Intro to Data Science:

Regression and Generalization

Ed Podojil Data Engineer/Scientist, Animoto

Warm Up: Probability

Pick a card game like blackjack or poker and determine how Bayes theorem could be used to improve your performance in the game

Try to come up with two or three examples specific to the game!

Post your answer to github. Try writing code that explains your probability to win given what is known about cards in play

Agenda

I. Intro to Regression

II. Regularization

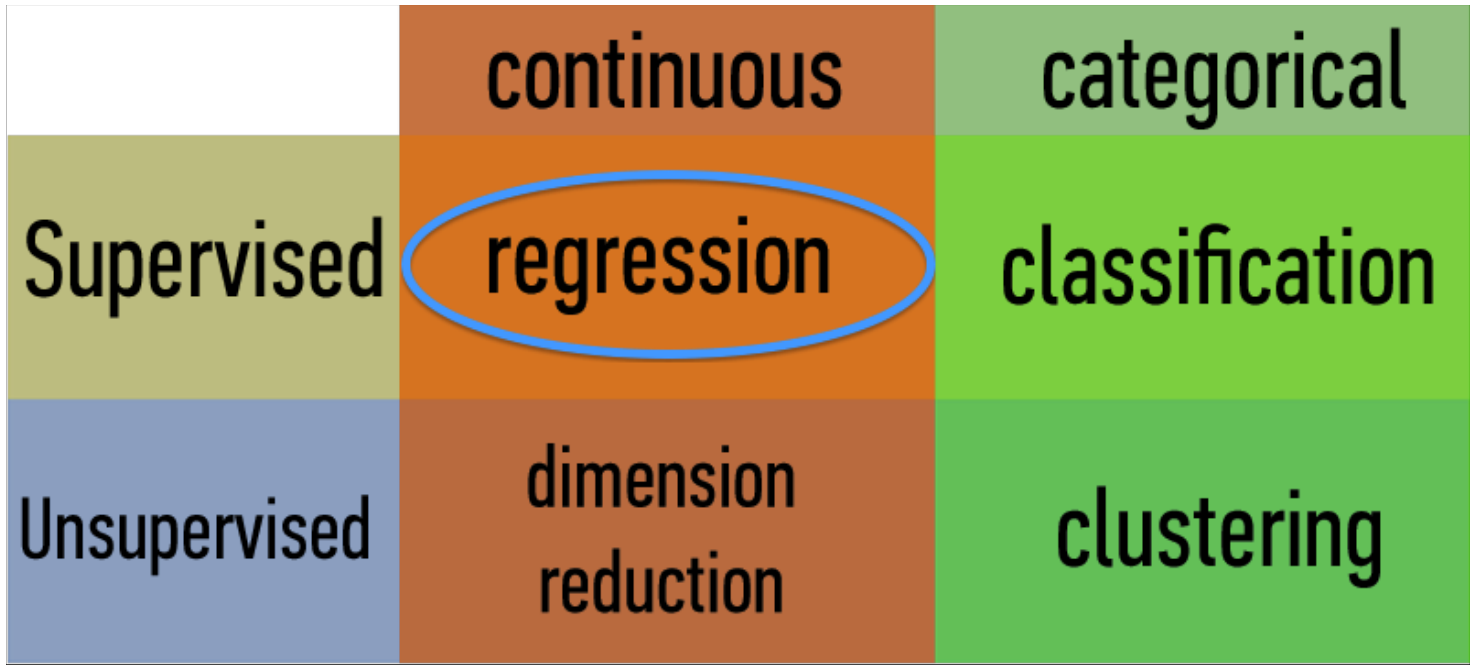
III. Implementing a Regularized Fit in R

I. Intro to Regression

Goals

- Define what a **regression model** is
- Understand the differences between a simple regression and a polynomial regression
- Understand the basics behind how regressions are fit


continuous
categorical
Supervised
Unsupervised


continuous
categorical
Supervised
regression
classification
Unsupervised
dimension reduction
clustering

What is a regression model?

A **functional relationship** between **input** & **response variables**

A **simple linear regression** model captures a linear relationship between a single input x and response variable y

$$y = \alpha + \beta x + \varepsilon$$

What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

Well, it looks a lot like something else we know...

$$y = ax + b$$

(line of best fit)

$$y = \alpha + \beta x + \varepsilon$$

y = response variable

x = input variable

α = intercept

β = regression coefficient

ε = residual (prediction error)

Common linear regression model data problems

I know the prices for all of these other apartments in my area. What could I get for mine?

