

Machine Learning

Steffen Rendle

(Slides mainly by Lars Schmidt-Thieme)

Information Systems and Machine Learning Lab (ISMLL)
Institute for Business Economics and Information Systems
& Institute for Computer Science
University of Hildesheim
<http://www.ismll.uni-hildesheim.de>

1. What is Machine Learning?

2. Overview

3. Organizational stuff

What is Machine Learning?

Computer Science: computers should solve problems that are too hard, time-consuming, simple, etc. for humans.

Machine Learning: computers **learn** by themselves how to solve the problems based on observations.

What is Machine Learning?

1. Information Systems: predict what a customer is interested in based on his actions in the past.

Your Recent History [\(What's this?\)](#)

Recently Viewed Items



[Raging Bull \(Special Edition\)](#)
DVD ~ Robert De Niro



[The Godfather \(Widescreen Edition\)](#) DVD ~ Marlon Brando

Continue shopping Customers Who Bought Items in Your Recent History Also Bought:



[Chinatown \(Special Collector's Edition\)](#) DVD ~ Jack Nicholson



[The Deer Hunter](#) DVD ~ Robert De Niro



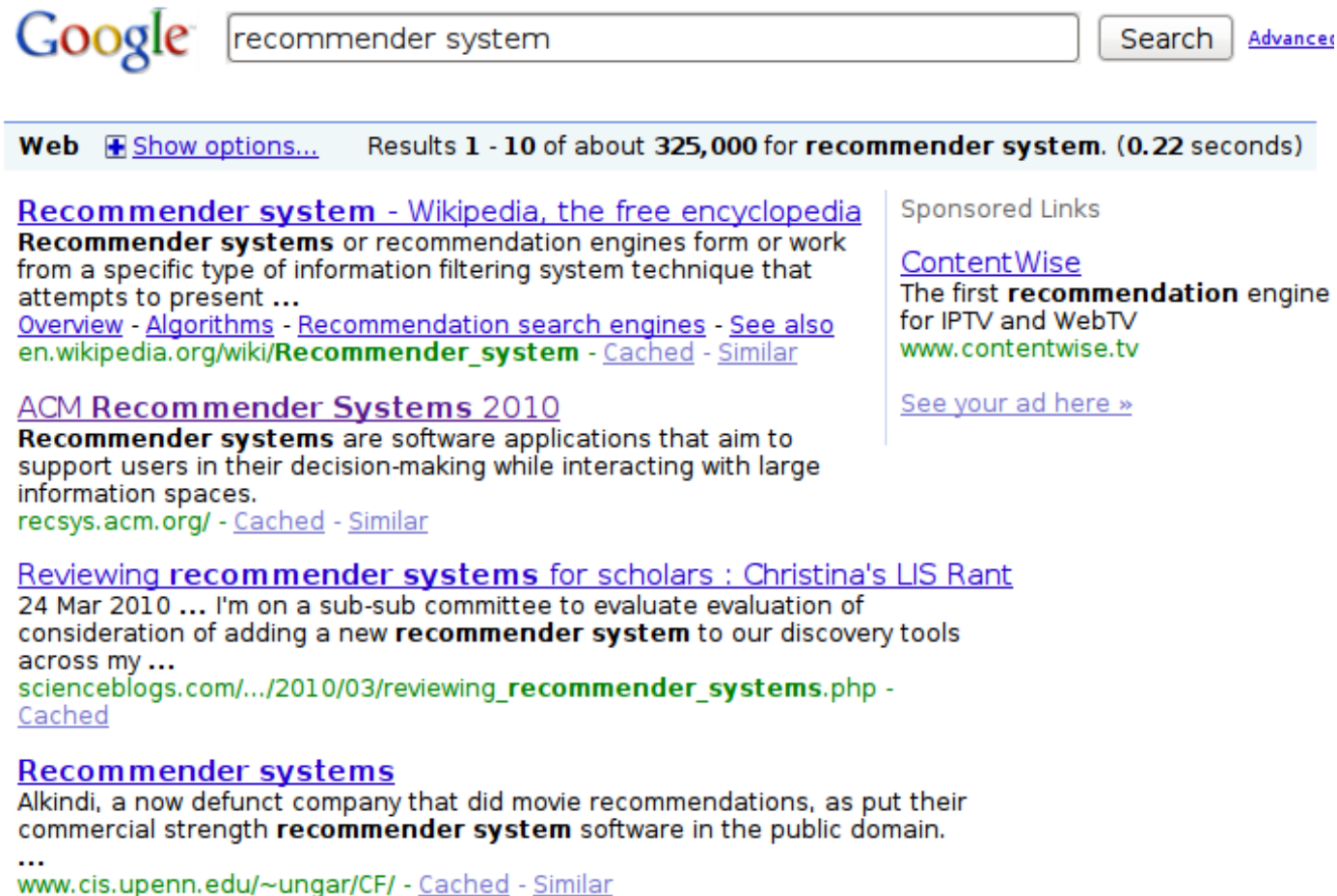
[Mean Streets \(Special Edition\)](#) DVD ~ Julie Andleman



[The Graduate](#) DVD ~ Dustin Hoffman

What is Machine Learning?

1. Information Systems: predict what pages are relevant for a search query based on the clicks in the past.



Google Search [Advanced](#)

Web [+ Show options...](#) Results **1 - 10** of about **325,000** for **recommender system**. (0.22 seconds)

[Recommender system - Wikipedia, the free encyclopedia](#)
Recommender systems or recommendation engines form or work from a specific type of information filtering system technique that attempts to present ...
[Overview](#) - [Algorithms](#) - [Recommendation search engines](#) - [See also](#)
en.wikipedia.org/wiki/Recommender_system - [Cached](#) - [Similar](#)

[ACM Recommender Systems 2010](#)
Recommender systems are software applications that aim to support users in their decision-making while interacting with large information spaces.
recsys.acm.org/ - [Cached](#) - [Similar](#)

[Reviewing recommender systems for scholars : Christina's LIS Rant](#)
24 Mar 2010 ... I'm on a sub-sub committee to evaluate evaluation of consideration of adding a new **recommender system** to our discovery tools across my ...
scienceblogs.com/.../2010/03/reviewing_recommender_systems.php - [Cached](#)

[Recommender systems](#)
Alkindi, a now defunct company that did movie recommendations, as put their commercial strength **recommender system** software in the public domain.
...
www.cis.upenn.edu/~ungar/CF/ - [Cached](#) - [Similar](#)

Sponsored Links
[ContentWise](#)
The first **recommendation** engine for IPTV and WebTV
www.contentwise.tv
[See your ad here »](#)

