

S&DS 265 / 565

Introductory Machine Learning

Stochastic Gradient Descent and Bias-Variance Tradeoffs

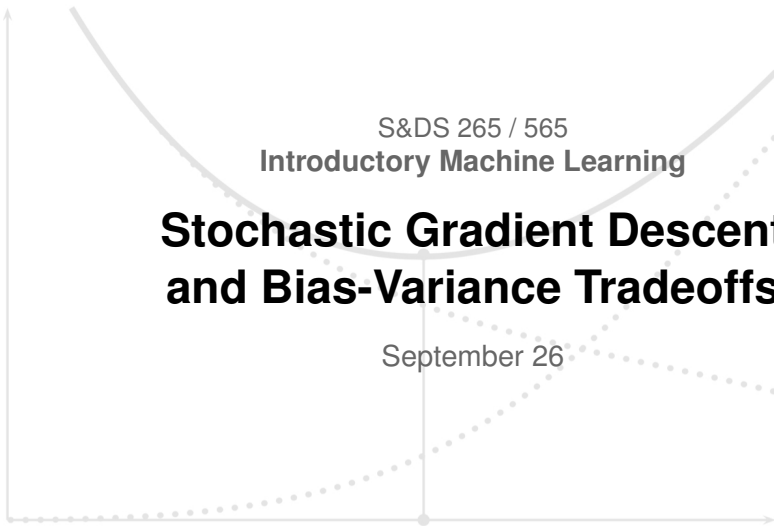
September 26

Risk

Bias squared

Variance








Yale



Goings on

- Assignment 2 is out; due next Thursday
- Quiz 2 this Thursday
- Midterm: October 17 (in class)

Readings

			examples		classification	
4	Sept 19, 21	Stochastic gradient descent	 SGD examples	Tue: Classification (continued) Thu: Stochastic gradient descent	ISL Section 6.2.2 ISL Section 10.7.2	Assn 1 in  Assn 2 out
5	Sept 26, 28	Bias and variance, cross-validation	 Bias-variance tradeoff  Covid trends (revisited)  California housing	Tue: Bias and variance Thu: Cross-validation	ISL Section 2.2 ISL Section 5.1	Quiz 2
6	Oct 3, 5	Tree-based	 Trees and forests Visualizing trees	Tue: Trees — —	ISL Sections 8.1, 8.2	Assn 2 in 

Sample quiz questions



Question 1

2 pts

Suppose that we have a training dataset with three predictor variables, and we fit a linear regression model using only two of the variables.

If we add in the third variable and refit the model, the **training error** (residual sum of squares) can only decrease or stay the same.

☐ True

☐ False

Sample quiz questions



Question 2

2 pts

Suppose that we have a training dataset with three predictor variables, and we fit a linear regression model using only two of the variables.

If we add in the third variable and refit the model, the **testing error** (residual sum of squares on test data) can only decrease or stay the same.

☐ True

☐ False

Sample quiz questions



A logistic regression model is built to predict if an individual will have greater than the median income ($Y = 1$) or less than the median income ($Y = 0$) from the number of months of education X they have received. The model takes the form

$$\mathbb{P}(Y = 1 | X = x) = \frac{e^{x/6 - 24}}{1 + e^{x/6 - 24}}$$

Sample quiz questions



Question 3

2 pts

Under this model, what is the decision boundary? Give your answer as a specific number of months; answer with only the number.

Sample quiz questions



Question 4

2 pts

Suppose person A has $X = 144$ months of education, and person B has $X = 156$ months of education. How much greater is the log-odds for person B than the log-odds for person A? Answer with only the number.

Outline for today

- Stochastic gradient descent (redux)
- Regularization
- Jupyter notebook example
- Bias-variance tradeoffs

SGD idea

- For each parameter β_j , see what happens to the loss if that parameter is increased a little bit.
- If the loss goes down (up), then increase (decrease) β_j proportionately
- Do this simultaneously for all of the parameters
- Rinse and repeat

Stochastic gradient descent

Initialize all parameters to zero: $\beta_j = 0, j = 1, \dots, p$.

Read through the data one record at a time, and update the model.

