CS 360: Machine Learning

Prof. Sara Mathieson

Fall 2019

HAVERFORD COLLEGE

Admin

- Lab 2 due **TONIGHT!**
- Lab 1 graded (I will put on Moodle soon)
- Office hours today **12:30-1:30pm (L310)**
- **Reading quiz: Thursday**
  - Duame 7.6 (2+ pages)
  - ISL 59-63 (4+ pages)

Outline for September 17

- Finish Decision Trees (recap continuous features)
- Learning problem so far + terminology
- Bias-Variance tradeoff
- Linear regression

Linear Regression Goals

- Regression as a way to study *expected loss* and the *bias-variance tradeoff*
- Review matrix algebra and expected values
- As an introduction to optimization (specifically *stochastic gradient descent*)

Outline for September 17

- Finish Decision Trees (recap continuous features)
- Learning problem so far + terminology
- Bias-Variance tradeoff
- Linear regression

Decision Trees: base cases summary

1) All examples have the same label ★
2) No more features remain to split on ★

Duame

3) Partition does not contain any examples
4) Maximum depth reached

5) (recommended) No features produce information gain

i.e. all have same remaining features but there is still label heterogeneity

Decision Trees: implementation ideas

1) Make sure you can accommodate more than two children (i.e. not a binary tree)

2) Make sure your prediction/classification algorithm is recursive

3) You can parse the feature name to figure out continuous/discrete and how to classify

```
age<=44.5
```

Continuous Features

	~~temp~~	$t \le 44$	$t \le 54$	$t \le 85$	play tennis
x_1	~~80~~	F			Y
x_2	~~48~~	F			Y
x_3	~~60~~	F	⋮		Y
x_4	~~48~~	F	⋮		Y
x_5	~~40~~	T			N
x_6	~~48~~	F			N
x_7	~~90~~	F			N

① sort

40	48	48	48	60	80	90
N	Y	Y	N	Y	Y	N

②

40	48	60	80	90
N	None	Y	Y	N

③ split when label changes

$\boxed{t \le 44}$ $\boxed{t \le 54}$ $\boxed{t \le 85}$

Outline for September 17

- Finish Decision Trees (recap continuous features)
- Learning problem so far + terminology
- Bias-Variance tradeoff
- Linear regression

Learning Problem so far

- Performance on training data <u>overestimates</u> accuracy
- We must use a <u>held aside</u> test set to evaluate
- Both training and testing data should be drawn from the same distribution
- Training/test data should be drawn from the same distribution as seen in deployment (ideally)

Loss Functions

- E.g., zero-one loss
  - Simple accuracy - is prediction right?
  - For binary or multi-class prediction

$$l(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{otherwise} \end{cases}$$

Slide: modified from Ameet Soni

Loss Functions

- E.g., zero-one loss
  - Simple accuracy - is prediction right?
  - For binary or multi-class prediction

$$l(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{otherwise} \end{cases}$$

- E.g., squared loss
  - For regression

$$l(y, \hat{y}) = (y - \hat{y})^2$$

Slide: modified from Ameet Soni

Loss Functions

- E.g., zero-one loss
  - Simple accuracy - is prediction right?
  - For binary or multi-class prediction

$$l(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{otherwise} \end{cases}$$

- E.g., squared loss
  - For regression

$$l(y, \hat{y}) = (y - \hat{y})^2$$

- Absolute loss (also for regression)

$$\ell(y, \hat{y}) = |y - \hat{y}|$$

Slide: modified from Ameet Soni

Formalizing the learning problem

- Given:
  - Loss function, ℓ
  - A sample of data D from an unknown distribution of all data $\mathcal{D}$
  - A hypothesis space $H = \{h | h : X \rightarrow Y\}$

Slide: Ameet Soni

Formalizing the learning problem

- Given:
  - Loss function, ℓ
  - A sample of data D from an unknown distribution of all data $\mathcal{D}$
  - A hypothesis space $H = \{h | h : X \rightarrow Y\}$
- Do:
  - Find a function $f(X) \rightarrow y$ that
  - minimize error over $\mathcal{D}$ with respect to ℓ

Slide: Ameet Soni

Why might learning fail?

