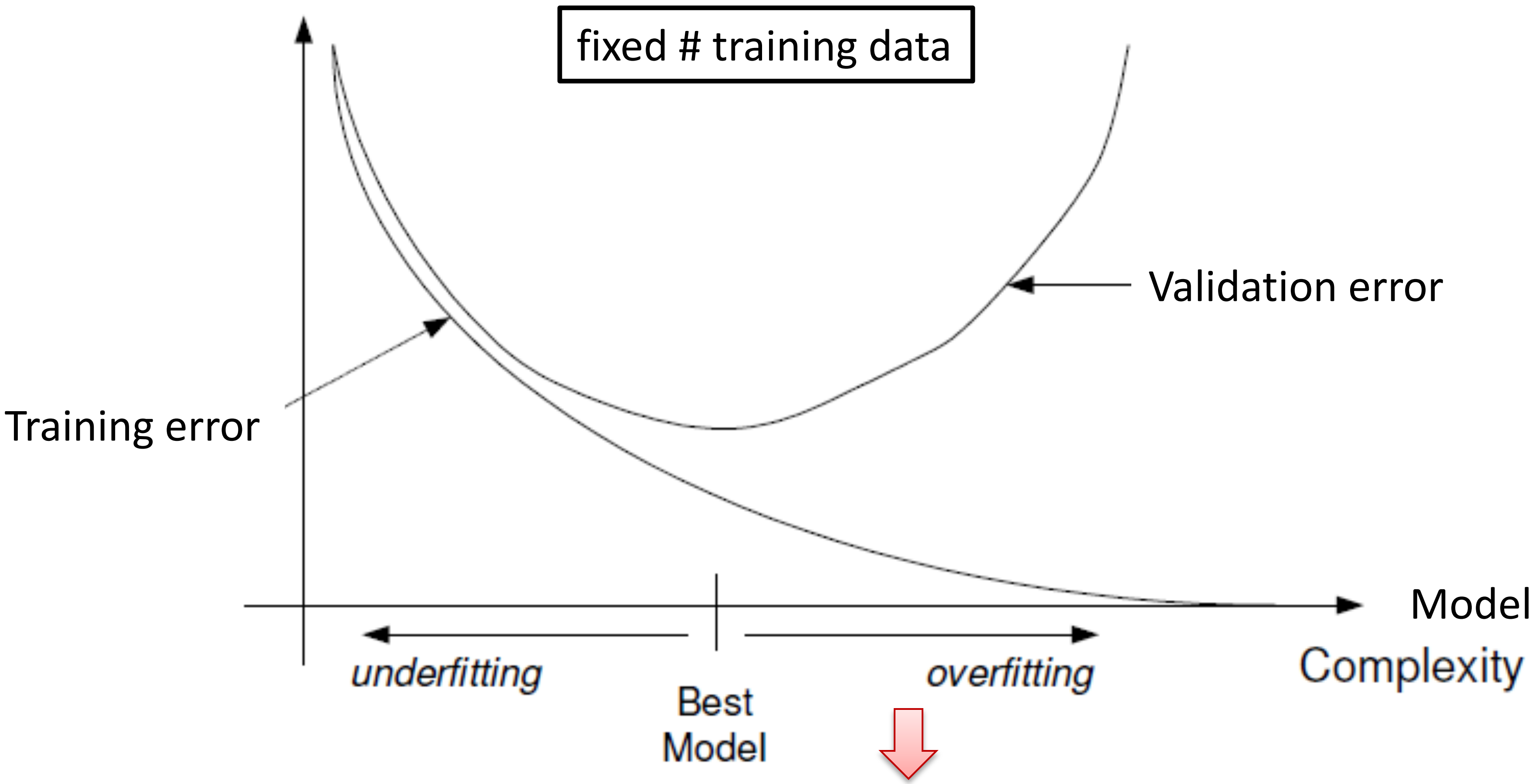
# Model selection

Aarti Singh

Machine Learning 10-701
Mar 15, 2023

# Training vs. Test Error




Training error is no longer a good indicator of test error

# Examples of Model Spaces

Model Spaces with varying complexity:

- Nearest-Neighbor classifiers with increasing neighborhood sizes $k = 1,2,3,\ldots$

