Practical Subgroup Discovery

Janis Kalofolias

based on work by Mario Boley

max planck institut
informatik

UNIVERSITÄT
DES
SAARLANDES

CISPA
HELMHOLTZ CENTER FOR
INFORMATION SECURITY

Algorithmic Data Analysis

Unsupervised Learning

Correlation

Machine Learning

Predictive Modelling

Big Data

Subgroup Discovery

Exploratory Data Analysis

Classification

Deep Learning

Regression

Data Mining

Descriptive Modelling

Statistical Learning Theory

Supervised Learning

Statistics

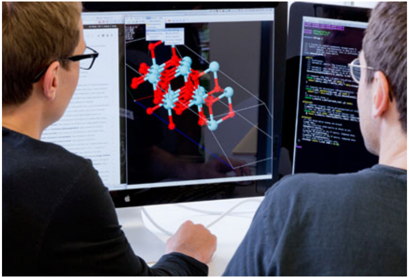
Neural Networks

Local Pattern Discovery

Ridge Regression

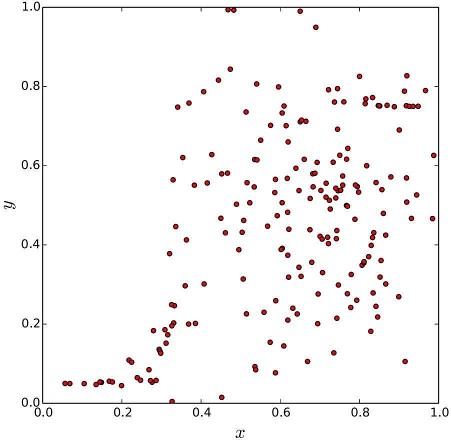
Two flavors of data science

Predictive modelling



Exploratory data analysis

Global versus local modelling

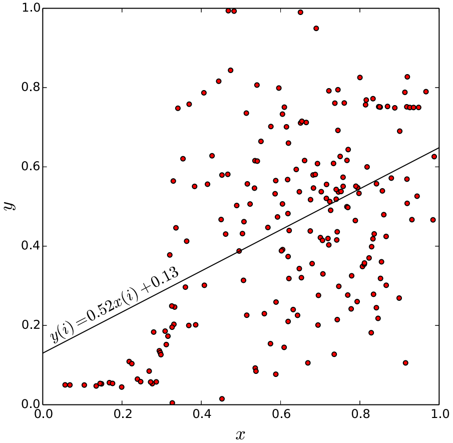


Given
Population $P = \{1, \ldots, n\}$
Target variable $y\colon P \to \mathbb{R}$
Description variable $x\colon P \to \mathbb{R}$

What can we tell about y?

(in terms of x)

Global versus local modelling



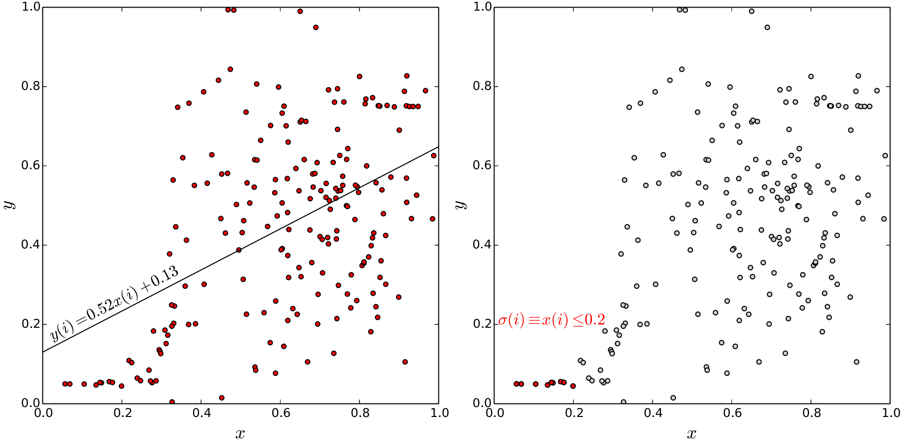
Given
Population $P = \{1, \dots, n\}$
Target variable $y: P \to \mathbb{R}$
Description variable $x: P \to \mathbb{R}$

