Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Prediction of Dialysis Length

Adrian Loy, Antje Schubotz

2 February 2017

Agenda

1. **Introduction**
   - Dialysis
   - Research Questions and Objectives
2. Methodology
   - MIMIC-III
   - Algorithms SVR and LPR
   - Preprocessing with *rapidminer*
   - Optimization Challenges
3. Preliminary Results
4. Discussion

How much do you know about dialysis?

A. I have never heard of it.

B. I have heard of it, but cannot explain it.

C. I can explain it in general.

D. I can explain it in detail.

How much do you know about Support Vector Machines (SVM)?

A. I have never heard of it.
B. I have heard of it, but cannot explain it.
C. I can explain it in general.
D. I can explain it in detail.

How much do you know about Support Vector Regression (SVR)?

A. I have never heard of it.

B. I have heard of it, but cannot explain it.

C. I can explain it in general.

D. I can explain it in detail.

Adrian Loy
Antje Schubotz

Do you know Polynomial Regression?

A. I have never heard of it.

B. I have heard of it, but cannot explain it.

C. I can explain it in general.

D. I can explain it in detail.

Do you know Local Polynomial Regression (LPR)?

A. I have never heard of it.

B. I have heard of it, but cannot explain it.

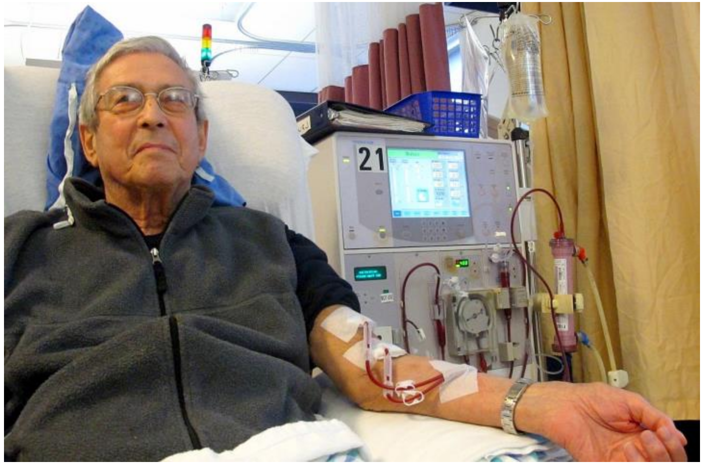
C. I can explain it in general.

D. I can explain it in detail.

Hemodialysis

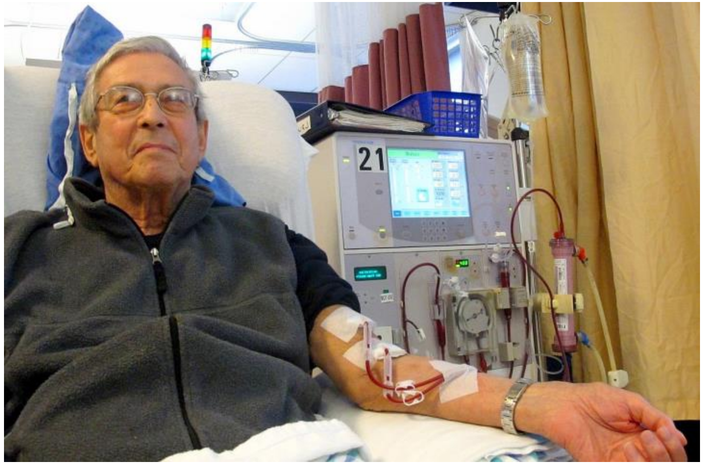
Hemodialysis

- If kidneys malfunction, there are a lot of substances in the blood that have to be removed
- This can be done with hemodialysis: Blood is pumped out of the body and runs next to a semi permutable membrane
- The small harmful substances diffuse through the membrane



Hemodialysis

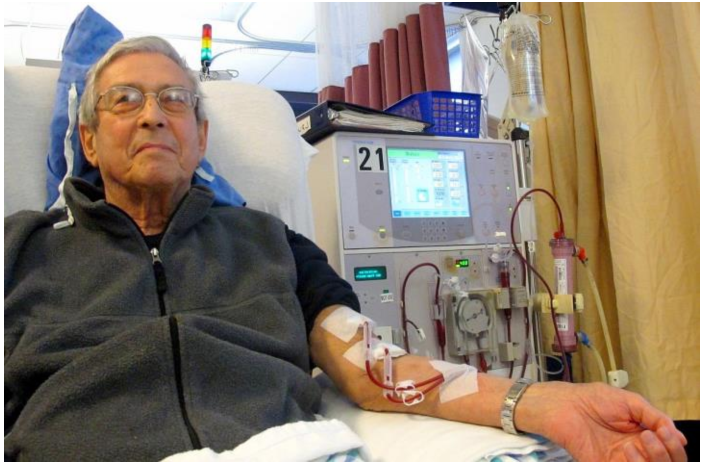
- Usually 3 times a week, 4-5 hours per session
- For older or injured people much longer with a lower rate