- 1 Read data item x
- 2 Make a prediction $\hat{y}(x)$
- 3 Observe the true response/label y
- 4 Update the parameters β so \hat{y} is closer to y

Stochastic gradient descent

Suppose we are doing *linear regression*. We initialize all parameters to zero: $\beta_j = 0, j = 1, \dots, p$.

We read through the data one record at a time, and update the model.

- 1 Read data item x
- 2 Make a prediction $\hat{y}(x) = \sum_{j=1}^p \beta_j x_j$
- 3 Observe the true response/label y
- 4 Update the parameters β so \hat{y} is closer to y

SGD: Computation for squared error

Change β_j by a little bit:

$$\beta_j \rightarrow \beta_j + \varepsilon$$

SGD: Computation for squared error

Change β_j by a little bit:

$$\beta_j \rightarrow \beta_j + \varepsilon$$

What happens to the squared error?

$$\begin{aligned}(y - \hat{y})^2 &\rightarrow (y - \hat{y} - \varepsilon x_j)^2 \\ &\approx (y - \hat{y})^2 + \underbrace{-2(y - \hat{y})x_j}_{\text{derivative of loss}} \varepsilon\end{aligned}$$

SGD: Computation for squared error

Change β_j by a little bit:

$$\beta_j \rightarrow \beta_j + \varepsilon$$

What happens to the squared error?

$$\begin{aligned}(y - \hat{y})^2 &\rightarrow (y - \hat{y} - \varepsilon x_j)^2 \\ &\approx (y - \hat{y})^2 + \underbrace{-2(y - \hat{y})x_j}_{\text{derivative of loss}} \varepsilon\end{aligned}$$

Use adjustment

$$\begin{aligned}\beta_j &\rightarrow \beta_j + \overbrace{-\eta \cdot \text{derivative of loss}}^{\varepsilon} \\ &= \beta_j + \eta \cdot 2(y - \hat{y})x_j\end{aligned}$$

SGD: Computation for squared error

Change β_j by a little bit:

$$\beta_j \rightarrow \beta_j + \varepsilon$$

What happens to the squared error?

$$\begin{aligned}(y - \hat{y})^2 &\rightarrow (y - \hat{y} - \varepsilon x_j)^2 \\ &\approx (y - \hat{y})^2 + \underbrace{-2(y - \hat{y})x_j}_{\text{derivative of loss}} \varepsilon\end{aligned}$$

Use adjustment

$$\begin{aligned}\beta_j &\rightarrow \beta_j + \overbrace{-\eta \cdot \text{derivative of loss}}^{\varepsilon} \\ &= \beta_j + \eta \cdot 2(y - \hat{y})x_j\end{aligned}$$

Squared error then decreases:

$$(y - \hat{y})^2 \approx (y - \hat{y})^2 - \eta \cdot \text{derivative of loss} \underline{\text{squared}}$$

SGD for general loss

Suppose $L(y, \beta^T x)$ is the loss for an input (x, y) , e.g., $(y - \beta^T x)^2$

SGD update, for a small step size $\eta > 0$:

$$\beta_j \leftarrow \beta_j - \eta \frac{\partial L(y, \beta^T x)}{\partial \beta_j}$$

$$\beta \leftarrow \beta - \eta \nabla_{\beta} L(y, \beta^T x) \quad (\text{vector notation})$$

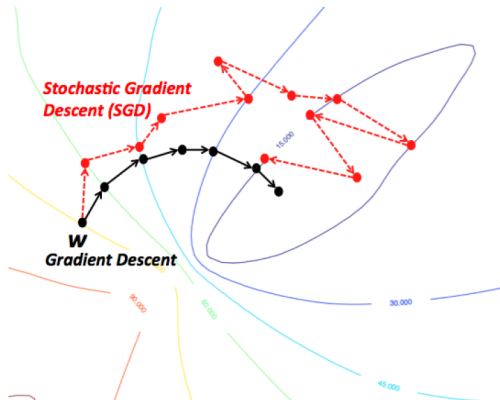
“Batch” gradient descent uses the entire training set in each step of gradient descent.

Stochastic gradient descent computes a quick approximation to this gradient, using only a single or a small “mini-batch” of data points

Batch vs. stochastic gradient descent

- The average derivative over a mini-batch can be thought of as a noisy version of the average derivative over the entire data set
- (Which can in turn be thought of as a sample estimate of a population)
- The stochastic gradient is computed more cheaply, and updating the parameters makes progress more quickly

Batch vs. stochastic gradient descent



<https://wikidocs.net/3413>

SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\beta^T x)}$$

Case checking:

- Suppose $y = 1$ and probability $p(x)$ is high?

SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\beta^T x)}$$

Case checking:

- Suppose $y = 1$ and probability $p(x)$ is high? *small change*
- Suppose $y = 1$ and probability $p(x)$ is small?

SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\beta^T x)}$$

Case checking:

- Suppose $y = 1$ and probability $p(x)$ is high? *small change*
- Suppose $y = 1$ and probability $p(x)$ is small? *big change* \uparrow
- Suppose $y = 0$ and probability $p(x)$ is small?

SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\beta^T x)}$$

Case checking:

- Suppose $y = 1$ and probability $p(x)$ is high? *small change*
- Suppose $y = 1$ and probability $p(x)$ is small? *big change* \uparrow
- Suppose $y = 0$ and probability $p(x)$ is small? *small change*
- Suppose $y = 0$ and probability $p(x)$ is big?

SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\beta^T x)}$$

Case checking:

- Suppose $y = 1$ and probability $p(x)$ is high? *small change*
- Suppose $y = 1$ and probability $p(x)$ is small? *big change* \uparrow
- Suppose $y = 0$ and probability $p(x)$ is small? *small change*
- Suppose $y = 0$ and probability $p(x)$ is big? *big change* \downarrow

SGD: choice of learning rate

Often the learning rate η_t is adjusted experimentally to get good convergence properties for a particular problem

We require that η_t decreases as t —the number of steps so far—increases

SGD: Regularization

A “ridge” penalty $\frac{1}{2}\lambda \sum_{j=1}^d \beta_j^2$ is easily handled.

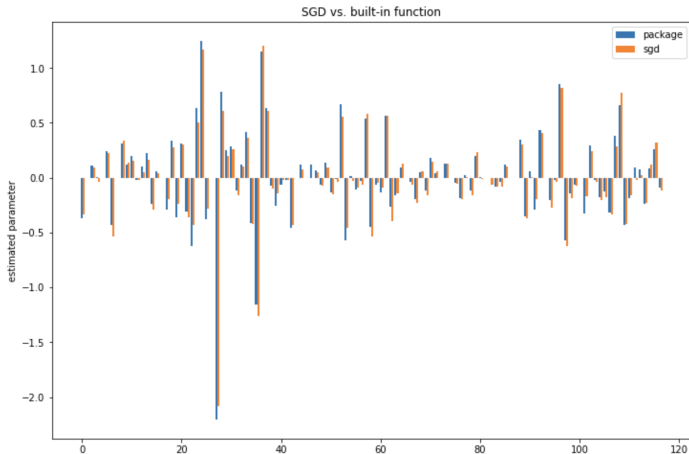
Gradient changes by an additive term $2\lambda\beta_j$. Update becomes

$$\begin{aligned}\beta_j &\longleftarrow \beta_j + \eta\{(y - p(x))x_j - \lambda\beta_j\} \\ &= (1 - \eta\lambda)\beta_j + \eta(y - p(x))x_j\end{aligned}$$

Check that this “does the right thing” whether β_j wants to be large positive or negative.

- *The penalty shrinks β_j toward zero*

Recall from demo



Each bar indicates an estimated parameter $\hat{\beta}_j$. The estimates from SGD are very similar to those obtained using the package.

Bias and variance

Bias: How much are we off—on average?

Variance: How variable are we—on average?

Bias and variance

$$\text{Bias: } \theta - \mathbb{E}\hat{\theta}$$

$$\text{Variance: } \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2$$

Bias and variance

Examples of $\theta, \hat{\theta}$:

Estimating height, population, election outcome, ad click rate...