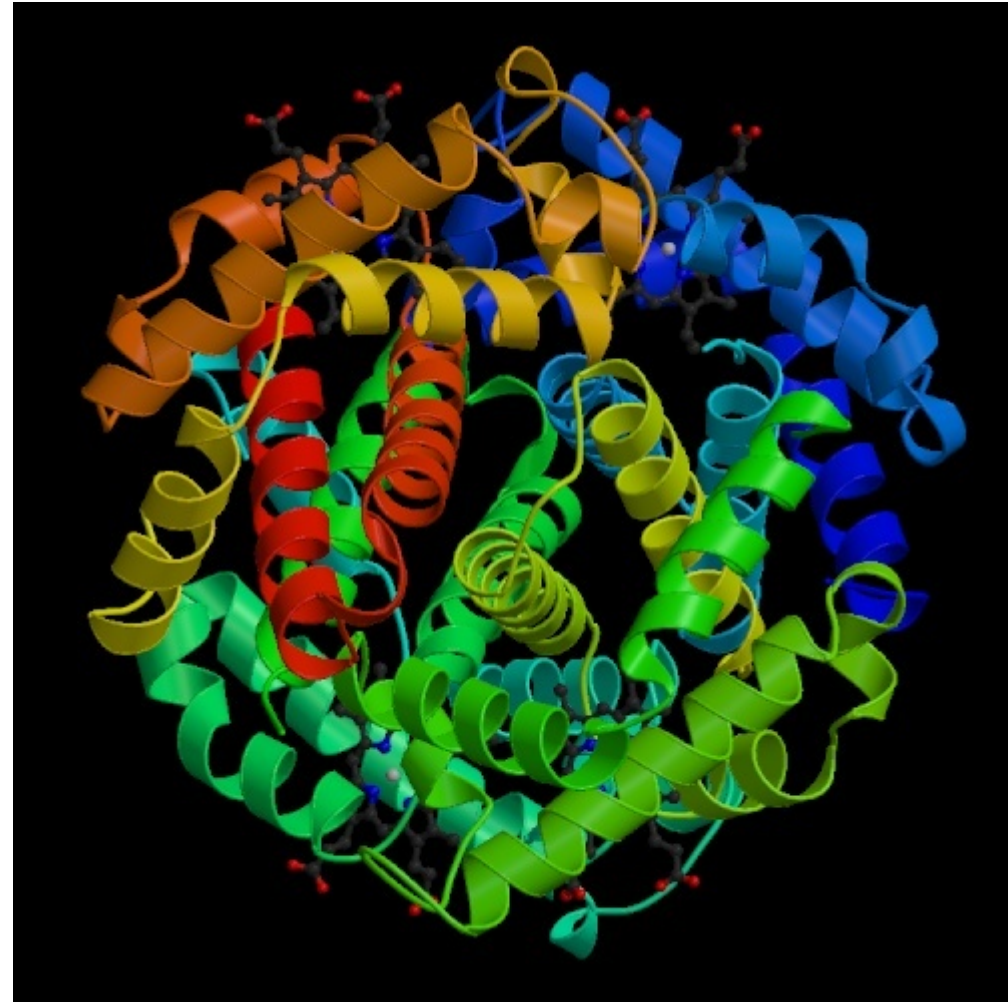
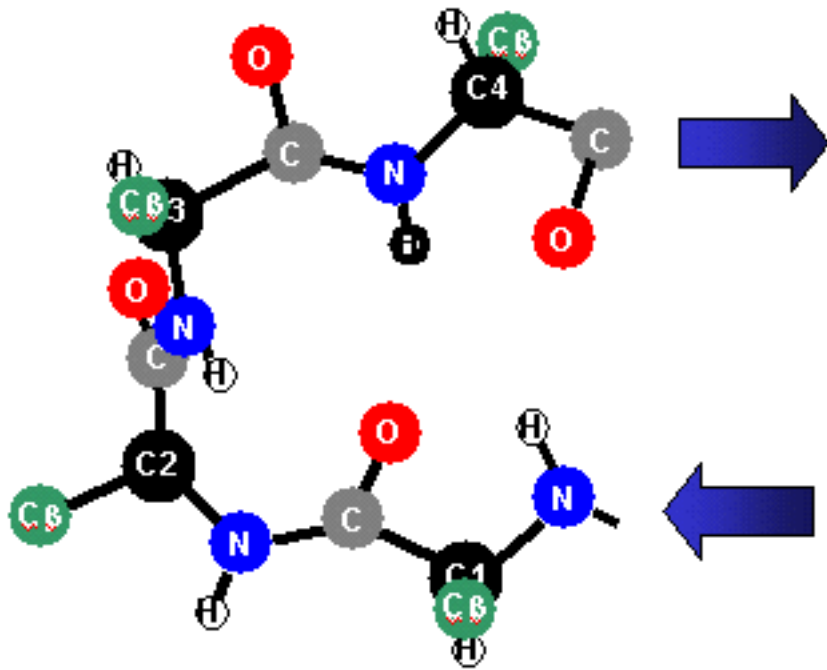
What is Machine Learning?

2. Robotics: Build a map of the environment based on sensor signals.



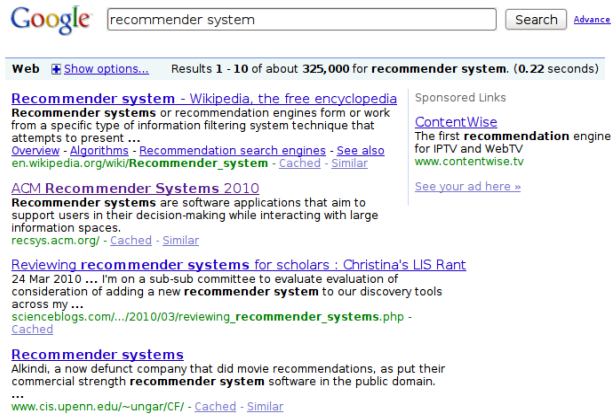
What is Machine Learning?

3. Bioinformatics: predict the 3d structure of a molecule based on its sequence.



What is Machine Learning?

Information Systems

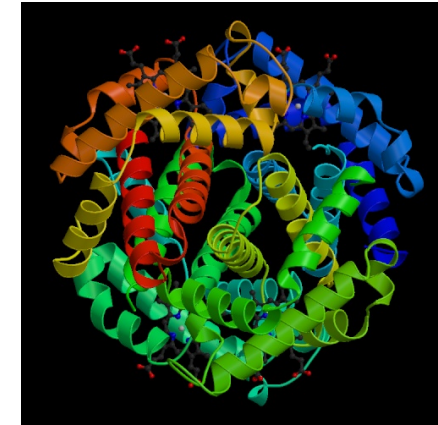


Google Search results for "recommender system". The search bar shows "recommender system" and "Search" with "Advanced" options. The results show approximately 325,000 results. The first result is a Wikipedia entry for "Recommender system". Other results include "ACM Recommender Systems 2010" and "Reviewing recommender systems for scholars".

Robotics



Bioinformatics



...and many more applications.

Machine Learning methods build on:

- Statistics
- Algorithms
- Optimization
- Numerics

One area of research, many names (and aspects)

machine learning

historically, stresses learning logical or rule-based models (vs. probabilistic models).

data mining

stresses the aspect of large datasets and complicated tasks.

knowledge discovery in databases (KDD)

stresses the embedding of machine learning tasks in applications, i.e., preprocessing & deployment; data mining is considered the core process step.

data analysis

historically, stresses multivariate regression methods and many unsupervised tasks.

pattern recognition

name preferred by engineers, stresses cognitive applications such as image and speech analysis.

applied statistics

stresses underlying statistical models, testing and methodical rigor.

1. What is Machine Learning?

2. Overview

3. Organizational stuff

Machine Learning Problems

1. Density Estimation
2. Regression / Supervised Learning
3. Classification / Supervised Learning
4. Optimal Control / Reinforcement Learning
5. Clustering / Unsupervised Learning
6. Dimensionality Reduction
7. Association Analysis

Machine Learning Problems

- 1. Density Estimation
 - 2. Regression
 - 3. Classification
 - 4. Optimal Control
 - 5. Clustering
 - 6. Dimensionality Reduction
 - 7. Association Analysis
- } Supervised Learning
- } Reinforcement Learning
- } Unsupervised Learning

1. Density Estimation

Example 1: duration and waiting times for eruptions of the “Old Faithful” geyser in Yellowstone National Park, Wyoming (Azzalini and Bowman 1990).

continuous measurement from August 1 to August 15, 1985:

- duration (in min.),
- waiting time (in min.)

duration:

