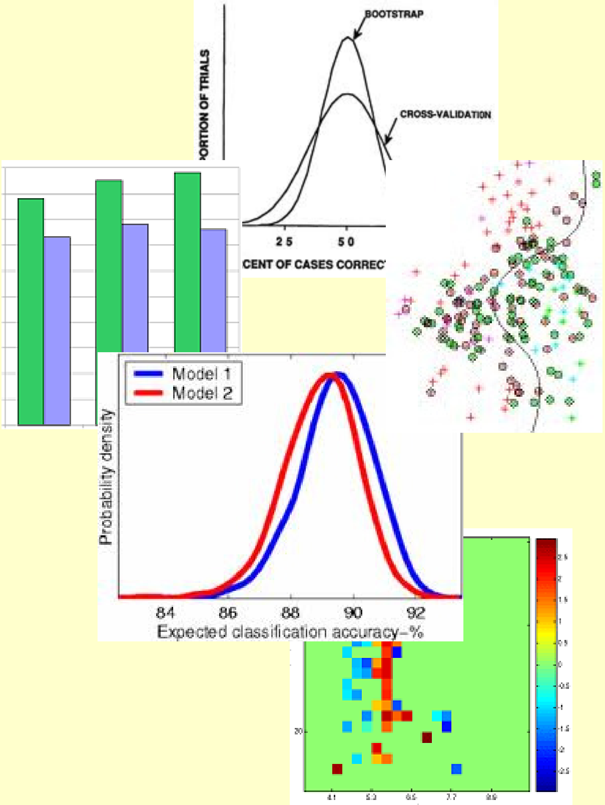


Validation Model of Classification

Ali Ridho Barakbah

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

Classification: Definition

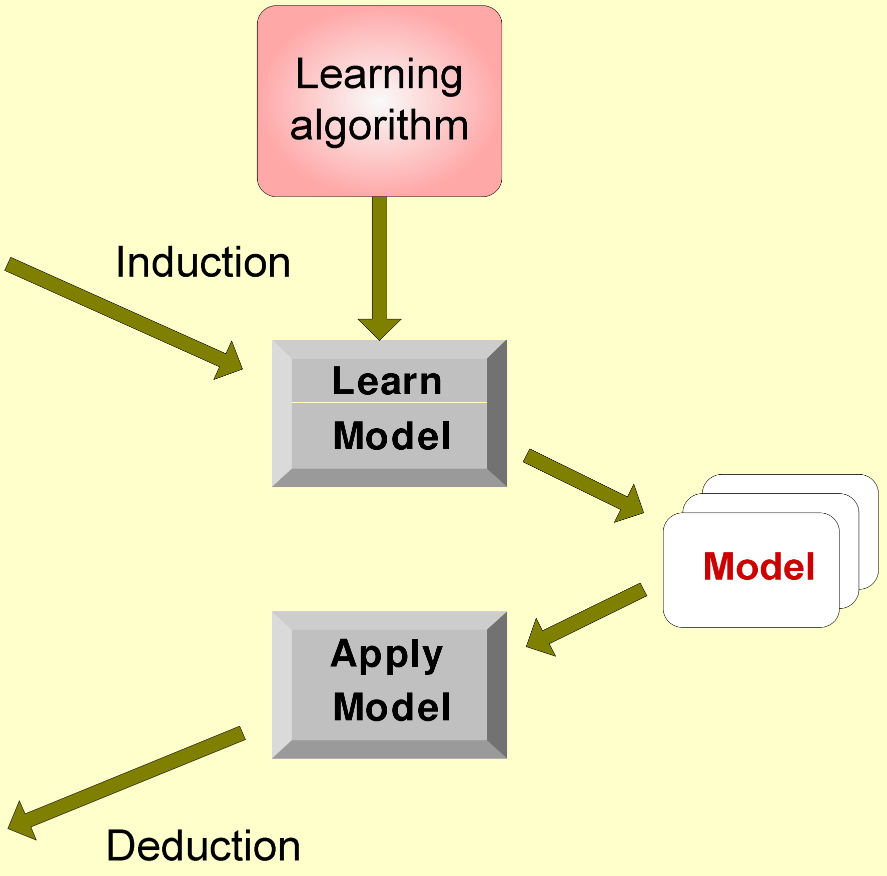
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Tan, Steinbach, Kumar, *Introduction to Data Mining*

