LECTURE 11:

LINEAR MODEL SELECTION PT. 2

October 18, 2017

SDS 293: Machine Learning

Announcements 1/2

CS Internship Lunch Presentations

Come hear where Computer Science majors interned in Summer 2017!

Employers range from companies in the tech industry to research labs.

All are welcome! Pizza lunch provided.

Thursday,
October 26th
12:10 - 1 pm
Ford Hall 241

Extra credit opportunity
Want to drop a missing lab? Attend and post to #talks!

Announcements 2/2

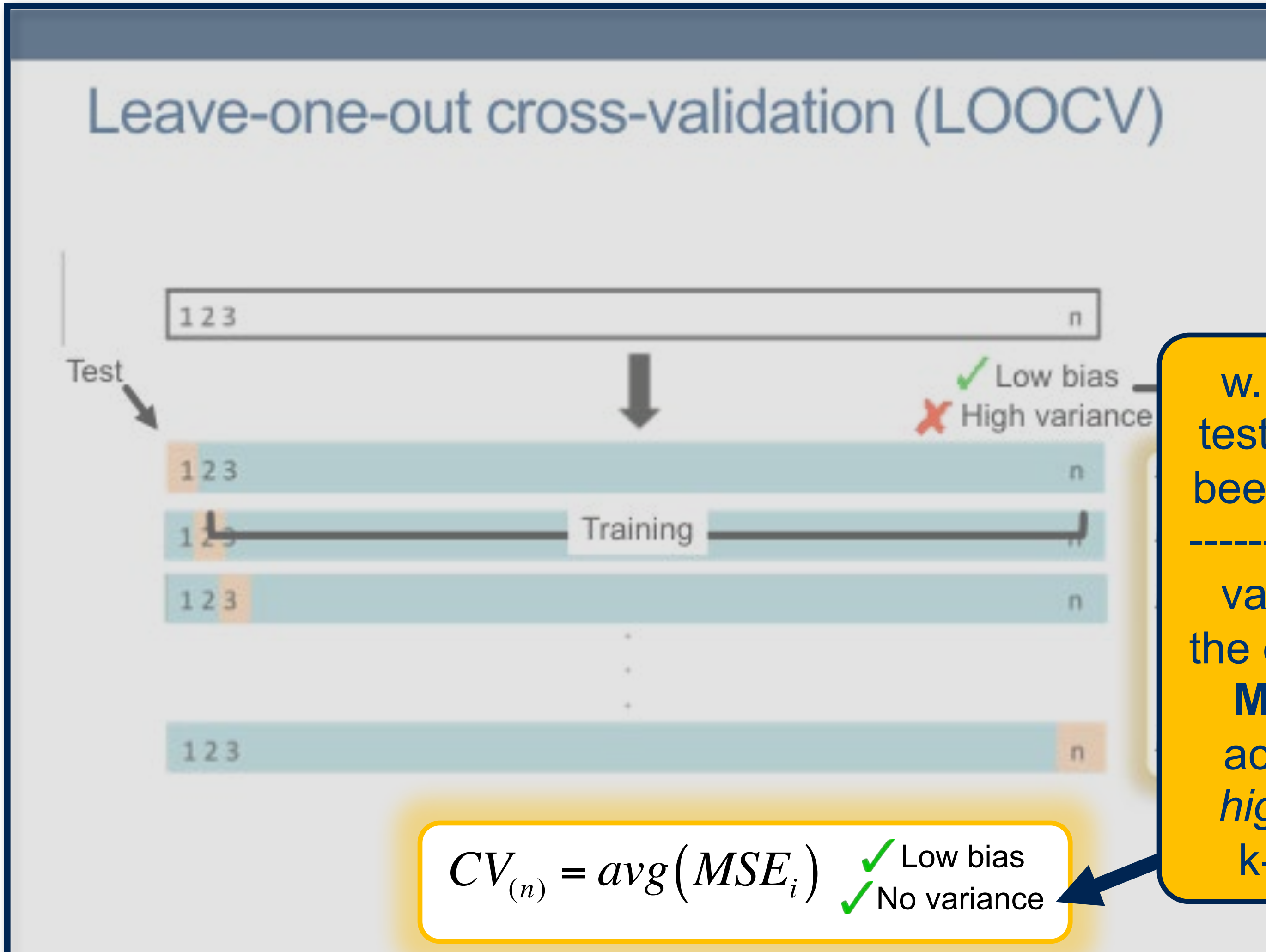
Computer Science

Silvana, Artemis, Marina and Kyra present their research posters at the Collaborations event, 4/22/17.

Presentation of the
CS Major & Minors

Monday @ lunch
Ford 240
FREE FOOD!