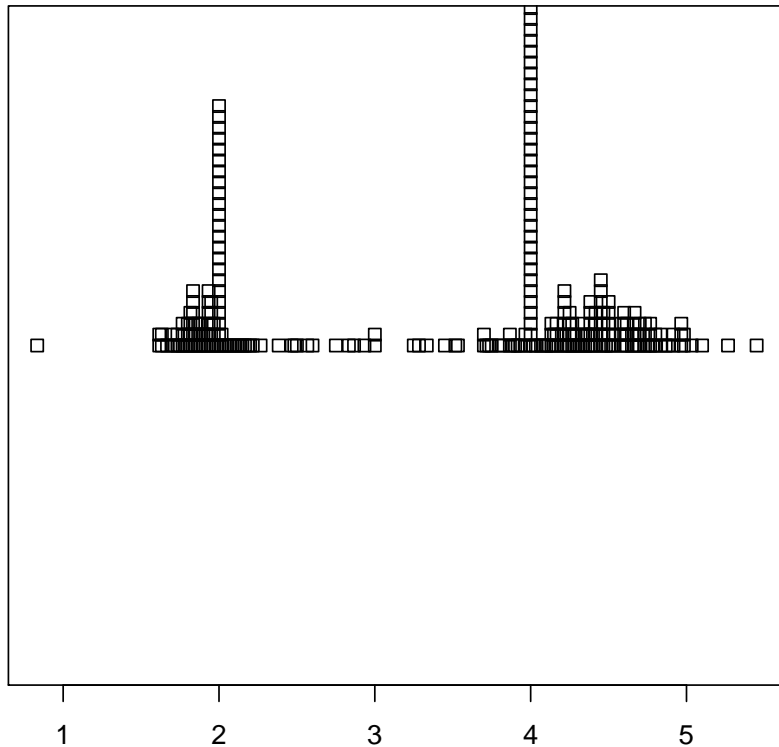
4.016667, 2.15, 4.0, 4.0, 4.0, 2.0,
4.383333, 4.283333, 2.033333,
4.833333, ...

What is a typical duration? waiting time?

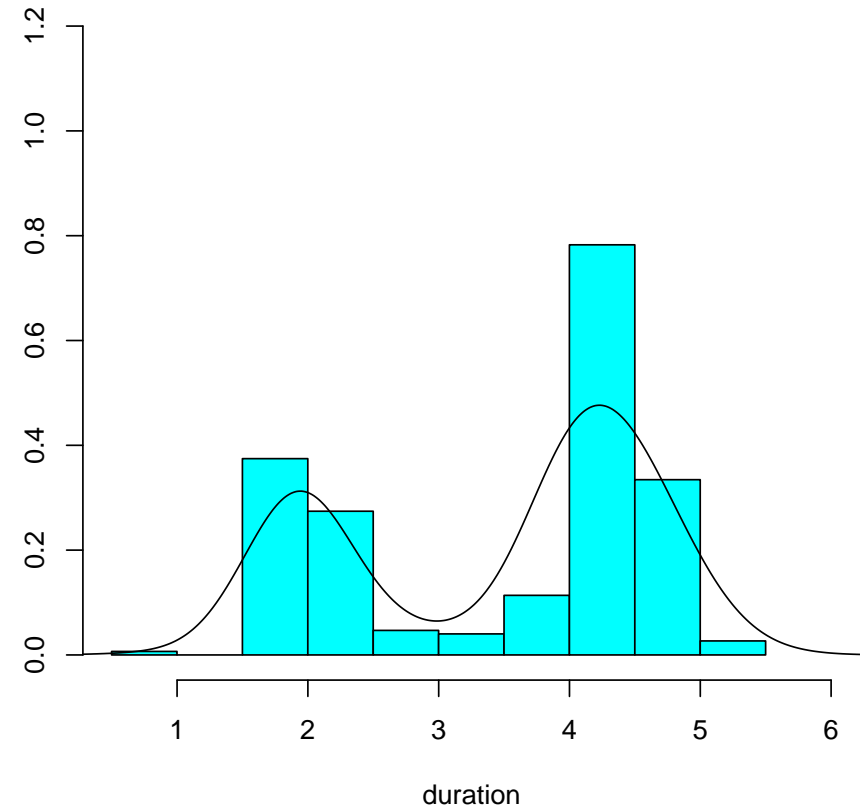


1. Density Estimation

durations: 4.016667, 2.15, 4.0, 4.0, 4.0, 2.0, 4.383333, 4.283333, ...

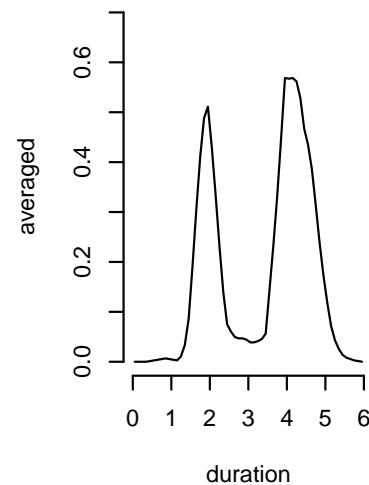
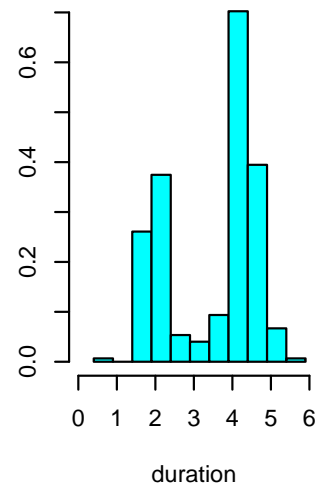
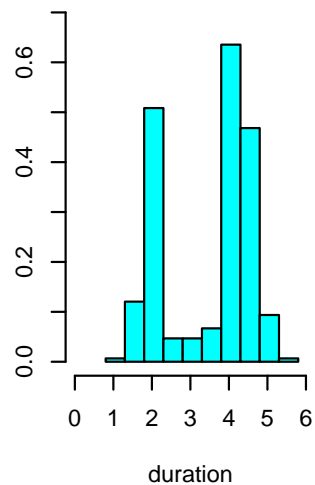
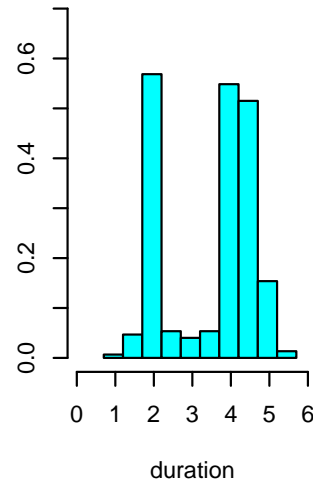
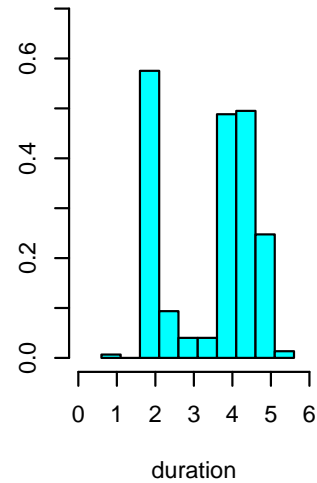
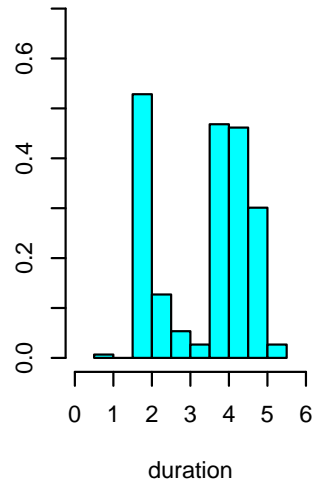