  Large neighborhood => complexity

- Decision Trees with increasing depth k or with k leaves

  Higher depth/ More # leaves => complexity

- Neural Networks with increasing layers or nodes per layer

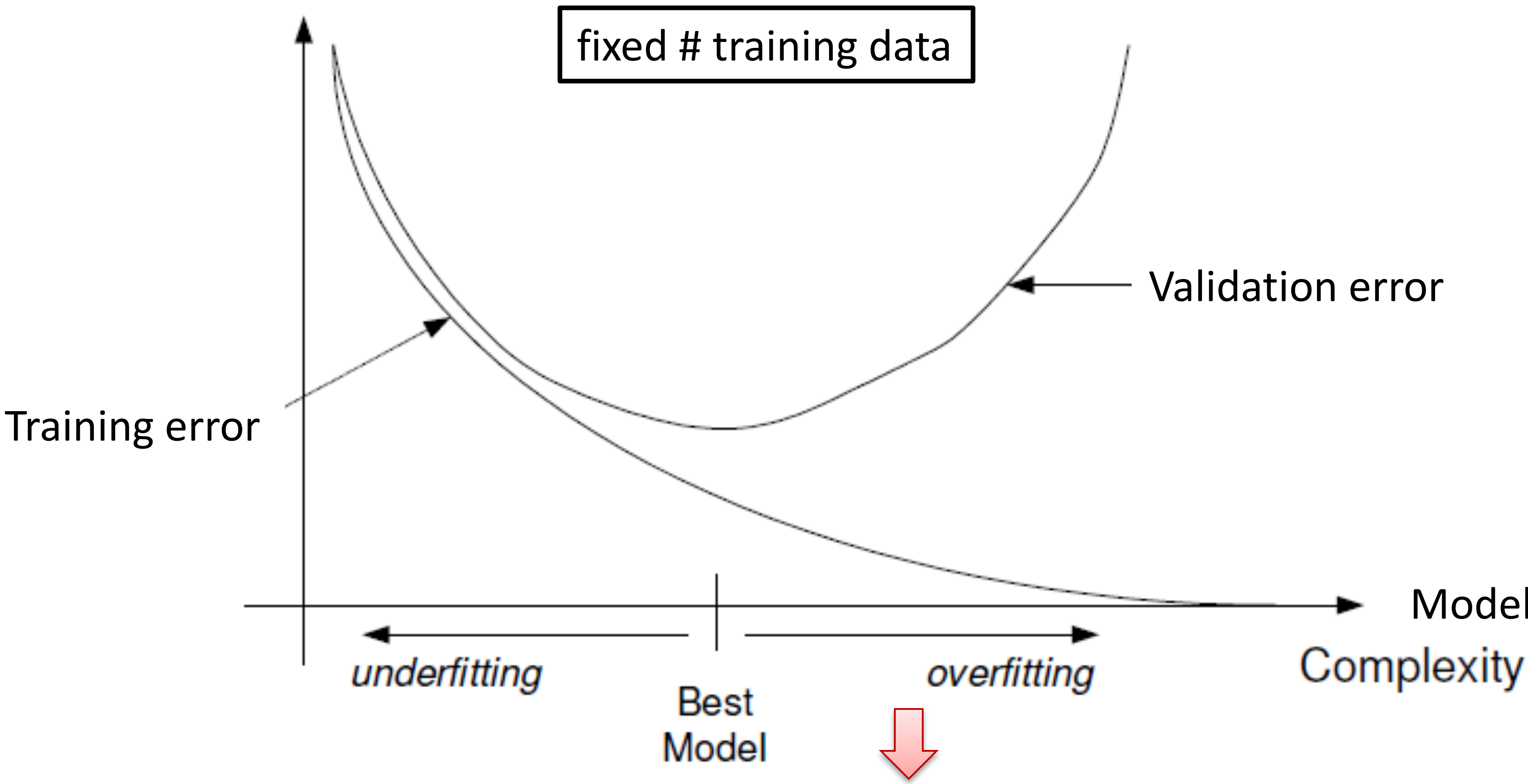
  More layers/Nodes per layer => complexity

- MAP estimates with stronger priors (larger hyper-parameters $\beta_H$, $\beta_T$ for Beta distribution or smaller variance for Gaussian prior)

  => complexity

**How can we select the right complexity model ?**

# Training vs. Test Error




Training error is no longer a good indicator of test error

# Bias-Variance Tradeoff

- Why does test/validation error go down then up with increasing model complexity?