Inductive Bias

Training Data

class A

class B

Testing Data

Figures 2.1 and 2.2 from Duame

Inductive Bias

Training Data

class A

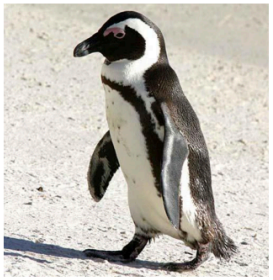
class B

Testing Data

A

A

B



B

A: “fly”

B: “no fly”

Figures 2.1 and 2.2 from Duame

Inductive Bias

Training Data

class A

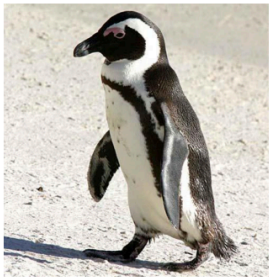
class B

Testing Data

A

B

B



A

A: “bird”

B: “mammal”

Figures 2.1 and 2.2 from Duame

Why might learning fail?

- Noise in the training data
  - Typos in a restaurant review

Why might learning fail?

- Noise in the training data
  - Typos in a restaurant review
- Available features are insufficient
  - x-ray does not capture the medical issue

Why might learning fail?

- Noise in the training data
  - Typos in a restaurant review
- Available features are insufficient
  - x-ray does not capture the medical issue
- "Correct" prediction is up to interpretation
  - Parental controls on web content

Why might learning fail?

- Noise in the training data
  - Typos in a restaurant review
- Available features are insufficient
  - x-ray does not capture the medical issue
- "Correct" prediction is up to interpretation
  - Parental controls on web content
- Learning algorithm cannot cope with the data

Adapted from Duame Chap 2

Hyperparameters

- Difficult to define precisely, but typically a parameter that controls other parameters

- What is one hyperparameter in decision trees?

Hyperparameters

- Difficult to define precisely, but typically a parameter that controls other parameters

- What is one hyperparameter in decision trees?

Max depth!

- We can't choose hyperparameters via test data (breaks cardinal rule!)

Hyperparameters

- Difficult to define precisely, but typically a parameter that controls other parameters

- What is one hyperparameter in decision trees?

Max depth!

- We can't choose hyperparameters via test data (breaks cardinal rule!)

- But we can use *validation data*

General approach to training

1. Split your data into 70% training data, 10% development data and 20% test data. (validation data)

2. For each possible setting of your hyperparameters:

   (a) Train a model using that setting of hyperparameters on the training data.

   (b) Compute this model's error rate on the development data.

3. From the above collection of models, choose the one that achieved the lowest error rate on development data.

4. Evaluate that model on the test data to estimate future test performance.

Outline for September 17

- Finish Decision Trees (recap continuous features)
- Learning problem so far + terminology
- Bias-Variance tradeoff
- Linear regression

Regression Setup

model: we see $y = f(x) + \varepsilon$ (don't know)

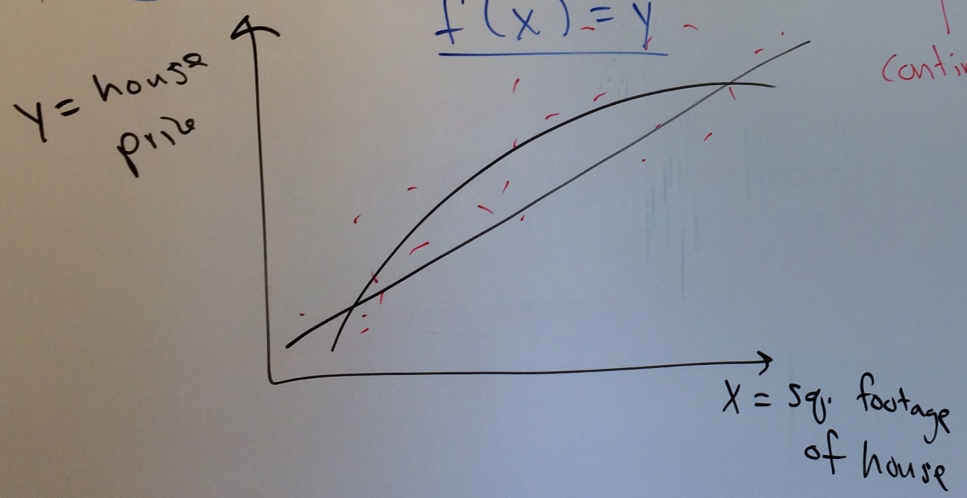
error term mean 0 (independent of x)

RV

RV

$\hat{f}$: we choose our estimate of f

$\hat{f}(x) = \hat{y}$



Expected Loss

want: $E_{(x,y)}[\ell(y, \hat{f}(x))]$ "error"

loss function.