By Anna Frodesiak - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=19317170

Benefits of predicting dialysis duration:

- Doctors refer to guidelines that are bases on empiric results
- Hospitals could better plan their occupancy rate
- Shorter sessions would reduce the infection risk and might lower the costs
- Could affect 100.000 patients by 2020



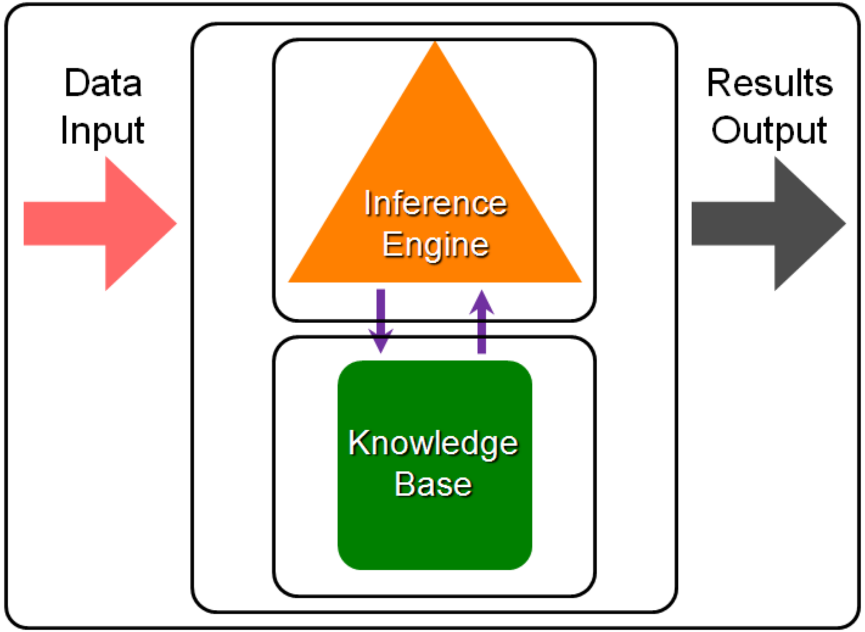
Research Objective

- Is it possible to predict the optimal duration of a dialysis session from various personal data collected in hospitals?

Our approach:

Perform regression on the duration
using SVR and LPR

- Extract data from a database with hospital data
- Perform some preprocessing



Bonney (2011)

- Compare results

Agenda

1. Introduction
   - Dialysis
   - Research Questions and Objectives
2. **Methodology**
   - MIMIC-III
   - Algorithms SVR and LPR
   - Preprocessing with *rapidminer*
   - Optimization Challenges
3. Preliminary Results
4. Discussion

MIMIC-III Database

MIMIC-III Database

- Openly available dataset developed by the MIT Lab for Computational Physiology
- Contains information about 60.000 intensive care admissions from 2001-2012
- Information includes:
  - Demographics
  - Vital signs
  - Laboratory test results
  - Medications
  - Diagnosis

Selected Features

- Which available information might effect dialysis length?
- We decided to include 15 features in our dataset:
  - Gender, Height, Weight, Age
  - Averages of blood lab values: Urea, Calcium, Sodium, Potassium, PH, Creatinine
  - Health scores: Elixhauser, Akin, EGFR
  - Duration

