

Pendahuluan*

Dr. rer. nat. Hendri Murfi

* Beberapa bagian dari slide ini adalah terjemahan dari slide Data Mining oleh I. H. Witten, E. Frank dan M. A. Hall

Intelligent Data Analysis (IDA) Group

Departemen Matematika, Universitas Indonesia – Depok 16424

Telp. +62-21-7862719/7863439, Fax. +62-21-7863439, Email. hendri@ui.ac.id

Data vs Informasi

- Interaksi dalam bermasyarakat menghasilkan data dalam jumlah yang sangat besar
 - Sumber: bisnis, ilmu pengetahuan, kedokteran, ekonomi, geografi, lingkungan, olah raga, dll
- Perlu teknik tertentu untuk mengekstrak informasi dari data mentah
 - Data: fakta-fakta yang tercatat
 - Informasi: pola-pola struktural yang ada pada data, misal keterhubungan (*relationships*) atau ketergantungan (*dependencies*)

Data vs Informasi

- Contoh 1: *in vitro fertilization*
 - Diberikan: embrio yang dideskripsikan oleh 60 fitur
 - Masalah: pemilihan embrio yang akan bertahan hidup
 - Data: fakta historis dari embrio dan hasil akhir
- Contoh 2: *cow culling*
 - Diberikan: sapi yang dideskripsikan oleh 700 fitur
 - Masalah: pemilihan sapi yang akan dimusnakan
 - Data: fakta historis dari sapi dan keputusan peternak

3

Machine Learning

Latar Belakang

- Memahami keterhubungan (*relationships*) dan ketergantungan (*dependencies*) dalam suatu koleksi data adalah suatu aspek yang sangat penting dalam menganalisa data untuk mengekstrak informasi yang ada pada data tersebut.
- Deskripsi struktural tersebut dapat digunakan untuk:
 - Memprediksi output pada suatu situasi yang baru
 - Memahami dan menjelaskan bagaimana prediksi tersebut dihasilkan
- Ketika tidak ada pendekatan pemodelan (*modelling approaches*) yang mudah untuk melakukan hal tersebut, maka metode-metode cerdas, dikenal juga dengan nama *machine learning*, menjadi solusi alternatif

4

Machine Learning

Deskripsi Struktural: Penentuan Jenis Contact Lens

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
...	...	...	...	...

If-then rules:

If tear production rate = reduced
then recommendation = none

Otherwise, if age = young and astigmatic = no
then recommendation = soft

5

Machine Learning

Deskripsi Struktural: Penentuan Jenis Contact Lens

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

Machine Learning

Deskripsi Struktural: Penentuan Jenis Contact Lens

If-then rules yang lengkap:

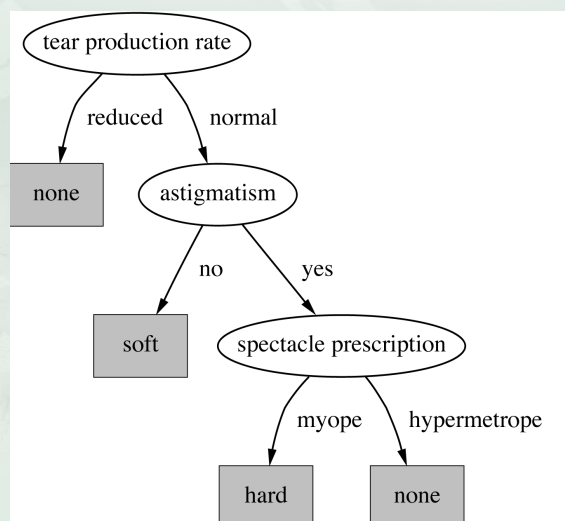
```
If tear production rate = reduced then recommendation = none
If age = young and astigmatic = no
  and tear production rate = normal then recommendation = soft
If age = pre-presbyopic and astigmatic = no
  and tear production rate = normal then recommendation = soft
If age = presbyopic and spectacle prescription = myope
  and astigmatic = no then recommendation = none
If spectacle prescription = hypermetrope and astigmatic = no
  and tear production rate = normal then recommendation = soft
If spectacle prescription = myope and astigmatic = yes
  and tear production rate = normal then recommendation = hard
If age young and astigmatic = yes
  and tear production rate = normal then recommendation = hard
If age = pre-presbyopic
  and spectacle prescription = hypermetrope
  and astigmatic = yes then recommendation = none
If age = presbyopic and spectacle prescription = hypermetrope
  and astigmatic = yes then recommendation = none
```