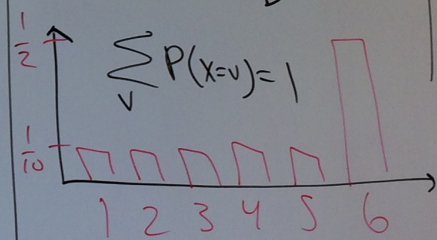
detour to expected value

$$E[X] = \sum_{v \in vals(X)} \underbrace{p(X=v)}_{\text{weight}} \cdot v$$

weighted die: D

$$\boxed{\begin{aligned} P(D=6) &= \tfrac{1}{2} \\ P(D \neq 6) &= \tfrac{1}{10} \end{aligned}}$$

$$\sum_v P(X=v) = 1$$



$$E[D] = \frac{1}{10}\cdot 1 + \frac{1}{10}\cdot 2 + \cdots \frac{1}{2}\cdot 6$$

$$\boxed{E[D] = 4.5}$$

$\rightarrow (x,y) \sim \mathcal{D} \leftarrow$ don't know this

$$\text{expected loss} = \sum_{(x,y)\in D} \overbrace{D(x,y)}^{\text{prob}}\,\overbrace{\ell(y, \hat{f}(x))}^{\text{loss}}$$

$$= \boxed{\frac{1}{n}\sum_{i=1}^{n} \underbrace{\ell(y_i, \hat{f}(x_i))}_{???}}$$

special case

D : • finite, discrete distribution

vals $\rightarrow \{(x_1, y_1), (x_2, y_2), \vdots, (x_n, y_n)\}$

• prob $\frac{1}{n}$ on each example

squared loss

$$\ell(y, \hat{y}) = (y - \hat{y})^2$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

mean squared error.

$$E[MSE] = \frac{1}{n}\sum_{i=1}^{n} E[(y_i - \hat{y}_i)^2]$$

generic y & $\hat{y}$ — true function

$$E[(y-\hat{y})^2] = E[(\underbrace{y - f}_{\text{"a"}} + \underbrace{f - \hat{f}}_{\text{"b"}})^2]$$

ε

noise

model issue

$$= \underbrace{Var(\varepsilon)}_{\text{irreducible error}} + \underbrace{E[(f - \hat{f})^2]}_{\text{reducible error}}$$

$(a+b)^2$

$= a^2 + b^2 + \cancel{2ab}^{\,0}$

here

$E[X+Y]$

$= E[X] + E[Y]$

$E[aX] = aE[X]$

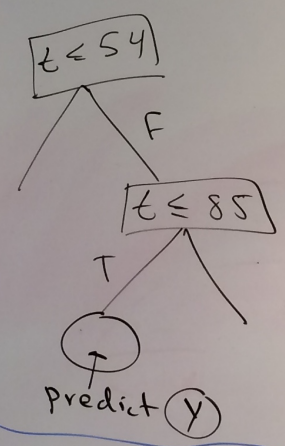
$\hat{f}(x) = \hat{y} \Rightarrow$ "$\hat{f} = \hat{y}$"

$y = f(x) + \varepsilon$

$y - f(x) = \varepsilon$

$E(X-\mu)^2 = Var(X)$

0 for ε



NNNNNNYNYNYYYNYYYYYY

Split

$E[XY] = E[X]E[Y]$

iff X & Y are independent.

$$E[(f - \hat{f})^2] = E[(\underbrace{f - E[\hat{f}]}_{\text{bias}} + \underbrace{E[\hat{f}] - \hat{f}}_{\text{var.}})^2]$$