strip chart

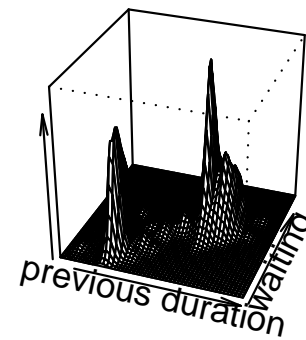
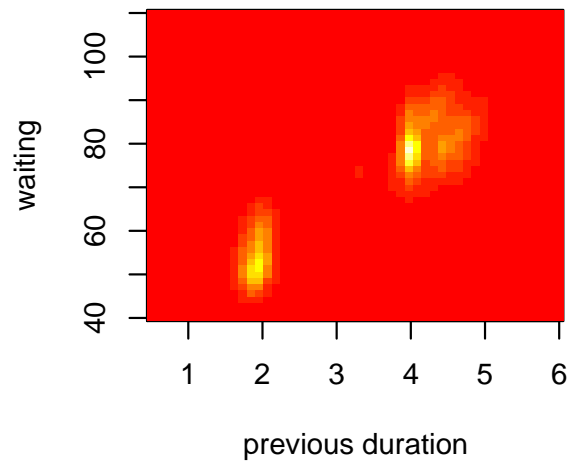
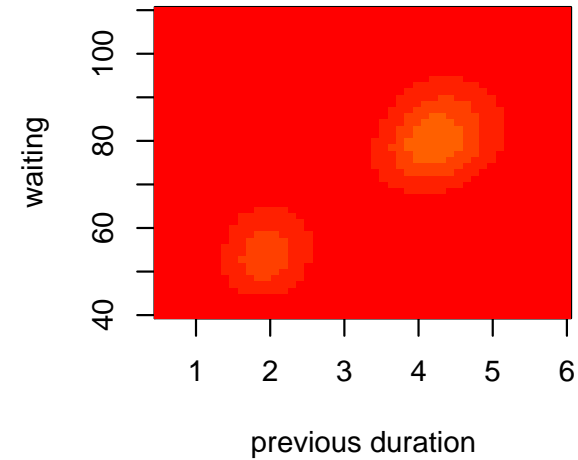
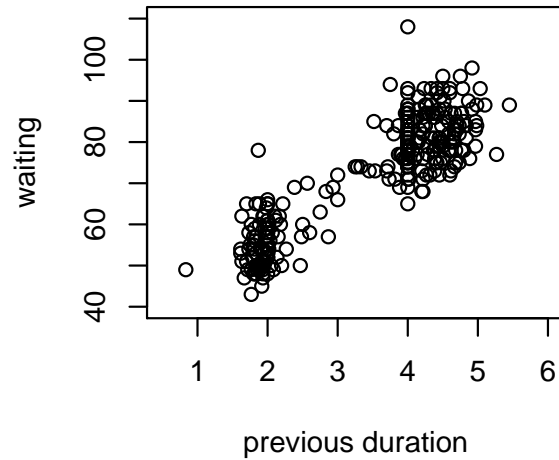


histogram

1. Density Estimation



1. Density Estimation



2. Regression

Example 2: how does gas consumption depend on external temperature?
(Whiteside, 1960s).

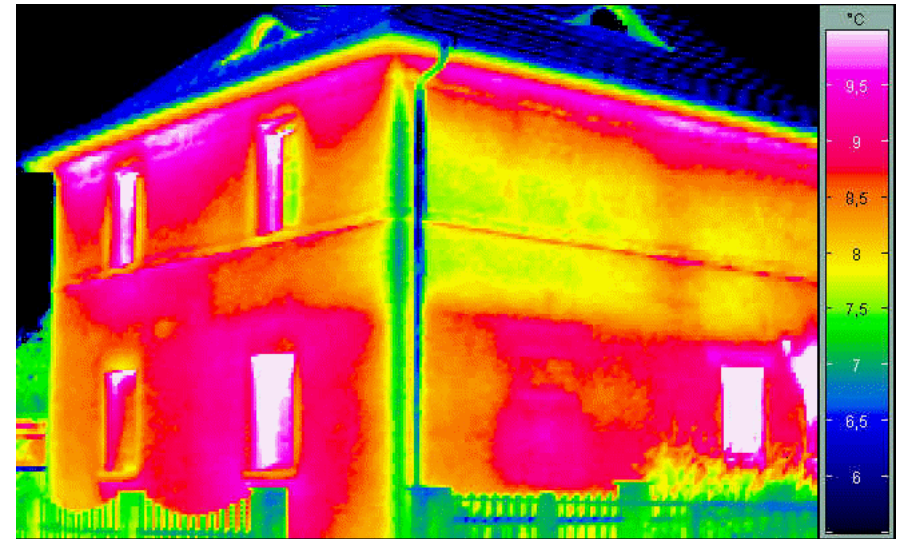
weekly measurements of

- average external temperature
- total gas consumption
(in 1000 cubic feet)

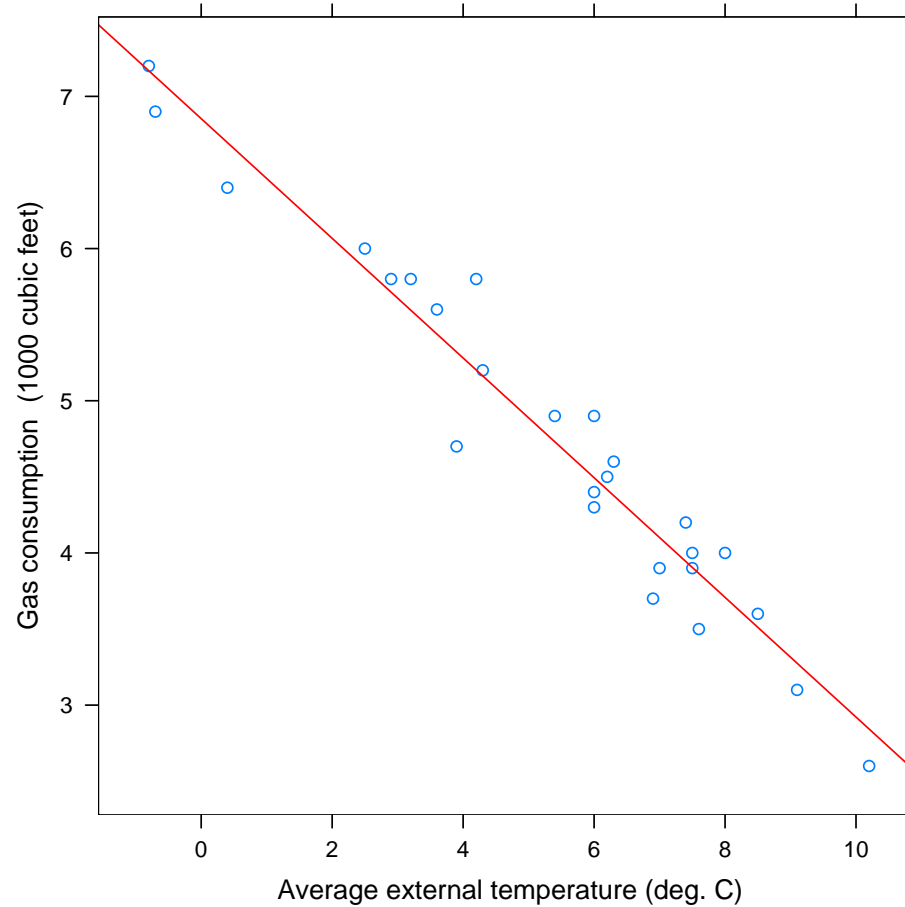
A third variable encodes two heating seasons, before and after wall insulation.

How does gas consumption depend on external temperature?

How much gas is needed for a given temperature ?

