CS189: Intro to Machine Learning
Summer 2018

Lecture 2: Hyperparameters, generalization, and regularization

Josh Tobin
UC Berkeley EECS

Outline for today

- Quick review
- Validation and generalization
- Ridge regression

Outline for today

- **Quick review**
- Validation and generalization
- Ridge regression

Four levels for ML problems

1. Data & application
2. Model
3. Optimization problem
4. Optimization algorithm

Four levels for ML problems (Linear regression)

1. Data & application — Matrix X and vector y
2. Model — Linearly *parameterized* function
3. Optimization problem — Minimize sum of squared diffs
4. Optimization algorithm — Normal equations

$$y \approx Xw$$

Goal: find the best w

Polynomial features

Original data

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Model

$$\hat{y}_i = \sum_{j=0}^{p} w_j x_i^j$$

Features

$$\{0, x, x^2, \cdots, x^p\}$$

Challenges with polynomial features

- How to pick polynomial degree?
- Sensitivity / numerical instability
- Computational challenges for high degree

Challenges with polynomial features

- **How to pick polynomial degree?**
- Sensitivity / numerical instability
- Computational challenges for high degree (not covered today)

Challenges with polynomial features

- **How to choose *hyperparameters***
  ~~How to pick polynomial degree?~~
- Sensitivity / numerical instability
- Computational challenges for high degree (not covered today)

Validation

How to choose hyperparameters?

- Unrealistic assumption: what if we have access to the underlying model?
- Slightly more realistic assumption: infinite noisy data
- Real world: validation sets