$f(x) - \hat{f}(x)$

$$= (\text{bias}(\hat{f}(x)))^2 + \text{Var}(\hat{f}(x))$$

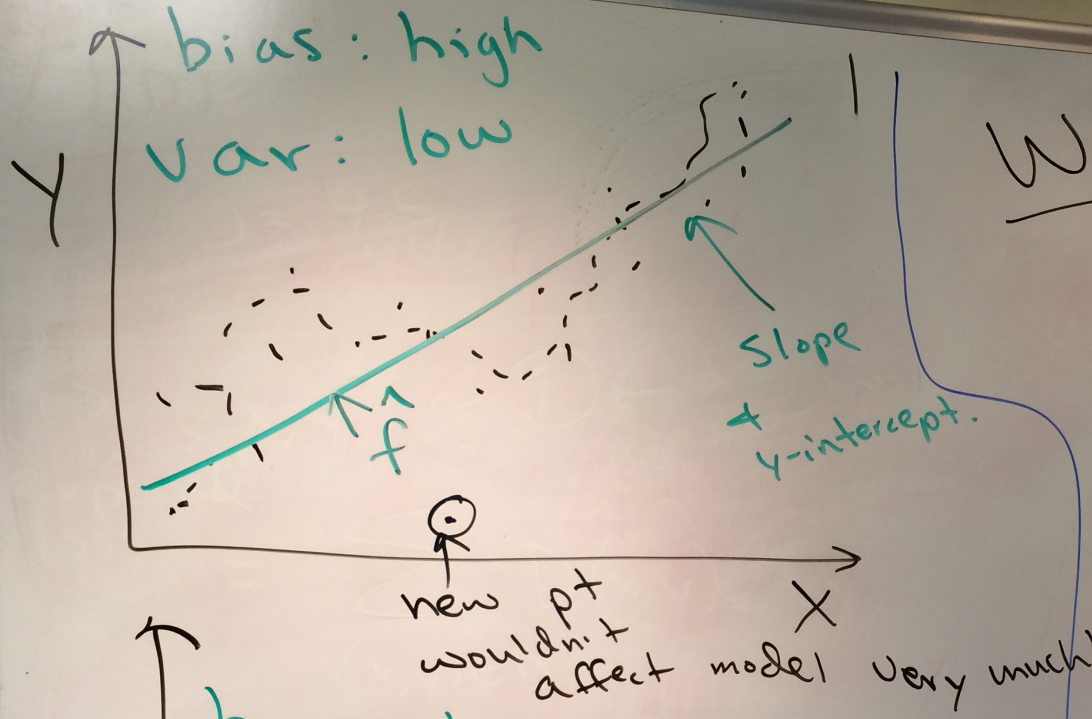
$$E[MSE] = \text{bias}(\hat{f})^2 + \text{Var}(\hat{f}) + \text{Var}(\varepsilon)$$

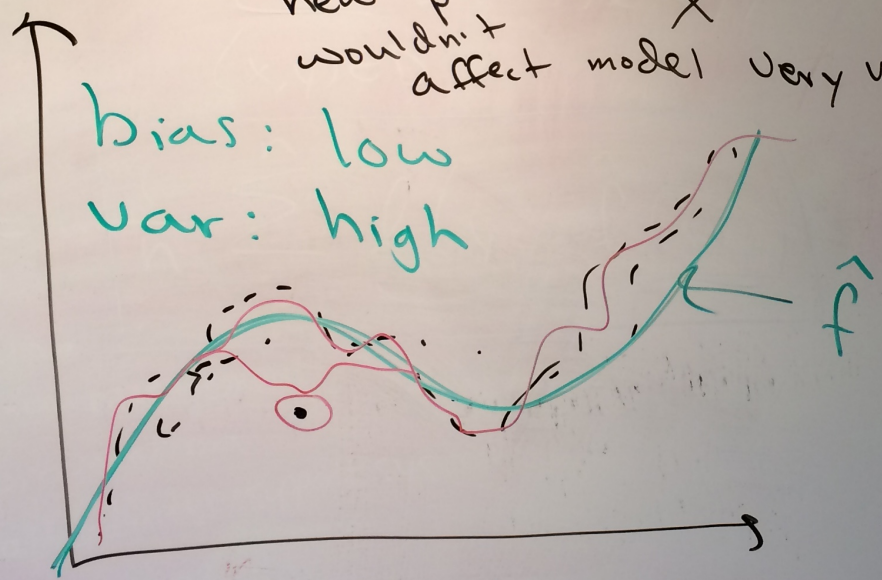
bias/variance tradeoff.

bias: error introduced by approximating a real-life problem

Variance: amount $\hat{f}$ would change if we trained on different data.