Selected Features

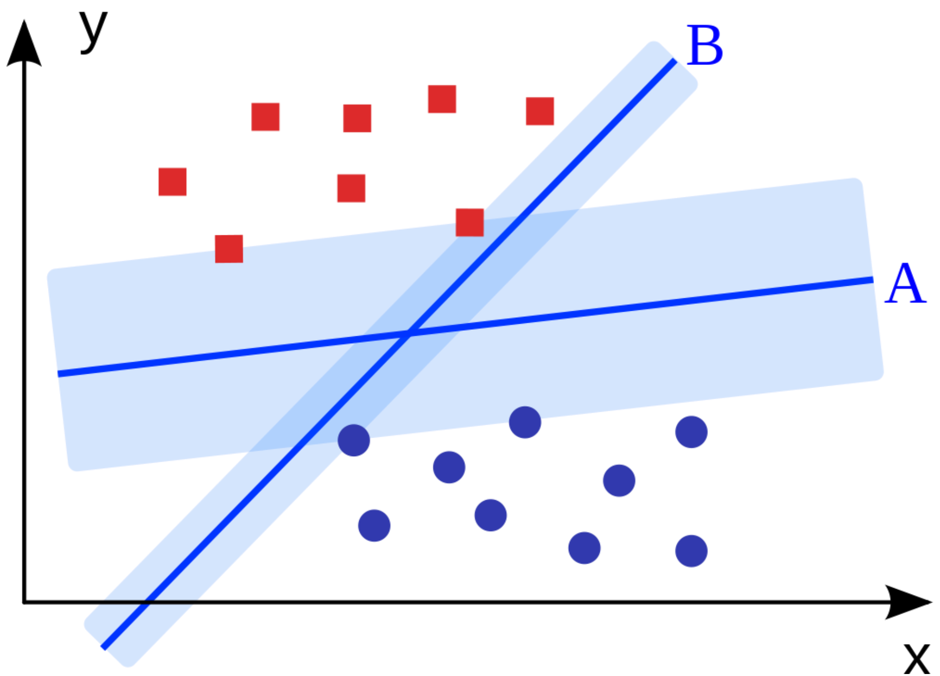
- MIMIC-III contains 2047 hemodialysis procedures
- Many outliers (age 300, duration 1min) and missing values
- Clean subset: Nearly no missing values, some outliers removed, only 76 data points

SVM and SVR

Support Vector Machines (SVM)

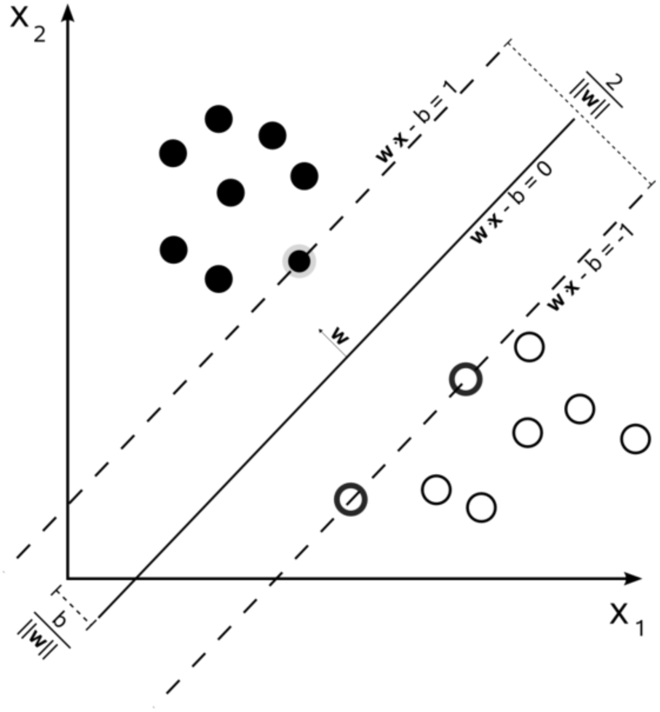
- State-of-the art for many classification problems
- Geometric model that finds a specific linear hyperplane that separates the feature space and the training data
- Can be tweaked to generate non-linear models (kernel-trick)
- Can be adapted to perform regression (SVR)

SVMs for Classification



SVMs for Classification

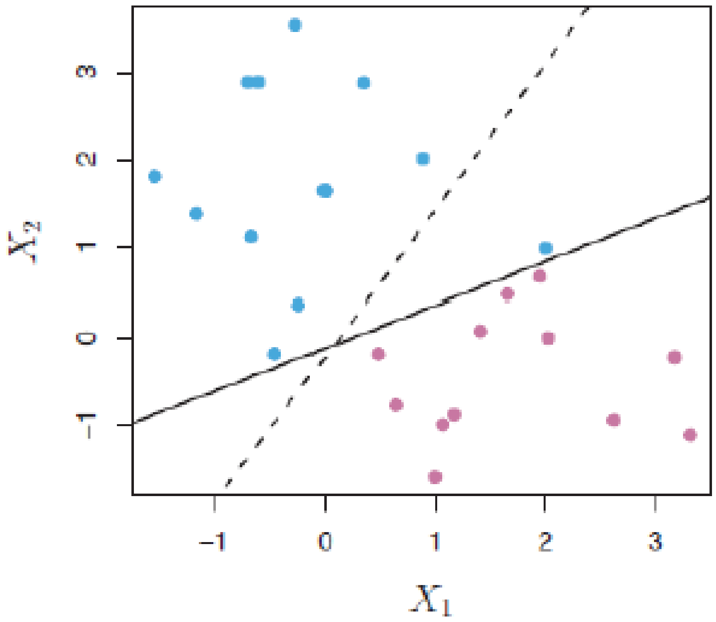
- Quadratic constrained minimization
- Problem:
  - $\text{argmin} \frac{1}{2} \|w\|^2$
  - with subject to:
  - $y_i(\langle w, x_i \rangle - t) \geq 1$



