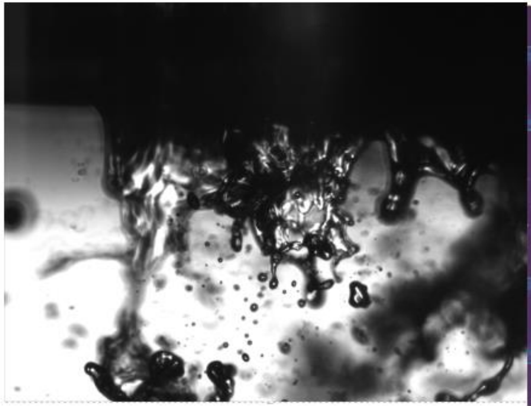
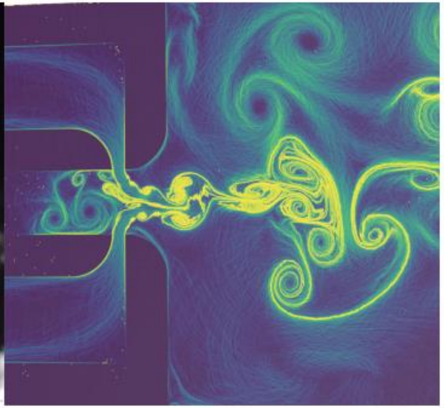
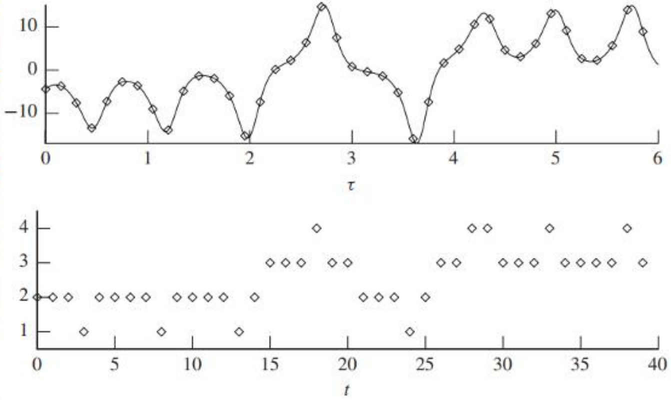
Data Driven Engineering II: Advaced Topics

Feature Engineering I

Institute of Thermal Turbomachinery
Prof. Dr.-Ing. Hans-Jörg Bauer





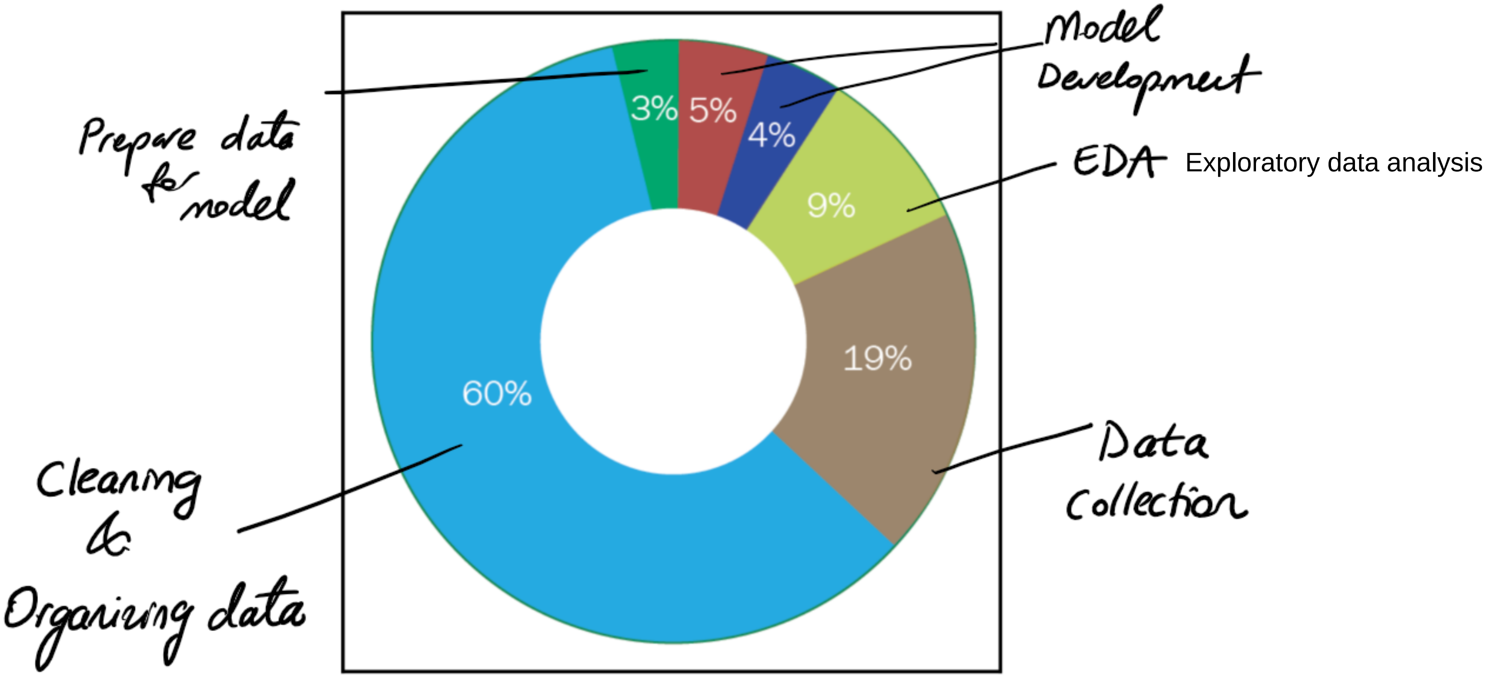


To Do List

- [ ] Check & think about projects
- [ ] Register via Ilias ⟹ latest 13th May
- [ ] Check dataset & materials

! Involve in projects ⟹ Register for HPC Access (ILIAS)

How do you spend your time?



Today's Agenda:

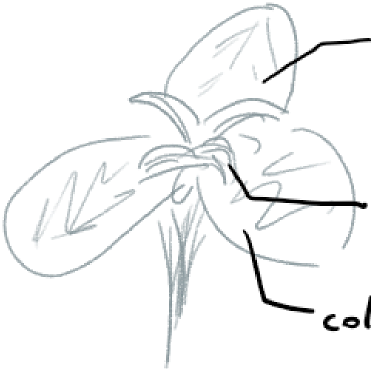
1) Feature Engineering (FE)