Find
Coefficients $\alpha, \beta \in \mathbb{R}$ such that objective function

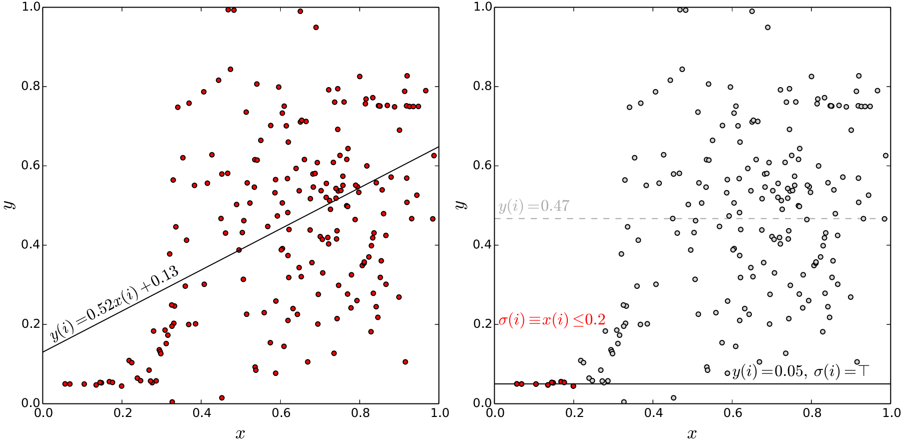
$$f(\alpha, \beta) = \frac{1}{n} \sum_{i \in P} (\alpha x(i) + \beta - y(i))^2 + \lambda \|\alpha, \beta\|_1$$

is minimal

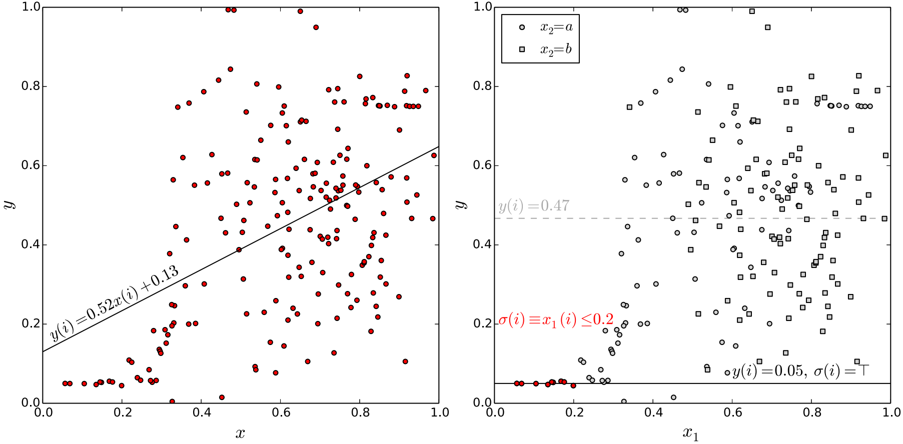
Global versus local modelling



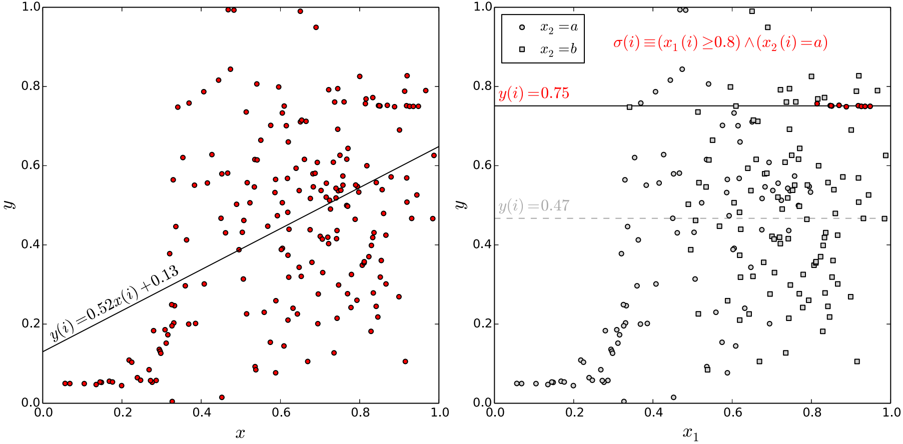
Global versus local modelling



Global versus local modelling



Global versus local modelling



Practical example

Background

Fraction of non-voters increased in German federal election 2009 to 0.28 (from 0.21 in 2005)

Question

“Where” did increase come from?

