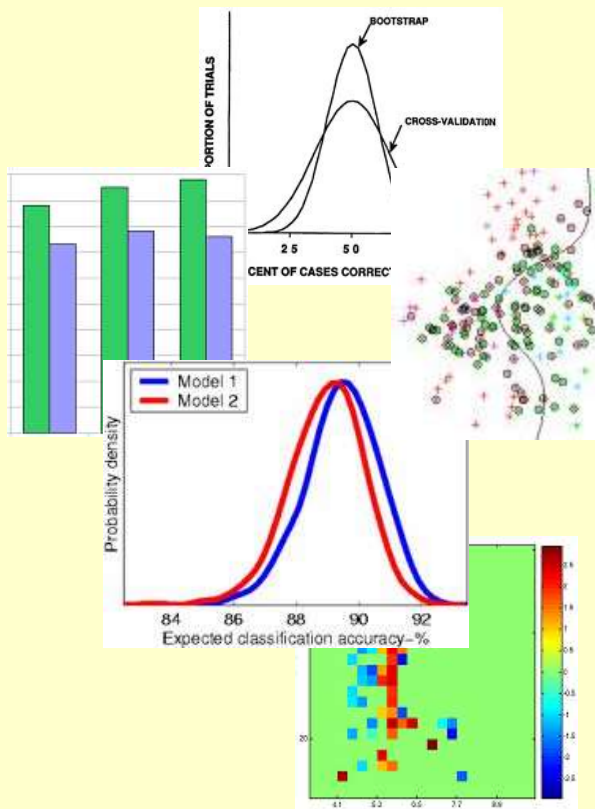# Validation Model & Classification Analysis

Ali Ridho Barakbah

Knowledge Engineering Research Group
Soft Computing Laboratory
Department of Information and Computer Engineering
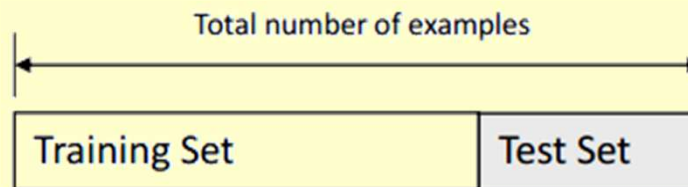Electronic Engineering Polytechnic Institute of Surabaya

# Validation Model of Classification

- Holdout method

- Random subsampling

- K-fold cross validation

- Leave-one-out cross validation

- Bootstrap

# Holdout Method

- Split dataset into two groups
    - Training set: used to train the classifier
    - Test set: used to estimate the error rate of the trained classifier



- The holdout method has two basic drawbacks
    - In problems where we have a sparse dataset we may not be able to afford the "luxury" of setting aside a portion of the dataset for testing
    - Since it is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an "unfortunate" split

Ricardo Gutierrez-Osuna, *Pattern Analysis*, CSE@TAMU

# Random Subsampling

- Random subsampling performs K data splits of the entire dataset
  - Each data split randomly selects a (fixed) number of examples without replacement
  - For each data split we retrain the classifier from scratch with the training examples and then estimate $E_i$ with the test examples

Total number of examples

Test example

Experiment 1

Experiment 2

Experiment 3
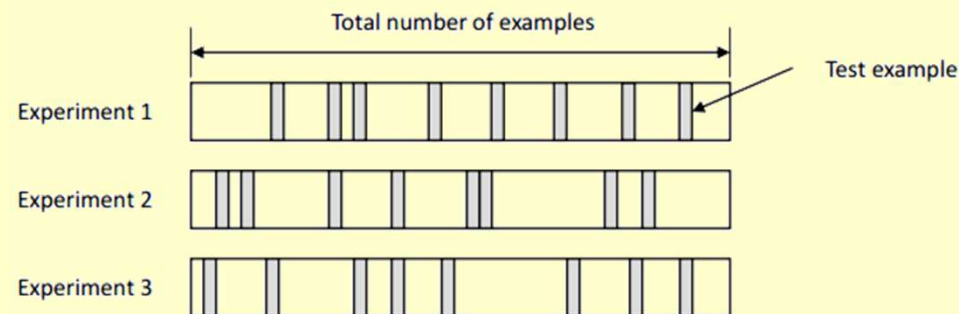
  - The true error estimate is obtained as the average of the separate estimates $E_i$
  - This estimate is significantly better than the holdout estimate