7

Machine Learning

Deskripsi Struktural: Penentuan Jenis Contact Lens

Decision Tree:



8

Machine Learning

Deskripsi Struktural: Prediksi Kinerja CPU

Contoh: 209 konfigurasi komputer yang berbeda

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performanc e
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

Model Linear (regresi):

$$\text{PRP} = -55.9 + 0.0489 \text{ MYCT} + 0.0153 \text{ MMIN} + 0.0056 \text{ MMAX} \\ + 0.6410 \text{ CACH} - 0.2700 \text{ CHMIN} + 1.480 \text{ CHMAX}$$

9

Machine Learning

Deskripsi Struktural: Representasi

- Pola-pola struktur yang ada pada suatu koleksi data (*knowledge*) dapat direpresentasi/diekspresikan dalam bentuk:
 - Tabel
 - Aturan (If-then Rules)
 - Pohon (Decision Trees)
 - Fungsi → Model Linear
 - Instance

10

Machine Learning

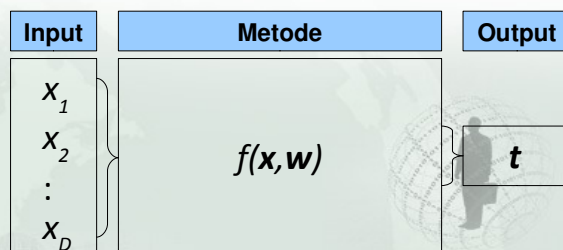
Definisi

- **Machine Learning** adalah metode yang dapat belajar dari data (*data-driven method*) sehingga menjadi cerdas. Cerdas dalam artian memiliki kemampuan generalisasi terhadap data baru yang belum dipelajari sebelumnya.
- Dalam tataran metode, belajar adalah proses penentuan nilai parameter-parameter dari metode tersebut.
- Pengembangan metode-metode machine learning menggunakan ide-ide dari: **Mathematics, Statistics, Computer Science, Computational Neuroscience, Cognitive Science, Psychology**

11

Machine Learning

Tahapan Umum Proses



Diberikan data pelatihan (*training data*), misal $\mathbf{x}_i$ dan/atau $\mathbf{t}_i$, $i = 1$ sd N

- **Preprocessing**: pemilihan/ekstraksi fitur dari data, misal $\mathbf{x}_i = (x_1, x_2, \dots, x_D)^T$
- **Learning**: penentuan parameter metode, misal $\mathbf{w}$, berdasarkan data pelatihan
- **Testing**: pengujian metode dengan data baru. Data penguji (*testing data*) tersebut harus dilakukan *preprocessing* yang sama dengan data pembelajaran sebelum dieksekusi oleh metode

12

Machine Learning

Preprocessing: Pemilihan Fitur

- Langka pertama dalam tahapan preprocessing adalah pemilihan fitur (*feature selection*) dari data. Selanjutnya, data dapat direpresentasi dalam bentuk vektor, matrik atau tensor.
- Misal: pada perekomendasi jenis *contact lens*, fitur-fitur tersebut adalah *age*, *spectacle prescription*, *astigmatism*, *tear production rate*. Sehingga, data dapat direpresentasikan dalam bentuk vektor, misal $\mathbf{x}_1 = (\text{Young}, \text{Myope}, \text{No}, \text{Reduced})^T$

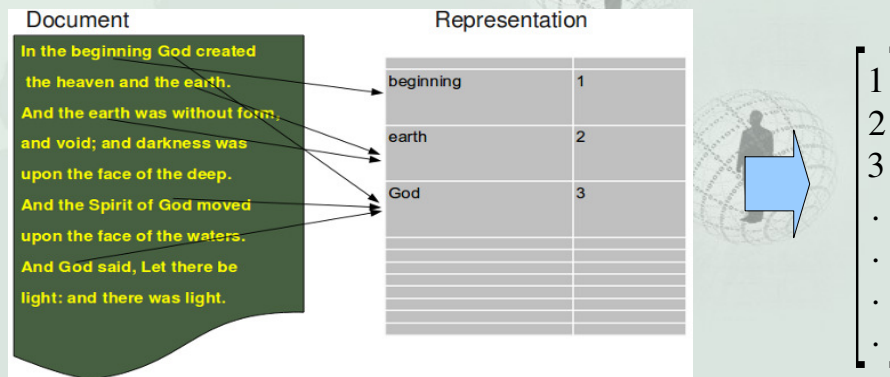
Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
...	...	...	...	...