2) Data Preparation $\Leftrightarrow$ FE

3) Continuous $X \Rightarrow y$

Introduction to features

- Feature := attributes defining properties of objects

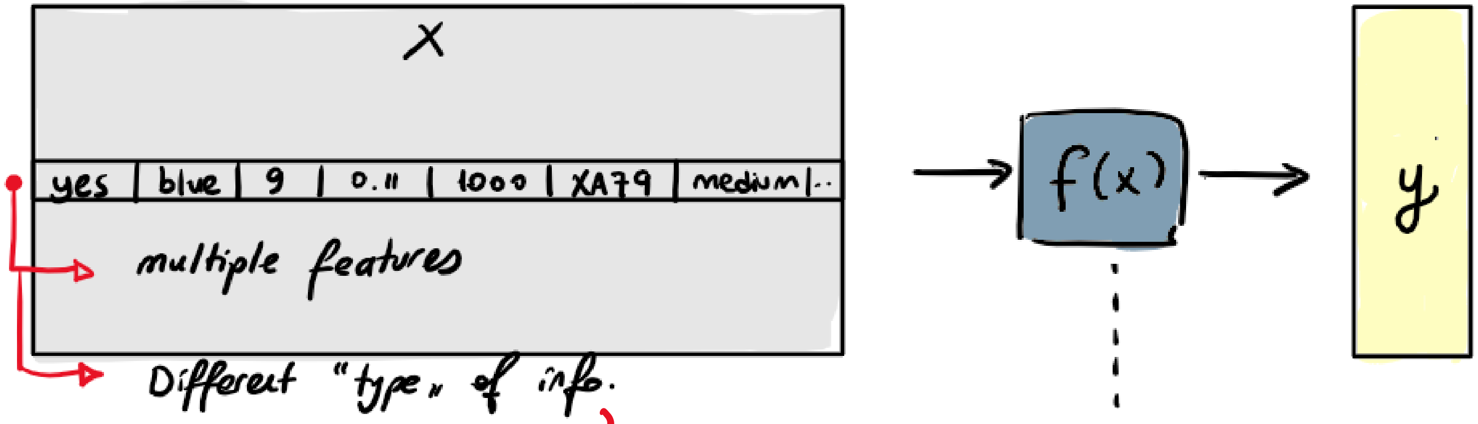


- Goal := how features are distributed among objects
  := how they are assoc. with a dep. var. (label)

! Success ⇒ Depends on how informative features are

Data Collection >>

Features in real life problems



$$y_p = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_0$$

?

Karlsruhe Institute of Technology

Features in real life problems

yes	blue	9	0.11	1000	XA79	medium	..

⇒ $x_1 :=$ Vector description of y_1
↳ converted to numbers

- Homogenous / Heterogenous in type

Categorical : feature value has no order in it.

e.g.// "color" := [blue, white, black, pink]
⇒ [0, 1, 2, 3] ⇒ vector space >> ordered !

Ordinal : Set of ordered values ⇒ different than cat. !

e.g.// "Doneness" → [Rare → Medium → well-done] ⇒ [0, 1, 2]

Categorical and ordinal features are two types of variables used in statistics and data analysis. They both describe non-numeric data, but they differ in terms of the information they convey and how they can be analyzed.

Categorical features:

- Categorical variables, also known as qualitative or nominal variables, represent distinct categories or groups. These categories have no inherent order or ranking.
- Examples include gender (male, female), colors (red, blue, green), or types of animals (cat, dog, bird).
- Categorical variables are typically used to describe the qualities or characteristics of a dataset, and they are often analyzed using techniques like frequency tables, bar charts, or chi-square tests.

Ordinal features:
-Ordinal variables, also known as ordinal data, represent categories with a specific order or ranking. These variables can be arranged in a meaningful sequence, but the differences between the levels are not quantifiable.
- Examples include survey responses (strongly disagree, disagree, neutral, agree, strongly agree), educational levels (high school, undergraduate, graduate), or customer satisfaction ratings (poor, average, good, excellent).
- Ordinal variables can be analyzed using methods similar to those used for categorical variables, but they can also be analyzed with techniques that account for the inherent order, such as the median, percentiles, or non-parametric statistical tests.

Features in real life problems

yes	blue	9	0.11	1000	XA79	medium | ..

⇨ x_1 := Vector description of y_1

↳ converted to numbers

- Homogenous / Heterogenous in type

Continuous: quantify the relationship of "nearbyness"

Ratio Scaled: $x_1 = 2$; $2 \times 2 \rightarrow x_1$

↳ Object property assoc. with feature x_1 (length) doubled.

$y_p = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_0$ ⇨ How to handle it?

Feature Engineering

- Operations we perform on object vectors during the whole pipeline.

FE

Feature Transformations

Here we are applying mathematical functions to our data. This is what we have learned in our Bachelors; Fourier transformations, differential equations, etc.

- Mapping functions
- Kernel trick
- Act. Functions
- Scaling
- Dim. Reduction
- Encoders

Feature Generation

- New features (not FT)
- MA
- Statistical feat.
- %, fract., ratios

Feature Selection

- Select a sub-set of original space.

