STATS 415: Resampling Methods

Prof. Liza Levina

Department of Statistics, University of Michigan

What are Resampling Methods?

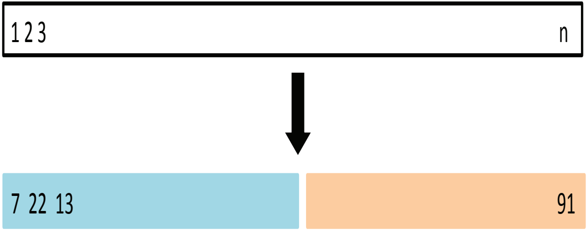
- Tools that involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain more information about the fitted model
  - Uncertainty estimation: estimate how much results vary from sample to sample bootstrap
  - Model assessment: estimate test error rates CV
  - Model selection: select the appropriate level of model flexibility CV
- They are computationally expensive! But we have powerful computers.
- Bootstrap: most often used for uncertainly estimation
- Cross-validation: most often used for model assessment and selection

Cross-Validation (CV)

- The validation set approach
- Leave-one-out cross-validation
- K-fold cross-validation
- Bias-variance trade-off for K-fold cross-validation
- Cross-validation for classification

The Validation Set Approach

- Suppose that we would like to find a set of variables that give the lowest test (not training) error.
- If we have enough data, we can randomly split the data into training and validation ("test stand-in") parts.
- We then use the training part to build each possible model and choose the model that gives the lowest error rate on validation data.



Example: Auto Data

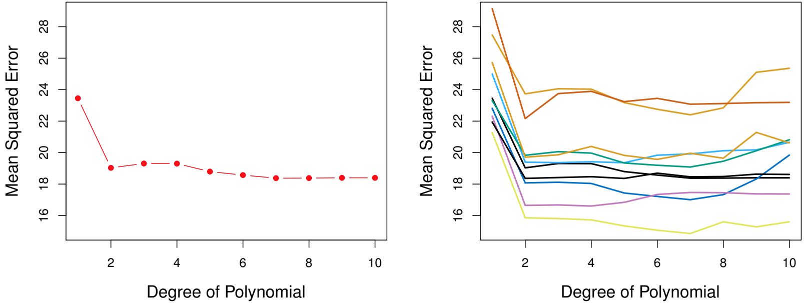
- Suppose that we want to predict “mpg” from “horsepower”
- Two models:

$$\begin{aligned} \text{mpg} &\sim \text{horsepower} \\ \text{mpg} &\sim \text{horsepower} + \text{horsepower}^2 \end{aligned}$$

- Which model is better?
  - Randomly split “Auto” data set into training and validation data (e.g. 196 observations each)
  - Fit both models using the training data set
  - Then evaluate both models using the validation data set
  - The model with the lower validation MSE is the winner.

Results: Auto Data

- Left: Validation error rate for a single split
- Right: Validation method repeated 10 times, each time with a new random split is done random
- All replications seem to suggest degree 2
- But there is a lot of variability among the MSE's. Not good for error estimation! We need more stable methods.



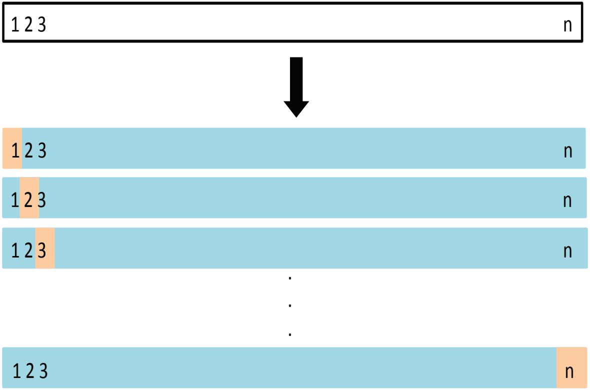
The Validation Set Approach

- Advantages: the validation set approach
  - Simple
  - Easy to implement
- Disadvantages:
  - The validation MSE can vary a lot from split to split
  - Only a subset of observations are used to fit the model (training data). Statistical methods tend to perform worse when trained on fewer observations.

Leave-One-Out Cross-Validation (LOOCV)

- Similar to the validation set approach, aiming to address disadvantages.
- Leave one point out
  - Training data: $n-1$ points
  - Validation data: 1 points
- Fit the model using the training data
- Compute the error for the one point you left out
- Repeat this process for every data point (n times)
- Estimate the overall MSE by averaging over all the splits