By Cyc – Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=3566688

Soft Margin SVM

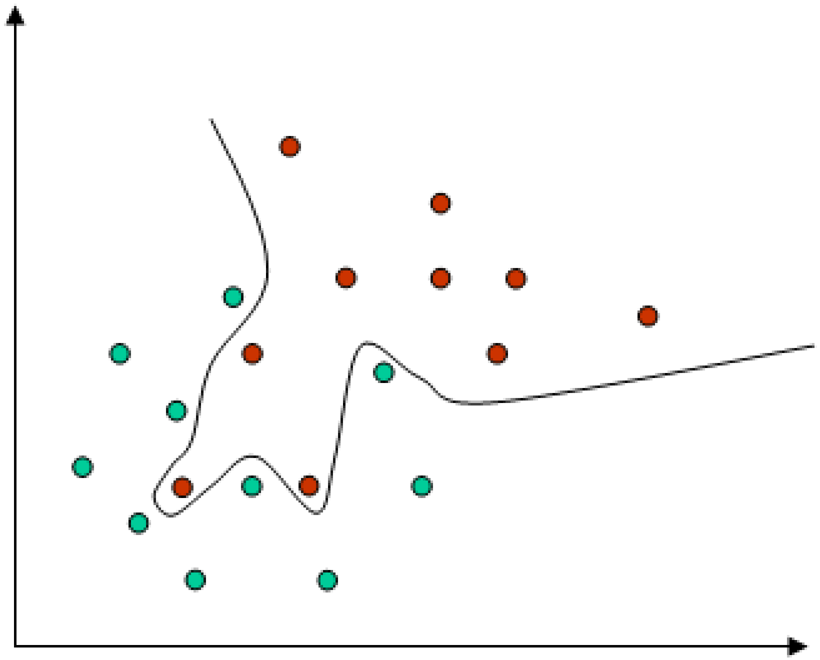
- often useful to allow some misclassification if it gives a plane with a bigger margin
- This can be achieved by introducing slack variables, that punish misclassification
- New optimization problem:
  - $\text{argmin} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i$
  - with subject to:
  - $y_i(\langle w, x_i \rangle - t) \geq 1 - \xi_i,$
  - $\xi_i \geq 0$



Prediction of Dialysis Length

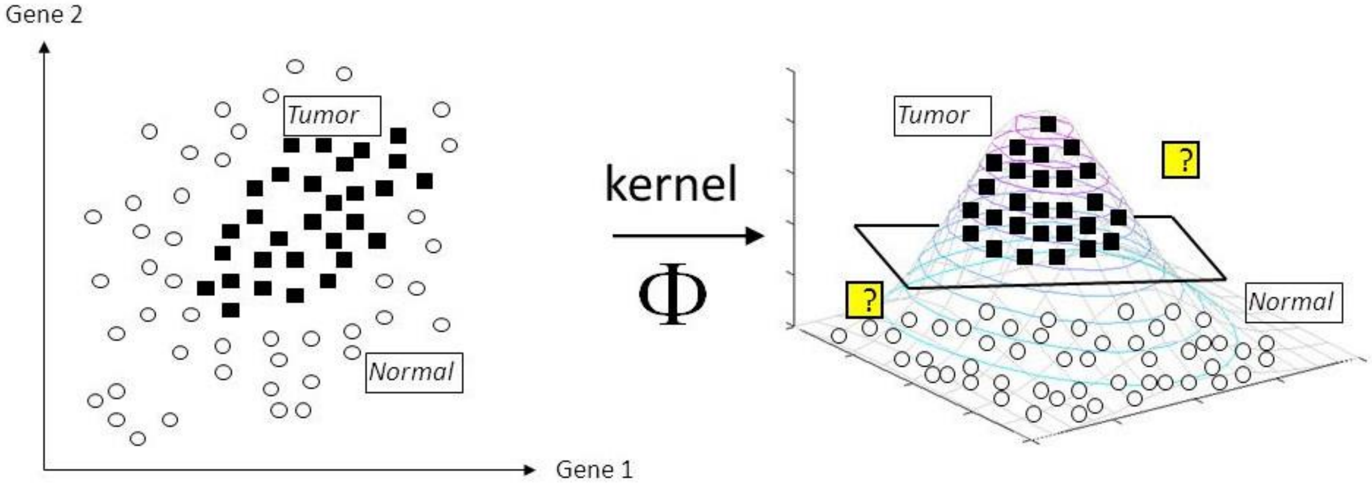
Adrian Loy
Antje Schubotz

What about data that cannot be separated linear?