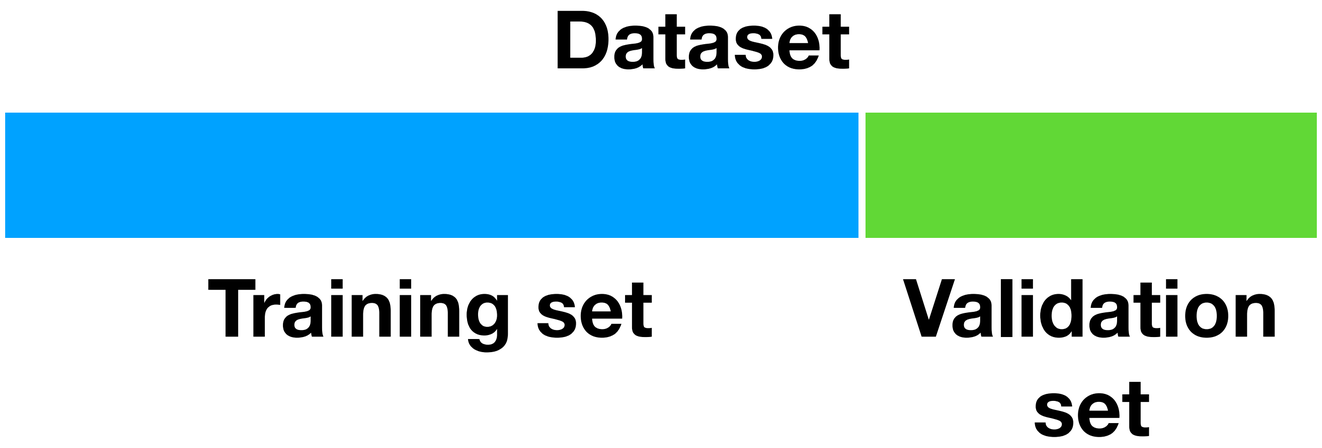
Validation sets

Dataset

Validation sets

Validation sets

Validation sets

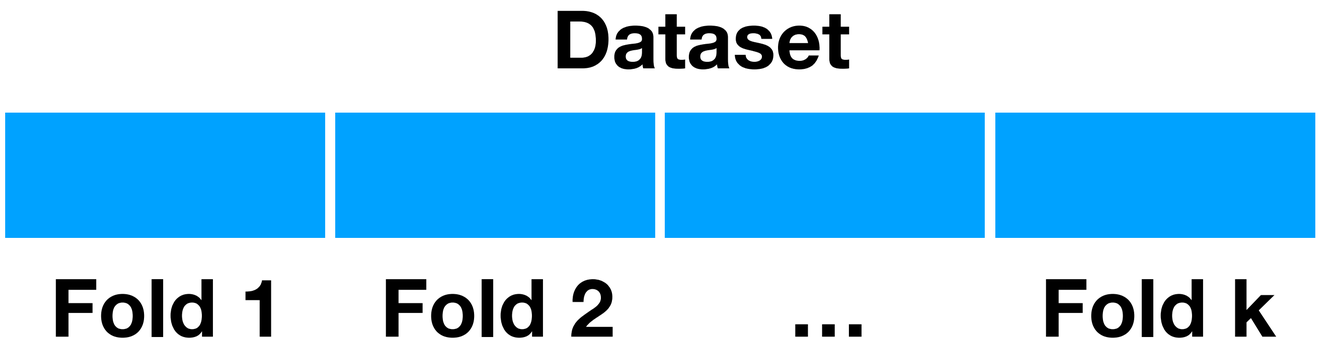


Validation sets

Making more with less: k-fold cross validation

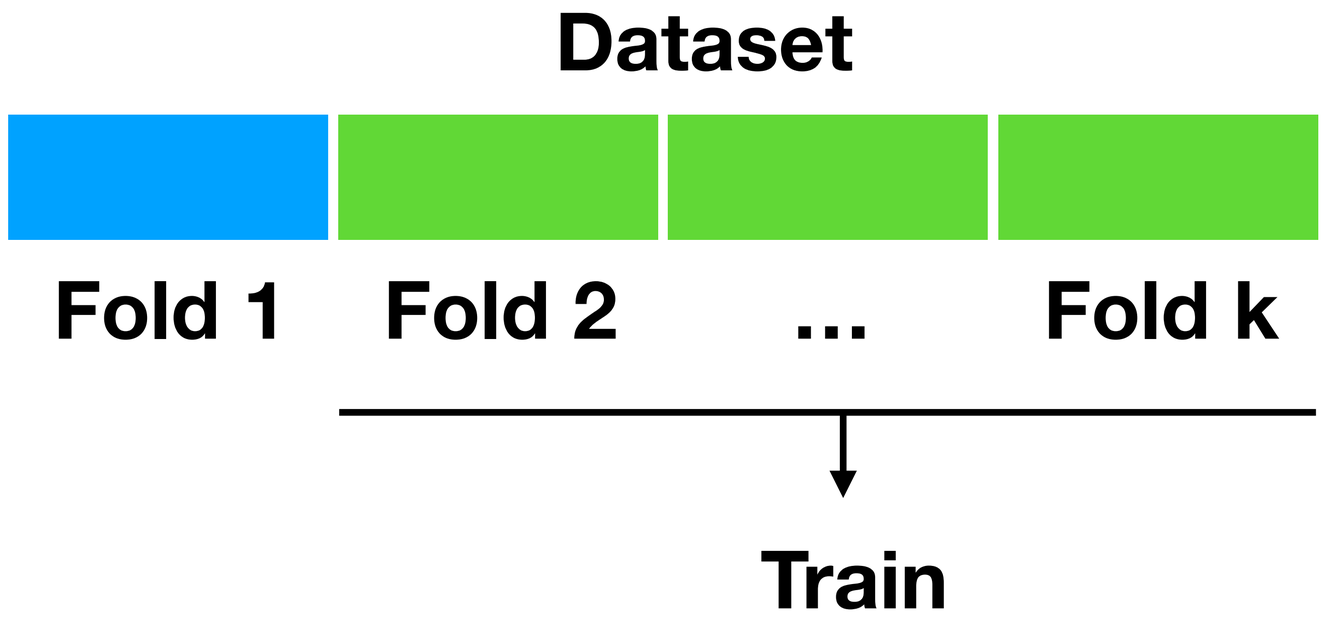
Dataset

Making more with less: k-fold cross validation

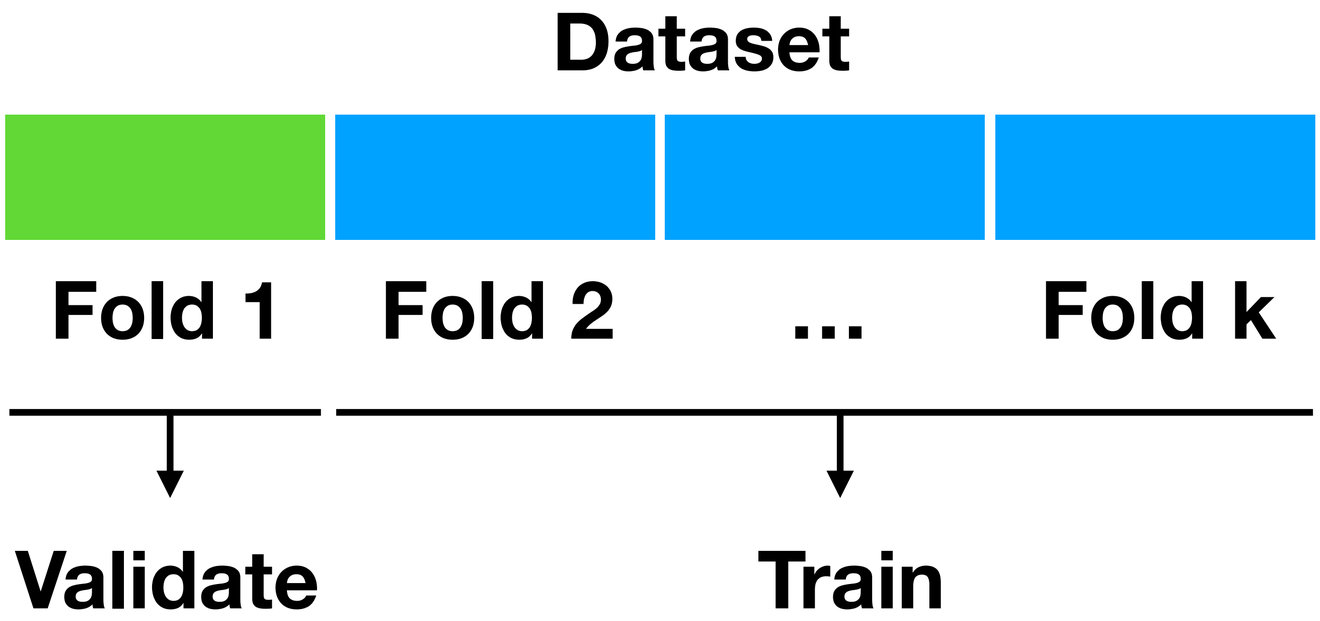


Model 1

Making more with less: k-fold cross validation