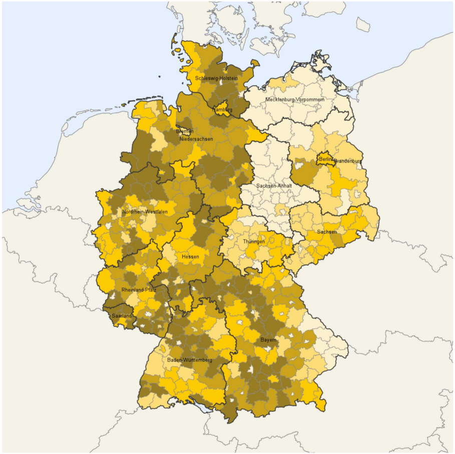
Data

Population: `admin. districts of Germany`

Target variable: `non-voter diff. 2005-2009`

Description attributes:

- Geographical (`region, state`)
- Demographic (`pop. density, highsch. degrees,` …)
- Economic (`GDP growth, web domains,`…)



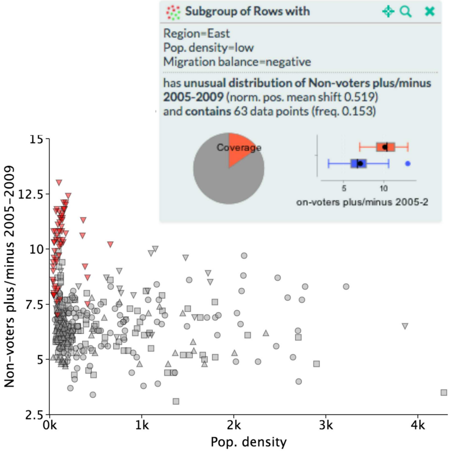
Practical example

Background
Fraction of non-voters increased in German federal election 2009 to 0.28 (from 0.21 in 2005)

Question
“Where” did increase come from?

Data
Population: `admin. districts of Germany`
Target variable: `non-voter diff. 2005-2009`
Description attributes:

- Geographical (`region, state`)
- Demographic (`pop. density, highsch. degrees`, ...)
- Economic (`GDP growth, web domains`,...)



Subgroup of Rows with

Region=East
Pop. density=low
Migration balance=negative

has unusual distribution of Non-voters plus/minus 2005-2009 (norm. pos. mean shift 0.519)
and contains 63 data points (freq. 0.153)

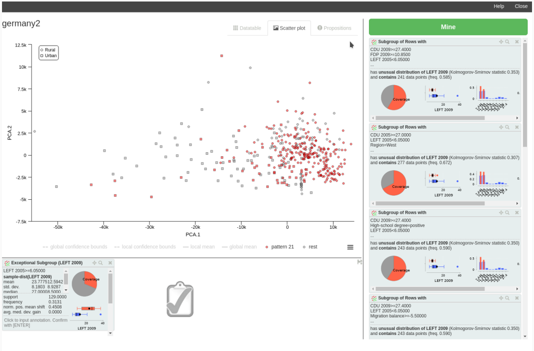
Tools and software

- VIKAMINE

  More for computer scientists: (communities, simpler patterns)
- RealKD/Creedo `https://bitbucket.org/realKD`

  Multiple algorithms, several objectives

  From a material scientist for material scientists



Exceptional Subgroup (LEFT 2009)

LEFT 2005>=6.05000

sample-dist(LEFT 2009)

```
mean       23.775512.5942
std. dev.  8.1803  8.9287
median     27.00008.5000
support               129.0000
frequency             0.3131
norm. pos. mean shift 0.4508
avg. med. dev. gain   0.0000
```

Click to input annotation. Confirm with [ENTER]

Subgroup of Rows with

CDU 2009>=27.4000
FDP 2009>=10.8500
LEFT 2005<6.05000
...