Two sources of error:

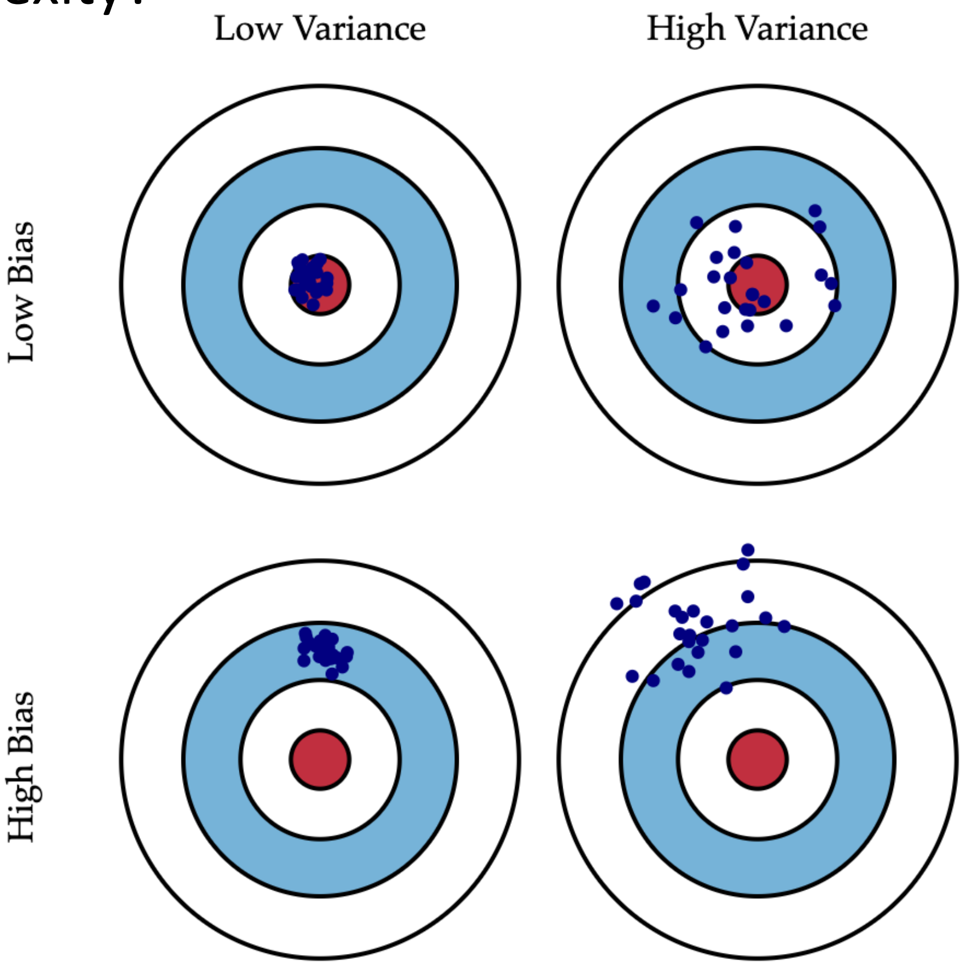
e.g. Regression

Bias

$|E[f_n] - f^*|$

Variance

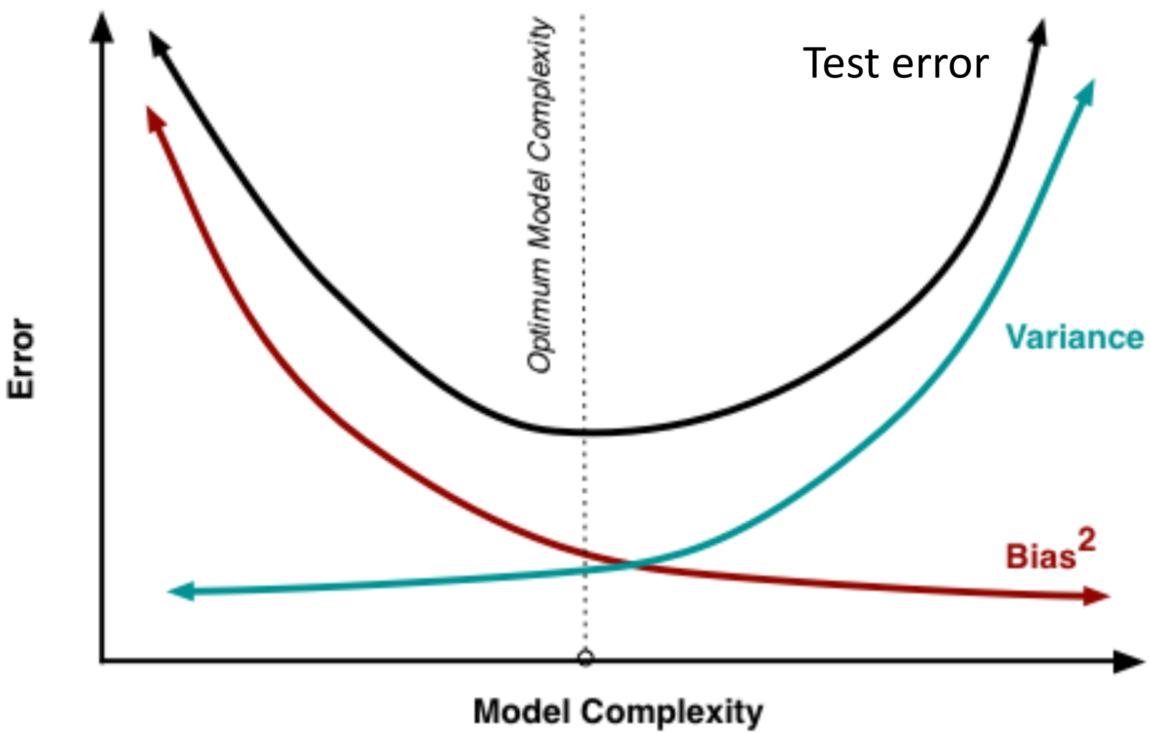
$E[|f_n - E[f_n]|^2]$



# Bias-Variance Tradeoff

- Why does test/validation error go down then up with increasing model complexity?

Mean square test error = Variance + $\text{Bias}^2$ + Irreducible error



# Judging Test error

- Training error of a classifier f

$$\frac{1}{n}\sum_{i=1}^{n} 1_{f(X_i)\neq Y_i}$$

Training Data

$$\{X_i, Y_i\}_{i=1}^{n}$$

- What about test error?

  Can't compute it.

- How can we know classifier is not overfitting?

  Hold-out or Cross-validation

# Hold-out method

Can judge test error by using an independent sample of data.

**Hold - out procedure:**

n data points available $D \equiv \{X_i, Y_i\}_{i=1}^n$

1) Split into two sets (randomly and preserving label proportion):

Training dataset $D_T = \{X_i, Y_i\}_{i=1}^m$

Validation/Hold-out dataset $D_V = \{X_i, Y_i\}_{i=m+1}^n$

often m = n/2

2) Train classifier on $D_T$. Report error on validation dataset $D_V$.
Overfitting if validation error is much larger than training error

# Hold-out method

Drawbacks:

- May not have enough data to afford setting one subset aside for getting a sense of generalization abilities
- Validation error may be misleading (bad estimate of test error) if we get an “unfortunate” split

Limitations of hold-out can be overcome by a family of sub-sampling methods at the expense of more computation.

