Chapter 11

k- Fold Cross Validation

- Comparison Technique
- Step
- Case

Sulidar Fitri, M.Sc

© Sulidar Fitri, Ms.C

REFERENCES

- *Jiawei Han and Micheline Kamber*. Data Mining: Concepts and Techniques. 2006. Department of Computer Science University of Illinois at Urbana-Champaign. www.cs.uiuc.edu/~hanj
- *Ian H. Witten, Eibe Frank, Mark A. Hall.* Data Mining Practical Machine Learning Tools and Techniques Third Edition.2011. Elsevier
- WEKA
- Any Online Resources

© Sulidar Fitri, Ms.C

Classification Measurements

1. Classification Accuracy (%) :
   - Testing accuracy

$$accuracy = \frac{T}{A} x100$$

where T and A are the number of true classified samples and total of samples in the data set respectively.

© Sulidar Fitri, Ms.C

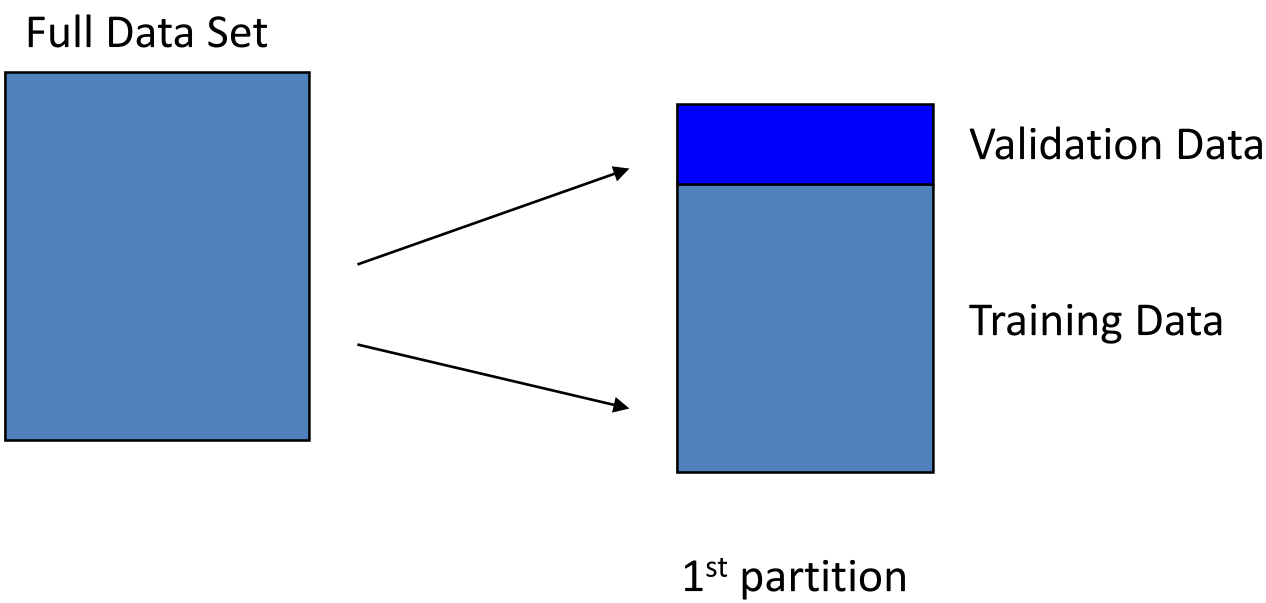
CROSS VALIDATION

- Merupakan metode statistic untuk meng-evaluasi dan membandingkan akurasi *Learning algorithm* dengan cara membagi dataset menjadi 2 bagian:
  - Satu bagian digunakan untuk training model,
  - Bagian yang lain untuk mem-validasi model

- Suatu dataset akan dibagi sesuai dengan banyaknya k, dan akan di test bergantian hingga seluruh bagian terpenuhi.

