Foundations of Machine Learning

Module 1: Introduction

Part D: Evaluation and Cross validation

Sudeshna Sarkar

IIT Kharagpur

Experimental Evaluation of Learning Algorithms

- Evaluating the performance of learning systems is important because:
  - Learning systems are usually designed to predict the class of "future" unlabeled data points.
- Typical choices for Performance Evaluation:
  - Error
  - Accuracy
  - Precision/Recall
- Typical choices for Sampling Methods:
  - Train/Test Sets
  - K-Fold Cross-validation

Evaluating predictions

- Suppose we want to make a prediction of a value for a target feature on example **x**:
  - y is the observed value of target feature on example ***x***.
  - $\hat{y}$ is the predicted value of target feature on example ***x***.
  - How is the error measured?

Measures of error

- Absolute error: $\frac{1}{n}|f(x) - y|$
- Sum of squares error: $\frac{1}{n}\sum_{i=1}^{n}(f(x) - y)^2$
- Number of misclassifications: $\frac{1}{n}\sum_{i=1}^{n}\delta(f(x), y)$

- $\delta(f(x), y)$ is 1 if $f(x) \neq y$, and 0, otherwise.

Confusion Matrix

True class → Hypothesized class ↓	Pos	Neg
Yes	TP	FP
No	FN	TN
	P=TP+FN	N=FP+TN

- Accuracy = (TP+TN)/(P+N)
- Precision = TP/(TP+FP)
- Recall/TP rate = TP/P
- FP Rate = FP/N

Sample Error and True Error

- The ***sample error*** of hypothesis *f* with respect to target function *c* and data sample *S* is:

$$error_s(f)= 1/n \ \Sigma_{x \in S}\delta(f(x),c(x))$$

- The ***true error*** (denoted $error_D(f)$) of hypothesis *f* with respect to target function *c* and distribution *D*, is the probability that *h* will misclassify an instance drawn at random according to *D*.

$$error_D(f)= Pr_{x \in D}[f(x) \neq c(x)]$$

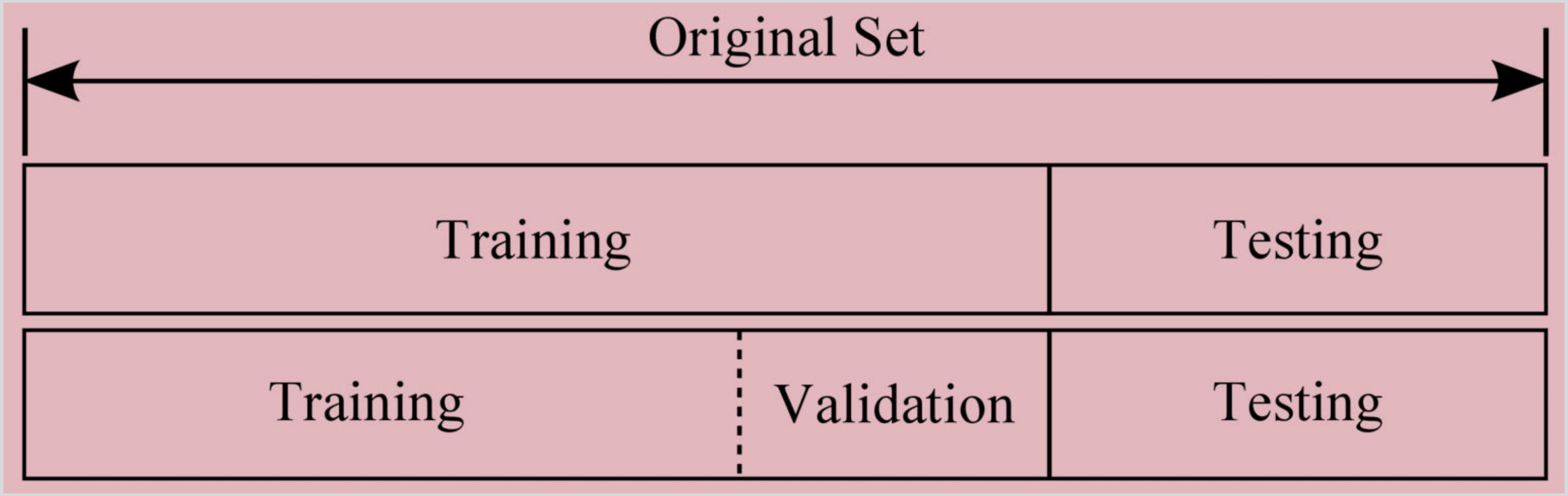
Why Errors

- Errors in learning are caused by:
  - Limited representation (representation bias)
  - Limited search (search bias)
  - Limited data (variance)
  - Limited features (noise)

Difficulties in evaluating hypotheses with limited data

- Bias in the estimate: The sample error is a poor estimator of true error
  - ==> test the hypothesis on an independent test set
- We divide the examples into:
  - **Training examples** that are used to train the learner
  - **Test examples** that are used to evaluate the learner
- Variance in the estimate: The smaller the test set, the greater the expected variance.