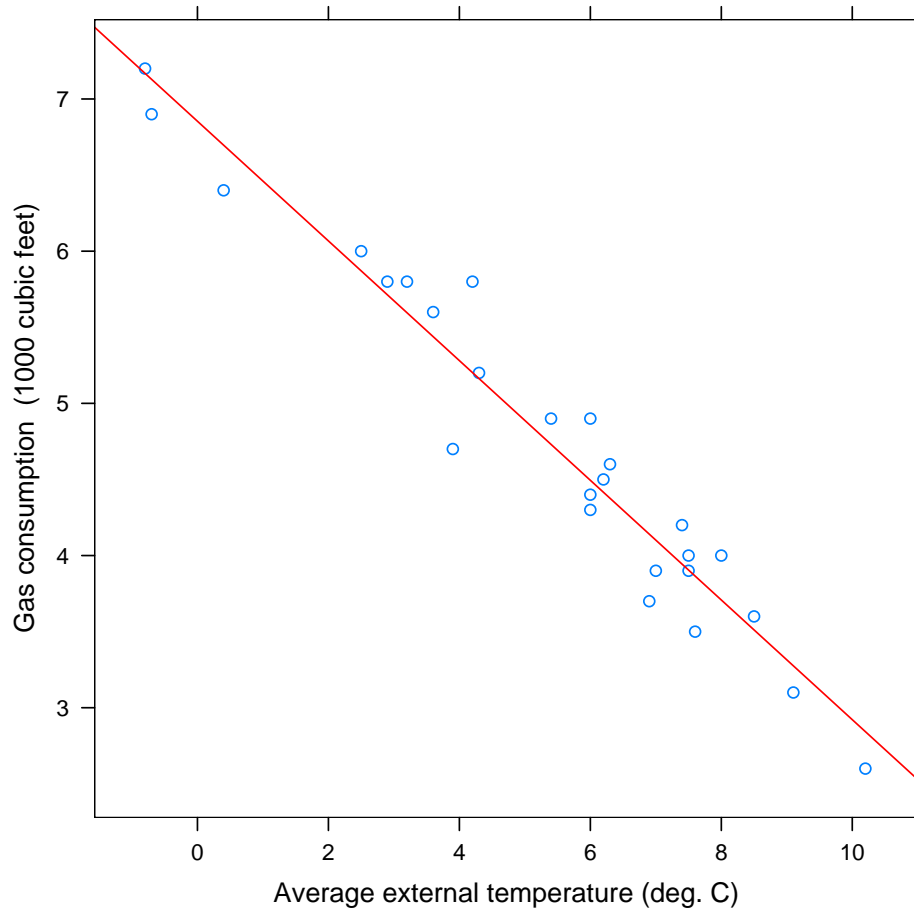


2. Regression

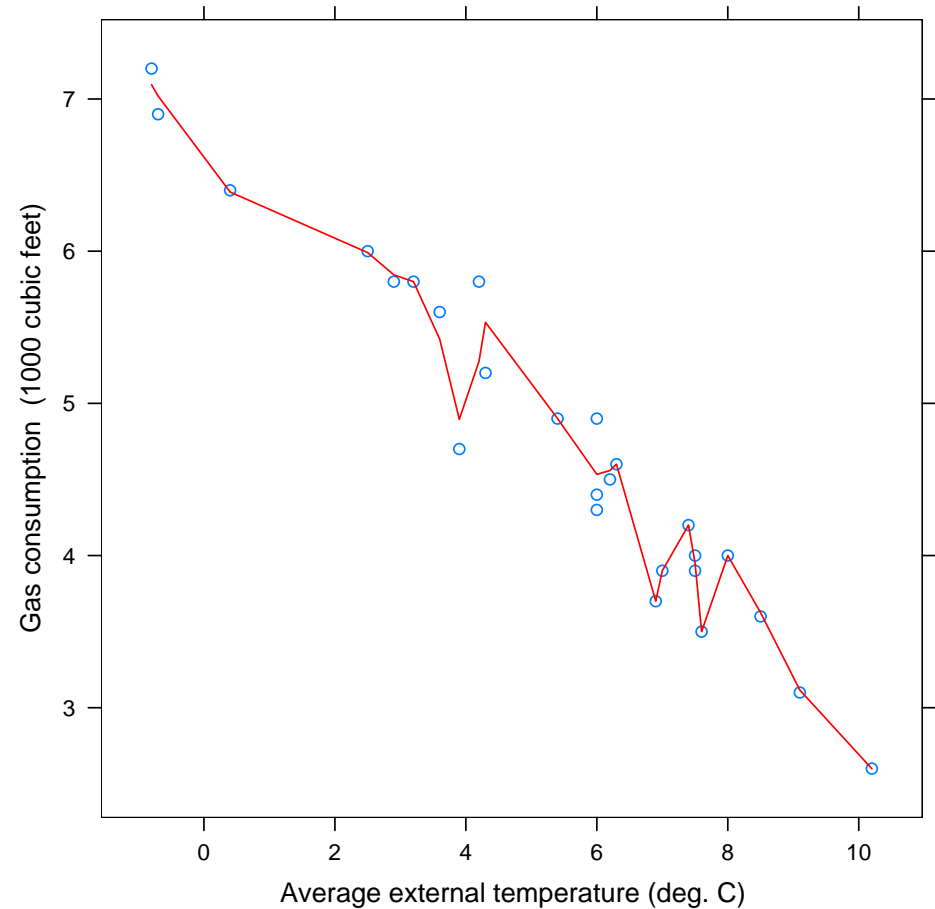


linear model

2. Regression



linear model



more flexible model

3. Classification / Supervised Learning

Example 3: classifying iris plants
(Anderson 1935).

150 iris plants (50 of each species):

- species: setosa, versicolor, virginica
- length and width of sepals (in cm)
- length and width of petals (in cm)



iris setosa



iris versicolor



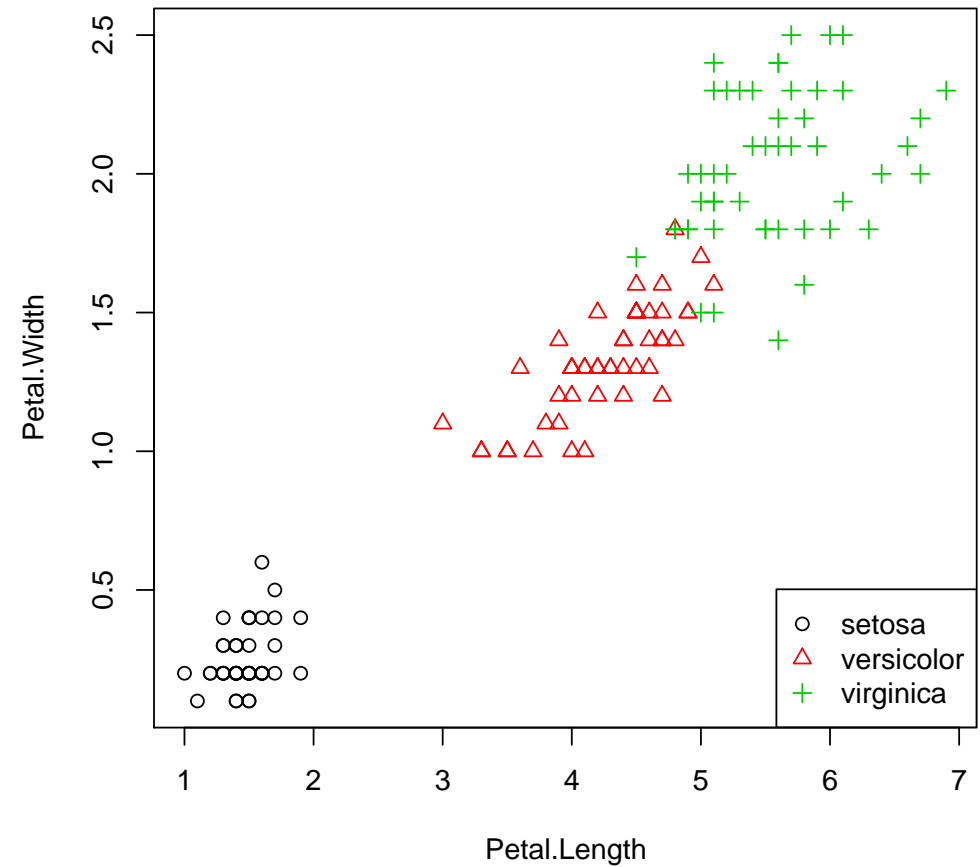
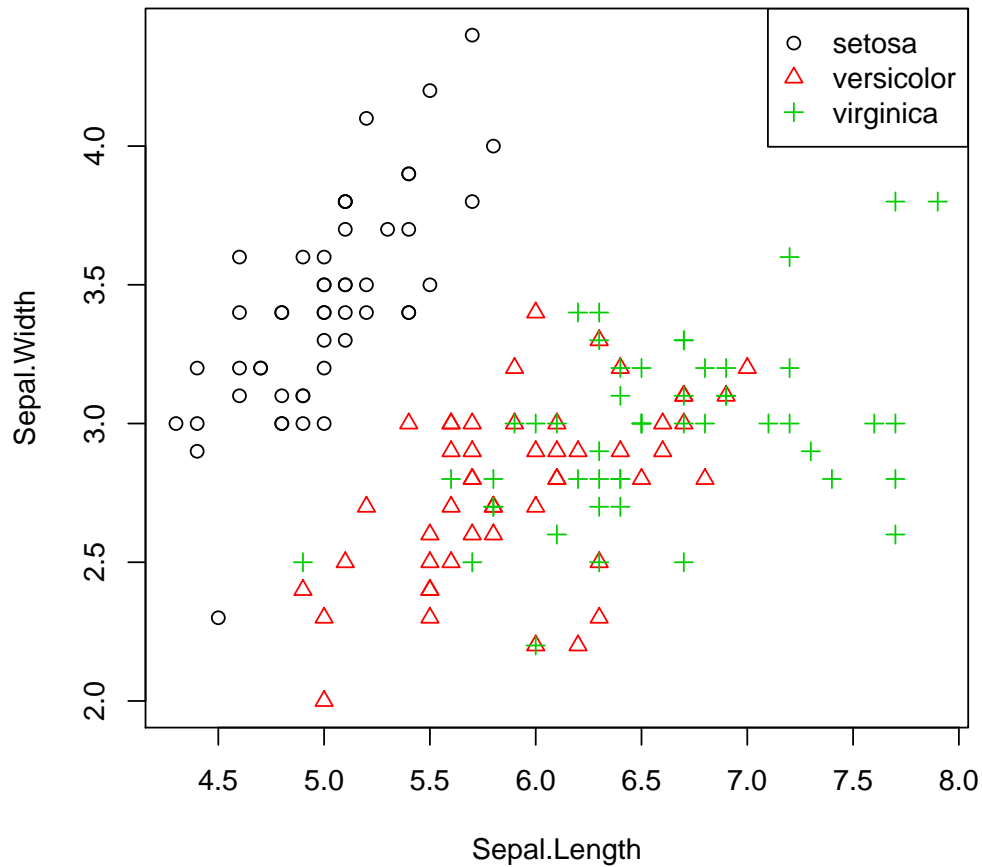
iris virginica

