


Matlab 12: Performance Evaluation of Classifiers



Cheng-Hsin Hsu

National Tsing Hua University

Department of Computer Science

Slides are based on the materials from Prof. Roger Jang

Introduction to Performance Evaluation

- Performance evaluation: An objective procedure to derive the performance index (or figure of merit) of a given model
- Typical performance indices
 - Accuracy
 - Recognition rate: ↗
 - Error rate: ↘
 - RMSE (root mean squared error): ↘
 - R-square ↗
 - Computation load: ↘

Synonyms

- Sets of synonyms to be used interchangeably (since we are focusing on classification):
 - Classifiers, models
 - Recognition rate, accuracy
 - Training, design-time
 - Test, run-time

Performance Indices for Classifiers

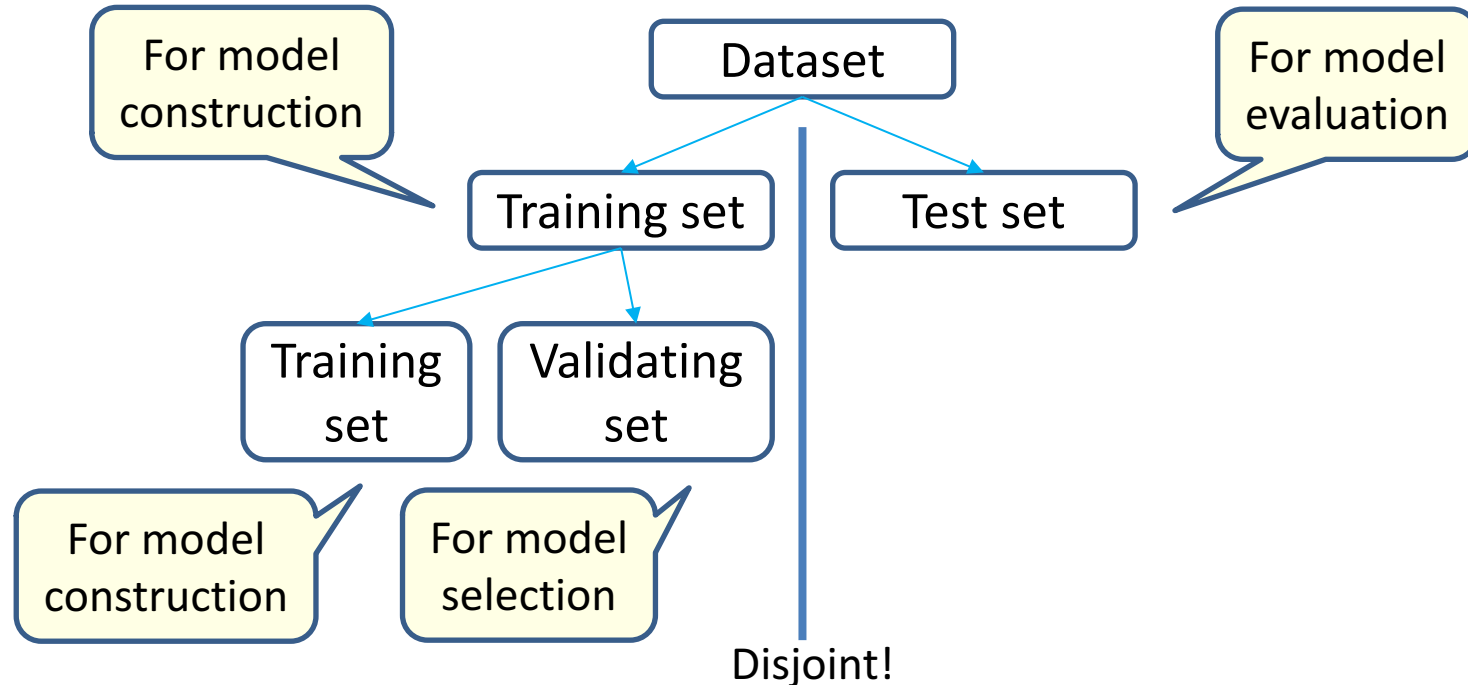
- Performance indices of a classifier
 - Recognition rate
 - How to have an objective method or procedure to derive it
 - Computation load
 - Design-time computation (training)
 - Run-time computation (test)
- Our focus
 - Recognition rate and the procedures to derive it
 - The estimated accuracy depends on
 - Dataset partitioning
 - Model (types and complexity)