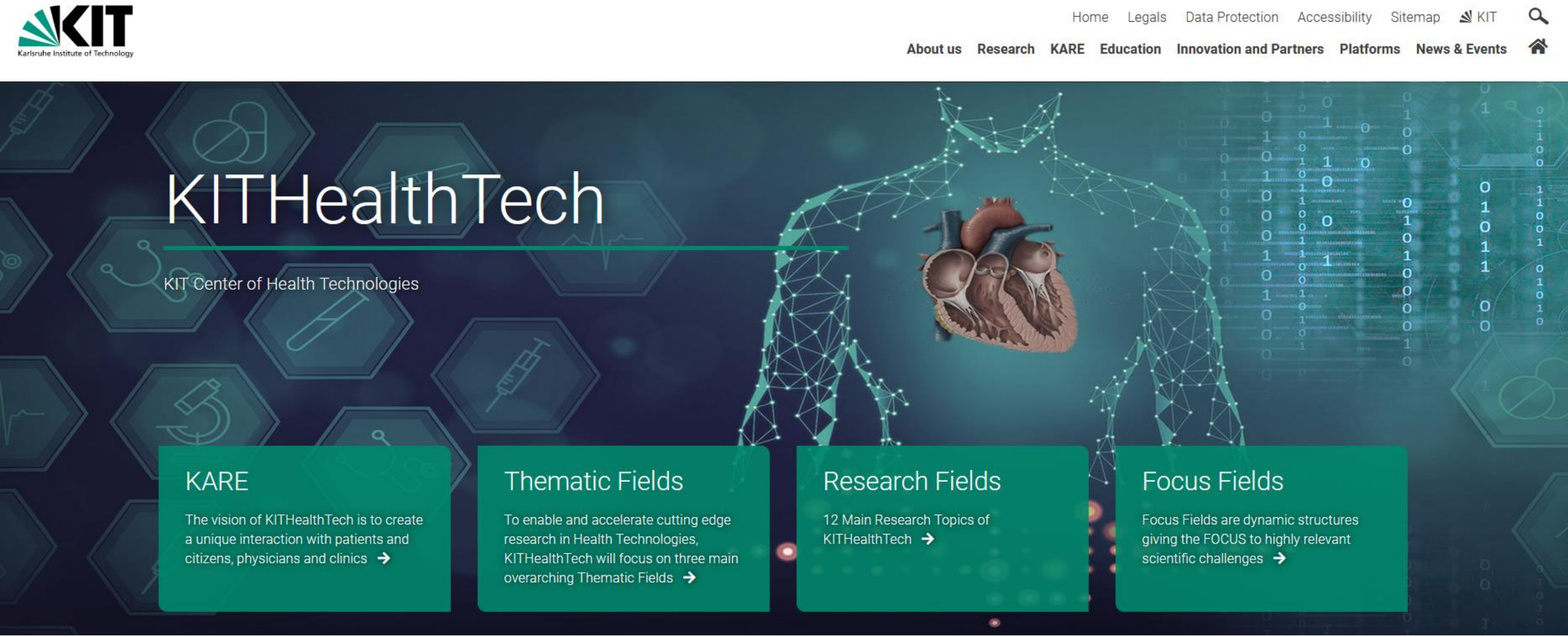
* No projection

Feature Analysis

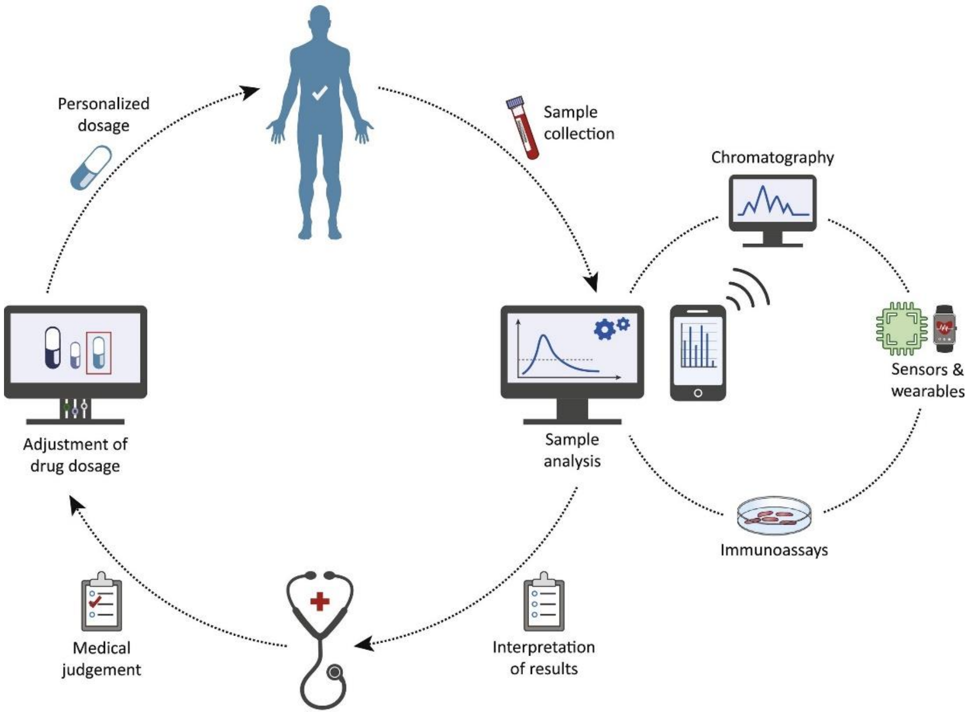
- Methods used to quantify the usefulness of features

Automation of the feature engineering

[x → expand → sub-sample ⇒ x^*]


Karlsruhe Institute of Technology
Home
Legals
Data Protection
Accessibility
Sitemap
KIT
About us
Research
KARE
Education
Innovation and Partners
Platforms
News & Events
KITHealthTech
KIT Center of Health Technologies
KARE
The vision of KITHealthTech is to create a unique interaction with patients and citizens, physicians and clinics
Thematic Fields
To enable and accelerate cutting edge research in Health Technologies, KITHealthTech will focus on three main overarching Thematic Fields
Research Fields
12 Main Research Topics of KITHealthTech
Focus Fields
Focus Fields are dynamic structures giving the FOCUS to highly relevant scientific challenges

Case Study: Th. Drug Management at ICU



https://doi.org/10.1016/j.tibtech.2020.03.001

i. Measuring this drug concentration in blood / plasma / non-invasive options
ii. Medical interpretation
iii. Dose regiment adjustment

Objective:

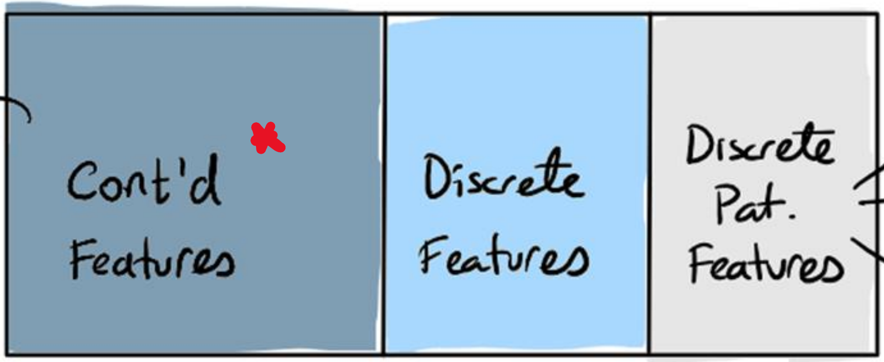
➢ Personalized dosage regimen

Challenge (assume measured):

➢ how to **represent** and **interpret** the **response** of the patient to the drug

Case Study: The Drug Management at ICU

every row is a patient with a day ID



Pat. Type?
Pip resistance?
can cause sepsis?

* 2386 instances ⇒ day info × patients
* [37 + 54 + 108] features

Numerical data : Easy to integrate ?!

[Q1] Do I know any features that are not related to my objective? (assumption ⟺ Domain knowledge)

[Q2] Can I filter uninformative columns?

↳ ∅ variance

$$y = w_1 x_1 + w_2 \underset{1}{x_2} + w_0 \Big\} = w_1 x_1 + w_0'$$

If there is little variance in the column, that column is acting as a variance.

if you have two features with the same or highly correlated information, they may introduce multicollinearity in your model. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, leading to unreliable and unstable estimates of the regression coefficients. This can make it difficult to determine the individual contribution of each feature to the model and can lead to overfitting.

→ "little variance" & duplicates

Notebook

Numerical data : Easy to integrate ?

[Q3] Do the magnitudes matter for the problem?

- (–) / (+) ⇒ is enough?

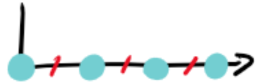
Is the magnitude important? For example, input of heat transfer is positive means that is going in.

- magnitude @ coarser granularity ?
  - --→ "counts" := # stations, daily visits etc.

We are not interested in the exact number of bus stations as 899 or 901, we need a rough number of 900. Thats when quantization comes in.

Quantization

* Group counts ⇒ "bins" ☐ Fixed-width binning



↳ continous var. ⇒ Discrete ☐ Quantile binning

Notebook

Numerical data : Easy to integrate !

[Q3] Do the magnitudes matter for the problem?

- Scale of features ⇒ models smooth func. of X — Need for Scaling
  - ↳ linear models; clustering; kNN, RBF, SVM ...
  - ↳ anything using distance
- ↳ Logical Functions ⇒ insensitive to scale — Not needed for Scaling
  - ↳ step functions; space partitioning (trees)
  - ↳ Be careful if domain shifts !