Illustrating Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

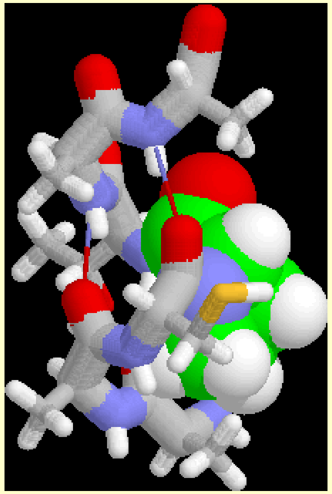


Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Examples of Classification Task

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
- Categorizing news stories as finance, weather, entertainment, sports, etc



Fase klasifikasi?

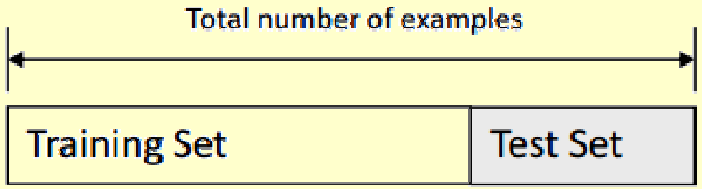
- Proses klasifikasi biasanya dibagi menjadi dua fase : learning dan test.
  - Fase learning → sebagian data yang telah diketahui kelas datanya diumpankan untuk membentuk model perkiraan.
  - Fase test → model yang sudah terbentuk diuji dengan sebagian data lainnya untuk mengetahui akurasi dari model tsb.
- Bila akurasinya mencukupi model ini dapat dipakai untuk prediksi kelas data yang belum diketahui.
- Klasifikasi dicirikan dengan data training mempunyai label, berdasarkan label ini proses klasifikasi memperoleh pola attribut dari suatu data.

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

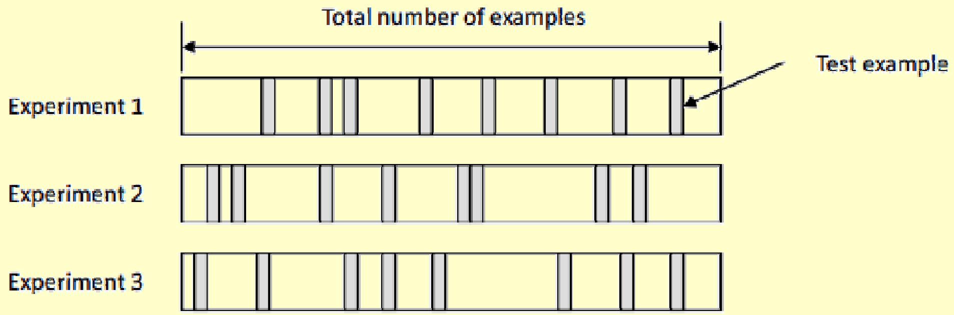


- The holdout method has two basic drawbacks
  - In problems where we have a sparse dataset we may not be able to afford the “luxury” of setting aside a portion of the dataset for testing
  - Since it is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen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



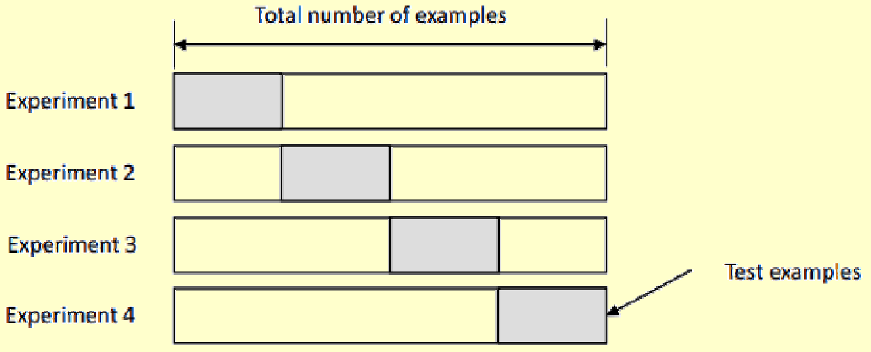
  - The true error estimate is obtained as the average of the separate estimates E_i
  - This estimate is significantly better than the holdout estimate