Methods for Performance Evaluation

- Methods to derive the recognition rates
 - Inside test (resubstitution accuracy)
 - One-side holdout test
 - Two-side holdout test
 - M-fold cross validation
 - Leave-one-out cross validation

Dataset Partitioning

- Data partitioning: to make the best use of the dataset
 - Training set
 - Training and test sets
 - Training, validating, and test sets



Inside Test (1/2)

- Dataset partitioning
 - Use the whole dataset for training & evaluation
- Recognition rate (RR)
 - Inside-test or resubstitution recognition rate

$$\text{Dataset } D = \{(\mathbf{x}_i, y_i) \mid i = 1, 2, \dots, n\}$$

$F_D(\bullet)$: Model identified by the dataset D

$$RR_{inside} = \frac{1}{|D|} \sum_{i=1}^n (y_i == F_D(\mathbf{x}_i))$$

Inside Test (2/2)

- Characteristics
 - Too optimistic since RR tends to be higher
 - For instance, 1-NNC always has an RR of 100%!
 - Can be used as the upper bound of the true RR.
- Potential reasons for low inside-test RR:
 - Bad features of the dataset
 - Bad method for model construction, such as
 - Bad results from neural network training
 - Bad results from k-means clustering

One-side Holdout Test (1/2)

- Dataset partitioning
 - Training set for model construction
 - Test set for performance evaluation
- Recognition rate
 - Inside-test RR
 - Outside-test RR

Dataset $D = A \oplus B$

$$A = \{(\mathbf{x}_i^A, y_i^A) \mid i = 1, 2, \dots, |A|\}, B = \{(\mathbf{x}_i^B, y_i^B) \mid i = 1, 2, \dots, |B|\}$$

$F_A(\bullet)$: Model identified by the dataset A

$$RR_{inside} = \frac{1}{|A|} \sum_{i=1}^{|A|} (y_i^A == F_A(\mathbf{x}_i^A))$$

$$RR_{outside} = \frac{1}{|B|} \sum_{i=1}^{|B|} (y_i^B == F_A(\mathbf{x}_i^B))$$

One-side Holdout Test (2/2)

- Characteristics
 - Highly affected by data partitioning
 - Usually Adopted when training (design-time) computation load is high (for instance, deep neural networks)

Two-sided Holdout Test (1/3)

- Dataset partitioning
 - Training set for model construction
 - Test set for performance evaluation
 - Role reversal

Dataset $D = A \oplus B$

$$A = \{(\mathbf{x}_i^A, y_i^A) \mid i = 1, 2, \dots, |A|\}, B = \{(\mathbf{x}_i^B, y_i^B) \mid i = 1, 2, \dots, |B|\}$$

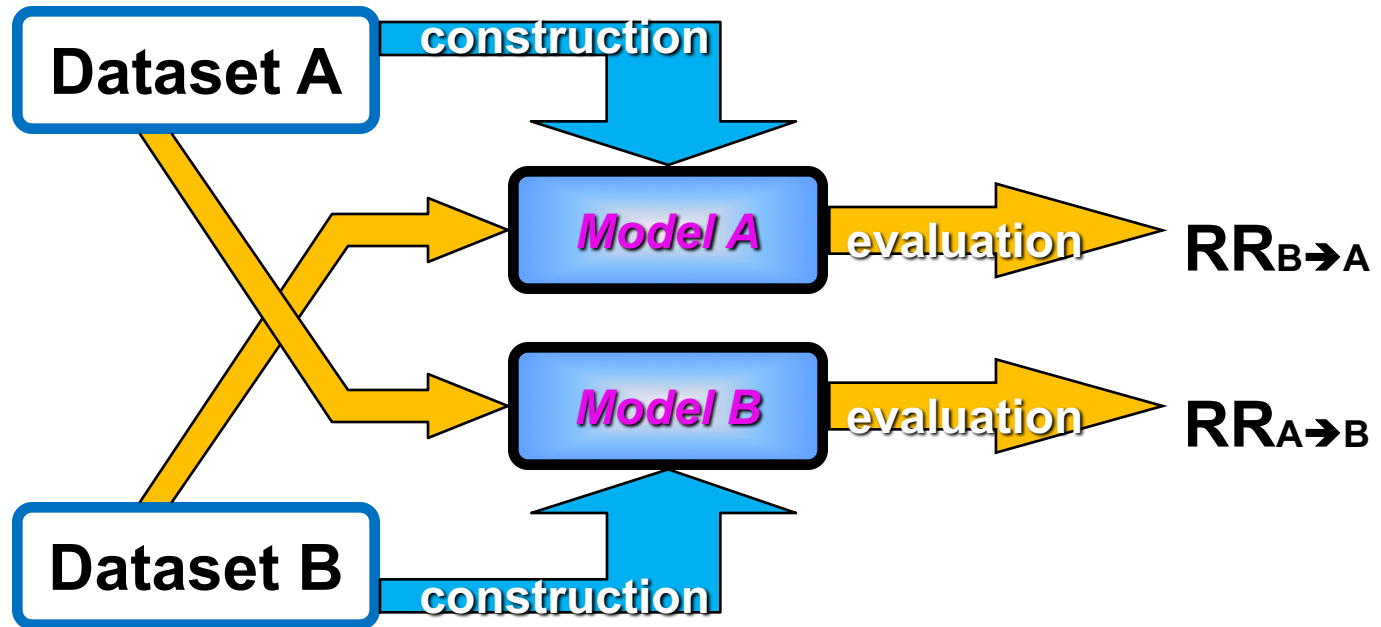
$F_X(\bullet)$: Model identified by the dataset X