© Sulidar Fitri, Ms.C

Disjoint Validation Data Sets



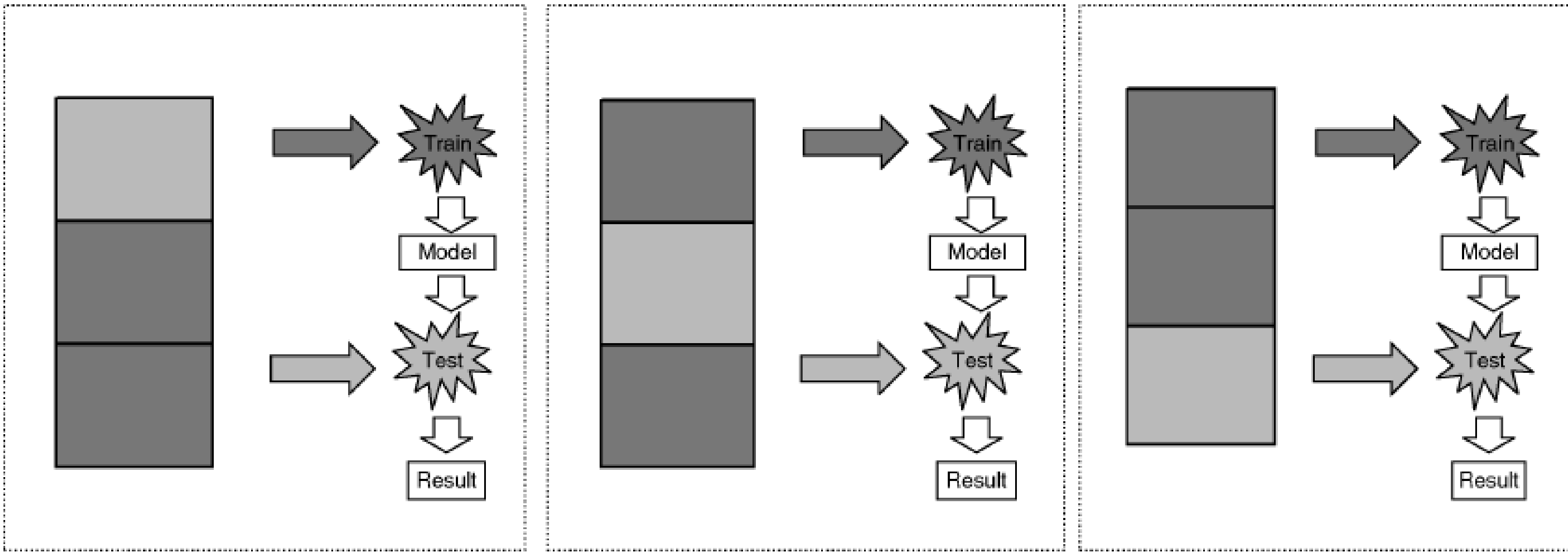
© Sulidar Fitri, Ms.C

k-fold Cross Validation

- In k-fold cross-validation the data is first partitioned into k equally (or nearly equally) sized segments or folds.
- Subsequently k iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining k - 1 folds are used for learning.
- Fig. 1 demonstrates an example with k = 3. The darker section of the data are used for training while the lighter sections are used for validation.
- In data mining and machine learning 10-fold cross-validation (k = 10) is the most common.

