Model Selection

CS 584 Data Mining (Spring 2022)

Prof. Sanmay Das

George Mason University

Slides are adapted from those of Malik Magdon-Ismail (for *Learning From Data)*, and also the available book slides developed by Tan, Steinbach and Kumar, with additional input from Prof. Huzefa Rangwala

Model Evaluation

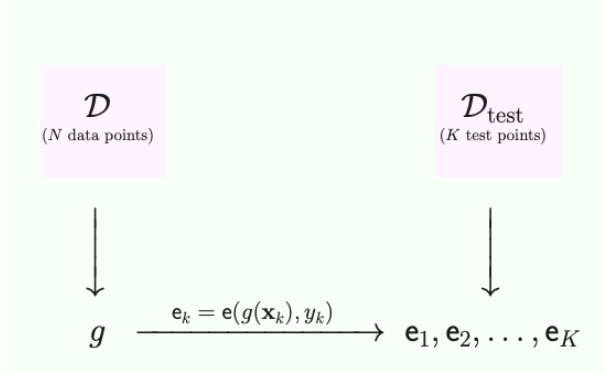
- Metrics for Performance Evaluation
  - How to evaluate the performance of a model?
- Methods for Performance Evaluation
  - How to obtain reliable estimates?
- Methods for Model Comparison
  - How to compare the relative performance among competing models?

Model Evaluation

- Metrics for Performance Evaluation
  - How to evaluate the performance of a model?
- Methods for Performance Evaluation
  - How to obtain reliable estimates?
- Methods for Model Comparison
  - How to compare the relative performance among competing models?

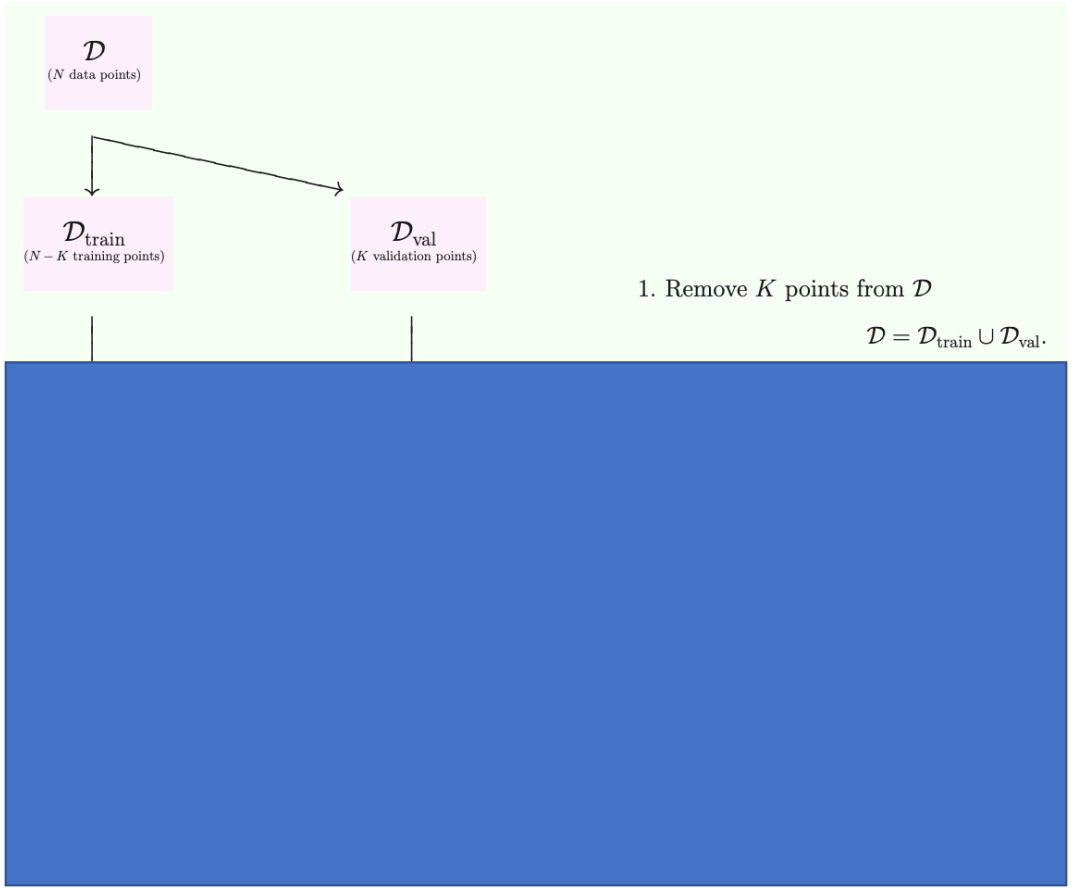
Key Idea

- I want to estimate *generalization error* using only *training data*
- Usually will use *test error* as a proxy
  - Aside: Why can't I just use training error?
  - Works better and better as I get more and more test data