Fixed-width binning (Group counts): This method divides the range of the continuous variable into fixed-width intervals or bins. The width of the intervals is determined by specifying the number of bins or by providing a specific width for each bin. Each data point is then assigned to one of the bins based on its value. The main advantage of fixed-width binning is its simplicity and ease of interpretation. However, it can be sensitive to the choice of bin width and may lead to uneven distribution of data points across bins if the data is not uniformly distributed.

Quantile binning (Continuous variable): This method divides the continuous variable into bins such that each bin contains approximately the same number of data points. To achieve this, the data is first sorted, and then the range is divided into intervals based on the quantiles (e.g., quartiles, deciles, percentiles). This ensures that each bin has an equal number of data points, leading to a more balanced distribution across bins. Quantile binning is particularly useful when dealing with skewed data or when you want to ensure that each bin has an equal number of observations, which can help with some statistical analyses.

Data Scaling

- min-max scaling

$$x^* = \frac{x - \min(x)}{\max(x) - \min(x)} \Rightarrow [0, 1]$$

- Standard Scaling

$$x^* = \frac{x - \text{mean}(x)}{\text{sqrt}(\text{var}(x))}$$

- ℓ-norm Scaling

$$\ell_2 \Rightarrow x^* = x / \|x\|_2 \Rightarrow x / \sqrt{x_1^2 + x_2^2 + \dots x_m^2}$$ for m examples

If X is sparse:

$[0 \dots 0 \dots 1 \dots 0]$

$\Downarrow$

$[0.1 \dots 0.4 \dots \dots 0.11]$

" Dense "

Notebook

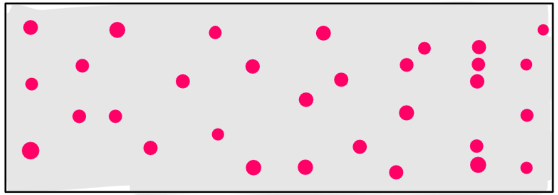
Numerical data : Easy to integrate !

Karlsruhe Institute of Technology

Data Cleaning

- Redundant Samples ✓
- Redundant Features ✓
- Outlier Detection
- Missing Values *

Handling the missing values: imputation



Fill up the column, (Which are the closest neighbors in different rows? And then compute the average and impute the missing value.

In every column you just train a bunch of regression models in each column.

- Simple Imputer ⟹ mean, median, most-frequent, constant
- kNN Imputer ⟹ use "n" neighbours to fill missing cells
  Distance-based
- Iterative Imputer ⟹ Loop over to solve multivar. reg. problem
  Distance based

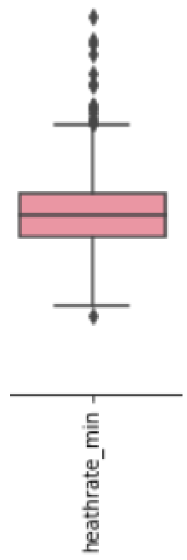
Notebook

Outlier Detection & Removal

- Distributions ~ Gaussian-like ; use SD as a cut-off
  ↳ 2 SD for small dataset ; 4 SD for large datasets

- IQR based >>  >> IQR = Q3 − Q1
  ↳ Character. length scale

- Automated Outlier Detection
  - Local outlier factor
  - Isolation forest



Local Outlier Factor

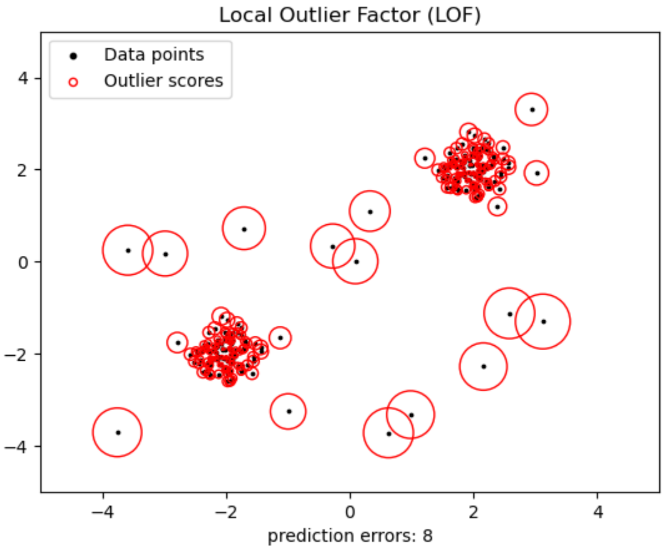
- Idea is similar to DBSCAN := Local density

- Compare the density wrt. k nearest neighbor

  → LOF << 1 → inliner ;

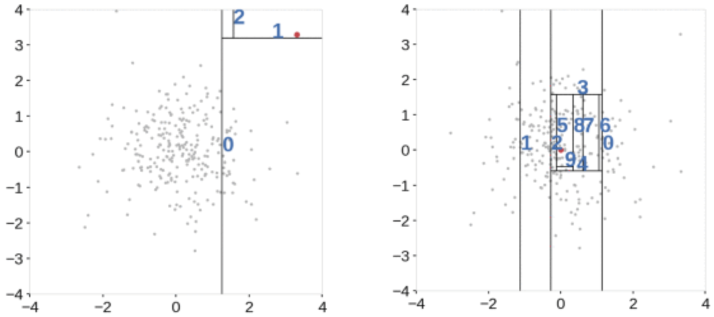
  → LOF ~ 1 → similar density

  → LOF >> 1 → Outlier

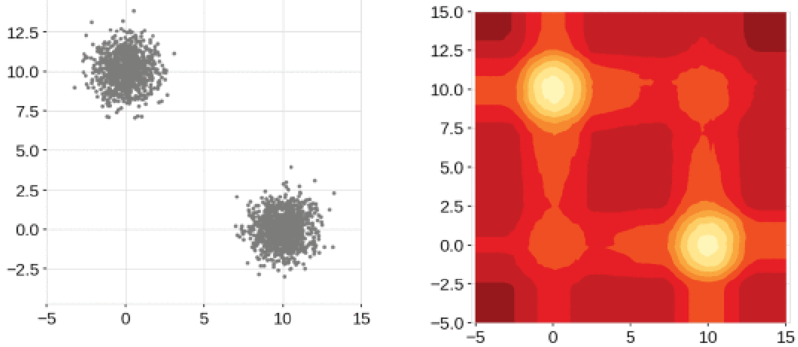


Isolation forest

- ☐ Anomalies are few in number.
  - --> very different in feature space.

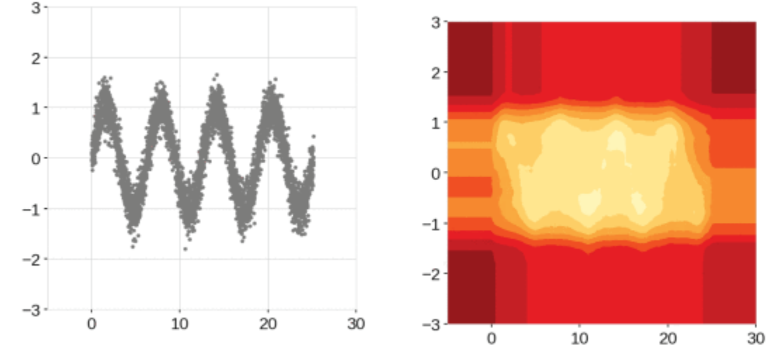


(a) Anomaly point (b) Nominal point



(a) Two normally distributed clusters

(b) Anomaly Score Map



(a) Sinusoidal data points with Gaussian noise.