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}_i$$


1 2 3
n
1 2 3
n
1 2 3
n
1 2 3
n
1 2 3
n

LOOCV vs. the Validation Set Approach

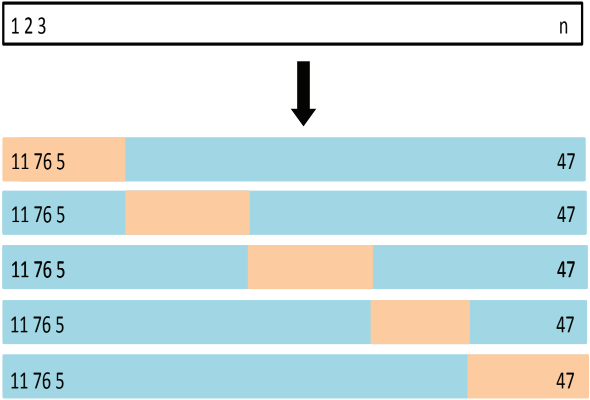
- LOOCV has less bias: We fit the model using training data that contains $n-1$ observations, i.e. almost all the data, and we do it n times; each point gets to "participate" in the training.
- No randomness: LOOCV will produce the same answer every time because every point is left out in turn, whereas the validation set approach depends on the random split.
- LOOCV is computationally intensive: We fit each model n times!

K-fold Cross-Validation

- A compromise between the validation set approach and LOOCV
- With K-fold cross-validation, we divide the data set into K different parts (e.g. $K = 5$, or $K = 10$, etc.)
- Remove one part, fit the model on the remaining $K - 1$ parts, and validate on the part that was removed.
- Repeat this K times, removing each part once.
- Estimate validation error by averaging the resuling K different MSE's:

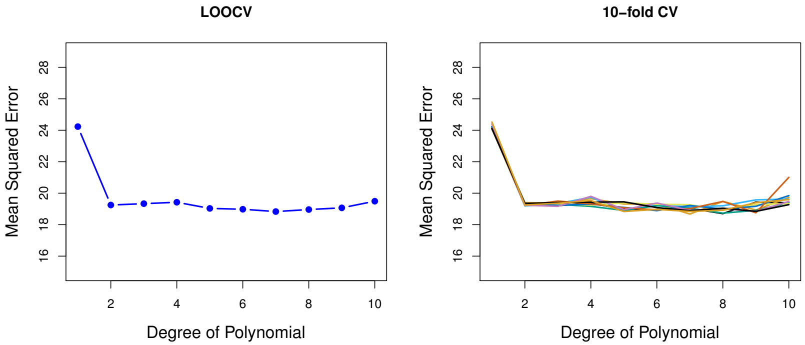
$$\mathrm{CV}_{(K)} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{MSE}_k$$

K-fold Cross-Validation



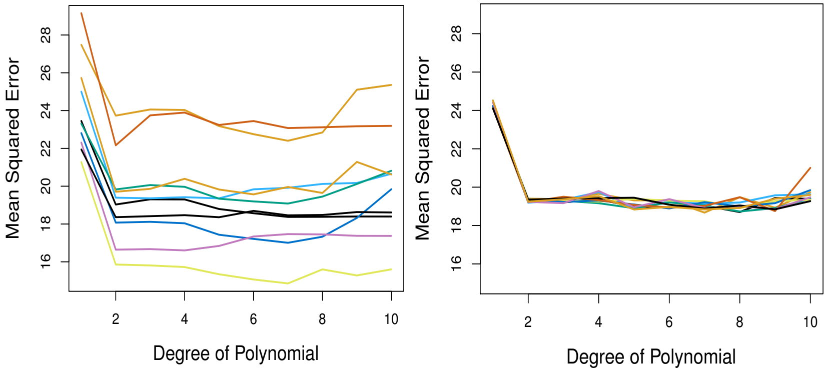
Auto Data: LOOCV vs. K-fold CV

- Left: LOOCV error curve
- Right: multiple realizations of 10-fold CV error curve
- LOOCV is a special case of K-fold, where $K = n$.
- They are both stable, but LOOCV is more computationally intensive.

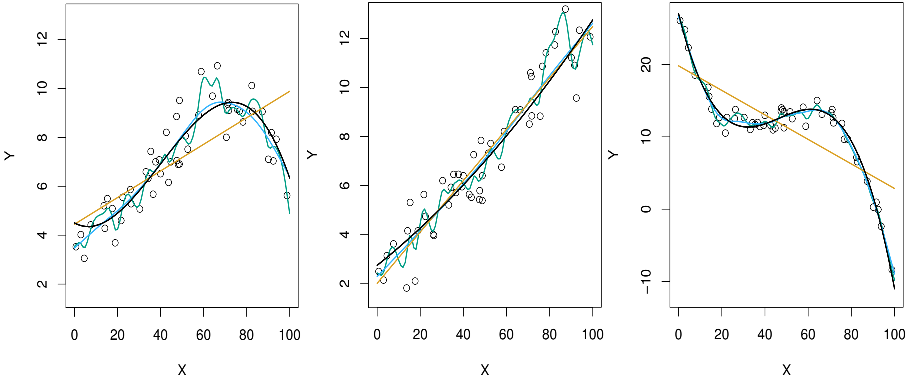


Auto Data: Validation Set Approach vs. K-fold CV

- Left: multiple realizations of the validation set error curve
- Right: multiple realizations of 10-fold cross-validation error curve