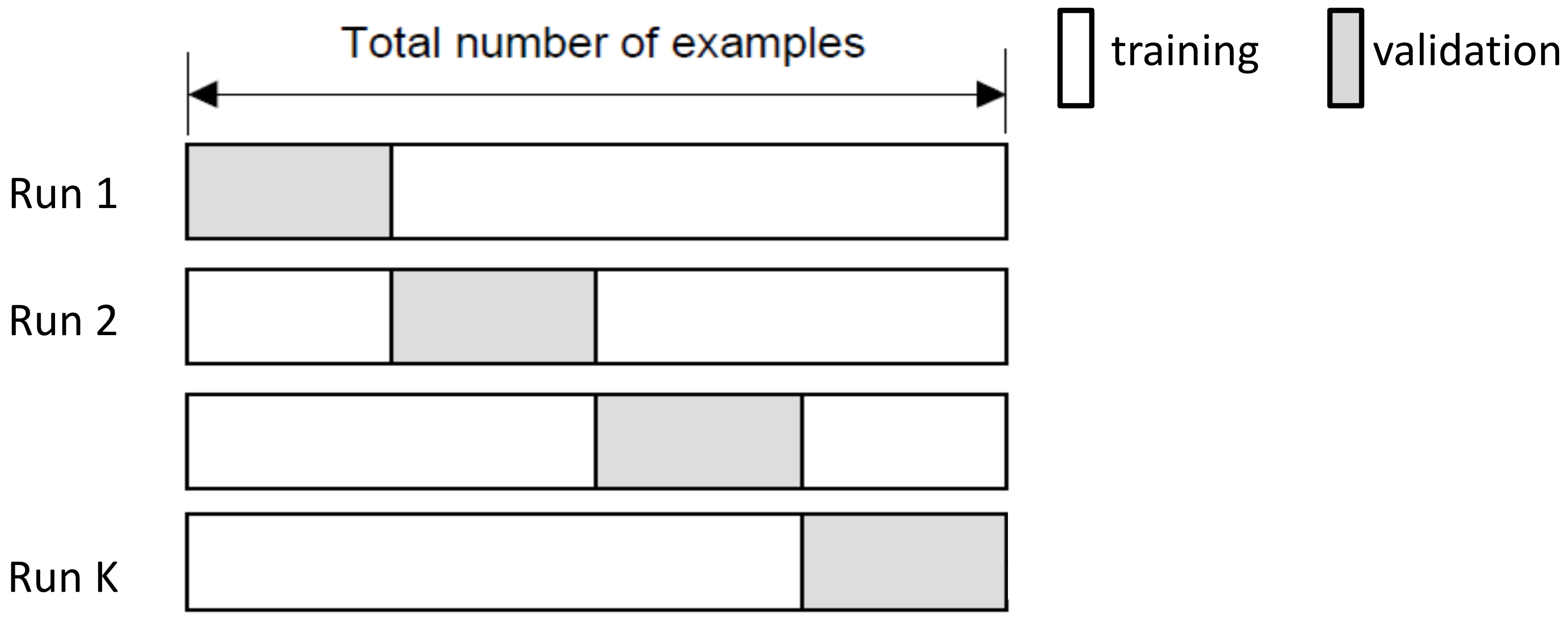
# Cross-validation

## K-fold cross-validation

Create K-fold partition of the dataset.

Do K runs: train using K-1 partitions and calculate validation error on remaining partition (rotating validation partition on each run).