(b) Anomaly Score Map

Notebook

Feature Engineering

- Operations we perform on object vectors during the whole pipeline.

FE

[Feature Transformations]

- Mapping functions
- Kernel trick
- Act. Functions
- Scaling
- Dim. Reduction
- Encoders

[Feature Generation]

- New features (not FT)
- MA
- Statistical feat.
- %, fract., ratios

[Feature Selection]

- Select a sub-set of original space.

* No projection

[Feature Analysis]

- Methods used to quantify the usefulness of features

Automation of the feature engineering

[x → expand → sub-sample ⇒ x*]

Feature Selection

Filtering

* Use feature analysis methods.
* Done once.
* "Supervised" Learn the relation of each individual feature with the labels
* "One feature at a time" ignore the rest of the features
* Fast
* Correl. ⇏ Causality

Embedded implicit/intrinsic

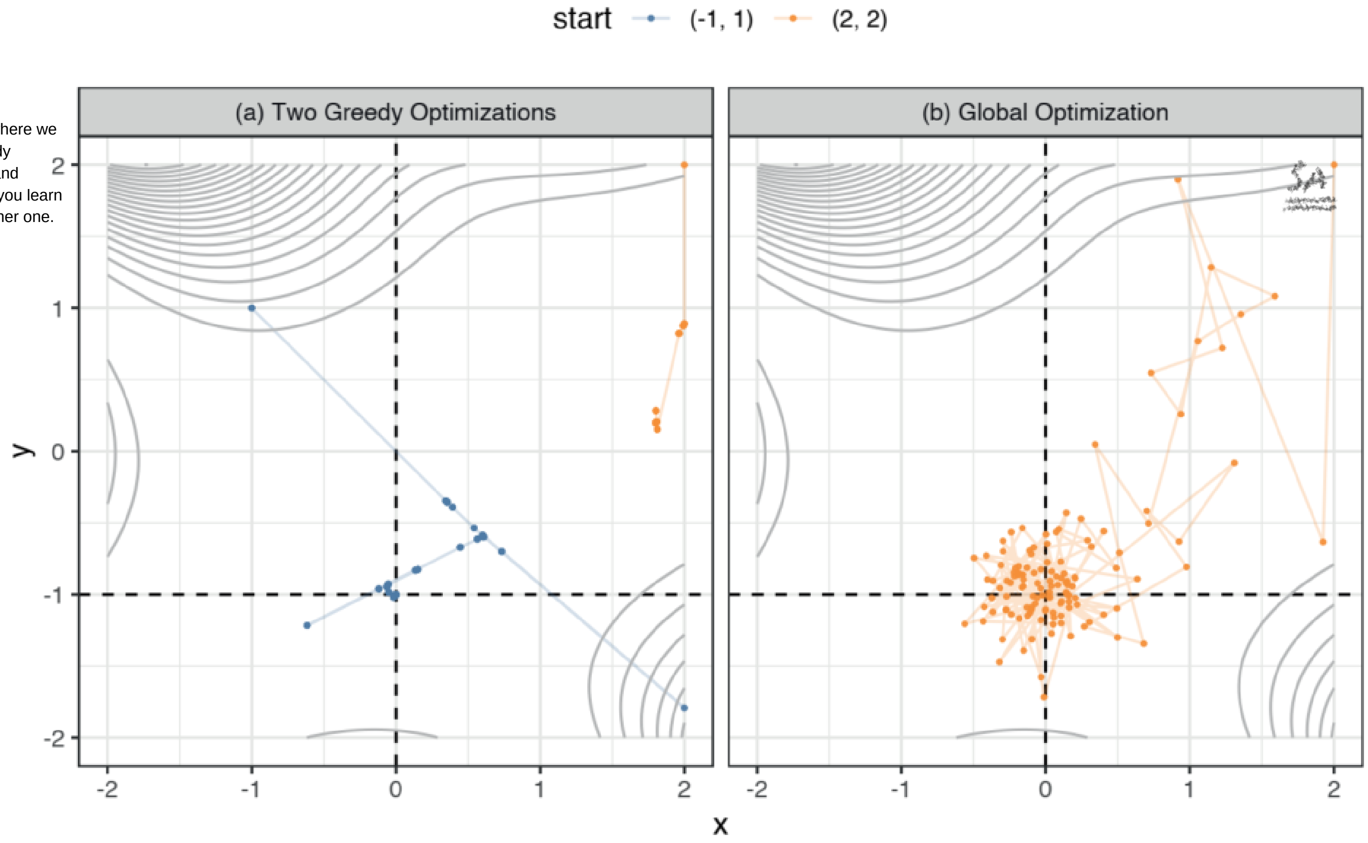
* Feature section is part of the model.
  - Tree-based models.
  - Rule-based models
  - Multivariate Adaptive Regression Spline (MARS)
  - Regularized ⇒ Lasso
* Model dependent

If you are using Lasso it checks the linearity of the model, not the non-linearity

