LECTURE 08:

CLASSIFICATION WRAP-UP

October 4, 2017
SDS 293: Machine Learning

Outline

- Lab: comparing classification approaches
- Evaluating models
  - Today: validation set
  - Monday: resampling

60 min: comparing classification methods

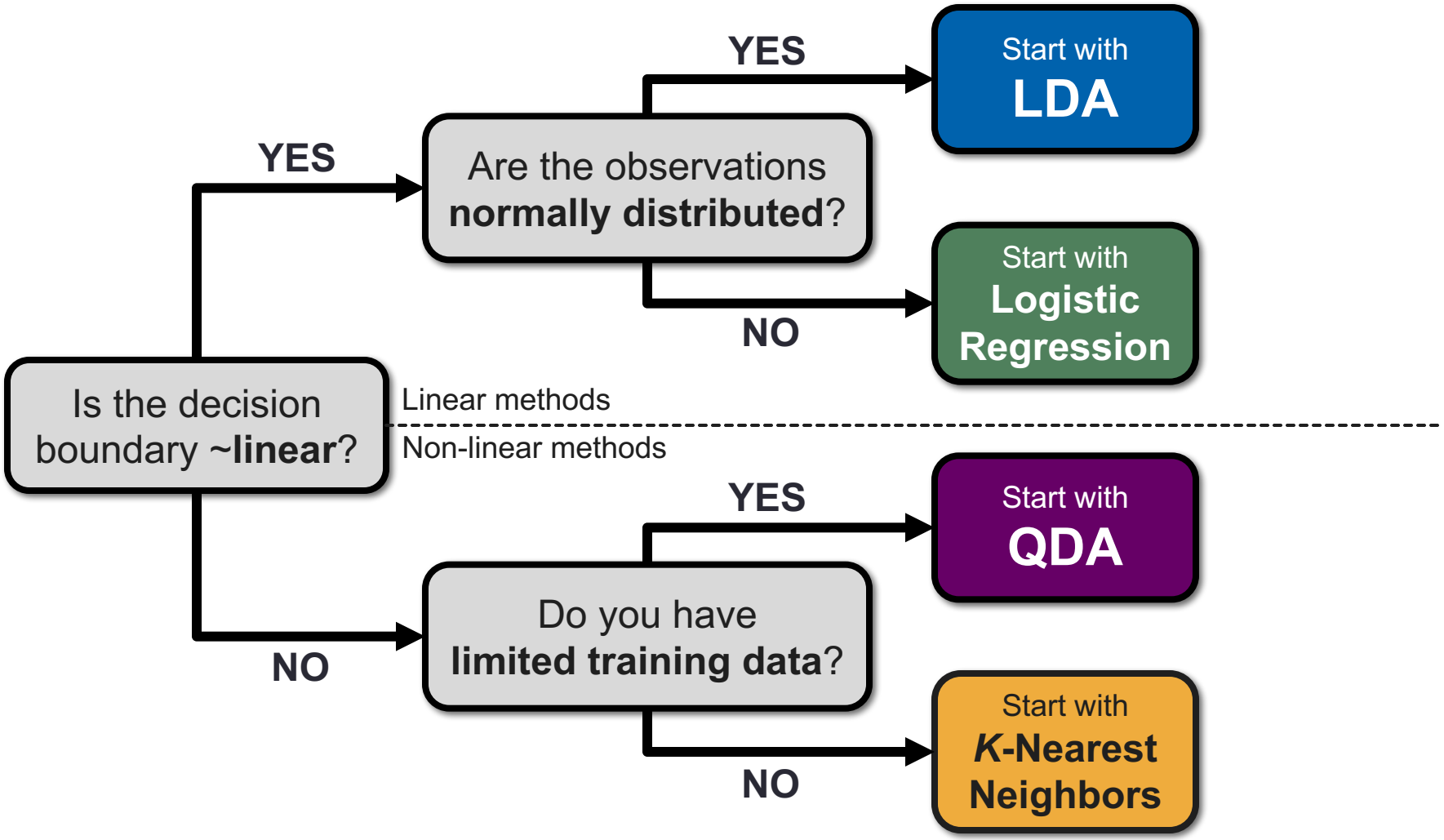
- Work in teams of 2-4 people (3 is ideal)
- **Goal:** to deepen our understanding of how the 4 classification methods we've discussed measure up
  - *K*-nearest neighbors
  - Logistic regression
  - LDA
  - QDA
- Instructions:

  http://www.science.smith.edu/~jcrouser/SDS293/labs/lab6.html
- You can check your intuition against p.153-4 of ISLR

Discussion: classification methods

	Logistic Regression	K-nearest neighbors	LDA	QDA
Scenario 1	(✓)		✓	
Scenario 2	(✓)		✓	
Scenario 3	✓		(✓)	✗
Scenario 4				✓
Scenario 5	✗	(✓)	✗	✓
Scenario 6	✗	✓	✗	✗

Rough guide to choosing a method



Test vs. training

- In all of these methods, we evaluate performance using **test error** rather than training error (why?)
- **Real life:** if the data comes in stages (i.e. trying to predict election results), we have a natural "test set"
- **Question:** when that's not the case, what do we do?
- **Answer**: set aside a "validation set"

What to set aside?