13

Machine Learning

Preprocessing: Pemilihan Fitur

- Pada beberapa jenis data, proses pemilihan fitur dan merepresentasikan data dalam bentuk vektor, matrik atau tensor membutuhkan mekanisme yang cukup kompleks, misal untuk data dalam bentuk gambar, video, suara, teks, dll



14

Machine Learning

Preprocessing: Ekstraksi Fitur

- Pemilihan fitur sangat mempengaruhi proses pembelajaran, baik dari segi kemudahan, kecepatan, maupun akurasi.
- Pemilihan fitur-fitur dari fitur-fitur yang ada dikenal juga dengan istilah ekstraksi fitur (*feature extraction*).
- Ekstraksi fitur erat kaitannya dengan mereduksi dimensi input (*dimensionality reduction*) untuk mengatasi masalah komputasi yang dikenal dengan istilah *curse dimensionality problem*

15

Machine Learning

Preprocessing: Verifikasi Data

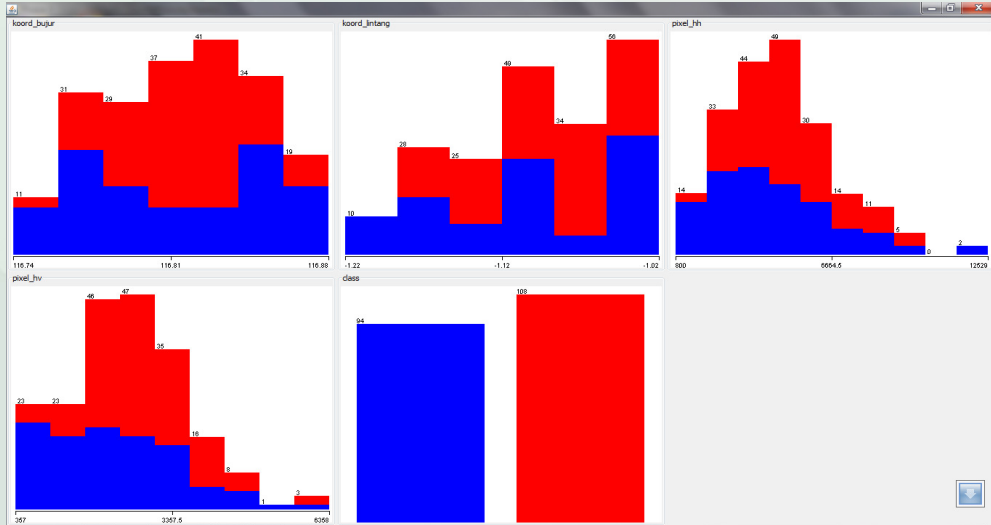
- Salah satu cara untuk melakukan verifikasi terhadap data yang ada adalah dengan melakukan visualisasi terhadap data tersebut
- Contoh sederhana adalah menggunakan histogram untuk menunjukkan distribusi data nominal, atau graph untuk data numerik
- Dengan visualisasi ini diharapkan kita akan dapat melihat outlier, yang menyatakan error pada data, atau kesalahan coding seperti tahun bernilai 9999, berat bernilai -1 kg.