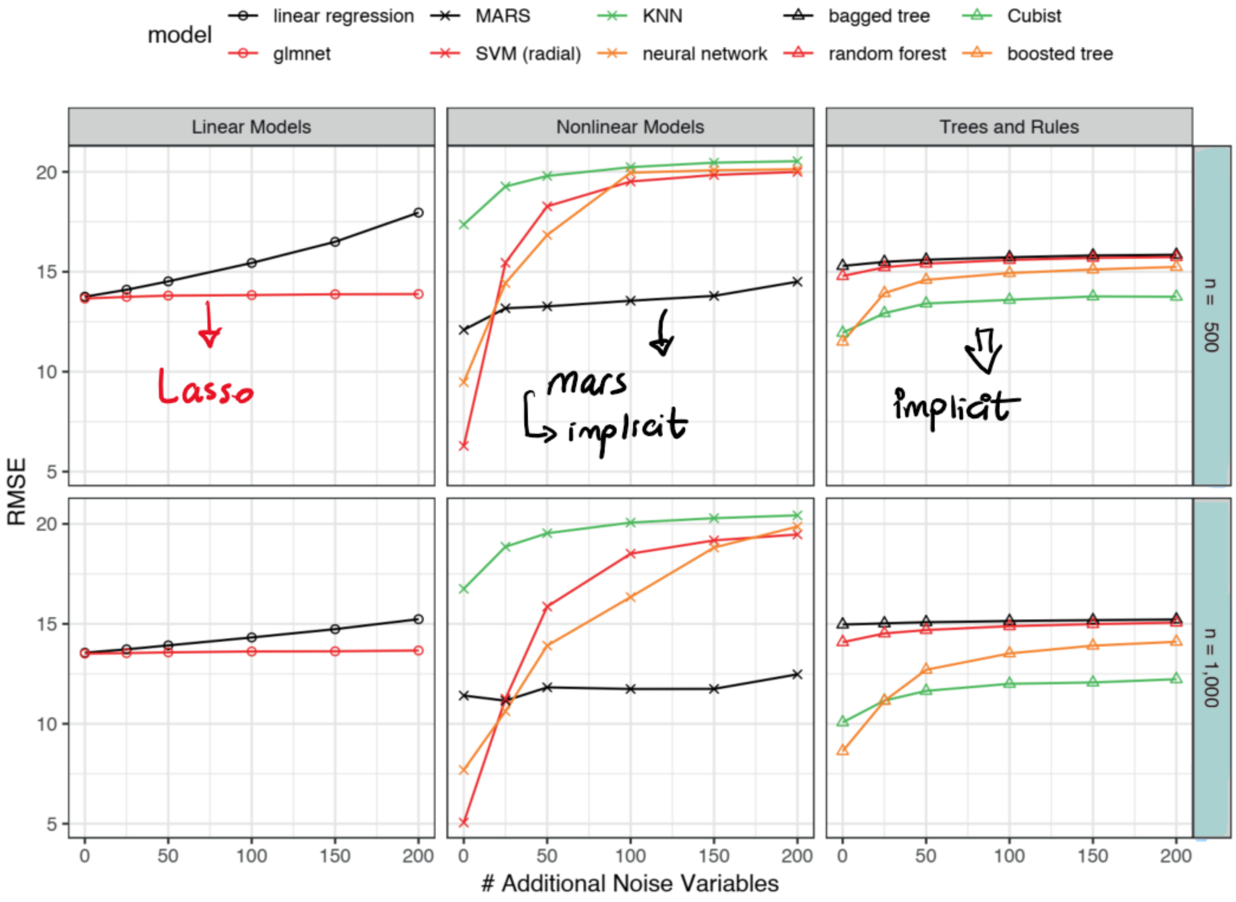
Wrappers

* use predictors.
* Iterative search
* Can be "greedy" or "non-greedy"
* Greedy → may trap in local minima.
  - RFE This is a Greedy Algorithm.
  - Genetic Algorithms

Two dimensional feature space where we know the local minima. At a greedy algorithm, you take one feature, and analyze the dependency on how you learn one feature with respect to the other one.



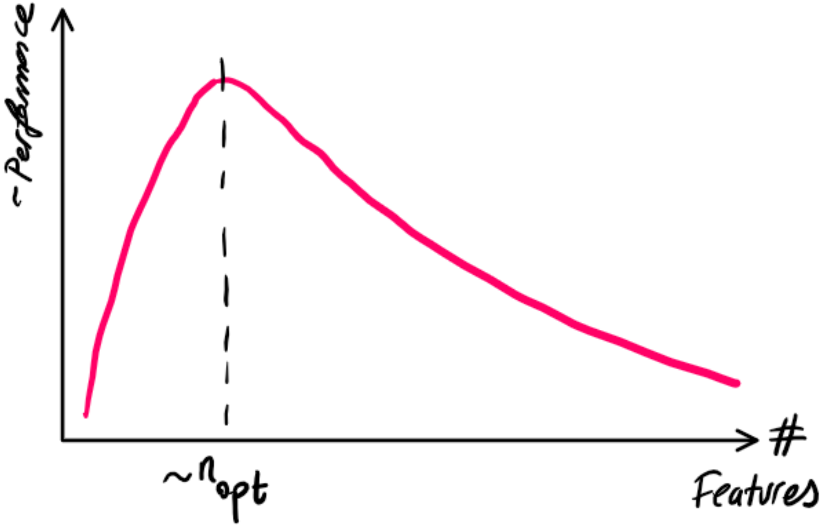
$$f(x, y) = [1 + (x + y + 1)^2(19 - 14x + 3x^2 - 14y + 6xy + 3y^2)] \times [30 + (2x - 3y)^2(18 - 32x + 12x^2 + 48y - 36xy + 27y^2)].$$



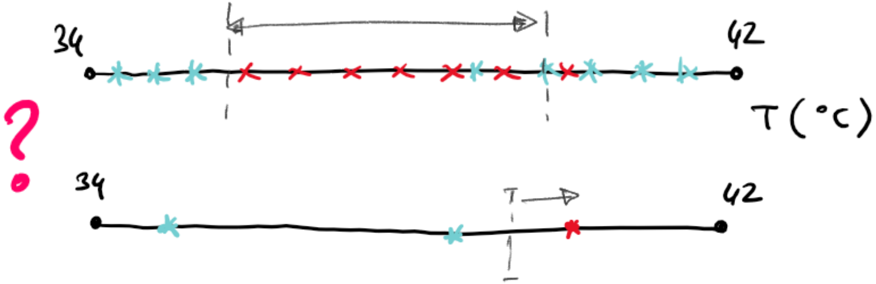
Feature Elimination helps !

More data may or may not help !

Curse of dimensionality



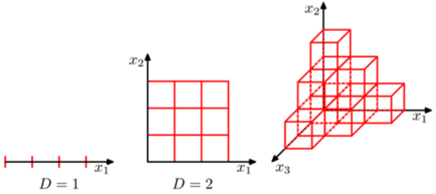
* Space gets large ⇒ more data
* "Distance" gets meaningless