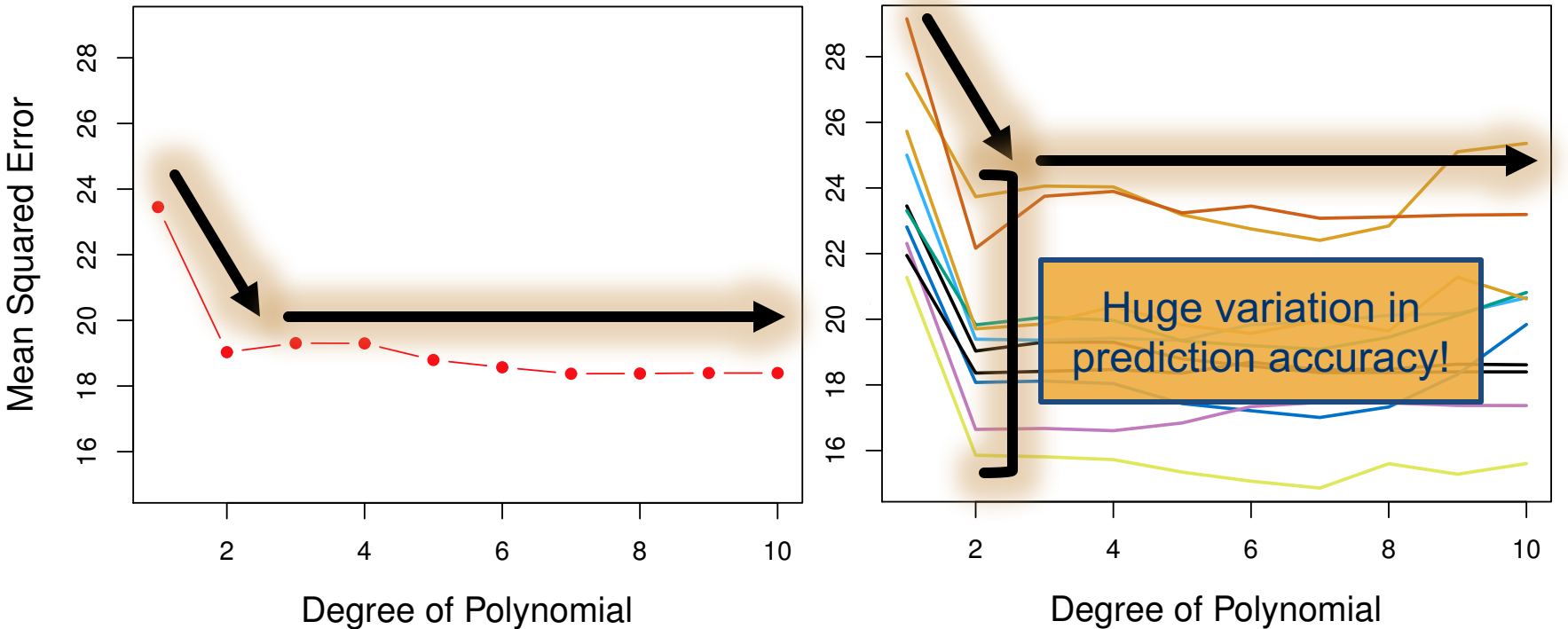
- `Smarket`: used years 2001-2004 to train, tested on 2005
- `Caravan`: used the first 1000 records as test
- `Carseats`:
  - some of you split the data in half
  - some sampled randomly
  - some split on one of the predictors (population>300, sales>7, etc.)

Question: did it make a difference?

Answer: absolutely

- **Example:** `Auto` dataset (392 observations)
- **Goal:** use linear regression to predict `mpg` using polynomial $f($`horsepower`$)$
- Randomly split into 2 sets of 196 obs. (test and training)

10x?



Issues with the “validation set” approach

1. The **test error rate** depends heavily on which observations are used for training vs. testing
2. We’re only training on a **subset** of the data (why is this a problem?)

We need a new approach...

Coming up

- Monday: **Resampling Methods**
- A2 due tonight
- A3 posted, due **Wednesday Oct. 11th by 11:59pm**