!





Want: low bias
low variance

but: as model flexibility **increases**

- Variance ↑
- bias ↓

Assessing Model Accuracy

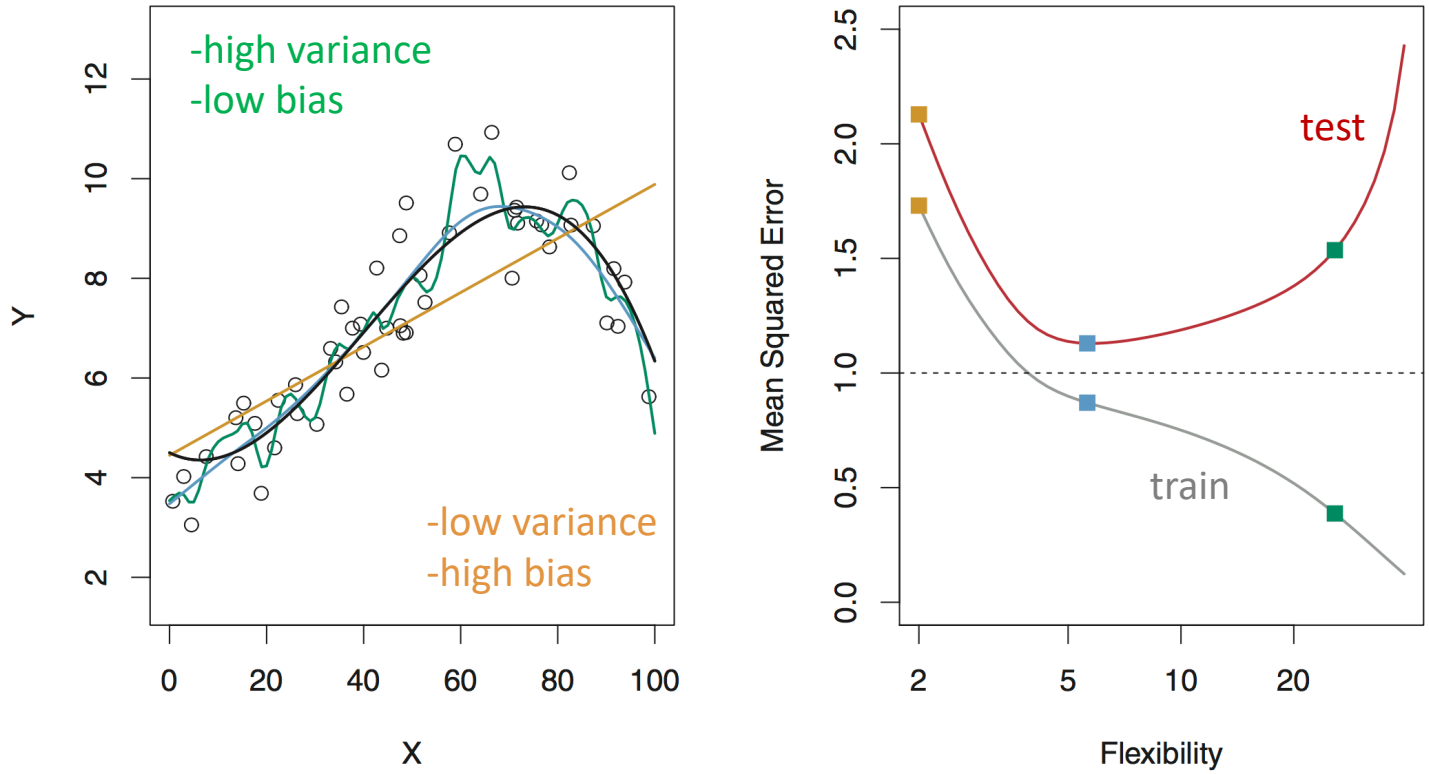


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

ISL, Fig 2.9

Outline for September 17

- Finish Decision Trees (recap continuous features)
- Learning problem so far + terminology
- Bias-Variance tradeoff
- Linear regression