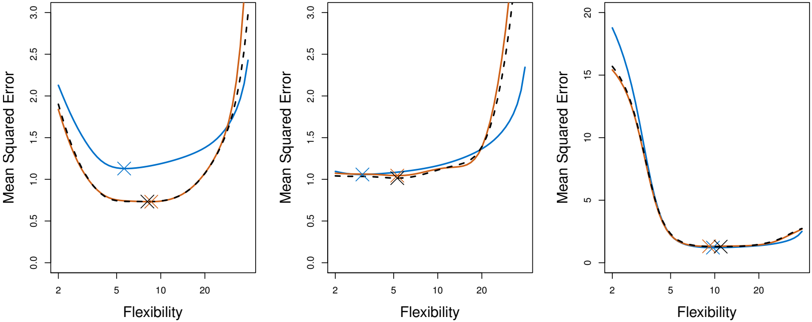


Three Simulated Datasets



K-fold Cross-Validation on Three Simulated Datasets

- Blue: True test MSE
- Black: LOOCV MSE
- Orange: 10-fold MSE



Bias-Variance Trade-off for K-fold CV

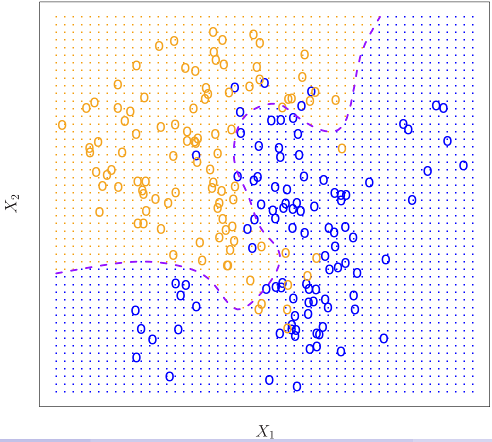
- Putting aside the computational cost, can we pick the best K for K-fold CV?
- K larger = bigger variance; K smaller = bigger bias
- Variance and bias refer to estimating the true test error, not to variability introduced by randomness of the splits (which goes down as K goes up)
- No hard rules
- In computer science / machine learning literature, LOOCV is common
- In statistics, 5-fold or 10-fold CV is standard
- Empirically, 5-fold, 10-fold, and LOOCV tend to work well on both bias and variance in most cases; the choice of K is thus not very important.

Cross-Validation for Classification

- So far, we have looked at CV for regression
- We can use cross-validation in a classification in a similar manner.
    - Divide data into K parts
    - Hold out one part, fit using the remaining data and compute the error rate on the held-out data
    - Repeat K times
    - CV error rate is the average of the K errors we have at the end, one from each ”fold”.

Example: Polynomial logistic regressoin

- The dataset is simulated.
- $p = 2$ predictors, $K = 2$ classes
- The purple dashed line is the Bayes' optimal classification boundary.



Polynomial logistic regression

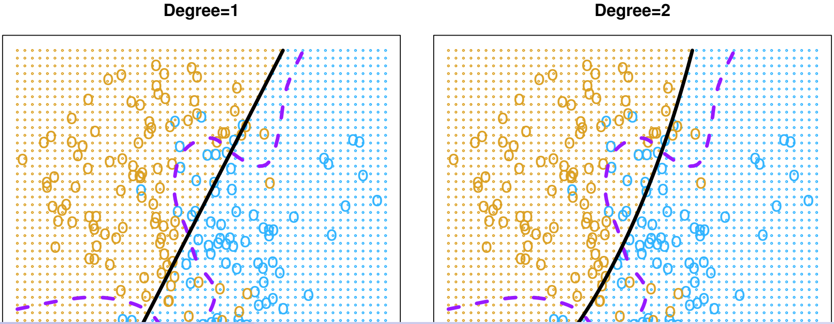
- Linear logistic regression (degree 1)

$$logit(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

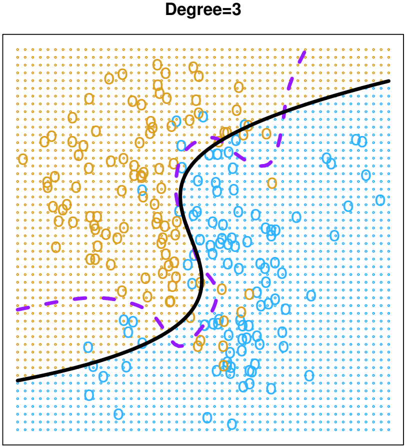
- Quadratic logistic regression (degree 2)

$$logit(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma_1 x_1^2 + \gamma_2 x_2^2 + \gamma_3 x_1 x_2$$

- Linear logistic regression is not able to fit the optimal boundary.
- Quadratic logistic regression does slightly better



- Higher degree terms improve the model fit:



CV to choose the degree

- Brown: Test error
- Blue: Training error
- Black: 10-fold CV error
- Can choose any other parameter this way, e.g., K for KNN