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

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



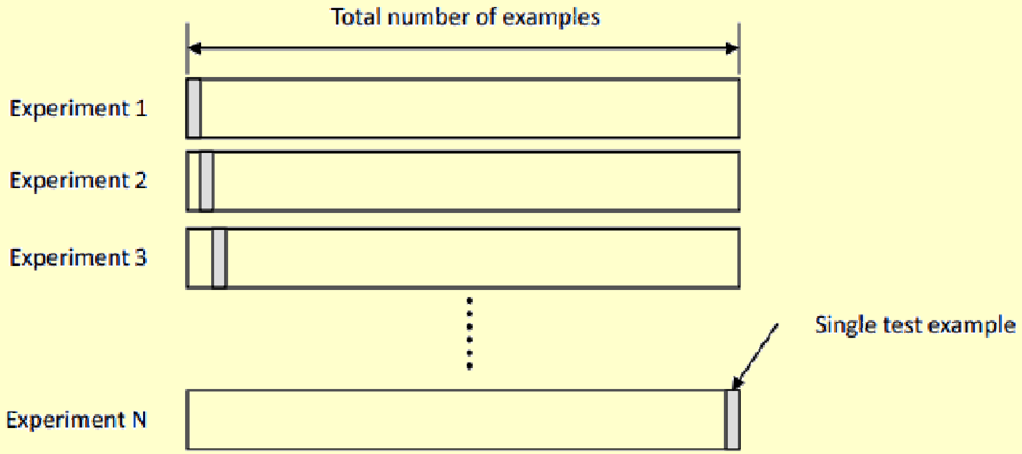
- K-Fold cross validation is similar to random subsampling
  - The advantage of KFCV is that all the examples in the dataset are eventually used for both training and testing
  - As before, the true error is estimated as the average error rate on test examples

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

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

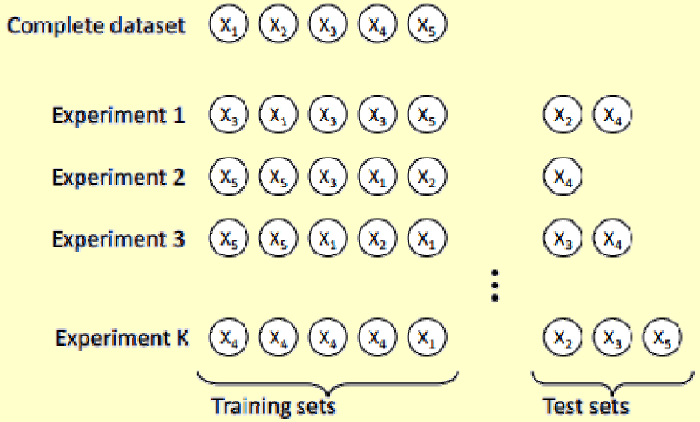
How many folds are needed?

- 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



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
	Training sets					Test sets		

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

Performance Analysis of Classification

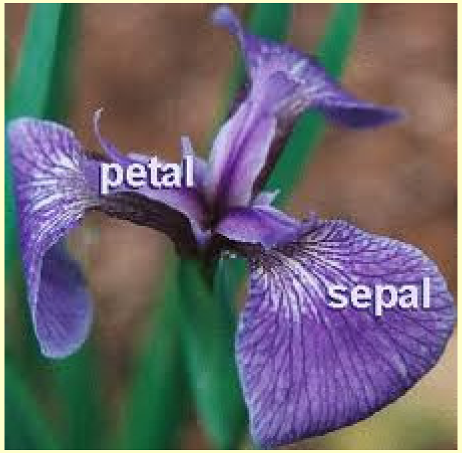
- Commonly used error rate/ratio
- Dataset → supervised
- Used to analyze precision of classification result from a classification algorithm

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

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.