$$RR_{inside} = \left(\sum_{i=1}^{|A|} (y_i^A == F_A(\mathbf{x}_i^A)) + \sum_{i=1}^{|B|} (y_i^B == F_B(\mathbf{x}_i^B)) \right) / (|A| + |B|)$$

$$RR_{outside} = \left(\sum_{i=1}^{|A|} (y_i^A == F_B(\mathbf{x}_i^A)) + \sum_{i=1}^{|B|} (y_i^B == F_A(\mathbf{x}_i^B)) \right) / (|A| + |B|)$$

Two-side Holdout Test (2/3)

- Two-side holdout test (two-fold cross-validation)



$$RR_{CV} = \frac{|A| * RR_{A \rightarrow B} + |B| * RR_{B \rightarrow A}}{|A| + |B|}$$

Outside test!

Two-sided Holdout Test (3/3)

- Characteristics
 - Better use of the dataset
 - Still highly affected by the partitioning
 - Suitable for models with high training (design-time) computation load

M-fold Cross Validation (1/3)

- Data partitioning
 - Partition the dataset into m folds
 - One fold for test, the other folds for training
 - Repeat m times

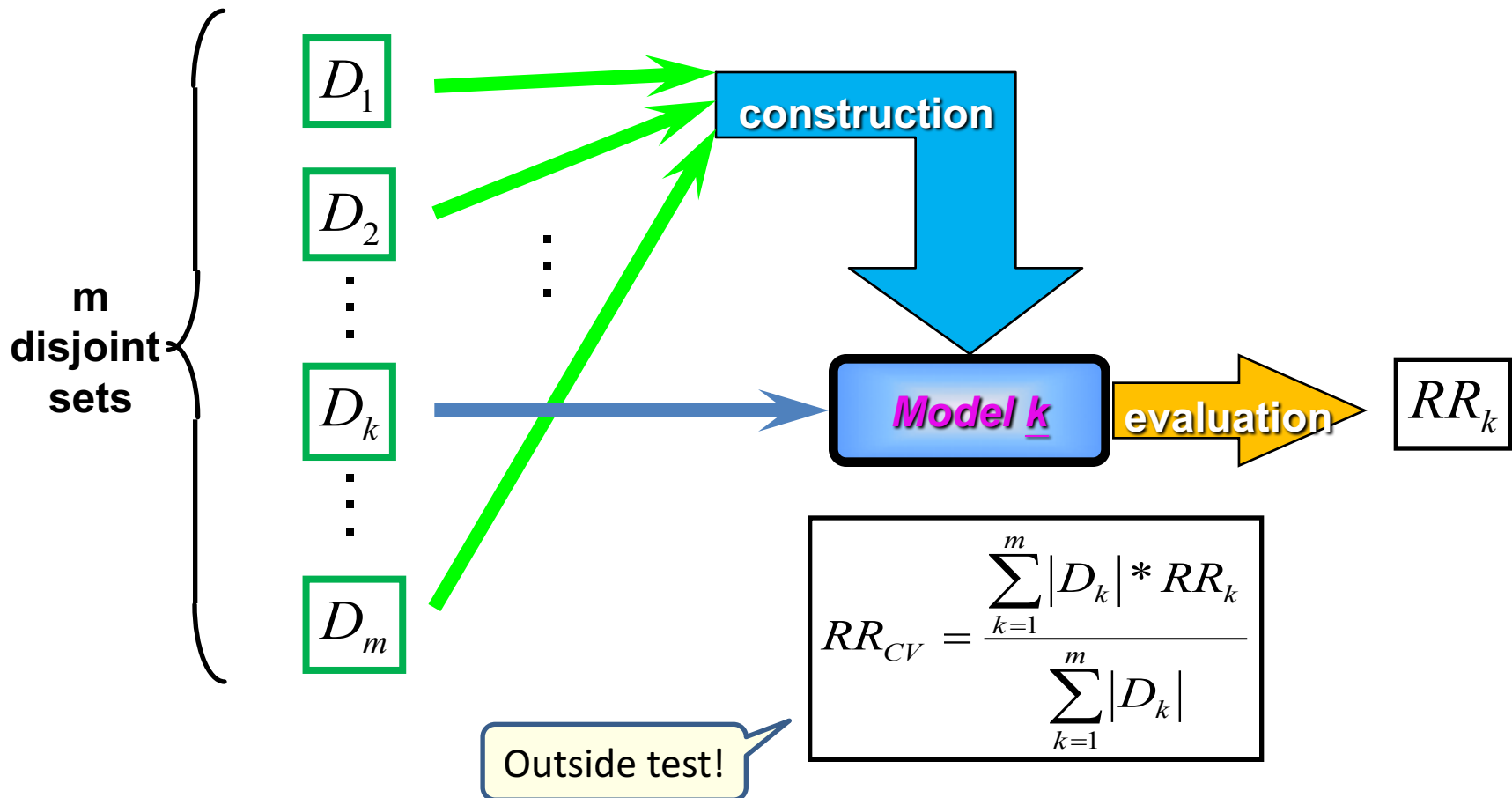
Dataset $D = D_1 \oplus D_2 \oplus \dots \oplus D_m$

$F_{D-D_i}(\bullet)$: Model identified by the dataset $D - D_i$

$$RR_{inside} = \left(\sum_{j=1}^m \sum_{(\mathbf{x}_i, y_i) \in D-D_j} (y_i == F_{D-D_j}(\mathbf{x}_i)) \right) / \left(\sum_{j=1}^m |D-D_j| \right)$$

$$RR_{outside} = \left(\sum_{j=1}^m \sum_{(\mathbf{x}_i, y_i) \in D_j} (y_i == F_{D-D_j}(\mathbf{x}_i)) \right) / \left(\sum_{j=1}^m |D_j| \right)$$

M-fold Cross Validation (2/3)



M-fold Cross Validation (3/3)

- Characteristics
 - When $m=2$ → Two-sided holdout test
 - When $m=n$ → Leave-one-out cross validation
 - The value of m depends on the computation load imposed by the selected model.

Leave-one-out Cross Validation (1/3)

