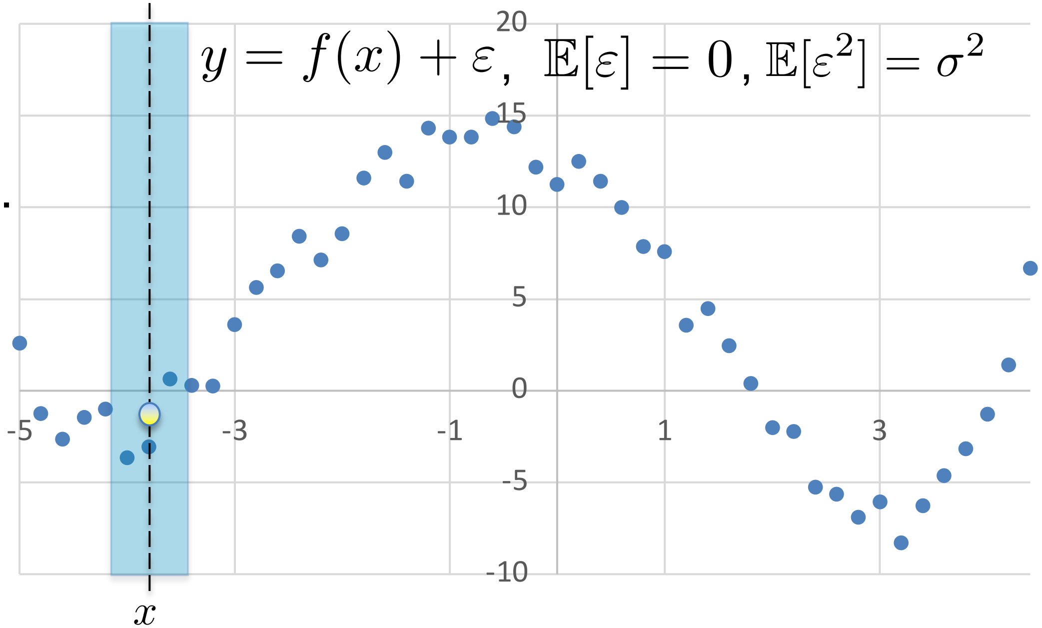
$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathbb{E}\left[(y - \mathbb{E}[y])^2\right] + \left(\mathbb{E}[y] - \mathbb{E}[\hat{f}(x)]\right)^2 + \underbrace{\mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right]}_{\text{estimator } \textbf{variance}}$$

Tufts

Bias vs. Variance Trade-off

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

ε_i i.i.d., $\mathbb{E}[\varepsilon_i] = 0\,,\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty\,.$



$$\hat{f}(x) = \frac{1}{k}\sum_{i \in N_k(x)} y_i$$

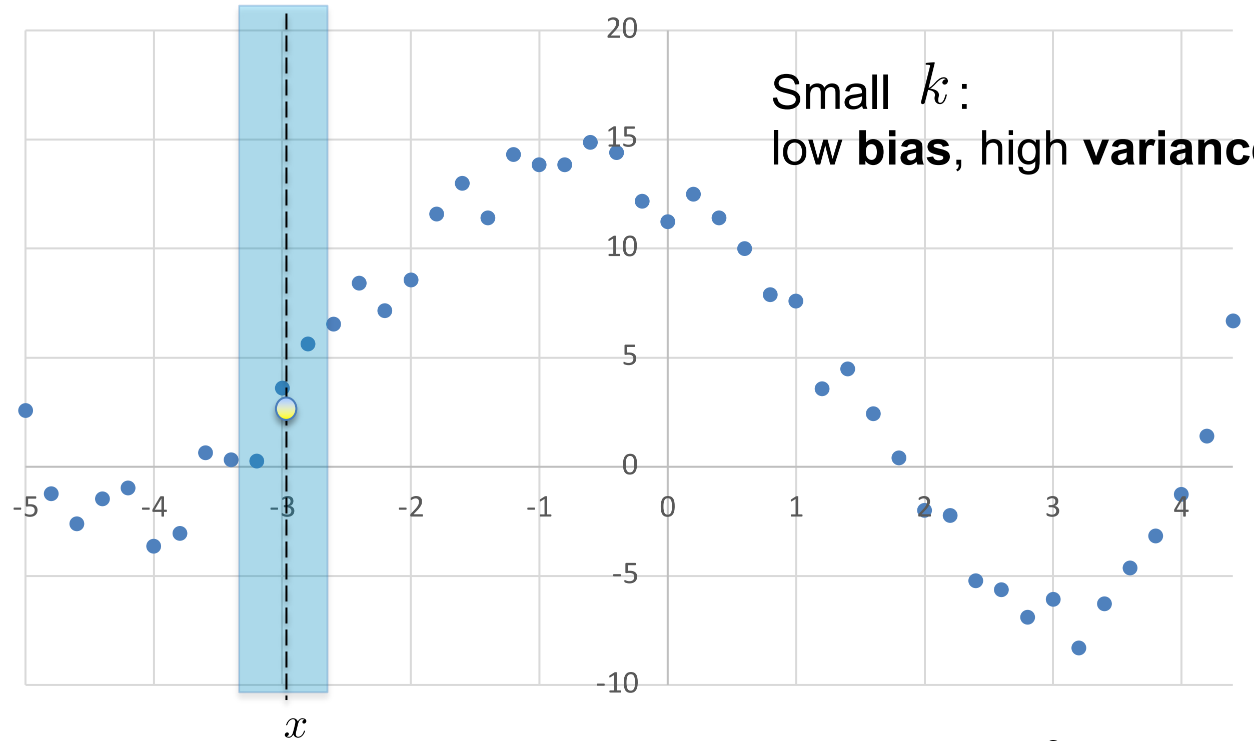
Expected Prediction Error (EPE):