Bias and variance

$$\text{Bias: } \theta - \mathbb{E}\hat{\theta}$$

$$\text{Variance: } \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2$$

Bias and variance

$$\text{Bias: } \theta - \mathbb{E}\hat{\theta}$$

$$\text{Variance: } \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2$$

- $\hat{\theta}$ is an estimate from a sample
- \mathbb{E} is the expectation (average) with respect to the sample
- So $\mathbb{E}\hat{\theta}$ is the average estimate
- We can only directly compute $\hat{\theta}$ for the sample we have
- *We don't know θ*

Bias and variance

In machine learning, bias and variance are two sides of a coin:
As squared bias goes up, variance goes down (and vice-versa)

Bias-variance tradeoff

$$\text{Risk} = \text{Bias}^2 + \text{Variance}$$

Bias-variance tradeoff

$$\mathbb{E}(\theta - \hat{\theta})^2 = \text{Bias}(\hat{\theta})^2 + \text{Variance}(\theta)$$

Bias-variance tradeoff

$$\mathbb{E}(\theta - \hat{\theta})^2 = (\theta - \mathbb{E}\hat{\theta})^2 + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2$$

Bias-variance tradeoff

Proof:

$$\mathbb{E}(\theta - \hat{\theta})^2 = \mathbb{E}(\theta - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \hat{\theta})^2$$

Bias-variance tradeoff

Proof:

$$\begin{aligned}\mathbb{E}(\theta - \hat{\theta})^2 &= \mathbb{E}(\theta - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \hat{\theta})^2 \\ &= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 - 2\mathbb{E}\left\{(\theta - \mathbb{E}\hat{\theta})(\hat{\theta} - \mathbb{E}\hat{\theta})\right\} + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2\end{aligned}$$

Bias-variance tradeoff

Proof:

$$\begin{aligned}\mathbb{E}(\theta - \hat{\theta})^2 &= \mathbb{E}(\theta - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 - 2\mathbb{E}\left\{(\theta - \mathbb{E}\hat{\theta})(\hat{\theta} - \mathbb{E}\hat{\theta})\right\} + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 - 2(\theta - \mathbb{E}\hat{\theta})\mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta}) + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2\end{aligned}$$

Bias-variance tradeoff

Proof:

$$\begin{aligned}\mathbb{E}(\theta - \hat{\theta})^2 &= \mathbb{E}(\theta - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 - 2\mathbb{E}\left\{(\theta - \mathbb{E}\hat{\theta})(\hat{\theta} - \mathbb{E}\hat{\theta})\right\} + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 - 2(\theta - \mathbb{E}\hat{\theta})\mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta}) + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2\end{aligned}$$

Bias-variance tradeoff

Proof:

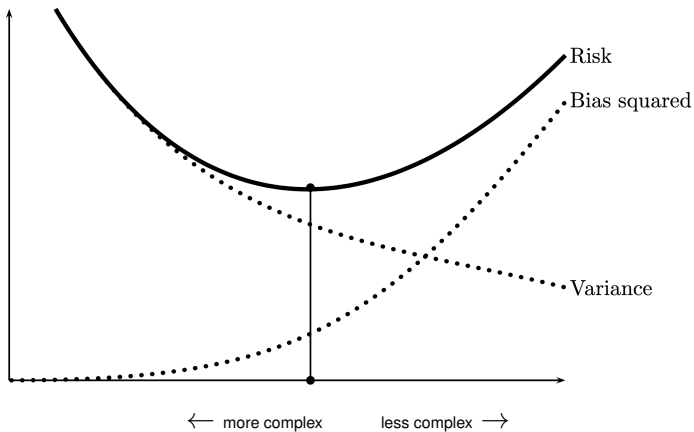
$$\begin{aligned}\mathbb{E}(\theta - \hat{\theta})^2 &= \mathbb{E}(\theta - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 - 2\mathbb{E}\left\{(\theta - \mathbb{E}\hat{\theta})(\hat{\theta} - \mathbb{E}\hat{\theta})\right\} + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 - 2(\theta - \mathbb{E}\hat{\theta})\mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta}) + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 \\&= \mathbb{E}(\theta - \mathbb{E}\hat{\theta})^2 + \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 \\&= \text{Bias}(\hat{\theta})^2 + \text{Variance}(\hat{\theta})\end{aligned}$$

Bias-variance tradeoff

If $Y = \theta + \text{noise}$, with $\mathbb{E}(\text{noise}) = 0$ and $\text{Var}(\text{noise}) = \sigma^2$,

$$\text{Risk} = \mathbb{E}[(Y - \hat{\theta})^2] = \text{Bias}^2 + \text{Variance} + \sigma^2$$