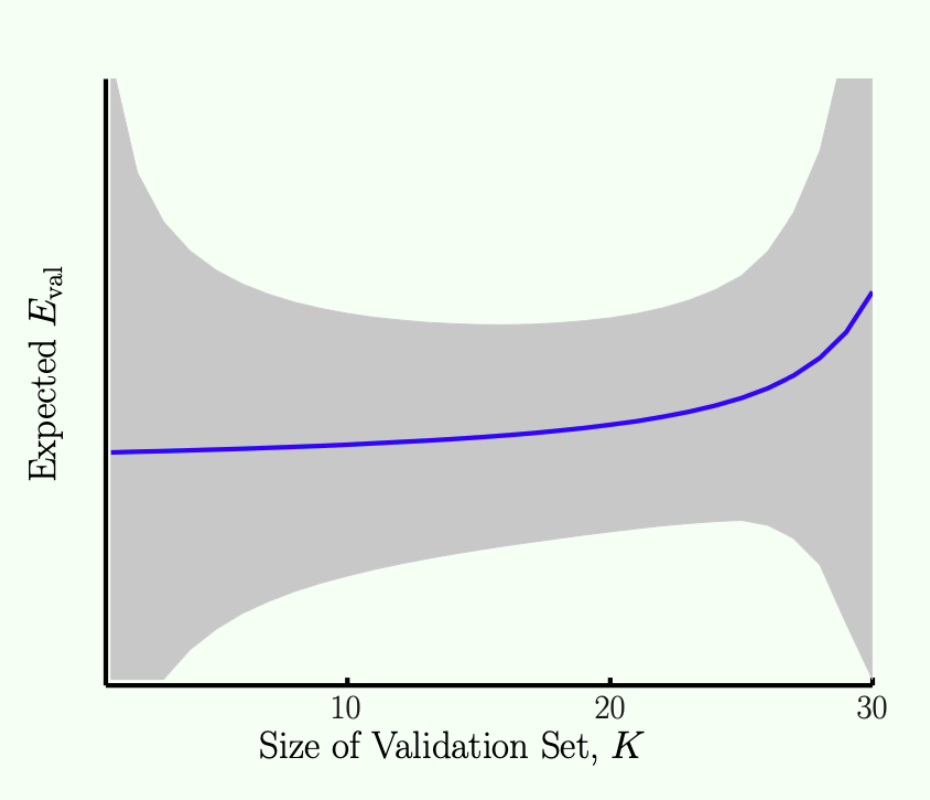


  - Unfortunately, and crucially!, I don't HAVE access to my test set.

The Validation Set



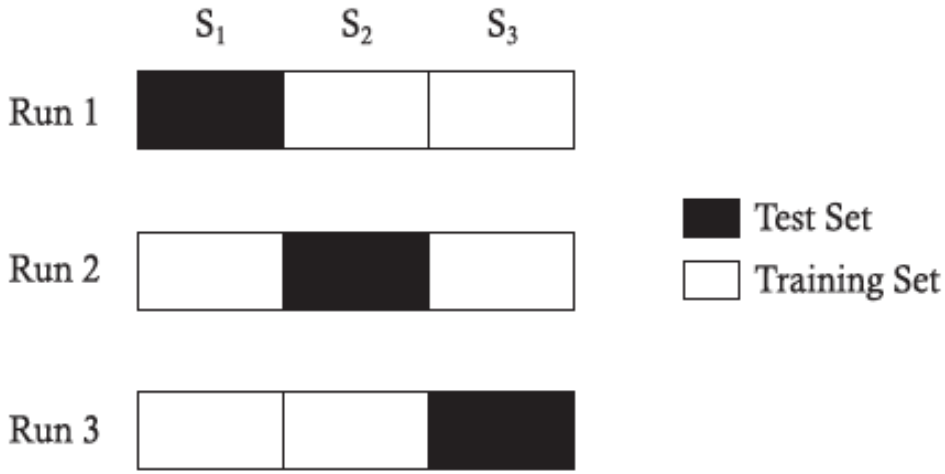
How Big Should the Validation Set Be?