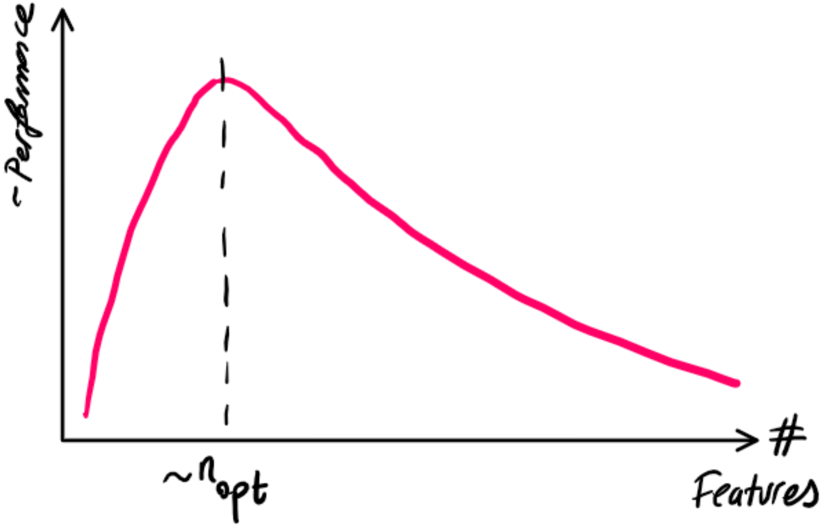
(1D) ⟹ 10 points

(2D) ⟹ 100 points

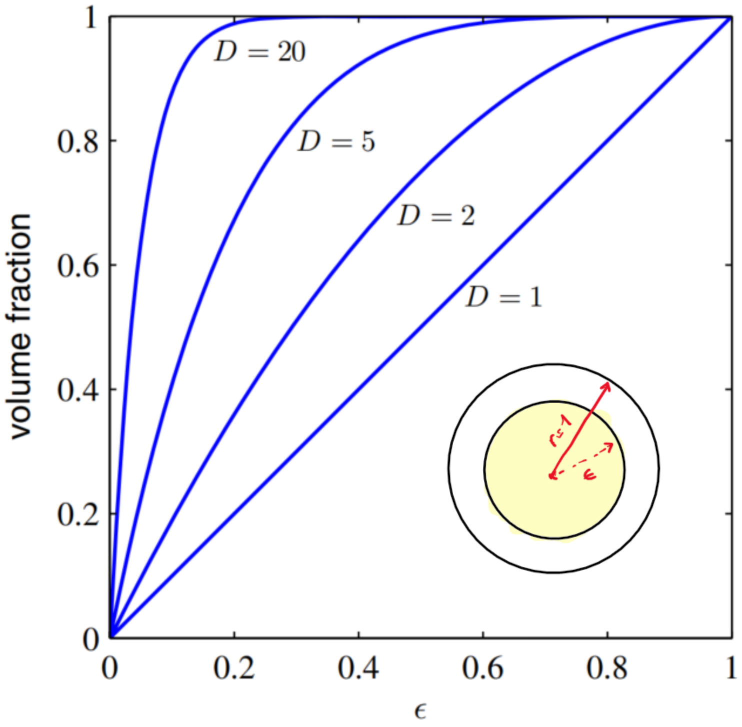
(3D) ⟹ 10^3 points



Curse of dimensionality



* Space gets large ⇒ more data
* "Distance" gets meaningless

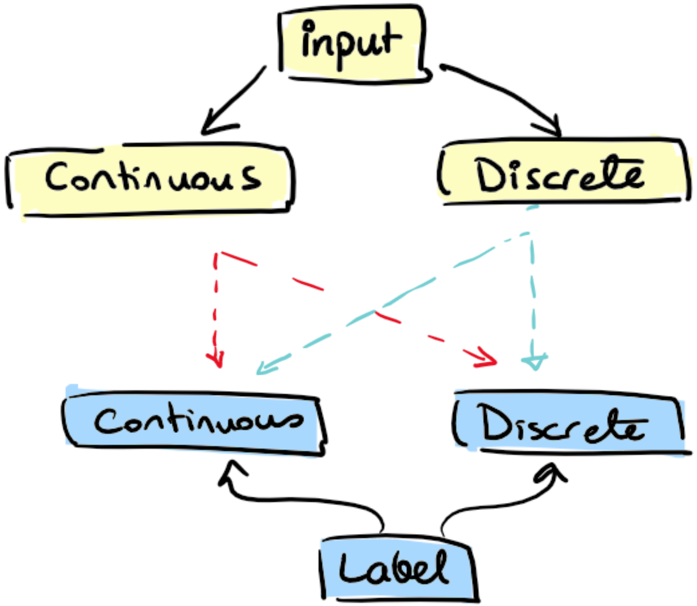


Feature Selection

- Goal := parsimonous model ⇨ Reduce model complexity
- ! Some models (SVM, ANN) sensitive to irrelevant features;
- ! Some models (LR/LogReg) vulnerable to correlated features.
- Curse of Dimensionality ⟺ Data Density to learn
- ! Which metric & method to use based on data type?

Feature Selection

- Goal := parsimonous model ⇨ Reduce model complexity



① Feature selection methods can depend on data type.

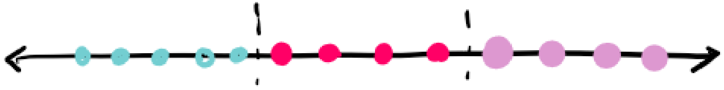
② Special care is needed for heterogenous feature space.

Case I: Numerical → Categorical

Anova - F Score

- Compares means & variances of categories.

e.g. $X_1 \Rightarrow$



values are different for each cat. value.

Mutual Information

- $MI_{ij} = \text{Entropy}(i) - \text{Entropy}(i \mid j)$
- Used for "feature & label" couple
- Generalize well to multiclass

Notebook

Case II: Numerical → Numerical

Pearson Correlation

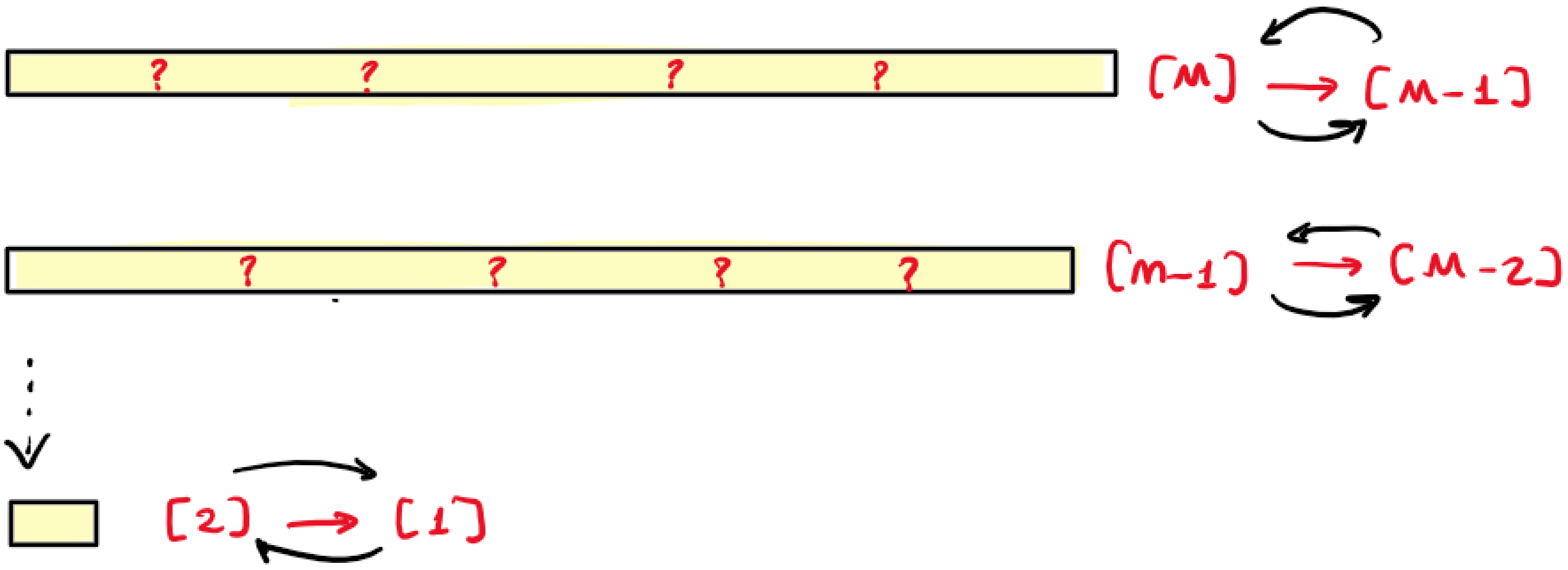
* what we have in base Corr. matrix.
* Lin. Corr. between feature & label.