See iris species database
(<http://www.badbear.com/signa>).

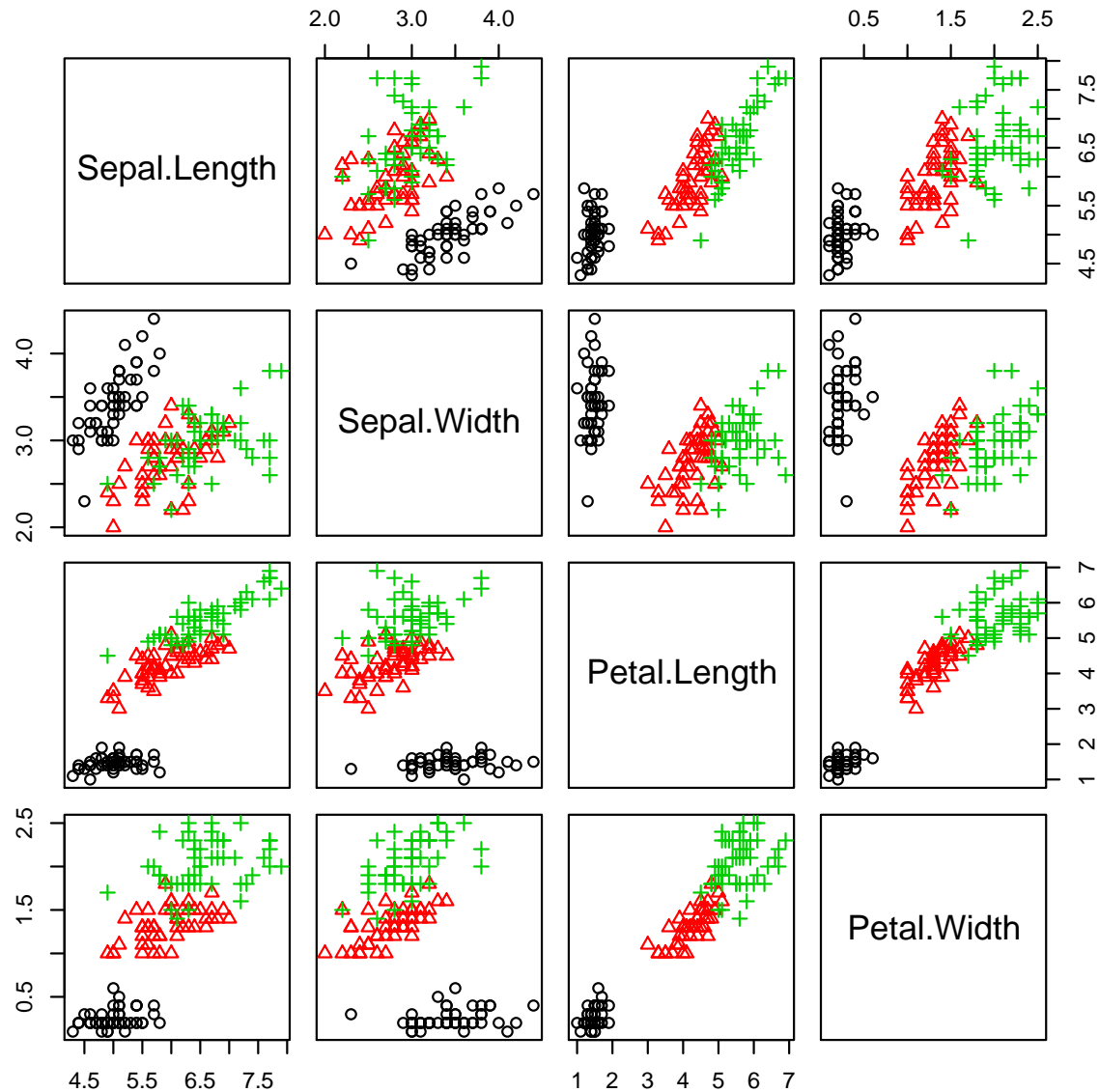
3. Classification / Supervised Learning

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.10	3.50	1.40	0.20	setosa
2	4.90	3.00	1.40	0.20	setosa
3	4.70	3.20	1.30	0.20	setosa
4	4.60	3.10	1.50	0.20	setosa
5	5.00	3.60	1.40	0.20	setosa
⋮	⋮	⋮	⋮	⋮	
51	7.00	3.20	4.70	1.40	versicolor
52	6.40	3.20	4.50	1.50	versicolor
53	6.90	3.10	4.90	1.50	versicolor
54	5.50	2.30	4.00	1.30	versicolor
⋮	⋮	⋮	⋮	⋮	
101	6.30	3.30	6.00	2.50	virginica
102	5.80	2.70	5.10	1.90	virginica
103	7.10	3.00	5.90	2.10	virginica
104	6.30	2.90	5.60	1.80	virginica
105	6.50	3.00	5.80	2.20	virginica
⋮	⋮	⋮	⋮	⋮	
150	5.90	3.00	5.10	1.80	virginica

3. Classification / Supervised Learning



3. Classification / Supervised Learning



3. Classification / Supervised Learning

Example 4: classifying email (lingspam corpus)

Subject: query: melcuk (melchuk)

does anybody know a working email
(or other) address for igor melcuk
(melchuk) ?

legitimate email (“ham”)

Subject: ‘

hello ! come see our naughty little
city made especially for adults
[http://208.26.207.98/freeweek/
enter.html](http://208.26.207.98/freeweek/enter.html) once you get here, you
won’t want to leave !

spam

How to classify email messages as spam or ham?

3. Classification / Supervised Learning

Subject: query: melcuk (melchuk)

does anybody know a working email
(or other) address for igor melcuk
(melchuk) ?

⇒

a	1
address	1
anybody	1
does	1
email	1
for	1
igor	1
know	1
melcuk	2
melchuk	2
or	1
other	1
query	1
working	1

3. Classification / Supervised Learning

lingspam corpus:

- email messages from a linguistics mailing list.
- 2414 ham messages.
- 481 spam messages.
- 54742 different words.
- an example for an early, but very small spam corpus.