- Data partitioning
 - When $m=n$ and $D_i = (\mathbf{x}_i, y_i)$

Dataset $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$

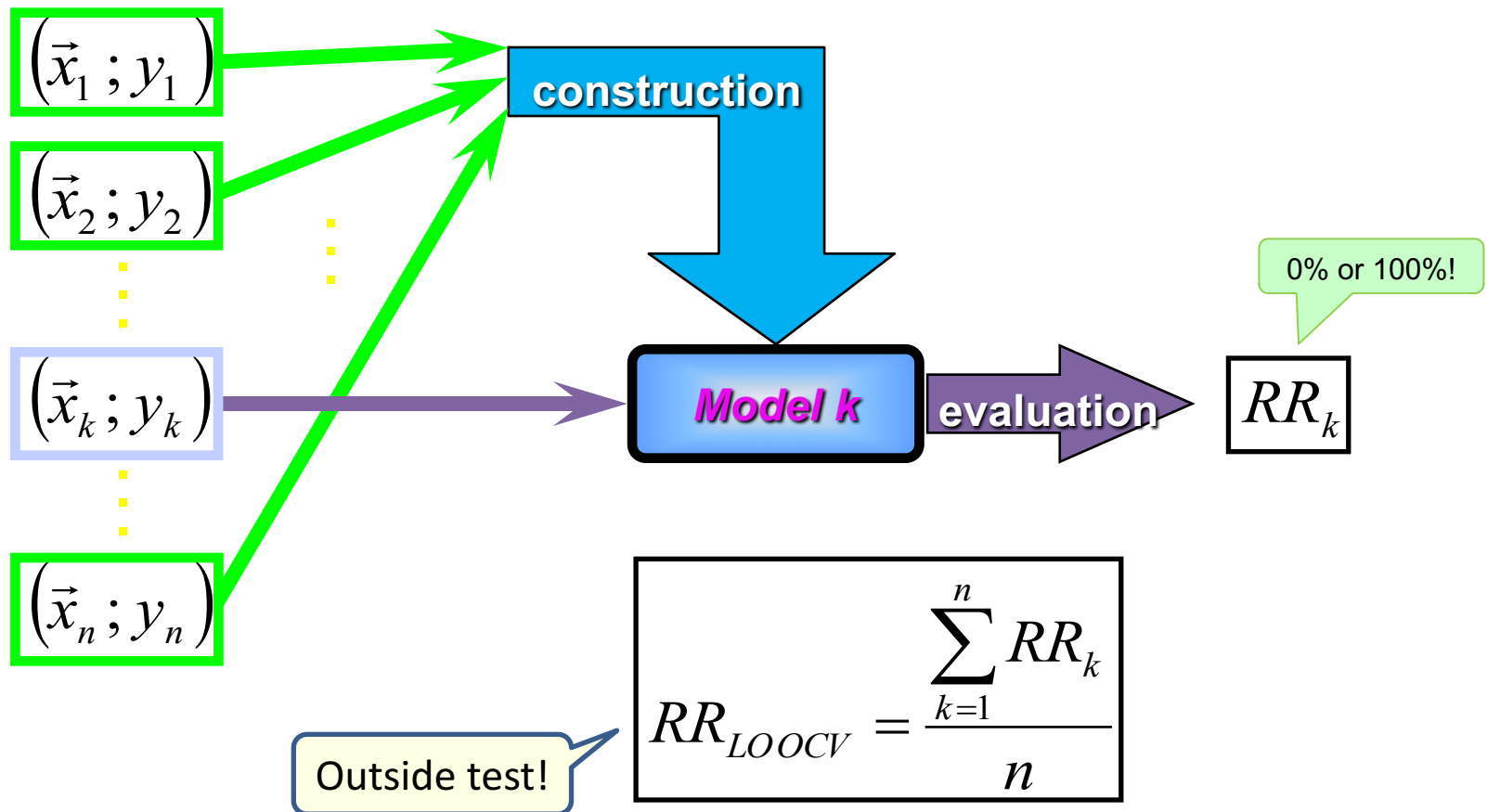
$F_{D-(\mathbf{x}_j, y_j)}(\bullet)$: Model identified by the dataset $D - (\mathbf{x}_j, y_j)$

$$RR_{inside} = \left(\sum_{j=1}^n \sum_{(\mathbf{x}_i, y_i) \in D - (\mathbf{x}_j, y_j)} (y_i == F_{D-(\mathbf{x}_j, y_j)}(\mathbf{x}_i)) \right) / (n * (n - 1))$$

$$RR_{outside} = \left(\sum_{i=1}^n (y_i == F_{D-(\mathbf{x}_i, y_i)}(\mathbf{x}_i)) \right) / n$$

Leave-one-out Cross Validation (2/3)

- Leave-one-out CV



Leave-one-out Cross Validation (3/3)

- General method for LOOCV
 - Perform model construction n times → Slow!
- To speed up the computation LOOCV
 - Construct a common part that is used repeatedly, such as
 - Global mean and covariance for QC

Applications and Misuse of Cross Validation (CV)

- Applications of CV
 - Input (feature) selection
 - Model complexity determination
 - Performance comparison among different models
- Caveat of CV
 - Do not try to boost validation RR too much, or you are running the risk of indirectly training on the left-out data!

Matlab #11 Homework (M11) 3%

- (1%) Compare One-side holdout test and Leave-one-out cross validation. What are their pros and cons?
- (2%) Pick your favorite classifier among the ones we have covered, and train it against the Wine dataset. Use the following methods to derive your recognition rates: (i) inside test, (ii) Two-side holdout test, (iii) 10-fold cross validation, and (iv) leave-one-out cross validation. Compare the resulting recognition rates and running time among them.

Questions?