$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathbb{E}\left[(y - \mathbb{E}[y])^2\right] + \left(\mathbb{E}[y] - \mathbb{E}[\hat{f}(x)]\right)^2 + \mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right]$$

$$= \sigma^2 + \left(f(x) - \frac{1}{k}\sum_{i \in N_k(x)} f(x_i)\right)^2 + \frac{\sigma^2}{k}$$

Tufts

Bias vs. Variance Trade-off

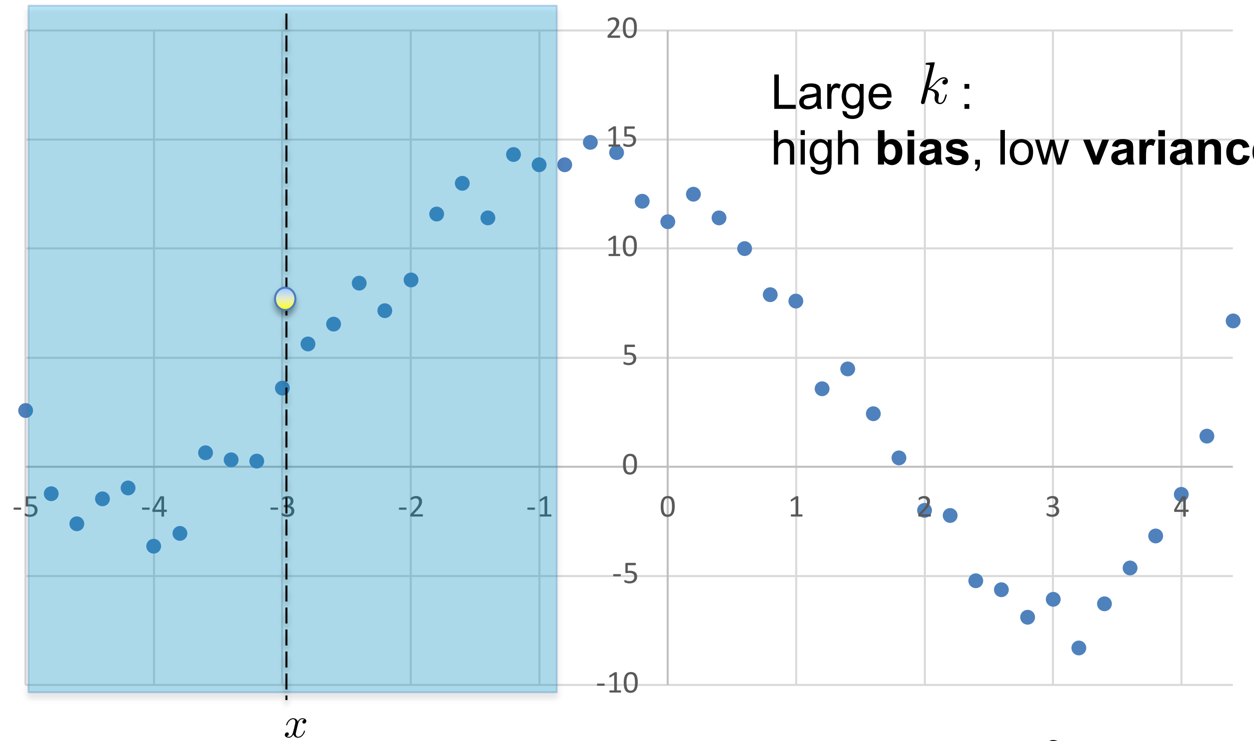


Small k:
low **bias**, high **variance**

EPE: $\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \sigma^2 + \underbrace{\left(f(x) - \frac{1}{k}\sum_{i \in N_k(x)} f(x_i)\right)^2}_{\text{estimator } \textbf{bias}} + \underbrace{\frac{\sigma^2}{k}}_{\text{estimator } \textbf{variance}}$

Tufts

Bias vs. Variance Trade-off



Large k :
high **bias**, low **variance**

EPE: $$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \sigma^2 + \underbrace{\left(f(x) - \frac{1}{k}\sum_{i \in N_k(x)} f(x_i)\right)^2}_{\text{estimator bias}} + \underbrace{\frac{\sigma^2}{k}}_{\text{estimator variance}}$$

Tufts

Bias vs. Variance Tradeoff