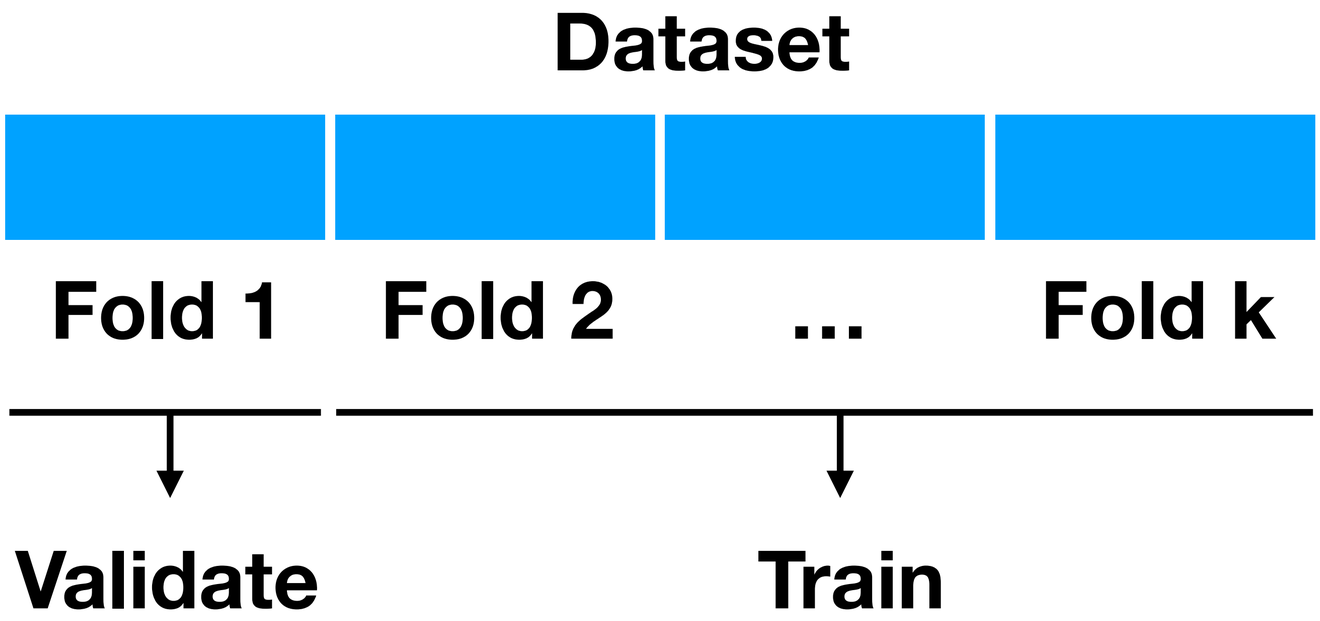


Making more with less: k-fold cross validation



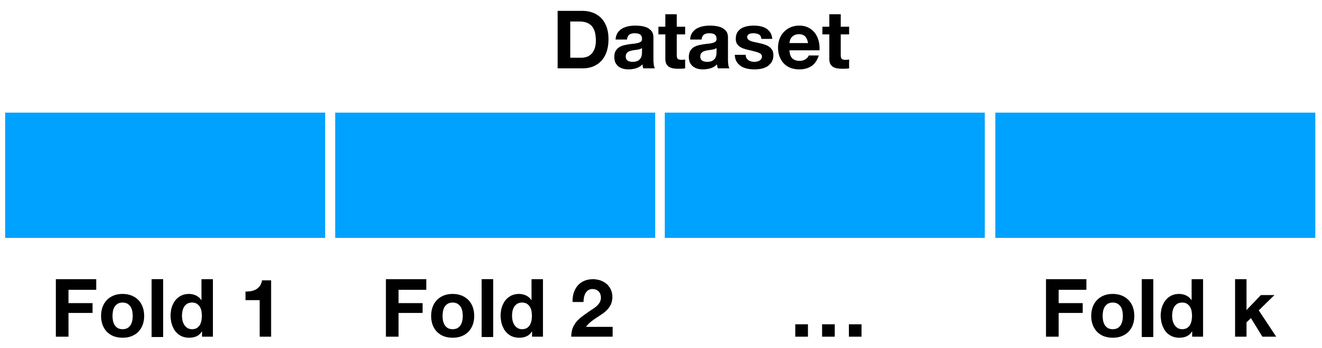
Model 1

Making more with less: k-fold cross validation



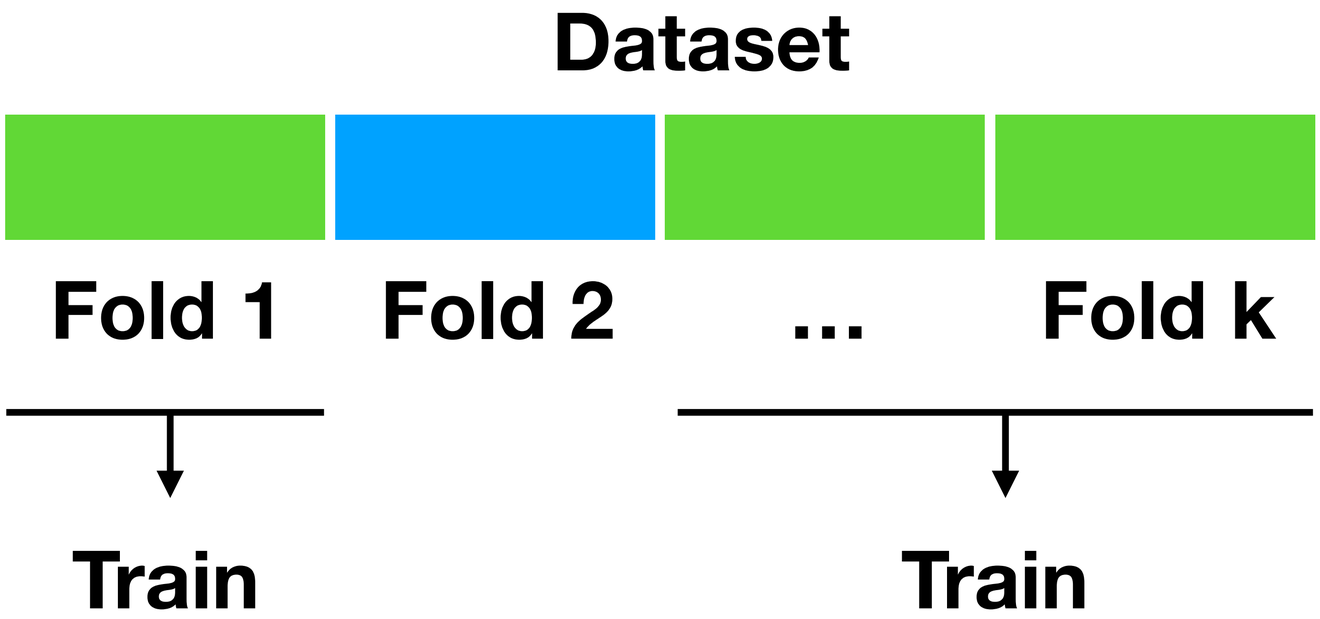
Model 1

Making more with less: k-fold cross validation



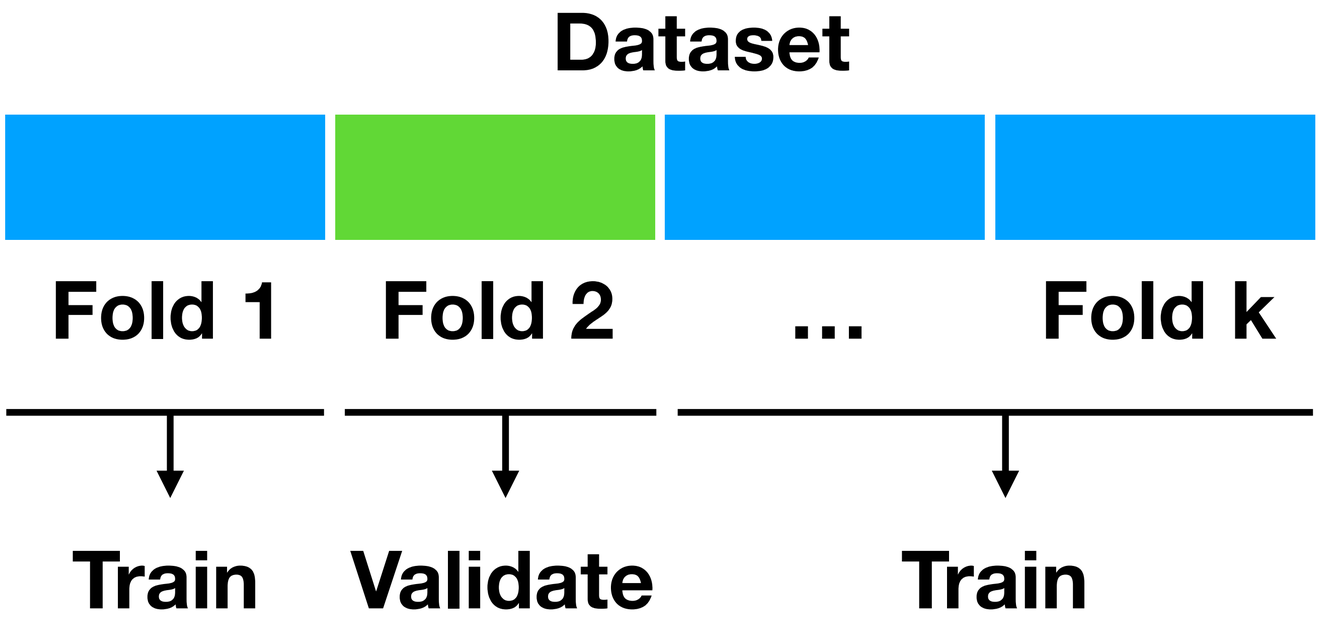
Model 2

Making more with less: k-fold cross validation



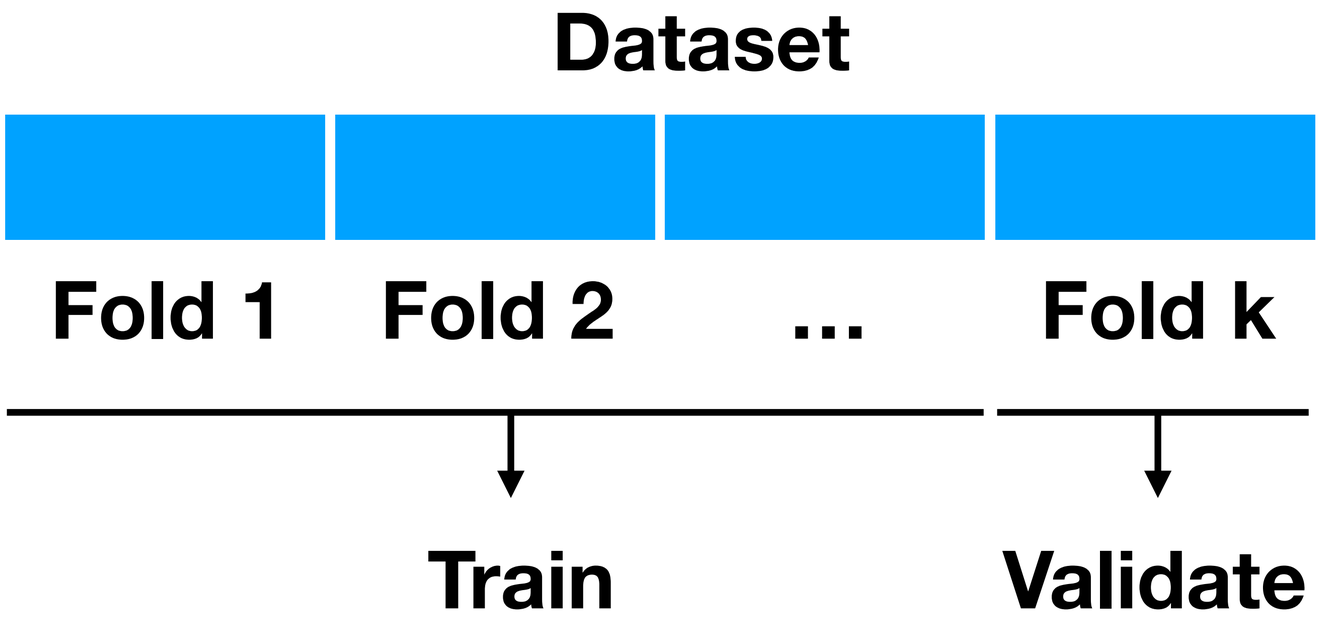