What's the relationship between total number of friends, posting activity, and the number of likes a new post would get on Facebook?

Careful! Time series data (believe it or not) does not always handle well with simple regression

Multiple Regression

We can extend this model to several input variables, giving us the multiple linear regression model:

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$

Linear regression involves several technical assumptions and is often presented with lots of mathematical formality.

The math is not very important for our purposes, but you should check it out if you get serious about solving regression problems.

How do we fit a regression model to a dataset?

Minimize the sum of the squared residuals (OLS)

In practice, we'd be considering the squared distances between estimation and the data represented

Our programs will do this (heck, Excel even does)

Studying regressions is serious work! Dive into more material if you find the problem interesting

II. Polynomial Regression

Goals:

- Determine an approach to polynomial regression
- Explain what multicollinearity is

Consider the following:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$$

It represents a nonlinear relationship. Is it a linear model?

Yes!

Polynomial regression allows us to fit very complex curves to data.

However, it poses one problem, particularly in comparison to **simple linear regression**

What's the difference between simple linear regression and polynomial regression?

What's the problem that simple regression doesn't have?

Multicollinearity

Multicollinearity is when predictor variables are **highly correlated** with each other

```
x <- seq(1, 10, 0.1)
cor(x^9, x^10)
[1] 0.9987608
```

This causes the model to break down because it can't tell the difference between predictor variables

How we fix multicollinearity

Replace the correlated predictors with uncorrelated predictors

$$y = \alpha + \beta_1 f_1(x) + \beta_2 f_2(x^2) + \ldots + \beta_n f_n(x^n) + \varepsilon$$

III. Regularization

Goals

- Review overfitting
- Learn methods to prevent overfitting
- Explain the signifiance of bias over variance

Review!

What's overfitting?

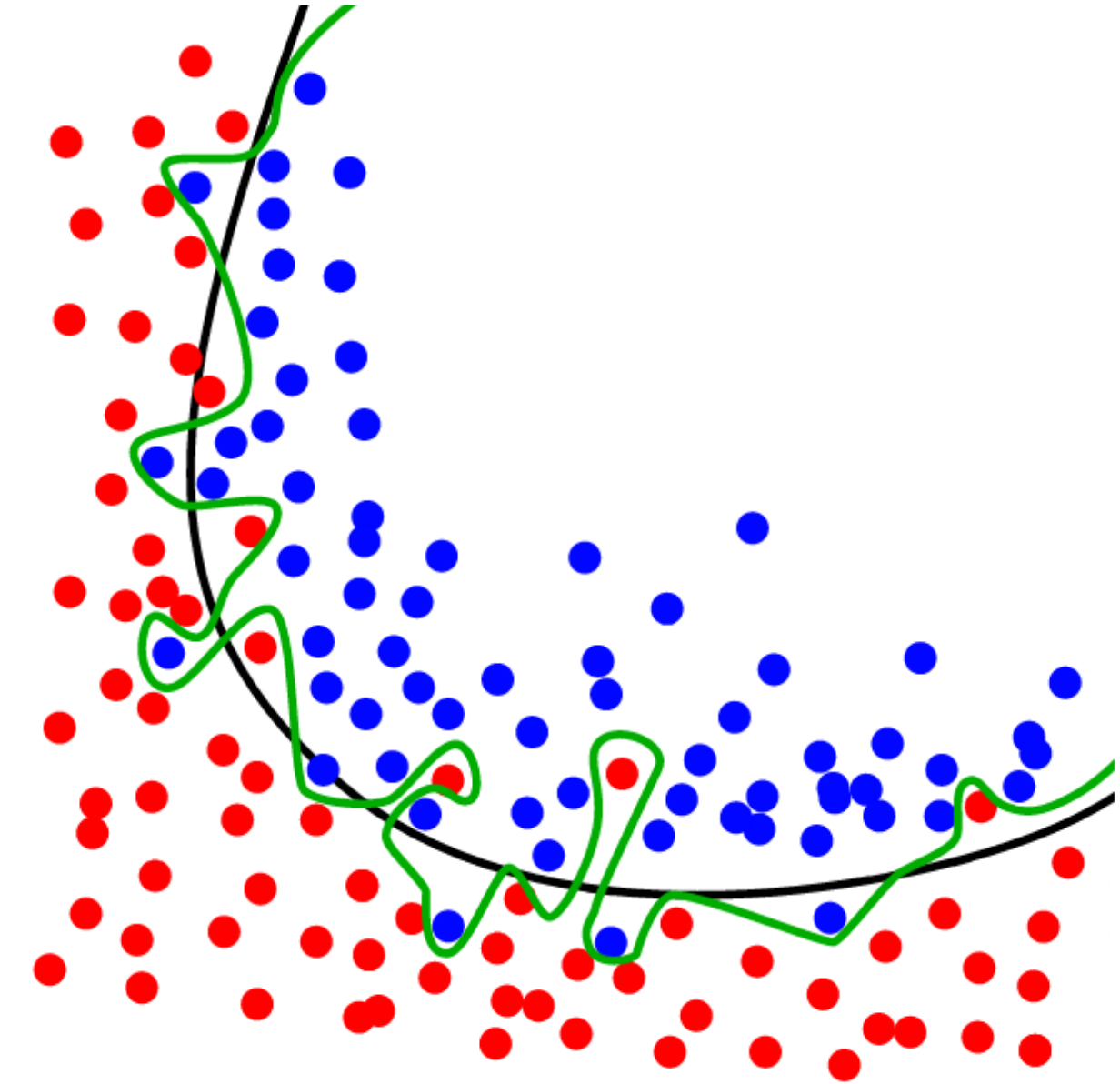
Definition upgrade: Overfitting occurs when a model matches the **noise** instead of the **signal**