Report average validation error

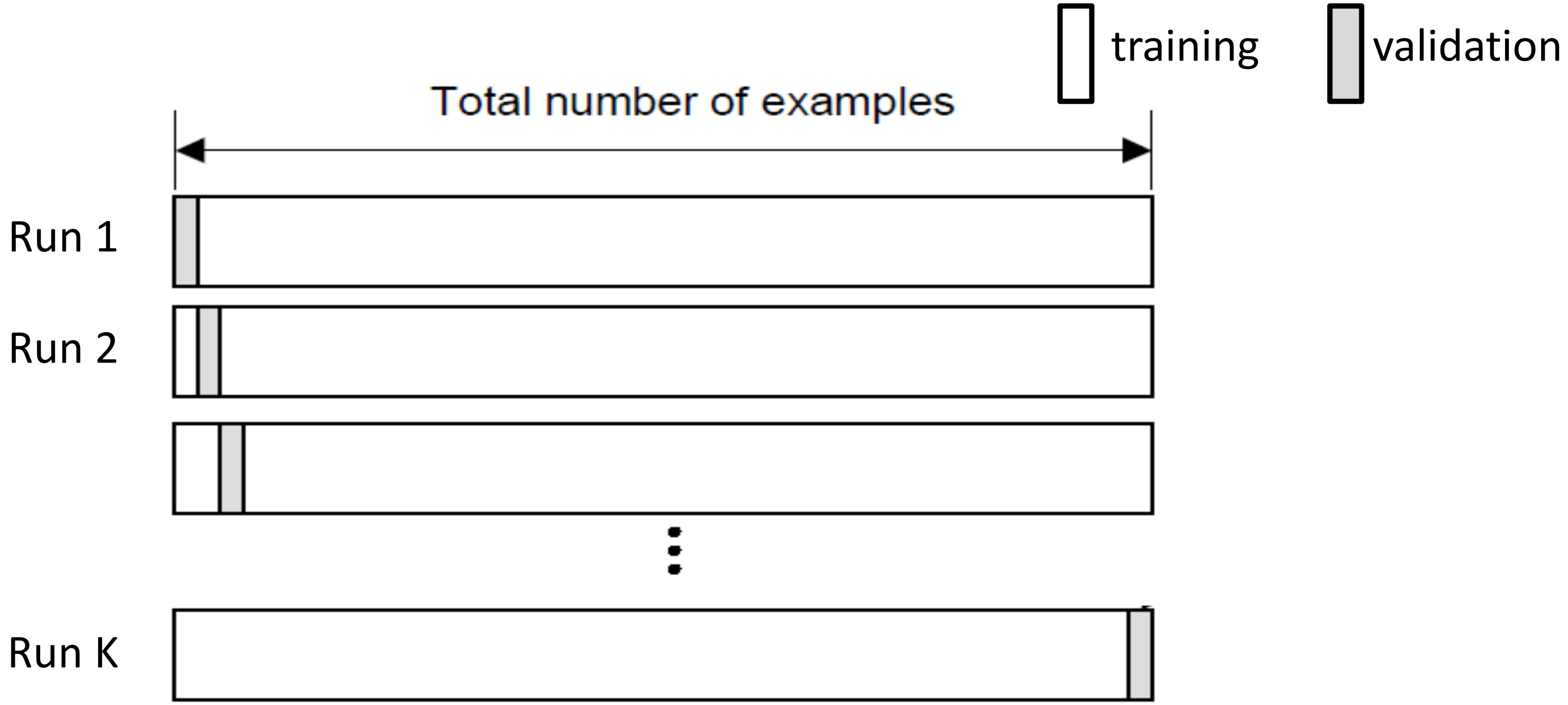


# Cross-validation

## Leave-one-out (LOO) cross-validation

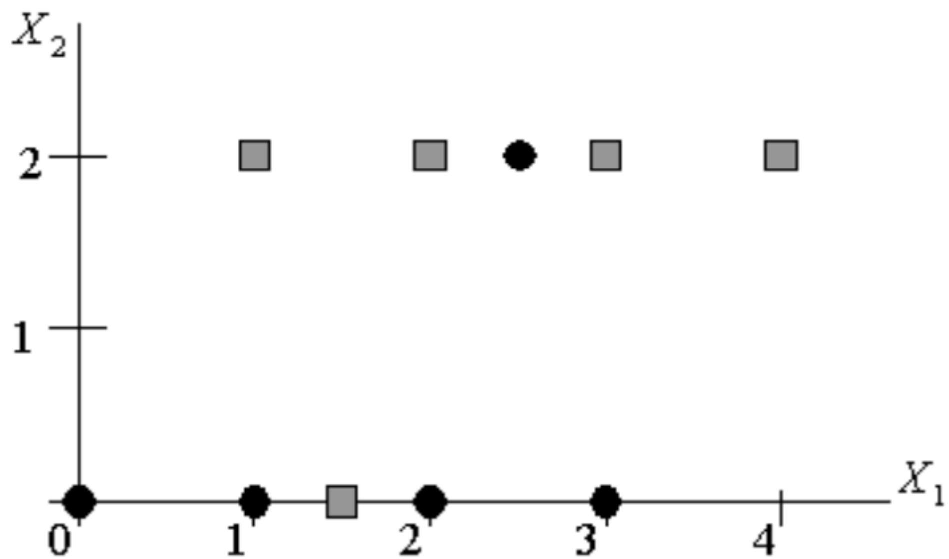
Special case of K-fold with K=n partitions

Equivalently, train on n-1 samples and validate on only one sample per run for n runs



# Cross-validation

What is the leave-one-out cross-validation error of the given classifiers on the following dataset?