- With small *K*, the hypothesis learned using the whole training set is close to the hypothesis learned using just the training portion of the training/validation split.
  - So, the estimate of generalization error is based on roughly the right hypothesis
  - But the variance in the estimate is high because the validation set is small
- With large K, the variance in the estimate is small, but the variance in what is learned is high, and that could be different from what you'd learn with the whole training set

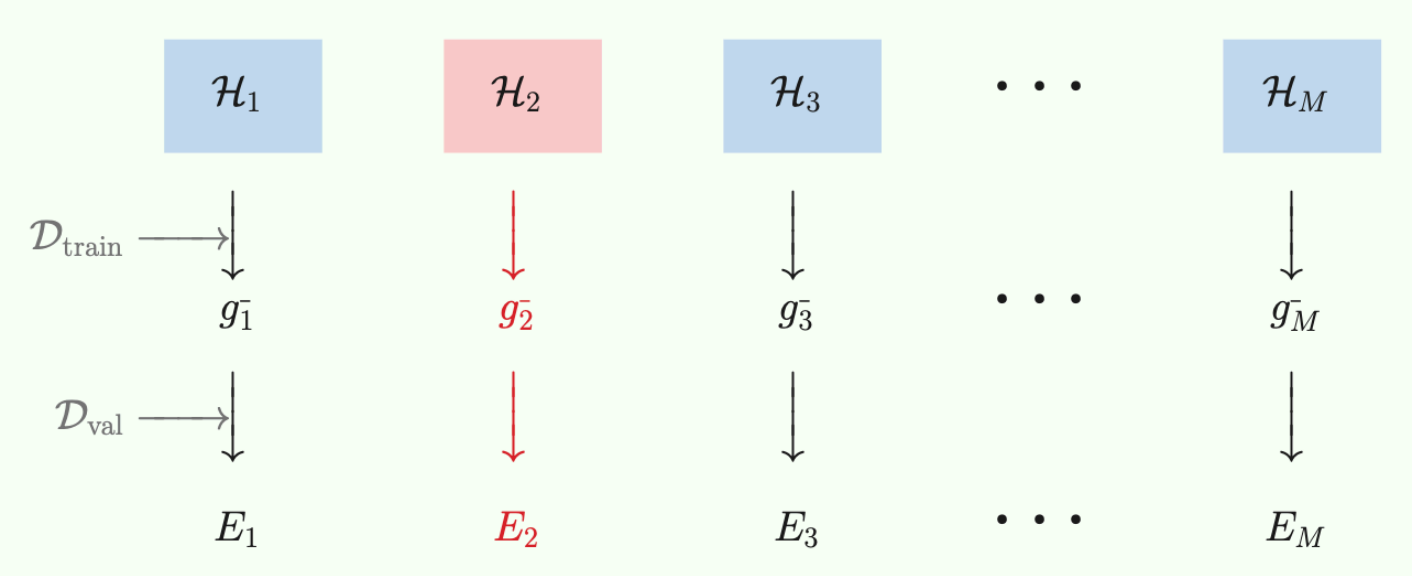
The Magic of Cross-Validation

- Incredibly efficient use of training data



- Can do leave-one-out, 5- or 10-fold CV
- Typically excellent estimate of test error for a single hypothesis

Model Selection Using Validation Error



- Pick the best model (e.g. value of *k*) -- the one with lowest validation error
- Then retrain that model on the entire training set (why?)
- Note that the error estimate is no longer valid! (Why?)

Model Evaluation

- Metrics for Performance Evaluation
  - How to evaluate the performance of a model?
- Methods for Performance Evaluation
  - How to obtain reliable estimates?
- Methods for Model Comparison
  - How to compare the relative performance among competing models?

Metrics for Performance Evaluation

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or to build models, scalability, etc.
- Confusion Matrix:

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Metrics for Performance Evaluation…

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

- Most widely-used metric:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

Cost Matrix

	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	**Class=Yes**	**Class=No**
	Class=Yes	C(Yes|Yes)	C(No|Yes)
	Class=No	C(Yes|No)	C(No|No)

C(i|j): Cost of misclassifying class j example as class i

Exercise

- Group 1
  - Construct a real-world scenario where the cost of false positives is much higher than the cost of false negatives
- Group 2
  - Construct a real-world scenario where the cost of false negatives is much higher than the cost of false positives