Data Cleaning Tutorial: Error Localization

Mark van der Loo

https://github.com/markvanderloo/UFPEL2019

Try the code

```
03valid/errorlocalization.R
```

Error localization

Error localization is a procedure that points out fields in a data set that can be altered or imputed in such a way that all validation rules can be satisfied.

Example

Ruleset

```
if (married == TRUE ) age >= 16
if (attends == "kindergarten") age <= 6
```

Data

age	married	attends
3	TRUE	kindergarten

Question

Which field or fields would you change?

Principle of Fellegi and Holt

Find the minimal (weighted) number of fields to adjust such that all rules, including implied rules, can be satisfied.

IP Fellegi and D Holt, JASA **71** 353 17–35 (1976).

Note

This should be used as a last resort, when no further information on the location of errors is available.

Implied rules?

```
turnover - total.cost == profit
               profit <= 0.6 * turnover
```

This implies (substituting profit):

```
           total.cost >= 0.4 * turnover
```

We need to take into account such *essentially new* rules: a rule set forms a system of rules and its implied rules. `errorlocate` takes this into account

Choosing weights

All weights equal (usually to one)

Least nr of variables adapted. In case of multiple solutions: choose randomly (e.g. by adding a small random perturbation to the weights).

Weights represent reliability

Heigher weight → variable is less likely chosen.

- Can be made to depend on 'outlierness', or expert judgement.
- Possible problem: minimal weights vs minimal nr of variables?

errorlocate

`errorlocate` formulates a Mixed Integer Problem with:

- `validate` rules set R as a hard constraints
- objective function: minimize

$$f(x_0, \delta) = \sum_i w_i \delta_i$$

  with $\delta_i \in \{0, 1\}$ and $\delta_i = 1$ if field i is an invalid value.
- Penalize the number of fields

locate_errors and replace_errors

Find the errors:

```
library(errorlocate)
errors <- locate_errors(data, rules)
```

Set the fields to NA:

```
data_errors_to_na <- replace_errors(data, rules)
```

Assignment

```r
# we first create a named weight vector with weight 1
weight <- rep(1, ncol(data_with_errors))
names(weight) <- names(data_with_errors)
```

- Set the weight of turnover to 10 and supply the weight to locate_errors
- Discuss the effect of setting te weight on turnover with your neighbor.
- Replace errors with `NA` using the `replace_errors` with the weights used above
- Store the results in "my_errors_located.csv".