16

Machine Learning

Preprocessing: Verifikasi Data

- Contoh visualisasi distribusi kelas dari suatu fitur

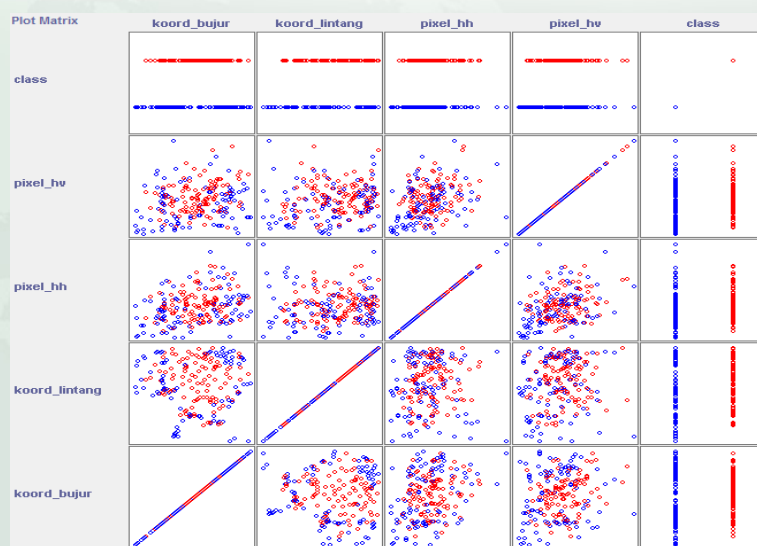


17

Machine Learning

Preprocessing: Verifikasi Data

- Contoh visualisasi distribusi kelas antar fitur



18

Machine Learning

Learning: Jenis-Jenis Learning

Diberikan data pelatihan $\mathbf{x}_i$, $i = 1$ sd N , dan/atau $\mathbf{t}_i$, $i = 1$ as N

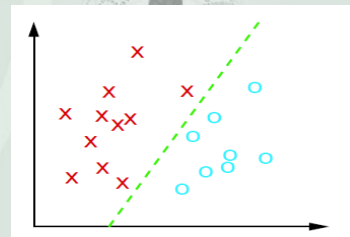
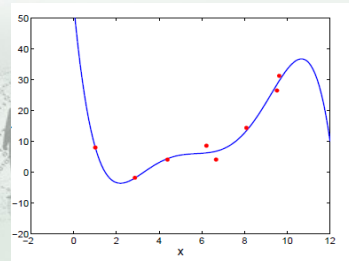
- **Supervised Learning.** Data pelatihan disertai target, yaitu $\{\mathbf{x}_i, \mathbf{t}_i\}$, $i = 1$ sd N . Tujuan pembelajaran adalah membangun model yang dapat menghasilkan output yang benar untuk suatu data input, misal untuk pengklasifikasian (*classification*), regresi (*regression*), ranking
- **Unsupervised Learning.** Data pelatihan tidak disertai target, yaitu $\mathbf{x}_i$, $i = 1$ sd N . Tujuan pembelajaran adalah membangun model yang dapat menemukan komponen/variabel/fitur tersembunyi pada data pelatihan, yang dapat digunakan untuk: pengelompokan (*clustering*), reduksi dimensi (*dimension reduction*), rekomendasi, dll

19

Machine Learning

Supervised Learning

- **Regresi**
 - Nilai output $\mathbf{t}_i$ bernilai kontinu (riil)
 - Bertujuan memprediksi output dengan akurat untuk data baru
 - Contoh: Prediksi kinerja CPU
- **Klasifikasi**
 - Nilai output $\mathbf{t}_i$ bernilai diskrit (kelas)
 - Bertujuan mengklasifikasi data baru dengan akurat
 - Contoh: Klasifikasi jenis *contact lens*

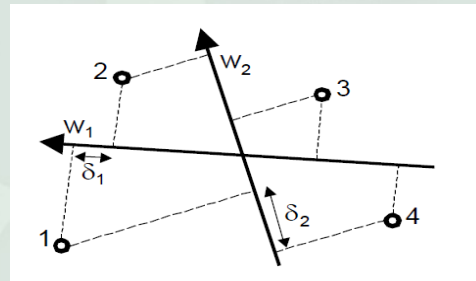


20

Machine Learning

Supervised Learning

- **Ranking / Regresi Ordinal**
 - Nilai output t_i bernilai kontinu (riil)
 - Bertujuan memprediksi ranking atau urutan dari data baru



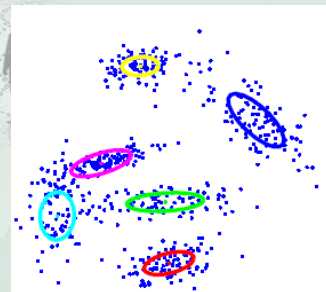
21

Machine Learning

Unsupervised Learning

Bertujuan untuk menemukan variabel/komponen/fitur tersembunyi pada data pembelajaran yang dapat digunakan untuk:

- *Clustering*
- *Dimension Reduction*
- *Recommendation*
- *Topic Modeling*



22

Machine Learning

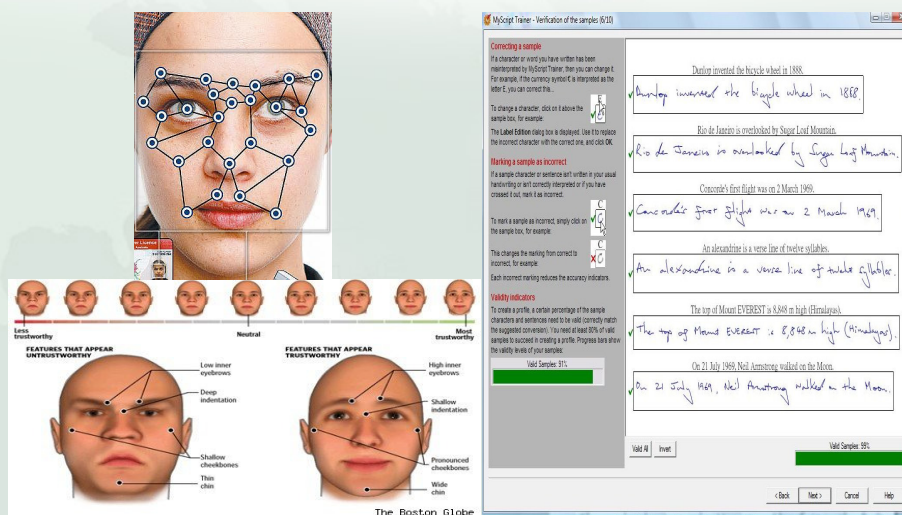
Testing: Generalization Capability

- Kinerja Metode diukur berdasarkan kemampuan generalisasinya, yaitu akurasi pada data yang tidak digunakan pada tahap pelatihan, yang disebut juga dengan istilah kapabilitas generalisasi (*generalization capability*).
 - Data dibagi menjadi dua bagian, yaitu data training dan data testing
 - Kapabilitas generalisasi diukur pada data yang tidak terlibat dalam pembentukan model, yaitu data testing. Kapabilitas generalisasi digunakan sebagai dasar tingkat kepercayaan pada hasil yang diberikan oleh suatu model
 - Selanjutnya, data training dan data testing digabung menjadi data training baru yang akan digunakan untuk membangun model baru untuk penggunaan aktual

23

Aplikasi

Computer Vision



24

Aplikasi

Speech Recognition



25

Aplikasi

Web: Learning to Rank, Recommender System

LEARNING TO RANK CHALLENGE from YAHOO!

Home Datasets Instructions Registration Submission Leaderboard FAQs Workshop

DATASETS

Datasets

The datasets come from web search ranking and are of a subset of what Yahoo! uses to train its ranking function. They consist of features vectors extracted from query-urls pairs along with relevance judgments. The relevance judgments can take 5 different values from 0 (irrelevant) to 4 (perfectly relevant). The queries, urls and features descriptions are not disclosed, only the feature values. There are two datasets for this challenge, each corresponding to a different country: a large one (labeled **set1**) and a small one (labeled **set2**). Both datasets are related, but also different to some extent. Each dataset is divided into 3 sets: training, validation, and test.

The statistics for the various sets are as follows:

	Set 1			Set 2		
	Train	Val	Test	Train	Val	Test
# queries	19,944	2,994	6,983	1,266	1,266	3,798
# urls	473,134	71,083	165,660	34,815	34,881	103,174
# features		519			596	

There are 700 features in total. Some of them are defined in set1 or set2 only, while some others are defined in both sets. When a feature is undefined for a set, its value is 0. All the features have been normalized to be in the [0,1] range.

Relevance labels on the validation and test sets are hidden and replaced by -1.