Bias-variance tradeoff



Example: Regularization

Suppose that $\mathbb{E}(Y) = \theta$ and we estimate $\hat{\theta}$ by minimizing

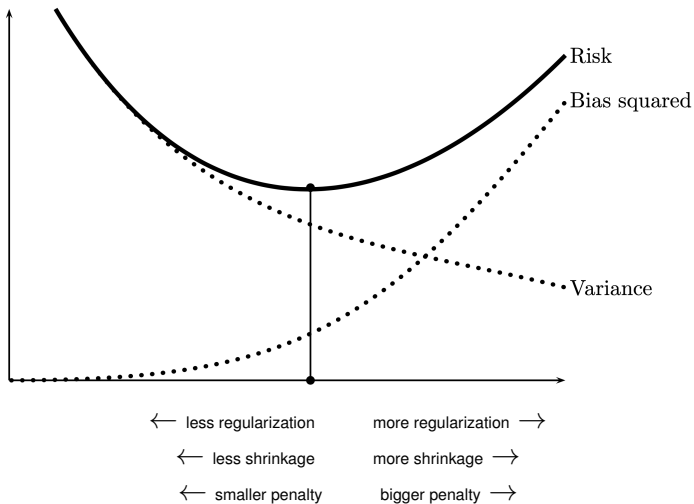
$$(Y - \theta)^2 + \lambda\theta^2$$

Then $\hat{\theta} = \frac{Y}{1+\lambda}$. What are the squared bias and variance?

$$\text{Bias}^2 = \theta^2 \left(\frac{\lambda}{1+\lambda} \right)^2$$

$$\text{Variance} = \left(\frac{1}{1+\lambda} \right)^2 \text{Variance}(Y)$$

Bias-variance tradeoff

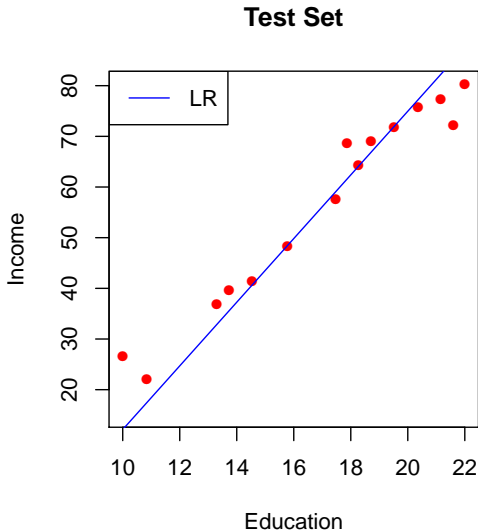


Intuition

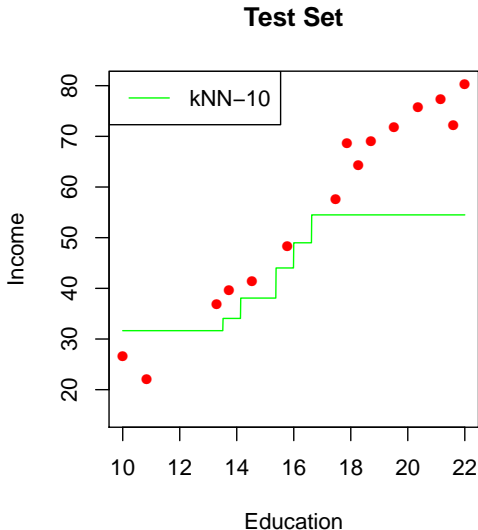
An informal description of the bias-variance tradeoff:

- As the fitted curve becomes more “wiggly” (complex) we’re seeing higher variance
- Averaging over different datasets, the “wiggles” average out to give a good fit to the data — this is low bias
- But we can’t do this averaging in practice, because we only have a single dataset