Noise: Extra "cruft" that doesn't contribute to a readable prediction

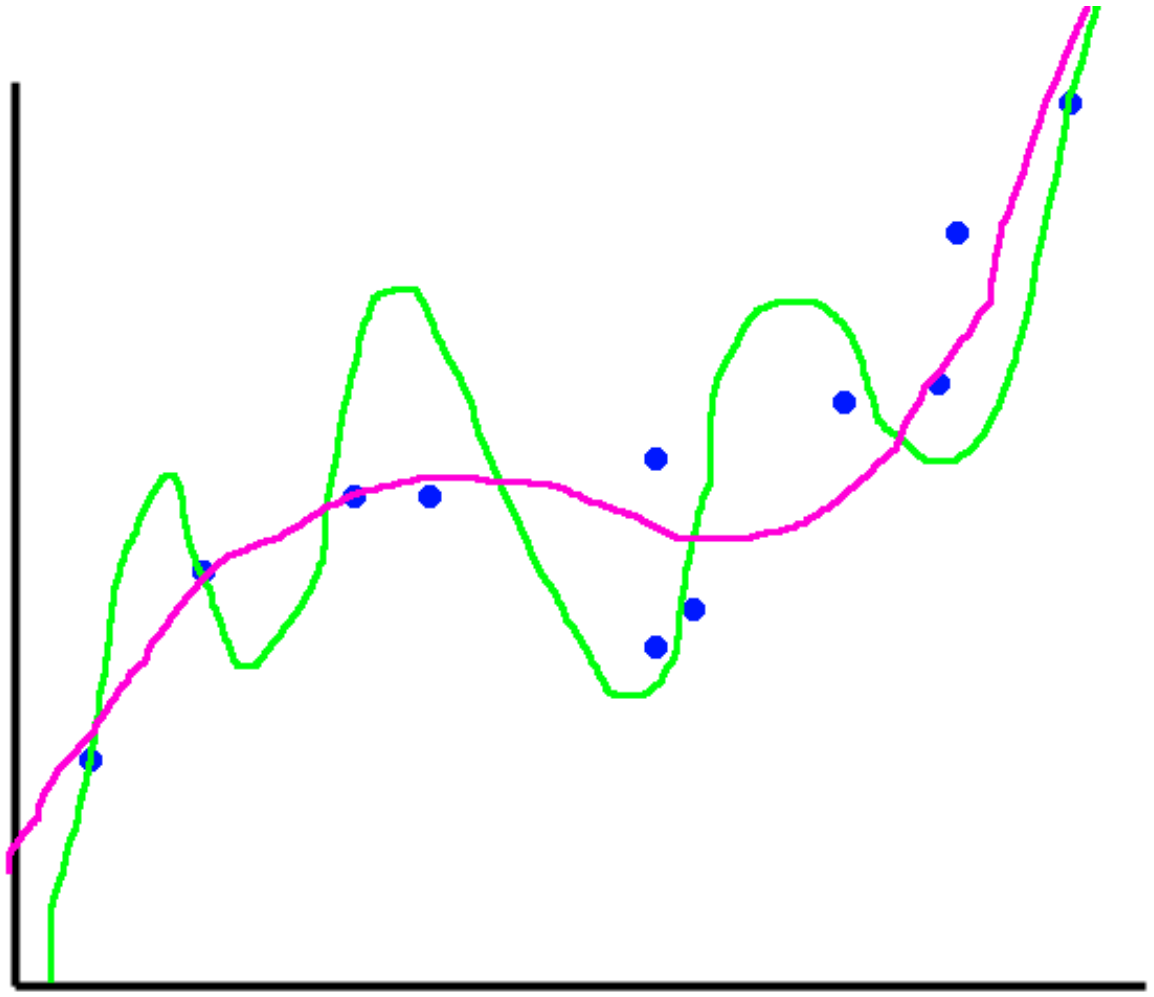
Signal: The clean, elegant interpretation based on sound science