Mutual information

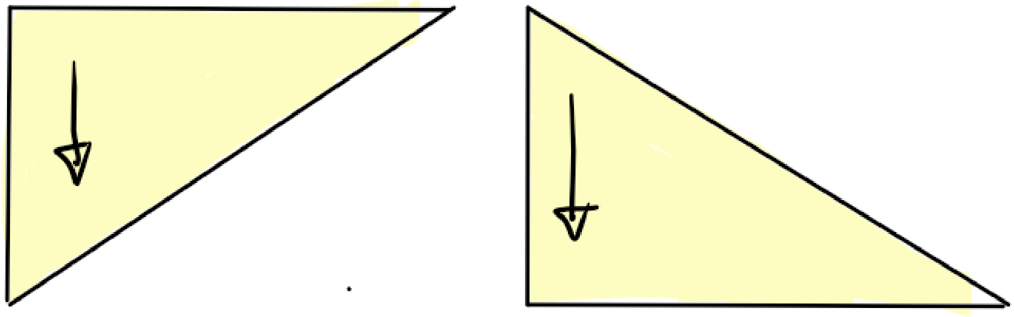
- MI_{ij} = Entropy(i) − Entropy (i | j)
- Used for "feature & label" couple
- Generalize well to multiclass

Notebook

Wrappers - I : Greedy Approach



Wrappers - I : Greedy Approach

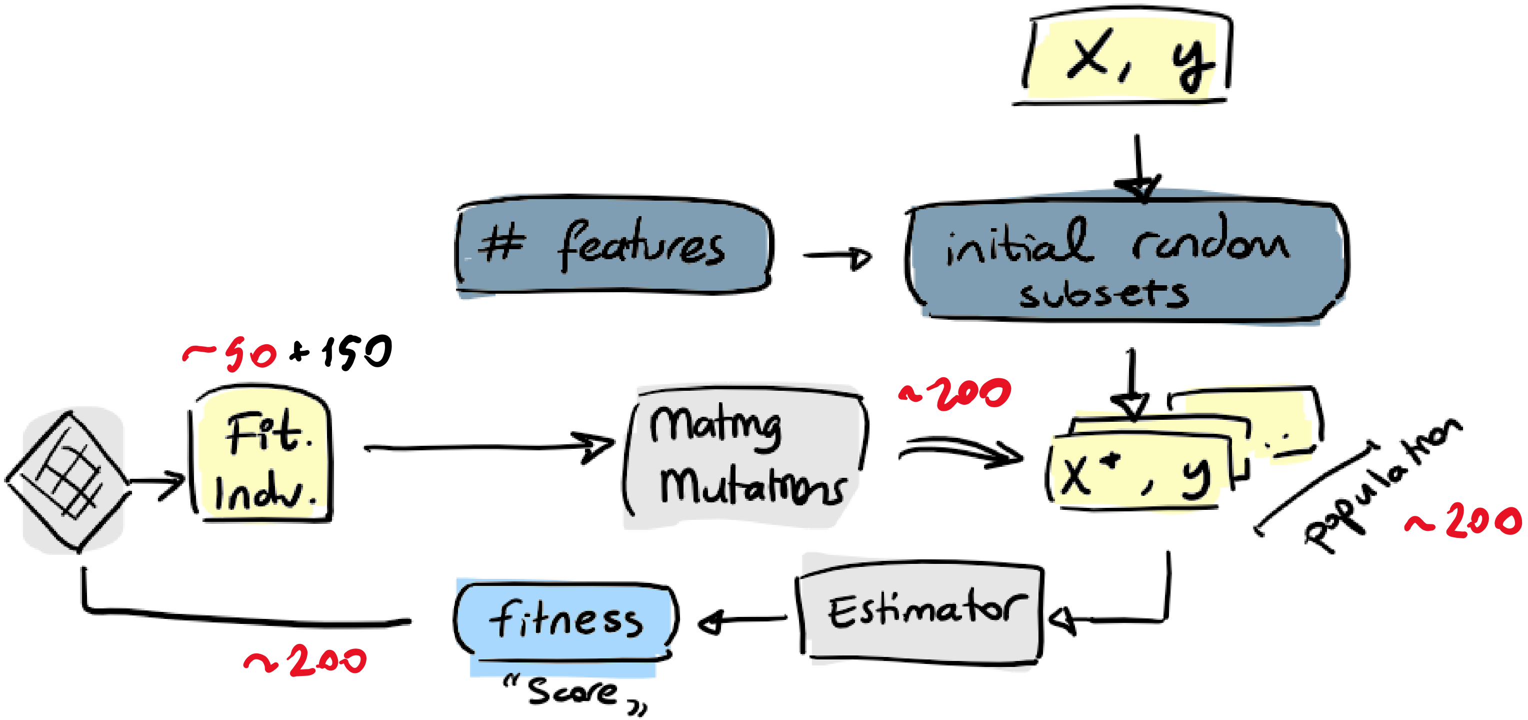


{NP-hard}

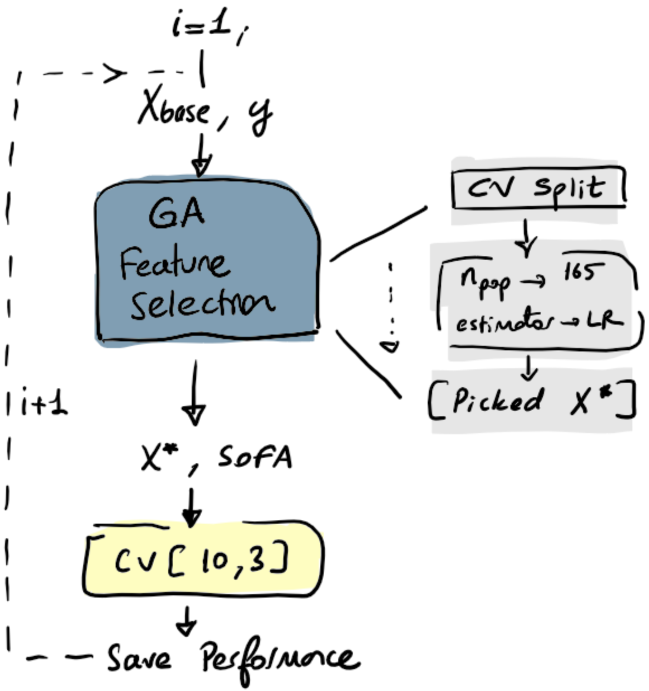
* Execute many times on random subsets ⇒ Check freq. of selection
* Stochastic search, simulated annealing, gradient decent, ...

Notebook

Genetic Algorithm for Feature Selection



Genetic Algorithm for Feature Selection



Simulated annealing algorithm

```
Create an initial random subset of features and specify the number of iterations;
for each iteration of SA do
    Perturb the current feature subset;
    Fit model and estimate performance;
    if performance is better than the previous subset then
        Accept new subset;
    else
        Calculate acceptance probability;
        if random uniform variable > probability then
            Reject new subset;
        else
            Accept new subset;
        end
    end
end
```