© Sulidar Fitri, Ms.C

k-fold Cross Validation

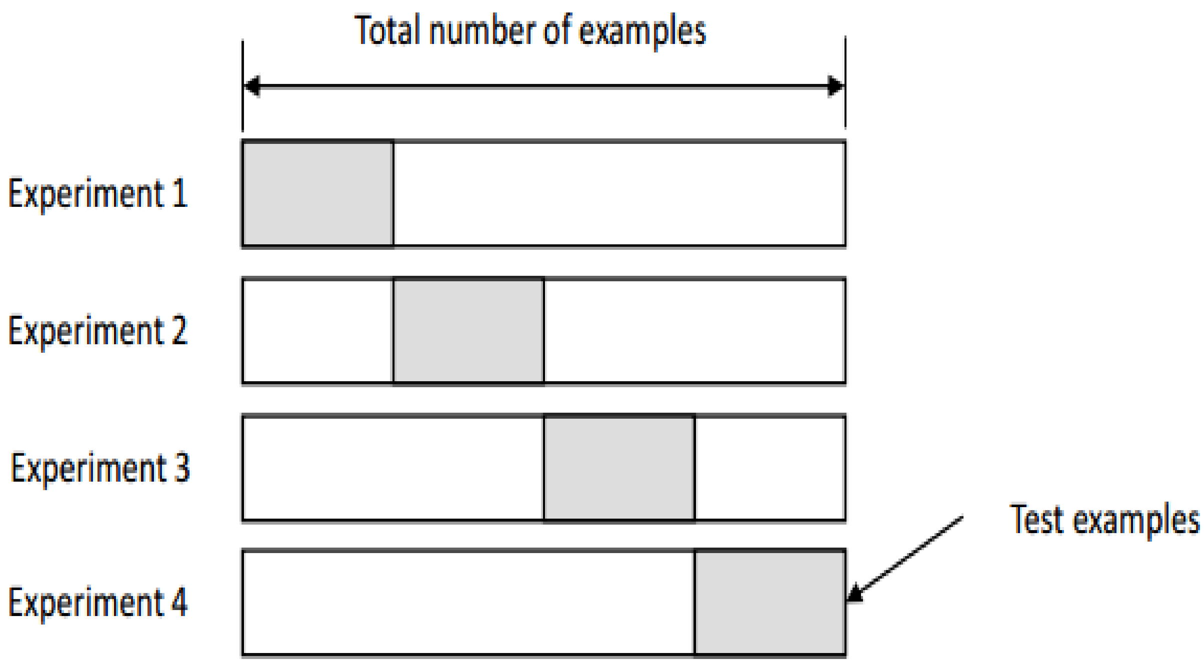


Cross-Validation. Figure 1. Procedure of three-fold cross-validation.

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 $K=4$



© Sulidar Fitri, Ms.C

Paired t-Test

- Collect data in pairs:
  - Example: Given a training set D_{Train} and a test set D_{Test}, train ***both*** learning algorithms on D_{Train} and then test their accuracies on D_{Test}.
- Suppose n paired measurements have been made
- Assume
  - The measurements are independent
  - The measurements for each algorithm follow a normal distribution
- The test statistic T_0 will follow a t-distribution with n-1 degrees of freedom

© Sulidar Fitri, Ms.C

Paired t-Test cont

Trial #	Algorithm 1 Accuracy X_1	Algorithm 2 Accuracy X_2
1	X_{11}	X_{21}
2	X_{12}	X_{22}
…	..	…
n	X_{1N}	X_{2N}

Assume: X_1 follows $N(\mu_1,\sigma_1)$
X_2 follows $N(\mu_2,\sigma_2)$

Let: $\mu_{D} = \mu_1 - \mu_2$

$D_i = X_{1i} - X_{2i}$ i=1,2,...,n

$$\overline{D} = \frac{1}{n}\sum_i X_{1i} - X_{2i}$$

$$S_D = STDEV(X_{1i} - X_{2i})$$

Null Hypothesis:

$H_0: \mu_D = \Delta_0$

Test Statistic:

$$T_0 = \frac{(\overline{D} - \Delta_0)\sqrt{n}}{S_D}$$

Rejection Criteria:

$H_1: \mu_D \neq \Delta_0$ $|t_0| > t_{\alpha/2,n-1}$

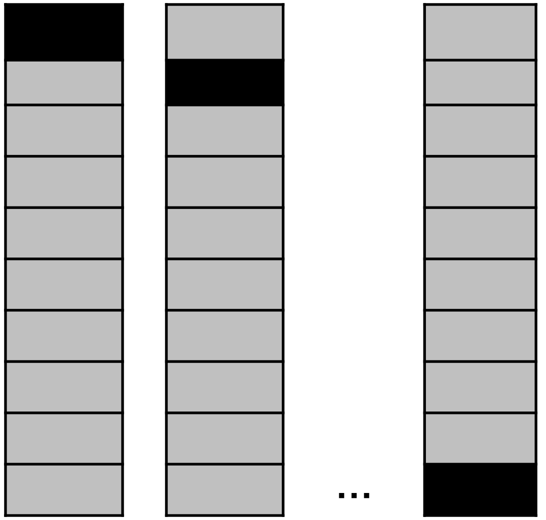
$H_1: \mu_D > \Delta_0$ $t_0 > t_{\alpha,n-1}$