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

Ricardo Gutierrez-Osuna, *Pattern Analysis*, CSE@TAMU

# K-fold Cross Validation

- Create a K-fold partition of the dataset
  - For each of *K* experiments, use *K* - 1 folds for training and a different fold for testing
  - This procedure is illustrated in the following figure for *K* = 4
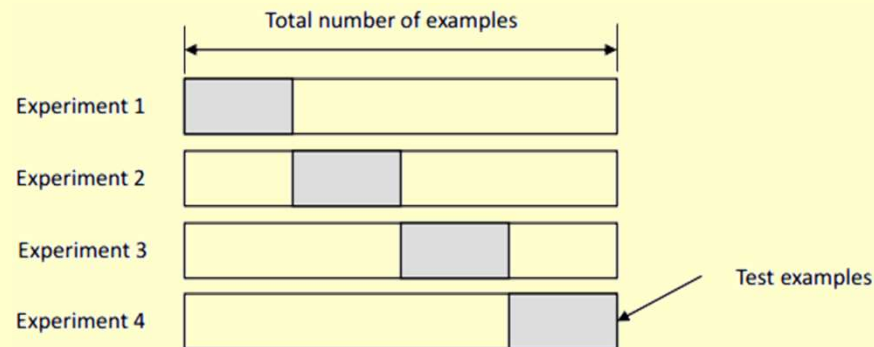


- K-Fold cross validation is similar to random subsampling
  - The advantage of KFCV is that all the examples in the dataset are eventually used for both training and testing
  - As before, the true error is estimated as the average error rate on test examples

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

Ricardo Gutierrez-Osuna, *Pattern Analysis*, CSE@TAMU

# Leave-one-out Cross Validation

- LOO is the degenerate case of KFCV, where K is chosen as the total number of examples
  - For a dataset with *N* examples, perform ?? Experiments
  - For each experiment use *N* - 1 examples for training and the remaining example for testing
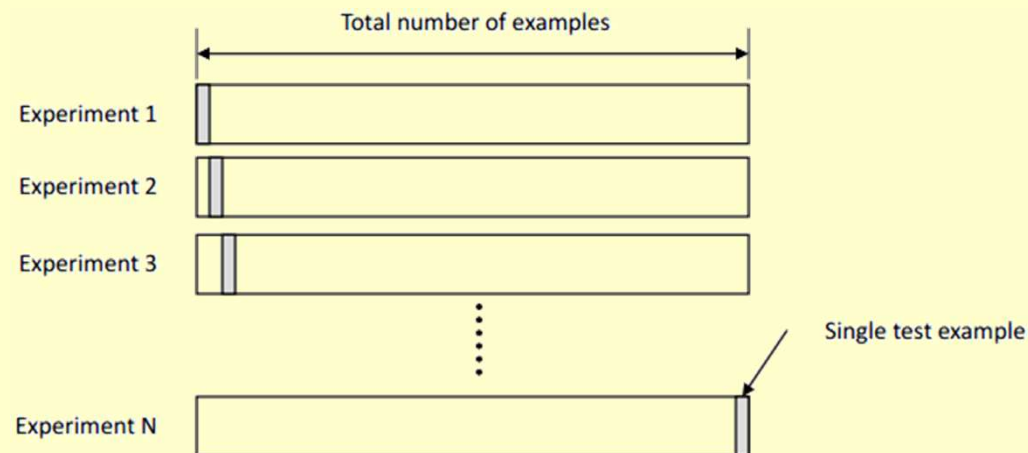


  - As usual, the true error is estimated as the average error rate on test examples

$$E = \frac{1}{N} \sum_{i=1}^{N} E_i$$

Ricardo Gutierrez-Osuna, *Pattern Analysis*, CSE@TAMU
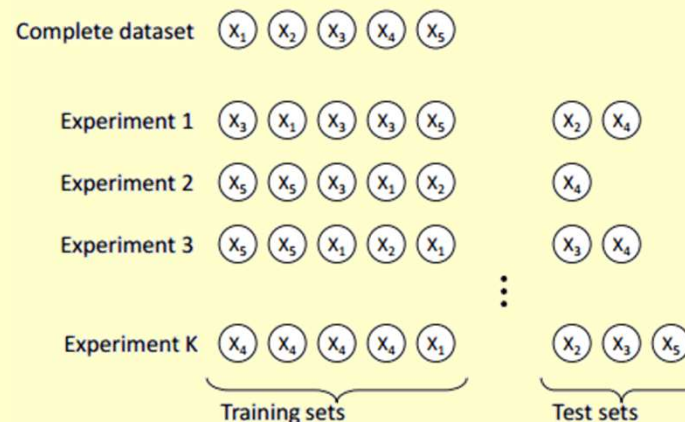
# How many folds are needed?