Validation set

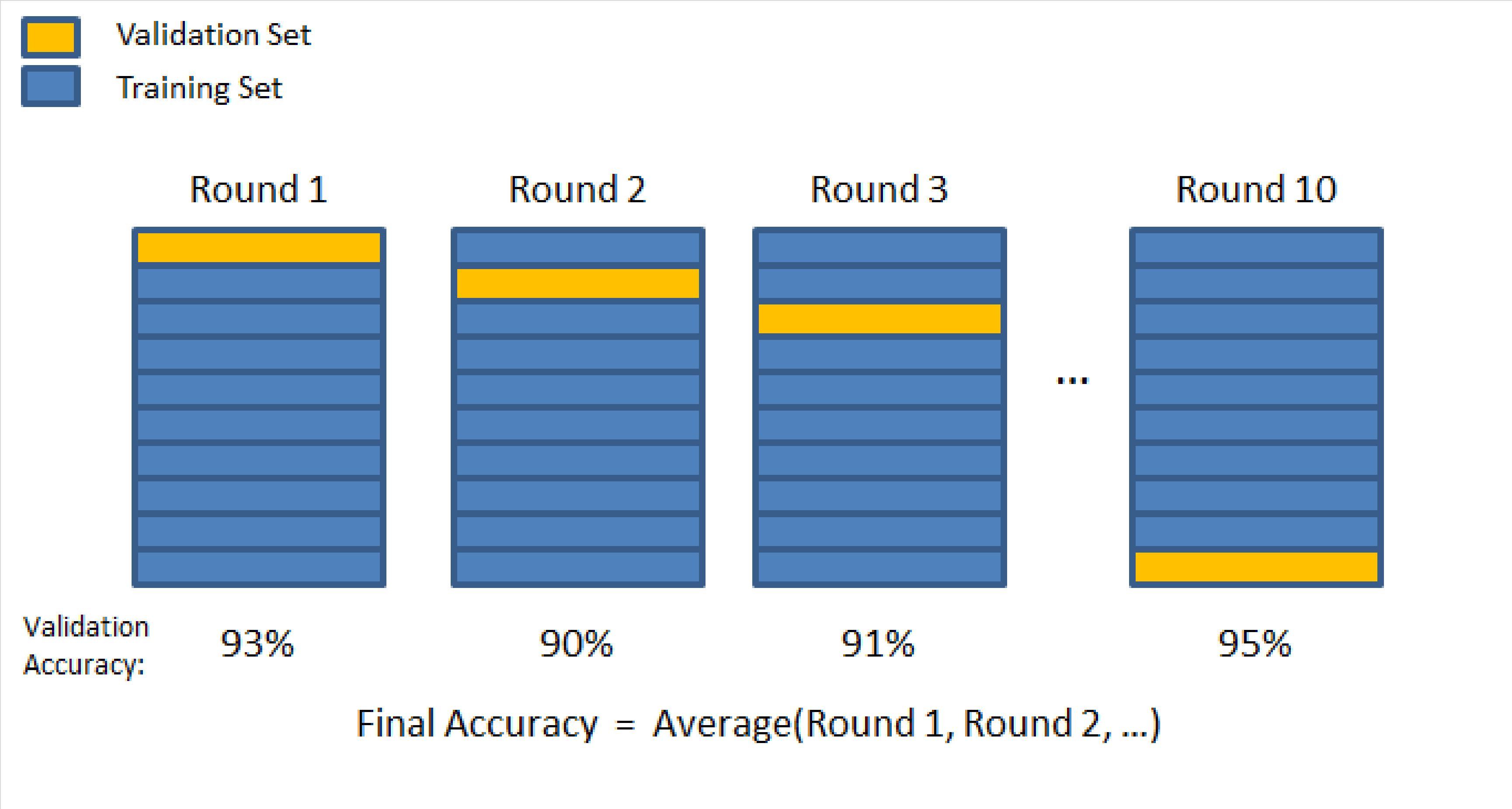


Validation fails to use all the available data

k-fold cross-validation

1. Split the data into k equal subsets
2. Perform k rounds of learning; on each round
   - 1/k of the data is held out as a test set and
   - the remaining examples are used as training data.
3. Compute the average test set score of the k rounds

K-fold cross validation



Trade-off

- In machine learning, there is always a trade-off between
  - complex hypotheses that fit the training data well
  - simpler hypotheses that may generalise better.
- As the amount of training data increases, the generalization error decreases.