3. Classification / Supervised Learning

All words that occur at least in each second spam or ham message on average (counting multiplicities):

	!	your	will	we	all	mail	from	do	our	email
spam	14.18	7.45	4.36	3.42	2.88	2.77	2.69	2.66	2.46	2.24
ham	0.38	0.46	1.93	0.94	0.83	0.79	1.60	0.57	0.30	0.39

	out	report	order	as	free	language	university
spam	2.19	2.14	2.09	2.07	2.04	0.04	0.05
ham	0.34	0.05	0.27	2.38	0.97	2.67	2.61

example rule:

if $\text{freq}("!\") \geq 7$ and $\text{freq}(\text{"language"}) = 0$ and $\text{freq}(\text{"university"}) = 0$ then spam,
else ham

Should we better normalize for message length?

5. Cluster Analysis

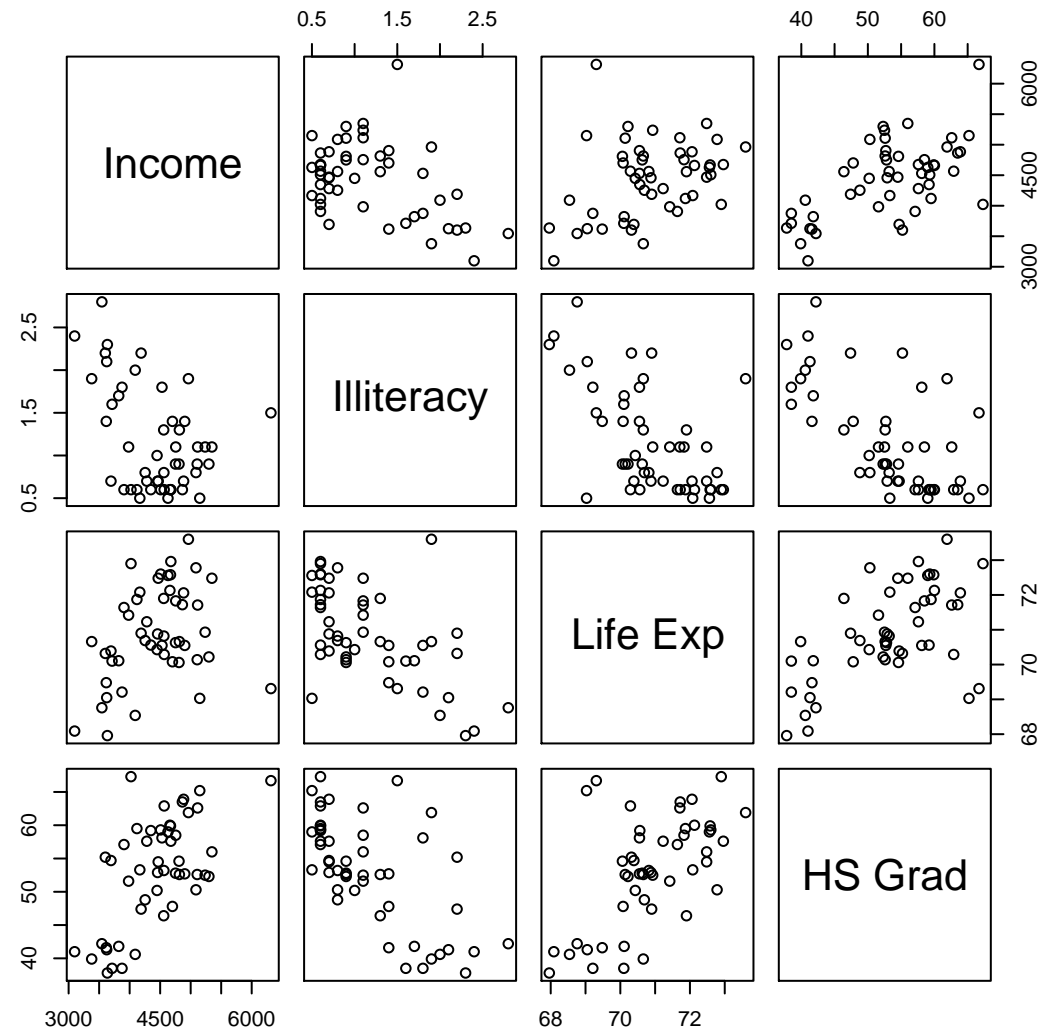
Finding groups of similar objects.

Example 6: sociographic data of the 50 US states in 1977.

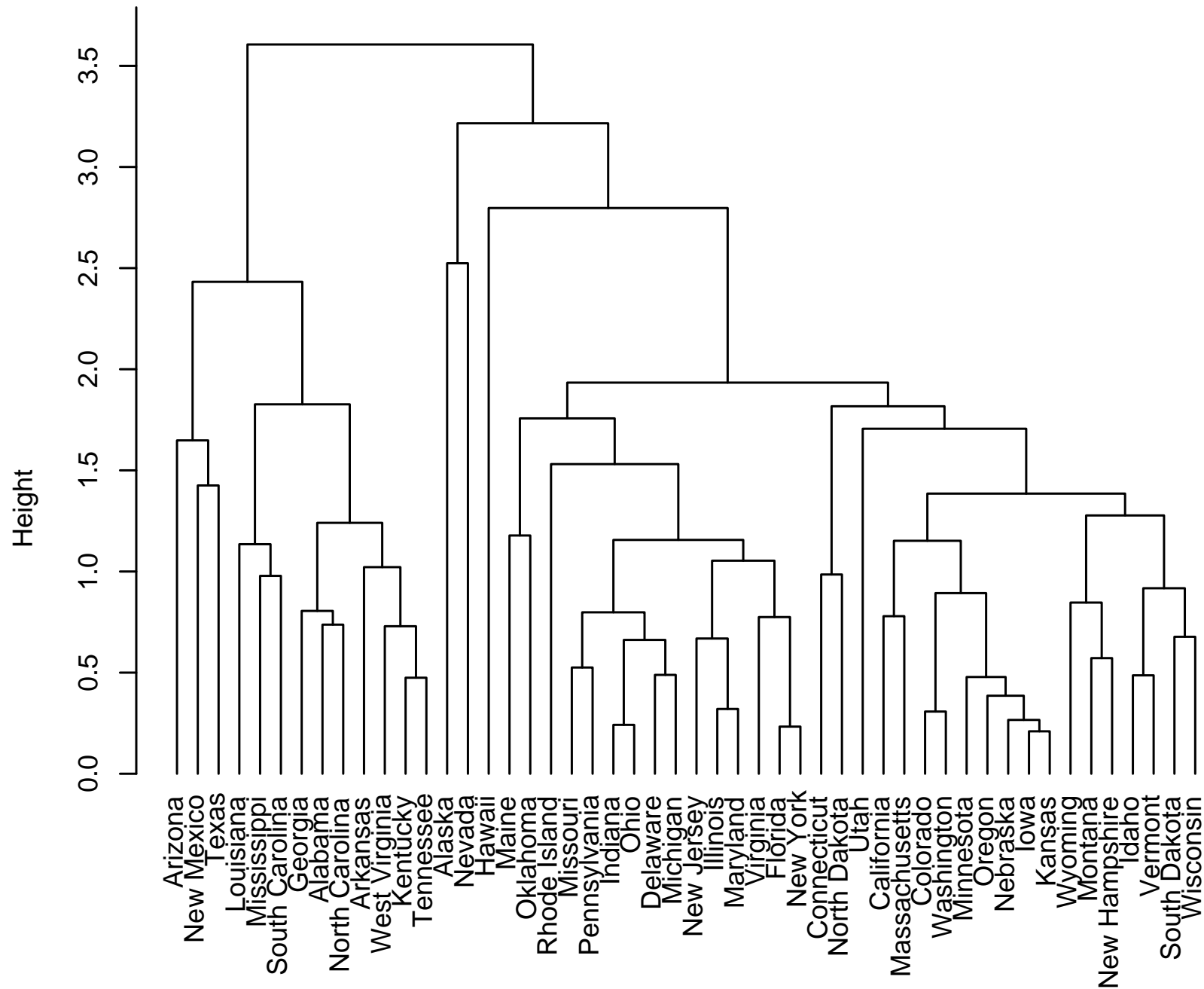
state dataset:

- income (per capita, 1974),
- illiteracy (percent of population, 1970),
- life expectancy (in years, 1969–71),
- percent high-school graduates (1970).

and some others not used here.