- Poll 1: Depth 1 Decision tree using best feature
- Poll 2: 1-NN classifier

# Cross-validation

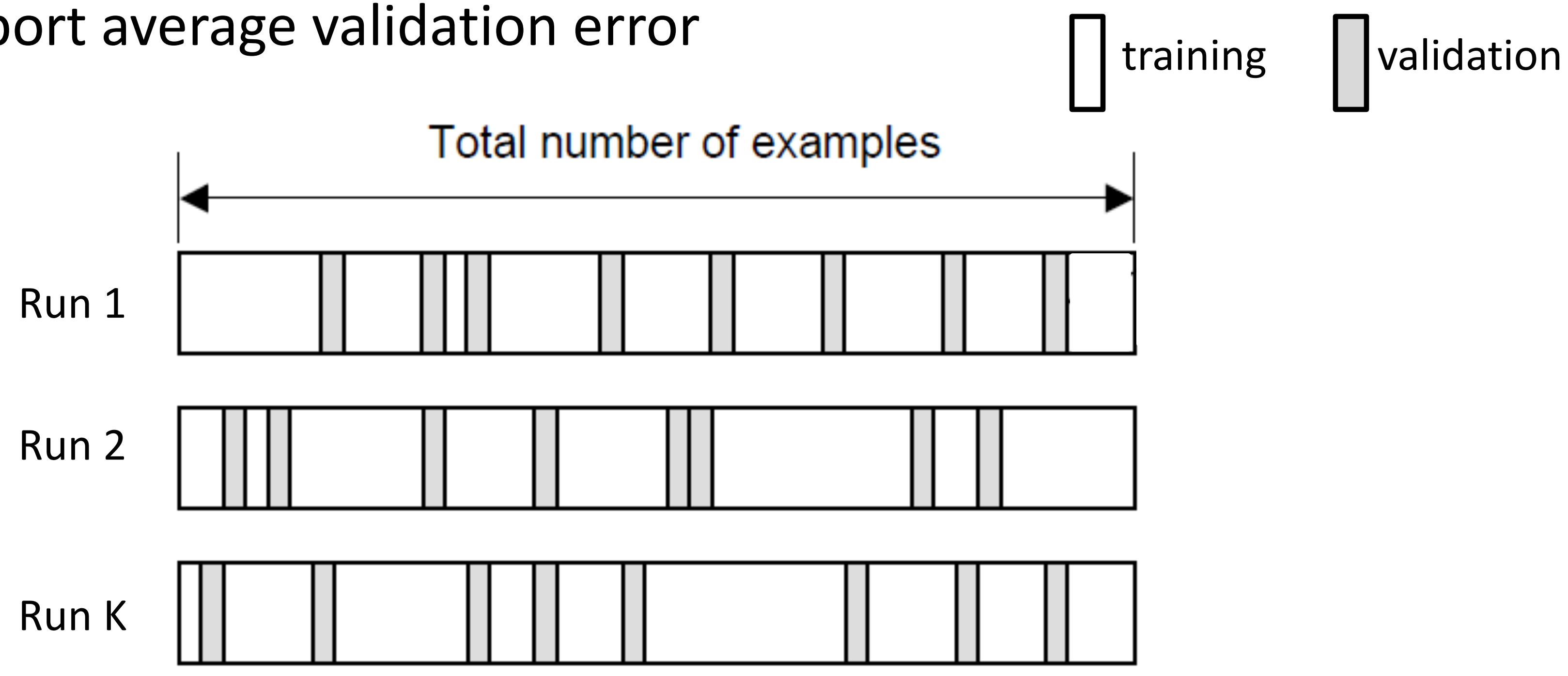
## Random subsampling

Randomly subsample a fixed fraction $\alpha n$ (0< $\alpha$ <1) of the dataset for validation.

Compute validation error with remaining data as training data.

Repeat K times

Report average validation error



# Practical Issues in Cross-validation

How to decide the values for $K$ and $\alpha$?

- Large K
  - \+ Validation error can approximate test error well
  - \- Observed validation error will be unstable (few validation pts)
  - \- The computational time will be very large as well (several runs)

- Small K
  - \+ The #runs and, therefore, computation time are reduced
  - \+ Observed validation error will be stable (many validation pts)
  - \- Validation error cannot approximate test error well

Common choice: K = 10, $\alpha$ = 0.1 ☺

# Model selection using Hold-out/Cross-validation

- Train models of different complexities and evaluate their validation error using hold-out or cross-validation

- Pick model with smallest validation error (averaged over different runs for cross-validation)

➢ When using hold-out or cross-validation for model selection, test error should be reported using independent data

# What you should know

- Estimating test error using
  - hold-out
  - cross-validation
- Bias-variance tradeoff
- Model selection using
  - hold-out
  - cross-validation