has **unusual distribution of LEFT 2009** (Kolmogorov-Smirnov statistic 0.353) and **contains** 241 data points (freq. 0.585)

Subgroup of Rows with

CDU 2005>=27.0000
LEFT 2005<6.05000
Region=West
...

has **unusual distribution of LEFT 2009** (Kolmogorov-Smirnov statistic 0.307) and **contains** 277 data points (freq. 0.672)

Subgroup of Rows with

CDU 2009>=27.4000
High-school degree=positive
LEFT 2005<6.05000
...

has **unusual distribution of LEFT 2009** (Kolmogorov-Smirnov statistic 0.350) and **contains** 243 data points (freq. 0.590)

Subgroup of Rows with

CDU 2009>=27.4000
LEFT 2005<6.05000
Migration balance>=-5.50000
...

has **unusual distribution of LEFT 2009** (Kolmogorov-Smirnov statistic 0.350) and **contains** 243 data points (freq. 0.590)

Getting RealKD/Creedo

A docker image is available

- Getting the image
  Option a) Build:

  ```
  docker build -t kalofoli/creedo-deps \
      git@github.com:kalofoli/docker.git#:creedo-deps
  docker build -t kalofoli/creedo \
      git@github.com:kalofoli/docker.git#:creedo
  ```

  Option b) Download: ($\approx$ 1.3GB)

  ```
  docker create -t kalofoli/creedo
  ```

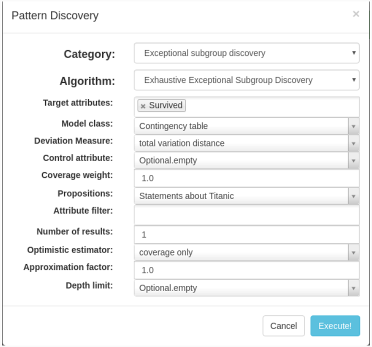
- Running docker

  ```
  docker run -it -p 8080:8080/tcp  kalofoli/creedo
  ```

Subgroup Discovery with Creedo

- Open browser at: `http://localhost:8080/Creedo`
- login with User: default and empty password
- (Optional) Upload xarf data file
- Click Analyze

Settings



Pattern Discovery

Category: Exceptional subgroup discovery

Algorithm: Exhaustive Exceptional Subgroup Discovery

Target attributes: × Survived

Model class: Contingency table

Deviation Measure: total variation distance

Control attribute: Optional.empty

Coverage weight: 1.0

Propositions: Statements about Titanic

Attribute filter:

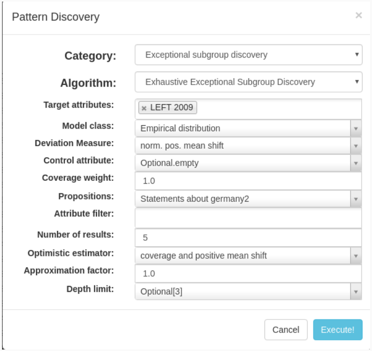
Number of results: 1

Optimistic estimator: coverage only

Approximation factor: 1.0

Depth limit: Optional.empty

Titanic



Pattern Discovery

Category: Exceptional subgroup discovery

Algorithm: Exhaustive Exceptional Subgroup Discovery

Target attributes: × LEFT 2009

Model class: Empirical distribution

Deviation Measure: norm. pos. mean shift

Control attribute: Optional.empty

Coverage weight: 1.0

Propositions: Statements about germany2

Attribute filter:

Number of results: 5

Optimistic estimator: coverage and positive mean shift

Approximation factor: 1.0

Depth limit: Optional[3]

Germany

Automating Experiments

Using RealKD Job descriptions

```
  {
  "type" : "productWorkScheme",
  "id" : "octet_binaries_fdd",
  "workspaces" : [ {
    "type" : "workspaceFromXarf",
    "id" : "binaries",
    "datafile" : "octet_binaries_2.1.1.xarf"
  } ],
  "computations" : [ {
    "type" : "functionalDependencyDiscovery",
    "id" : "titanic_functional_pattern_discovery",
    "target" : "sign_delta_e",
    "num_res" : 3,
    "alpha" : 1.0
  } ],
  "computationTimeLimit" : 3600
}
```