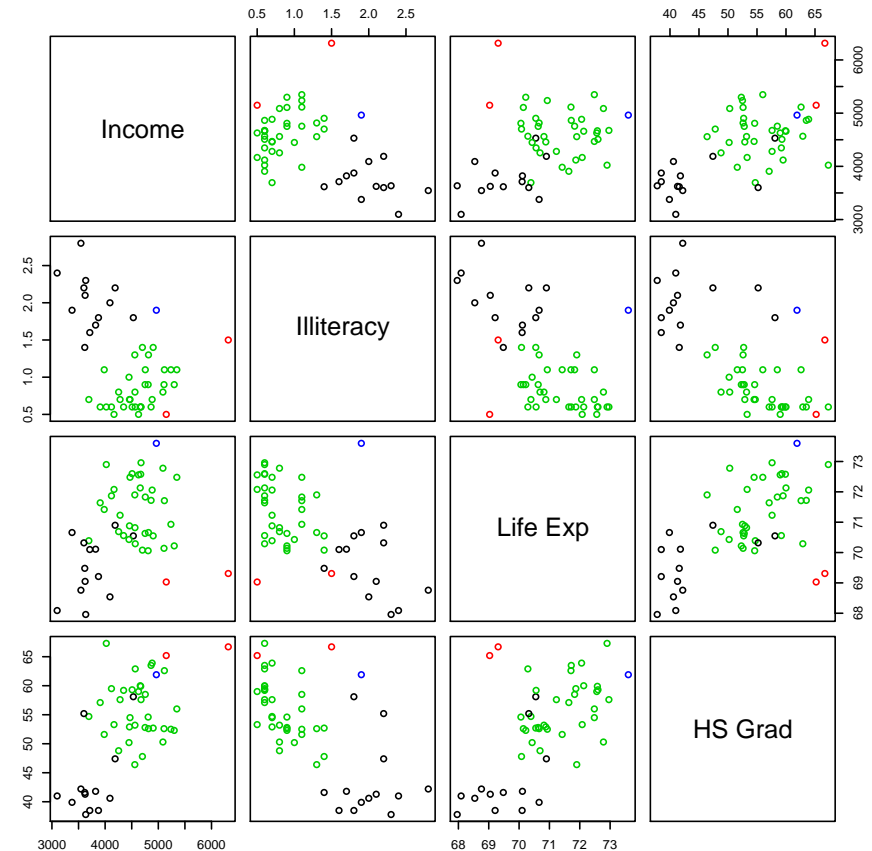
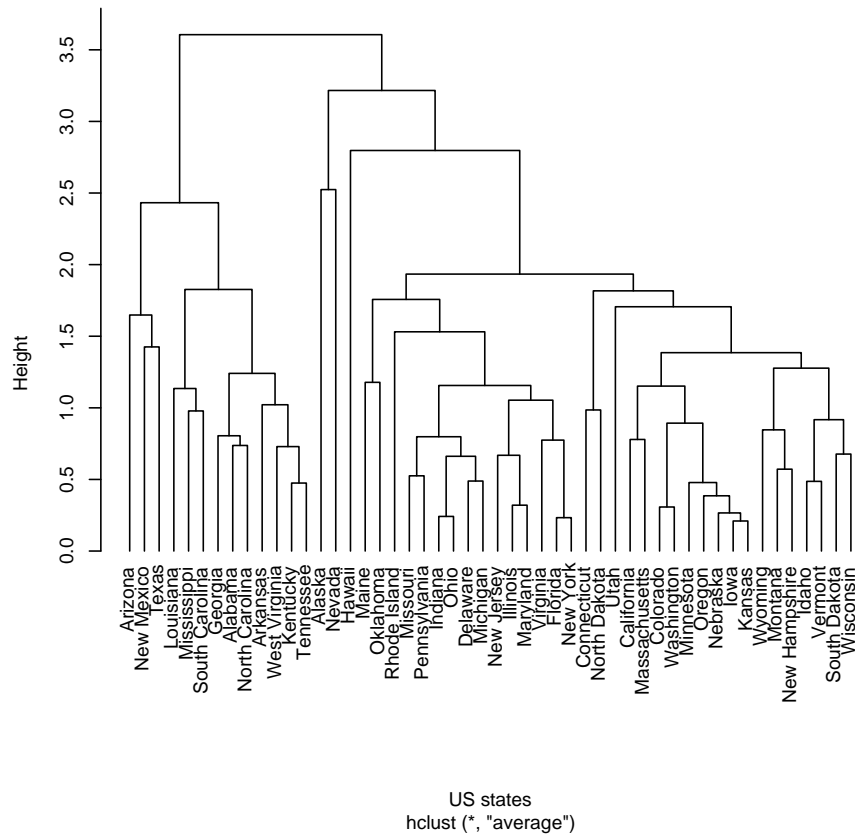


5. Cluster Analysis



5. Cluster Analysis

Cluster Dendrogram



black: Arizona et al., red: Alaska & Nevada, green: California et al., blue: Hawaii.

7. Association Analysis

Association rules in large transaction datasets:

- look for products frequently bought together (**frequent itemsets**).
- look for rules in buying behavior (**association rules**)

Examples:

- {beer, pampers, pizza} (support=0.5)
 {bread, milk} (support=0.5)
- If beer and pampers, then pizza (confidence= 0.75)
 If bread, then milk (confidence=0.75)

cid	beer	bread	icecream	milk	pampers	pizza
1	+	-	-	+	+	+
2	+	+	-	-	+	+
3	+	-	+	-	+	+
4	-	+	-	+	-	+
5	-	+	+	+	-	-
6	+	+	-	+	+	-

1. What is Machine Learning?

2. Overview

3. Organizational stuff

Exercises and tutorials

- There will be a weekly sheet with two exercises available online **each Tuesday/ Wednesday** (after the last lecture of the week).
1st sheet will be online tomorrow morning.
- Solutions to the exercises can be submitted until **next Tuesday/ Wednesday** (before the last lecture of the week)
1st sheet is due Wed. 3.11.
- Exercises will be corrected.
- Tutorials **each Monday 12–14**,
1st tutorial at Mon. 1.11.

Exam and credit points

- There will be a written exam at end of term (2h, 4 problems).
- The course gives 8 ECTS (3+2 SWS).
- The course can be used in
 - Wirtschaftsinformatik MSc / Informatik / Gebiet KI & ML
 - IMIT MSc. (neu) / Informatik / Gebiet KI & ML
 - IMIT MSc. (alt) / IT Machine Learning,
 - IMIT MSc. (alt) / BW Business Intelligence,
 - as well as in any BSc program.

Some books

- Christopher M. Bishop (2007):
Pattern Recognition and Machine Learning, Springer.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman (²2009):
The Elements of Statistical Learning, Springer.
Also available online as PDF at <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>
- Richard O. Duda, Peter E. Hart, David G. Stork (²2001):
Pattern Classification, Springer.

Some First Machine Learning / Data Mining Software

- R (v2.10.0, 26.10.2009; <http://www.r-project.org>).
- Weka (v3.6.1, 5.6.2009; <http://www.cs.waikato.ac.nz/ml/>).
- SAS Enterprise Miner (commercially).

Public data sets:

- UCI Machine Learning Repository
(<http://www.ics.uci.edu/mlearn/>)
- UCI Knowledge Discovery in Databases Archive
(<http://kdd.ics.uci.edu/>)