- With a large number of folds
  - \+ The bias of the true error rate estimator will be small (the estimator will be very accurate)
  - – The variance of the true error rate estimator will be large
  - – The computational time will be very large as well (many experiments)

- With a small number of folds
  - \+ The number of experiments and, therefore, computation time are reduced
  - \+ The variance of the estimator will be small
  - – The bias of the estimator will be large (conservative or larger than the true error rate)

- In practice, the choice for K depends on the size of the dataset
  - – For large datasets, even 3-fold cross validation will be quite accurate
  - – For very sparse datasets, we may have to use leave-one-out in order to train on as many examples as possible

- A common choice for is K=10

Ricardo Gutierrez-Osuna, *Pattern Analysis*, CSE@TAMU

# Bootstrap

- The bootstrap is a resampling technique with replacement
  - From a dataset with ?? Examples
    - Randomly select (with replacement) ?? examples and use this set for training
    - The remaining examples that were not selected for training are used for testing
    - This value is likely to change from fold to fold
  - Repeat this process for a specified number of folds (??)
  - As before, the true error is estimated as the average error rate on test data

| | Training sets | | | | | Test sets | | |
|---|---|---|---|---|---|---|---|---|
| Complete dataset | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | | | |
| Experiment 1 | $X_3$ | $X_1$ | $X_3$ | $X_3$ | $X_5$ | $X_2$ | $X_4$ | |
| Experiment 2 | $X_5$ | $X_5$ | $X_3$ | $X_1$ | $X_2$ | $X_4$ | | |
| Experiment 3 | $X_5$ | $X_5$ | $X_1$ | $X_2$ | $X_1$ | $X_3$ | $X_4$ | |
| Experiment K | $X_4$ | $X_4$ | $X_4$ | $X_4$ | $X_1$ | $X_2$ | $X_3$ | $X_5$ |

Ricardo Gutierrez-Osuna, *Pattern Analysis*, CSE@TAMU

# Performance Analysis of Classification

- Commonly used error rate/ratio

- Dataset → supervised

- Used to analyze precision of classification result from a classification algorithm

$$Error = \frac{missclassified}{Number\ of\ data} \times 100\%$$

# Tugas Klasifikasi dengan k-NN

- Case : klasifikasi bunga Iris

- Source : UCI Repository

- Number of attributes : 4

- Number of instances : 150

- Number of classes : 3
  - Iris Setosa (50 instances)
  - Iris Versicolour (50 instances)
  - Iris Virginica (50 instances)

**Bunga iris**





Iris Setosa



Iris Versicolor



Iris Virginica

# Assigment

- Lakukan performance analysis pada data Iris dengan menggunakan validation model:
  - Holdout method
  - Random subsampling
  - K-fold cross validation
  - Leave-one-out cross validation
  - Bootstrap
- Lakukan klasifikasi masing-masing data uji coba dan hitunglah error ratio-nya.
- Hitunglah error ratio rata-rata pada semua data uji coba (dalam persen).

# Metode klasifikasi

- Lakukan percobaan dengan melibatkan beberapa metode klasifikasi:
  - 1-NN
  - 3-NN
  - 5-NN

# References

- Tom Michael, Machine Learning, McGraw-Hill publisher, 1997.
- Ali Ridho Barakbah, *Machine Learning*, Lecture Handout, Electronic Engineering Polytechnic Institute of Surabaya.
- Ricardo Gutierrez-Osuna, *Pattern Analysis*, Lecture Courses, Department of Computer Science and Engineering - Texas A&M University.
- Tan, Steinbach, Kumar, *Introduction to Data Mining*, Pearson publisher, 2005.
- UCI Repository, Iris Dataset.