$H_1: \mu_D < \Delta_0$ $t_0 < -t_{\alpha,n-1}$

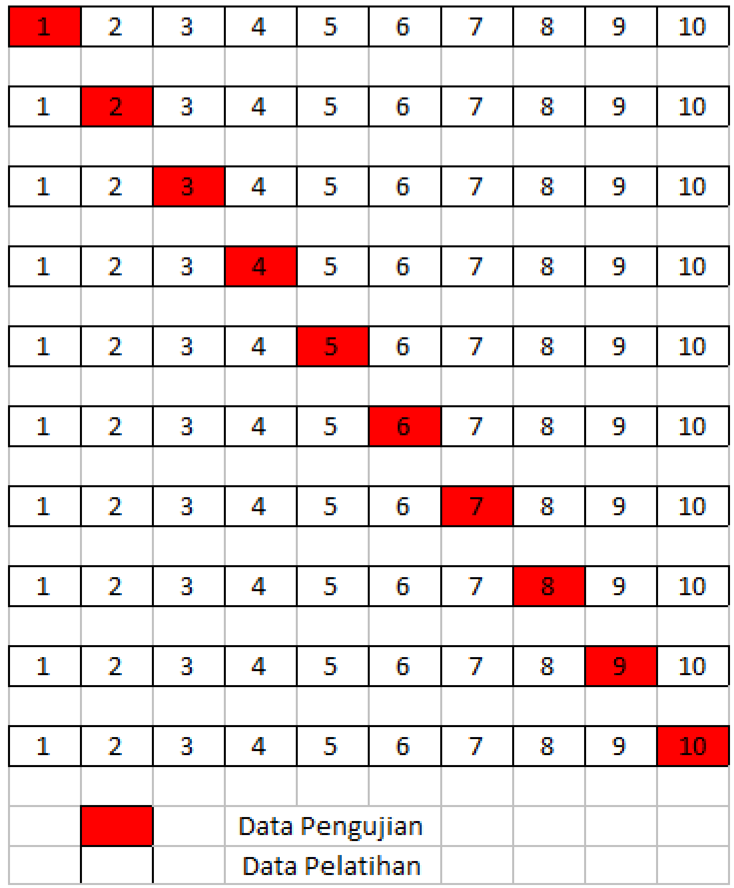
© Sulidar Fitri, Ms.C

Cross Validated t-test

- Paired t-Test on the 10 paired accuracies obtained from 10-fold cross validation
- Advantages
  - Large train set size
  - Most powerful (Diettrich, 98)
- Disadvantages
  - Accuracy results are not independent (overlap)
  - Somewhat elevated probability of type-1 error (Diettrich, 98)



© Sulidar Fitri, Ms.C



Gambar 1 – Skema 10 fold CV

© Sulidar Fitri, Ms.C

Student's distribution

- With small samples (k < 100) the mean follows *Student's distribution with k–1 degrees of freedom*
- Confidence limits:

Assuming we have 10 estimates

9 degrees of freedom

$Pr[X \geq z]$	z
0.1%	4.30
0.5%	3.25
1%	2.82
5%	1.83
10%	1.38
20%	0.88

normal distribution

$Pr[X \geq z]$	z
0.1%	3.09
0.5%	2.58
1%	2.33
5%	1.65
10%	1.28
20%	0.84

© Sulidar Fitri, Ms.C

© Sulidar Fitri, Ms.C

k-fold Cross Validation

- LOOCV is computationally intensive, so we can run k-fold Cross Validation instead
- With k-fold Cross Validation, we divide the data set into K different parts (e.g. K = 5, or K = 10, etc.)
- We then remove the first part, fit the model on the remaining K-1 parts, and see how good the predictions are on the left out part (i.e. compute the MSE on the first part)
- We then repeat this K different times taking out a different part each time
- By averaging the K different MSE's we get an estimated validation (test) error rate for new observations

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i .$$

Distribution of the differences

- Let $m_d = m_x - m_y$
- The difference of the means (m_d) also has a Student's distribution with k–1 degrees of freedom
- Let σ_d^2 be the variance of the difference
- The standardized version of m_d is called the t-statistic:

$$t = \frac{m_d}{\sqrt{\sigma_d^2 / k}}$$

- We use t to perform the t-test

© Sulidar Fitri, Ms.C

Contoh:

Jika dimiliki data :

210, 340, 525, 450, 275

maka variansi dan standar deviasinya :

mean = (210, 340, 525, 450, 275)/5 = 360

variansi dan standar deviasi berturut-turut :

X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
210	-150	22500
275	-85	7225
340	-20	400
450	90	8100
525	165	37225
	0	65450

$$=> \text{Variansi} : s^2 = \frac{65450}{4} = 16362{,}5$$

$$\text{deviasi standar} : s = \sqrt{16362{,}5} = 127{,}916$$

© Sulidar Fitri, Ms.C

Performing the test

- Fix a significance level
  - If a difference is significant at the α% level, there is a (100-α)% chance that the true means differ
- Divide the significance level by two because the test is two-tailed
- Look up the value for z that corresponds to $\alpha/2$
- If $t \leq -z$ or $t \geq z$ then the difference is significant
  - I.e. the *null hypothesis* (that the difference is zero) can be rejected

© Sulidar Fitri, Ms.C

Do Comparison !!

- Lakukan penghitungan perbandingan akurasi dari dua algorithma!

© Sulidar Fitri, Ms.C

Any Queries ?