Correcting a fuzzy statement



$$CV_{(n)} = avg\left(MSE_i\right)$$

✓ Low bias
✓ No variance

w.r.t. which test case has been "left out"

----

variance of the **estimated MSE** may actually be *higher* than k-fold CV

Outline

- Model selection: alternatives to least-squares
- Subset selection
  - Best subset
  - Stepwise selection (forward and backward)
  - Estimating error using cross-validation
- Shrinkage methods
  - Ridge regression and the Lasso
  - Dimension reduction
- Labs for each part

Flashback: subset selection

- **Big idea:** if having too many predictors is the problem maybe we can get rid of some

- Three methods:
  - **Best subset:** try all possible combinations of predictors
  - **Forward**: start with no predictors, greedily add one at a time
  - **Backward**: start with all predictors, greedily remove one at a time

"greedy" = Add/remove whichever predictor improves your model **right now**

Flashback: comparing methods

	Best Subset Selection	Forward Selection	Backward Selection
How many models get compared?	2^p	$1+\frac{p(p+1)}{2}$	$1+\frac{p(p+1)}{2}$
Benefits?	Provably optimal	Inexpensive	Inexpensive; doesn't ignore interaction
Drawbacks?	Exhaustive search is expensive	Not guaranteed to be optimal; ignores interaction	Not guaranteed to be optimal; breaks when $p>n$

Flashback: choosing the optimal model

- We know measures of training error (RSS and R^2) aren't good predictors of test error (what we actually care about)
- Two options:
  - We can **indirectly** estimate test error by making an adjustment to the training error to account for the bias:

$$R^2_{adj} \quad C_p \quad AIC \quad BIC$$

Pros: inexpensive to compute

Cons: makes additional assumptions about the model

  - We can **directly** estimate the test error, using either a validation set approach or a cross-validation approach

Validation set: how would this work?

From the kitchen of: Grandma SDS

Recipe for: Best Subset Selection

First divide the data into training and test sets

Preheat the null model M_0 with no predictors.* on the **training set**

1. For $k = 1, 2, \ldots, p$:
   a. Fit all the models that contain exactly k predictors.
   b. Keep only the model with the smallest training error. Call it M_k.

2. ~~Estimate the error,~~ and select a single "best" model from among $M_0 \ldots M_p$

^ Calculate the error rate on **the test set**

Discussion: potential problems?

Only training on a subset of the data means our model is **less accurate**

From the kitchen of: *Grandma SDS*

Recipe for: *Best Subset Selection*

First divide the data into training and test sets

Preheat the null model M_0 with no predictors.* on the **training set**

1. For $k = 1, 2, \ldots, p$:
   a. Fit all the models that contain exactly k predictors.
   b. Keep only the model with the smallest training error. Call it M_k.

2. ~~Estimate the error,~~ and select a single "best" model from among $M_0 \ldots M_p$

^ Calculate the error rate on **the test set**

Cross-validation: how would this work?

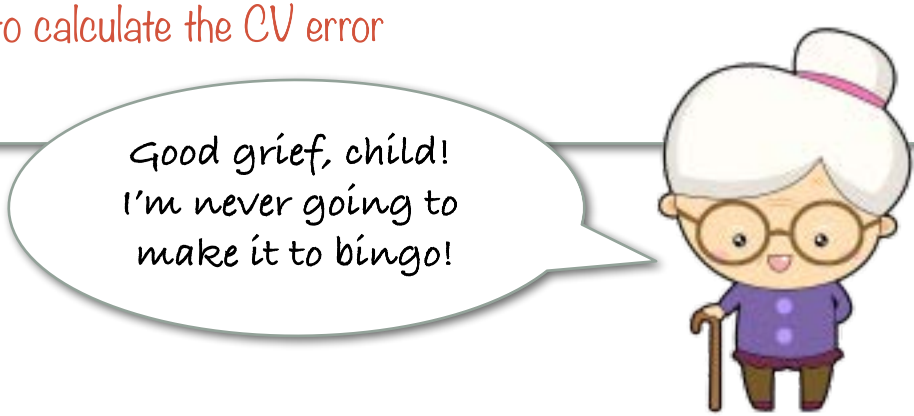
From the kitchen of: Grandma SDS

Recipe for: Best Subset Selection

Preheat the null model M_0 with no predictors.

1. For $k = 1, 2, \ldots, p$:
   a. Fit all the models that contain exactly k predictors.
   b. Keep only the model with the smallest training error. Call it M_k.

2. ~~Estimate the error,~~ and select a single "best" model from among $M_0 \ldots M_p$
^ Use k-fold cross-validation to calculate the CV error



Time to get our hands dirty

Lab: subset selection using validation

- To do today's lab in R: <nothing new>
- To do today's lab in python: <nothing new>
- Instructions and code for part 1:

  http://www.science.smith.edu/~jcrouser/SDS293/labs/lab9.html
- Full version can be found beginning on p. 248 of ISLR
- For part 2:
  - Apply these techniques to a dataset of your choice
  - You're welcome (encouraged?) to work in teams!

Coming up

- Reminder: A4 due tonight by 11:59pm
- Monday: “shrinkage methods”
  - ridge regression
  - the lasso