Model 2

Making more with less: k-fold cross validation



Model 2

Making more with less: k-fold cross validation



Model k

Making more with less: k-fold cross validation

K-fold cross validation

- for all hyperparameters:
    for i=1:k:
        train model on all folds except i
        evaluate performance on fold i

- Choose the hyperparameter with the best *average* val performance

- k is itself a hyperparameter, but usually set to 10 or 4

Challenges with polynomial features

- How to choose hyperparameters?
- **Sensitivity / numerical instability**
- Computational challenges for high degree (not covered today)

Regularization

Ridge regression

$$\hat{w} = (X^T X + \lambda I)^{-1} X^T y$$

Hacky motivation

- “Correct” for small eigenvalues in X^T X by adding a multiple of the identity

Optimization motivation

	OLS	Ridge
Optimization problem	$\min_w \|\|Xw - y\|\|_2^2$	$\min_w \|\|Xw - y\|\|_2^2 + \lambda \|\|w\|\|_2^2$
Solution	$\hat{w} = (X^T X)^{-1} X^T y$	$\hat{w} = (X^T X + \lambda I)^{-1} X^T y$

Optimization motivation

Hyperparameter

$$\min_w ||Xw - y||_2^2 + \lambda ||w||_2^2$$

Training error

Keep weights small

$$\hat{w} = (X^T X + \lambda I)^{-1} X^T y$$

Next time

Review of probability

Probabilistic interpretation of regression