Goals of Inference

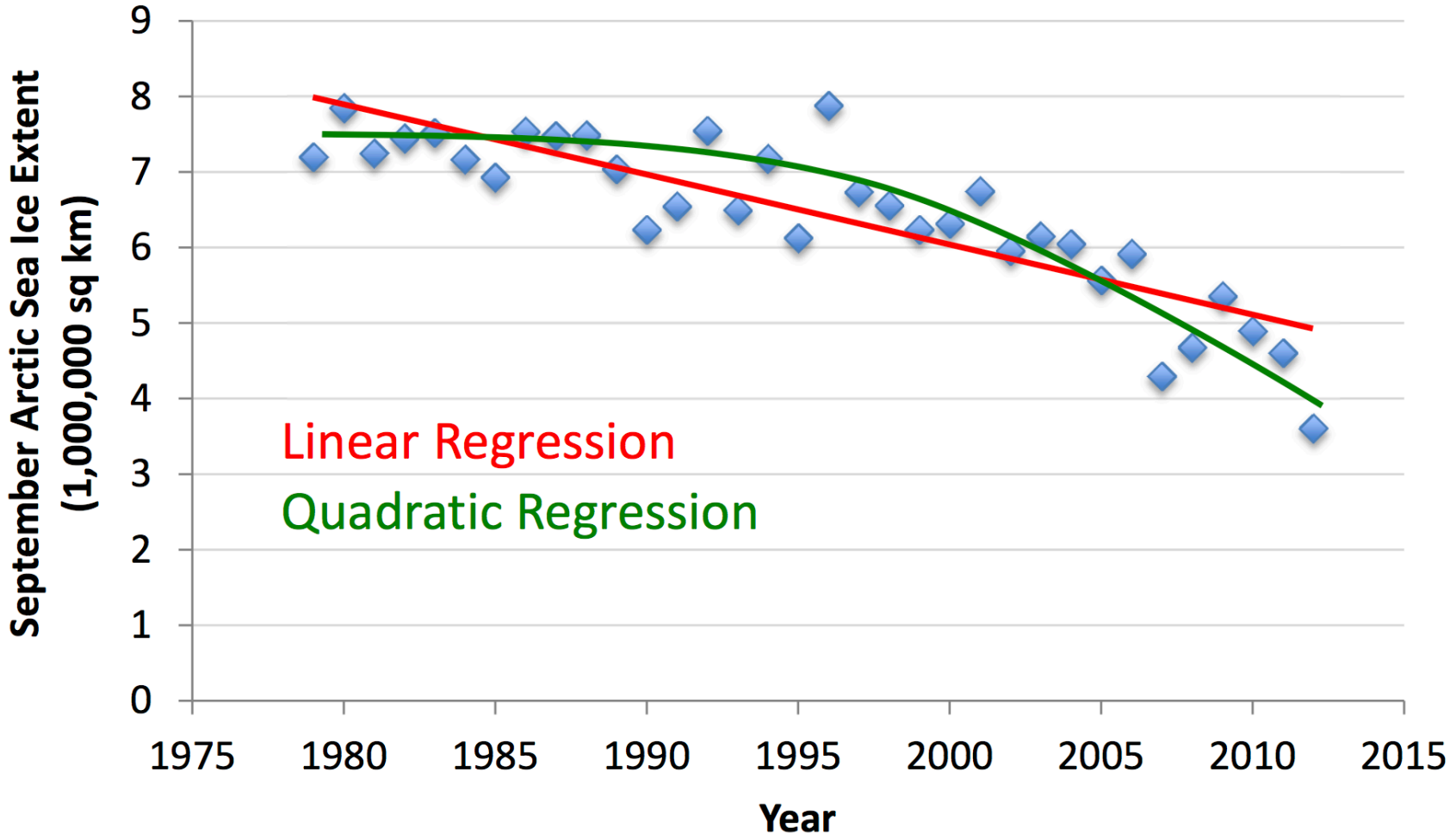
1) Which of the features/explanatory variables/predictors (x) are associated with the response variable (y)?

2) What is the relationship between x and y?

3) Is a linear model enough?

4) Can we predict y given a new x?

Regression Example



Slide: modified from Jessica Wu

[Data from G. Witt. Journal of Statistics Education, Volume 21, Number 1 (2013)]