Happens when our model becomes too complex

Overfitting example (Classification)



Overfitting example (Regression)



Defining Model Complexity

One way to define the complexity of a model is to define complexity as a function of the size of the **coefficients**.

Regularization

Regularization refers to the method of preventing overfitting by explicitly controlling model complexity.

We cover two different regularization functions

L1: $y = \Sigma \beta_i x_i + \varepsilon \; st. \; \Sigma |\beta_i| < s$

L2: $y = \Sigma \beta_i x_i + \varepsilon \; st. \; \Sigma \beta_i^2 < s$

More importantly (Ignoring the math for now...)

Use **L1 Generalization** when we have small data but many features

Use **L2 Generalization** in most other cases

If interested, there are many papers written about L1 and L2 generalization!

Bias and Variance

Bias refers to predictions that are systematically inaccurate.

Variance refers to predictions that are generally inaccurate.

Generalization error can be broken down into these two parts: bias and variance.

Bias Vs Variance

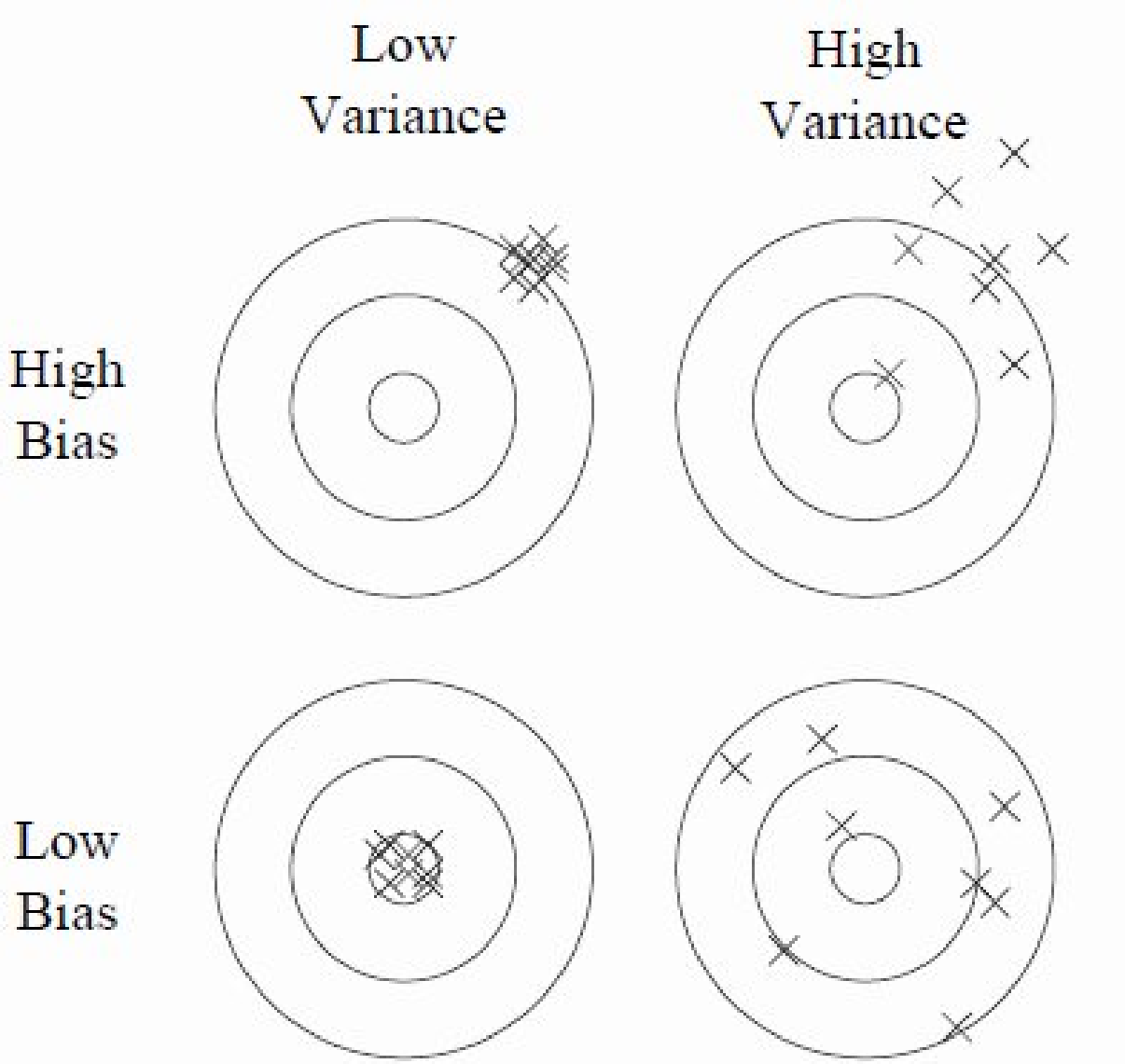
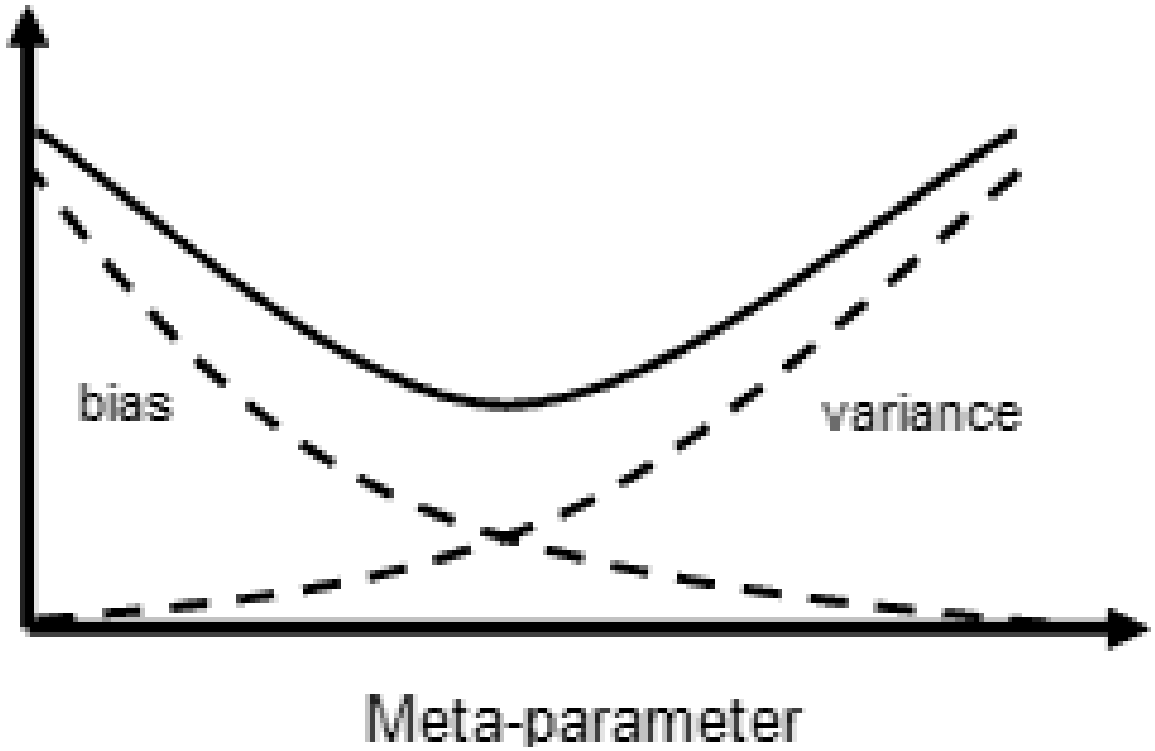


Figure 1: Bias and variance in dart-throwing.

Generalization creates bias-variance tradeoff



We should take advantage of generalization to trade off variance for bias in our fit

This will help our linear regression model create a better overall fit!

Class break

Ex: polynomial regression & regularization

Goals

- observe multicollinearity in naïve polynomial fit (lm())
- perform polynomial fit using orthogonal basis functions (poly())
- observe overfitting in polynomial fit of high degree (poly())
- perform regularized fit to control overfitting (glmnet())

Final Discussion

Questions to ponder about:

How has your understanding of linear models and linear regression changed since first class?

How is a regression different from a classification problem? How is it similar?

Do you think we could use regression for classification problems? Why or why not?