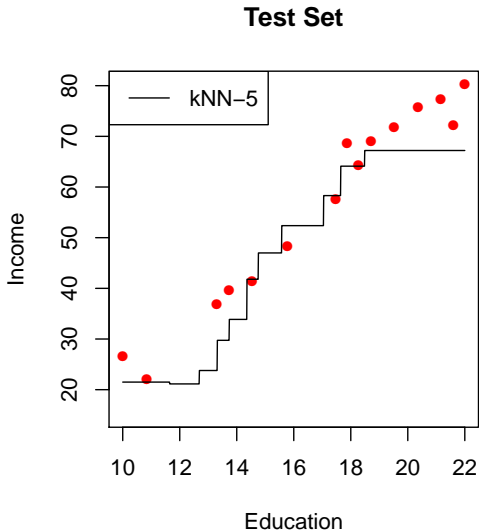
Recall: k -nearest neighbors



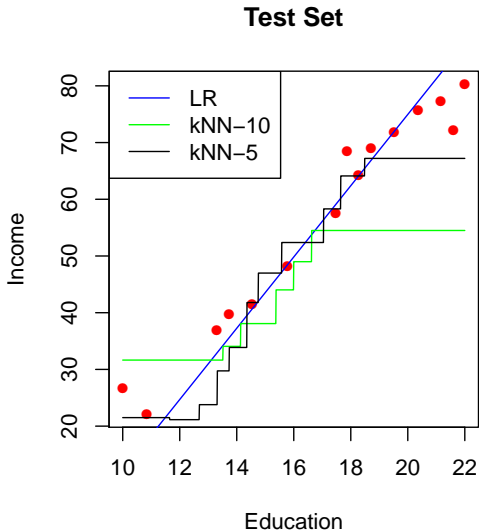
Recall: k -nearest neighbors



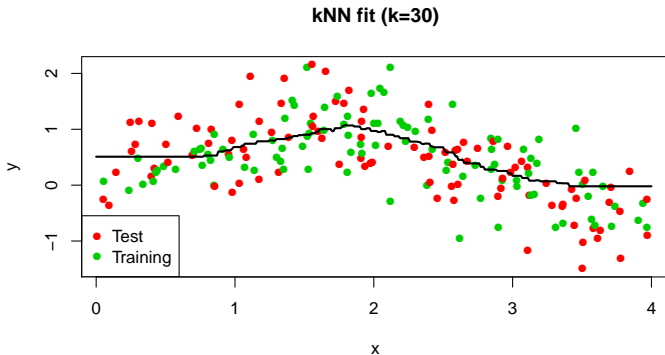
Recall: k -nearest neighbors



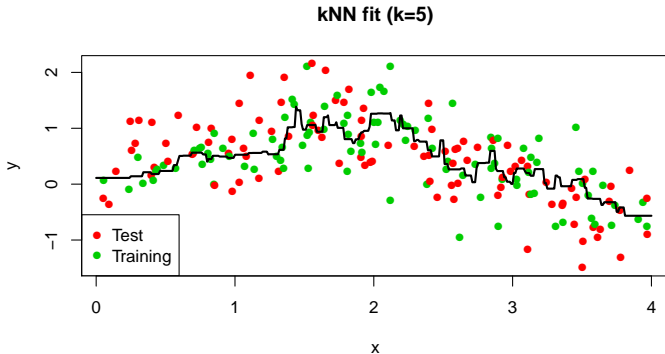
Recall: k -nearest neighbors



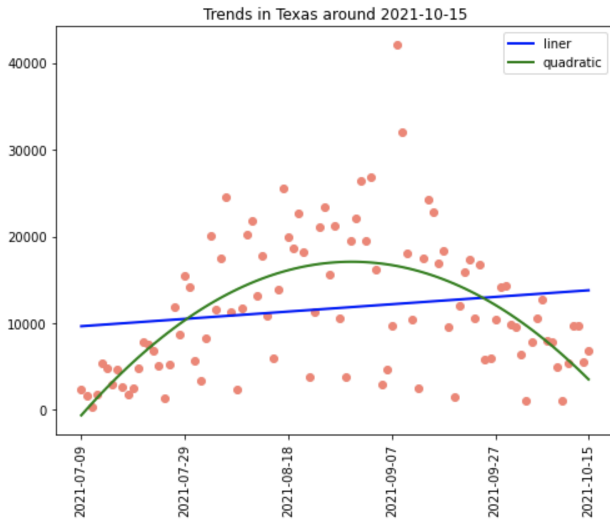
Recall: k -nearest neighbors



Recall: k -nearest neighbors



Recall: Covid example



What did we learn today?

- In SGD, a parameter is updated according to how much the loss changes when that parameter is changed by a little bit
- Mean squared error splits into squared bias plus variance
- As model complexity increases, squared bias decreases while variance increases