Format

The format for each of the 6 files is the same as the one used in SVMlight

```
<line> .-<relevance> <qid> <feature><value> ... <feature><value>
<relevance> .-<integer>
<qid> .-<positive integer>
<feature> .-<positive integer>
<value> .-<float>
```

NETFLIX

Netflix Prize

Home Rules Leaderboard Register Update Submit Download

Netflix

Hitman (2007)

You rated this movie: 3.0 stars
Average of 94,456 ratings: 3.6 stars

DETAILS

UR Unrated: This movie has not been rated by the MPAA.
PARENTS: for 16+ (more)

At A Glance Friends Member Reviews Critics

MORE LIKE THIS

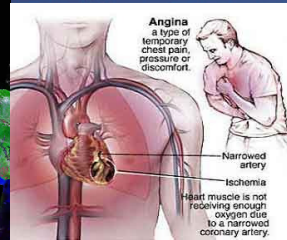
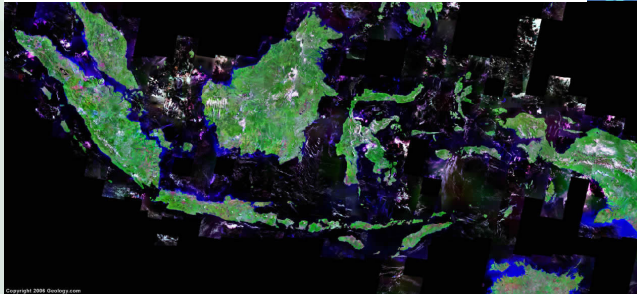
Here are some other movies you might enjoy...

Hitman **House of Eliott** **Coolest Whipper with Cesar Milan Season 2**

26

Aplikasi

Environment, Finance, Diseases, Agriculture



27

Topik

- **Supervised Learning.** Linear Model; Neural Networks; Metode Kernel, Radial Basis Function Network; Support Vector Machine: Classification, Regression, Ranking, Ordinal Regression, Density Estimation
- **Unsupervised Learning.** K-Means; Principal Component Analysis (PCA); Independent Component Analysis (ICA); Latent Semantic Analysis (LSA); Matrix Factorization
- **Semisupervised Learning**
- **Reinforcement Learning**

28

Perangkat Lunak

- WEKA
 - Dikembang dengan Java
 - Dapat digunakan melalui GUI maupun command-line dari Java
 - Open source dapat di unduh di <http://weka.pentaho.com/>
- scikit-learn
 - Dikembangkan dengan Python
 - Dapat digunakan hanya melalui command-line dari Python
 - Open source dapat di unduh di <http://www.scikit-learn.org>

29

Perangkat Lunak

- LIBSVM
 - Dikembang dengan C/C++
 - Dapat digunakan melalui command-line dari DOS, Java, Python, Matlab, R, WEKA, dll
 - Free dapat di unduh di <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
- SVMlight
 - Dikembangkan dengan C/C++
 - Dapat digunakan melalui command-line dari DOS, Java, Python, Matlab, dll
 - Free dapat diunduh di <http://svmlight.joachims.org/>

30

Buku Referensi

- Witten, I. H., et. al., *Data Mining: Practical Machine Learning Tools and Techniques*, Elsevier Inc., 2011
- Bishop, C. H., *Pattern Recognition and Machine Learning*, Springer, 2006
- Schoelkopf and Smola, *Learning with Kernels*, MIT Press, 2002
- C. E. Rasmussen & C. K. I. Williams, *Gaussian Process for Machine Learning*, MIT Press, 2006
- Vapnik, *The Nature of Statistical Learning Theory*, Spinger, 2000
- Jolliffe, *Principal Component Analysis*, Springer, 2000
- Hyvaerinen, Karhunen & Oja, *Independent Component Analysis*, Wiley, 2001
- Landauer, T. K., McNamara, D.S., Dennis, S. and Kintsch W. (Eds), *Handbook of Latent Semantic Analysis*, Mahwah NJ: Lawrence Erlbaum Associates, 2007

31

Grup Riset

- Menjadi topik penelitian pada banyak grup riset, misal:
 - Google (<http://research.google.com>)
 - Microsoft (<http://research.microsoft.com>)
 - Yahoo (<http://research.yahoo.com/>)
 - TU Berlin (<http://www.ml.tu-berlin.de/>,
<http://www.ni.tu-berlin.de/>, <http://www.dai-labor.de/en/>)

32



Konferensi dan Jurnal

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Asian Conference on Machine Learning (ACML)
- Conference on Neural Information Processing System (NIPS)
- Annual Conference on Learning Theory (COLT)
- Journal on Machine Learning Research (www.jmlr.org)
- Machine Learning diterbitkan oleh Springer