Advanced R Programming - Lecture 7

Leif Jonsson

Linköping University

leif.jonsson@ericsson.com
leif.r.jonsson@liu.se

October 4, 2016

Today

- Machine Learning
- Supervised learning in R
- Probability in R
- Big data
- Data munging

Questions since last time?

Machine learning?

Automatically detect patterns in data

Machine learning?

Automatically detect patterns in data

Predict future observation

Machine learning?

Automatically detect patterns in data

Predict future observation

Decision making under uncertainty

Types of Machine learning

Supervised learning

Types of Machine learning

Supervised learning

Unsupervised learning

Types of Machine learning

Supervised learning

Unsupervised learning

Reinforcement learning

Supervised learning

(also called predictive learning)

response variable

covariates/features

training set

$$D = (x_i, y_i)_{(i=1)}^{N}$$

Supervised learning types

If y_i is categorical:
classification

If y_i is real:
regression

Unsupervised learning

(also called knowledge discovery)

dimensionality reduction

latent variable modeling

$$D = (x_i)_{(i=1)}^{N}$$

clustering, PCA, discovering of graph structures

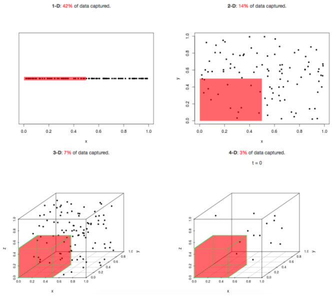
data visualization

Curse of dimensionality

The more variables the larger distance between datapoints

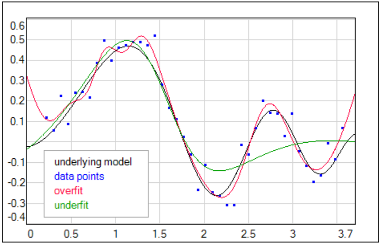
Curse of dimensionality

The more variables the larger distance between datapoints



source

Bias and variance in ML



Underfit = high bias, low variance
Overfit = low bias, high variance

Model selection

bias and variance - tradeoff

Model selection

bias and variance - tradeoff

hyper parameters

Model selection

bias and variance - tradeoff

hyper parameters

generalization error

Model selection

bias and variance - tradeoff

hyper parameters

generalization error

validation set/cross validation

Predictive modeling pipeline

1. Set aside data for test (estimate generalization error)
2. Set aside data for validation (if hyperparams)
3. Run algorithms
4. Find best/optimal hyperparameters (on validation set)
5. Choose final model
6. Estimate generalization error on test set

No free lunch theorem

different models work in different domains

No free lunch theorem

different models work in different domains

accuracy-complexity-intepratability tradeoff

No free lunch theorem

different models work in different domains

accuracy-complexity-intepratability tradeoff

No free lunch theorem

different models work in different domains

accuracy-complexity-intepratability tradeoff

...but more data always wins

the caret package

package for supervised learning

the caret package

package for supervised learning

does not contain methods - a framework

the caret package

package for supervised learning

does not contain methods - a framework

compare methods on hold-out-data

the caret package

package for supervised learning

does not contain methods - a framework

compare methods on hold-out-data

http://topepo.github.io/caret/

the caret package

package for supervised learning

does not contain methods - a framework

compare methods on hold-out-data

http://topepo.github.io/caret/

specific algorithms are part of other courses

the caret package

Prefix	Description	Example
r	Random draw	rnorm
d	Density function	dbinom
q	Quantile function	qbeta
p	CDF	pgamma

Big data

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...
- Dan Ariely

Big data is relative...

... to computational complexity

$$O(N) \quad 10^{12}$$

Big data is relative...

... to computational complexity

$$
\begin{array}{ll}
O(N) & 10^{12} \\
O(N^2) & 10^{6}
\end{array}
$$

Big data is relative...

... to computational complexity

$O(N)$	10^{12}
$O(N^2)$	10^6
$O(N^3)$	10^4

Big data is relative...

... to computational complexity

$O(N)$	10^{12}
$O(N^2)$	10^6
$O(N^3)$	10^4
$O(2^N)$	50

Big data is relative...

... to computational complexity

$$\begin{array}{ll} O(N) & 10^{12} \\ O(N^2) & 10^6 \\ O(N^3) & 10^4 \\ O(2^N) & 50 \end{array}$$

We need algorithms that scale!

Big data is relative...

... to computational complexity

$O(P^2 * N)$ Linear regression

Big data is relative...

... to computational complexity

$O(P^2 * N)$	Linear regression
$O(N^3)$	Gaussian processes

Big data is relative...

... to computational complexity

$O(P^2 * N)$	Linear regression
$O(N^3)$	Gaussian processes
$O(N^2)/O(N^3)$	Support vector machines

Big data is relative...

... to computational complexity

$O(P^2 * N)$	Linear regression
$O(N^3)$	Gaussian processes
$O(N^2)/O(N^3)$	Support vector machines
$O(T(P * N * log(N))$	Random forests

Big data is relative...

... to computational complexity

$O(P^2 * N)$	Linear regression
$O(N^3)$	Gaussian processes
$O(N^2)/O(N^3)$	Support vector machines
$O(T(P * N * log(N))$	Random forests
$O(I * N)$	Topic models

Big data in R

R stores data in RAM

Big data in R

R stores data in RAM

integers
4 bytes
numerics
8 bytes

Big data in R

R stores data in RAM

integers
4 bytes
numerics
8 bytes

A matrix with 100m rows and 5 cols with numerics

$$100000000 * 5 * 8/(1024^3) \approx 3.8$$

How to handle

Handle chunkwise
Subsampling
More hardware
C++/Java backend (dplyr)
Reduce data in memory
Database backend

If not enough

Spark and SparkR

Fast cluster computations for ML /STATS

introduction to Spark

Tidy data

Theoretical approach to data handling

Tidy data and **messy** data

Tidy data

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

Tidy data

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

Examples: `iris` and `faithful`

Why tidy?

80 % of Big Data work is data munging

Why tidy?

80 % of Big Data work is data munging

Analysis and visualization is based on tidy data

Why tidy?

80 % of Big Data work is data munging

Analysis and visualization is based on tidy data

Performant code

Why tidy?

80 % of Big Data work is data munging

Analysis and visualization is based on tidy data

Performant code

Data analysis pipeline

Messy data $\rightarrow$ Tidy data $\rightarrow$ Analysis

Messy data

1. Column headers are values, not variable names. (`AirPassengers`)

Messy data

1. Column headers are values, not variable names. (`AirPassengers`)
2. Multiple variables are stored in one column. (`mtcars`)

Messy data

1. Column headers are values, not variable names. (`AirPassengers`)
2. Multiple variables are stored in one column. (`mtcars`)
3. Variables are stored in both rows and columns. (`crimetab`)

Messy data

1. Column headers are values, not variable names. (`AirPassengers`)
2. Multiple variables are stored in one column. (`mtcars`)
3. Variables are stored in both rows and columns. (`crimetab`)
4. Multiple types of observational units are stored in the same table.

Messy data

1. Column headers are values, not variable names. (`AirPassengers`)
2. Multiple variables are stored in one column. (`mtcars`)
3. Variables are stored in both rows and columns. (`crimetab`)
4. Multiple types of observational units are stored in the same table.
5. A single observational unit is stored in multiple tables.

dplyr

Verbs for handling data

Highly optimized C++ code (backend)

Handling larger datasets in R
(no copy-on-modify)

dplyr+tidyr

the cheatsheet

The End... for today.
Questions?
See you next time!