Methodology: Assessment and Cross-Validation

Mario Parente

First story

- USPS uses a classifier to distinguish 4 from 9
- Pays $1 for every mistake
- How much money should it budget for 2015?
- **Model assessment**: estimate prediction error on future unseen data (generalization)

Second story

- USPS uses regularized logistic regression to prevent overfitting in its classifier
- What value of λ will lead to the model with the least prediction error?
- **Model selection**: compare prediction error of many models to select the best one

Two goals

Model assessment: estimate prediction error on future unseen data (generalization)

Model selection: compare prediction error of many models to select the best one

Two goals

Model assessment: estimate prediction error on future unseen data (generalization)

Model selection: compare prediction error of many models to select the best one

Can't do either of these with data used to train the model

Data-Generating Mechanism

- Assumption: training data representative of future unseen data
- Formally, training examples and future test examples drawn *independently* from same probability distribtuion $\mathcal{P}$

$$(\mathbf{x}^{(i)}, y^{(i)}) \sim \mathcal{P}$$
$$(\mathbf{x}, y) \sim \mathcal{P}$$

- How to think of this
  - huge bag of input-output pairs $(\mathbf{x}, y)$ ("nature")
  - m training examples pulled out randomly
  - future data drawn also pulled out randomly (e.g. one set for validation, one set for generalization)

In an Ideal World

If we are "data rich", this is what we would do:

- **Validation set**: labeled data reserved to compare models
- **Test set**: labeled data reserved to assess future performance

E.g., 50/25/25 split

In an Ideal World

If we are "data rich", this is what we would do:

- **Validation set**: labeled data reserved to compare models
- **Test set**: labeled data reserved to assess future performance

E.g., 50/25/25 split

Warning: Terminology of validation/test not always consistently used

The Dilemma: Train vs. Test Size

What if you only have 100 training examples? 50? 10?

The dilemma

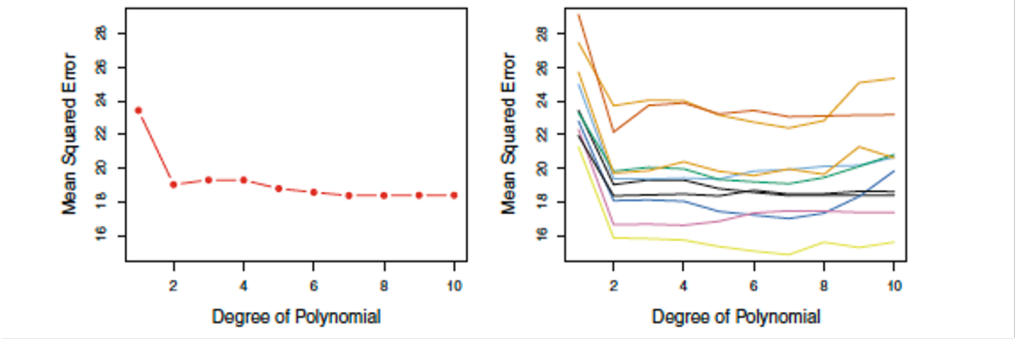
- More training data $\rightarrow$ more accurate classifier
- More test data $\rightarrow$ better estimate of generalization accuracy

Training-Validation

(Assume assessment for now... how much will USPS pay?)

Don't have a separate large test to estimate generalization error.

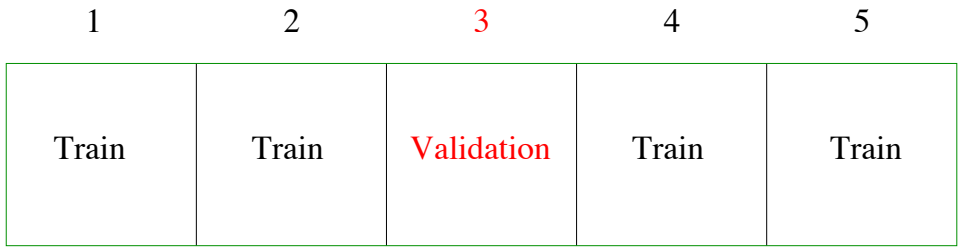
- Split the data in Training and Validation (e.g 50/50 or 70/30)
- Estimate accuracy (or error) on validation set
- We want validation error $\approx$ generalization error, but there is bias, since we trained with smaller training set. There is variance, since the samples are chosen randomly
- tends to overestimate true error



Cross-Validation

(Assume assessment for now... how much will USPS pay?)

- Split data in k equal-sized "folds" (usually 2, 5, 10)
- For each fold, test on that fold while training on all others:



- Estimate accuracy by averaging over all folds
- lower bias (each training set contains (k-1)n/k observations, generally$>$ validation approach)
- lower variance, since average error over K moderately correlated folds

Special case K=N leave-one-out CV (lower bias and higher variance than k-fold CV)

Example

5-fold cross-validation

	Train folds	Test folds	Accuracy
12345	2,3,4,5	1	85%
12345	1,3,4,5	2	83%
12345	1,2,4,5	3	91%
12345	1,2,3,5	4	88%
12345	1,2,3,4	5	84%

Average accuracy = 88.2%

Discussion

What if you need to do both model comparison and assessment?

Discussion

What if you need to do both model comparison and assessment?

Fancier methods:

- One fold for validation (e.g. train/valid/test $= 3/1/1$)
- Nested cross-validation

Discussion

What if you need to do both model comparison and assessment?

Fancier methods:

- One fold for validation (e.g. train/valid/test $= 3/1/1$)
- Nested cross-validation

Warning: There is no single agreed-upon methodology that is always best. Methods are applied somewhat flexibly. It's best to understand the *principles* so you can judge what is (or is not) appropriate.