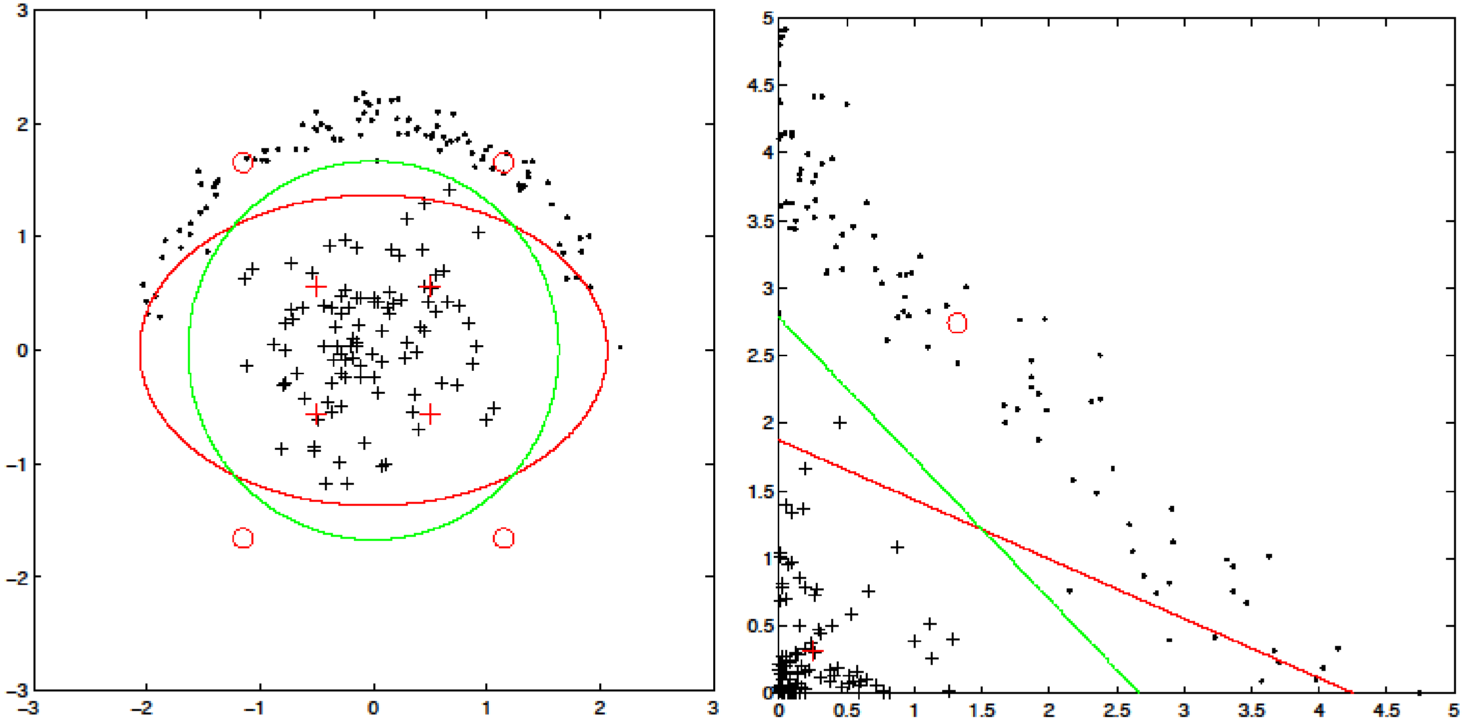
The Kernel Trick

- A transformation of the feature space into a higher dimensional space can help
- But: We need to choose the right kernel



By Teresa Powley – http://slideplayer.com/slide/1579281/

The Kernel Trick



The Kernel Trick

- The dual formulation of the minimization only contains dot products of the data points, no other norms
- This means we don't have to do a full feature transformation, we can just replace the dot product with the dot product in the transformed space
- Some popular kernels:
  - The polynomial kernel: $k(x, y) = (x \cdot y + 1)^d$
  - The radial kernel: $e^{-g \cdot \|x-y\|^2}$

http://www.svms.org/parameters/

Prediction of Dialysis Length

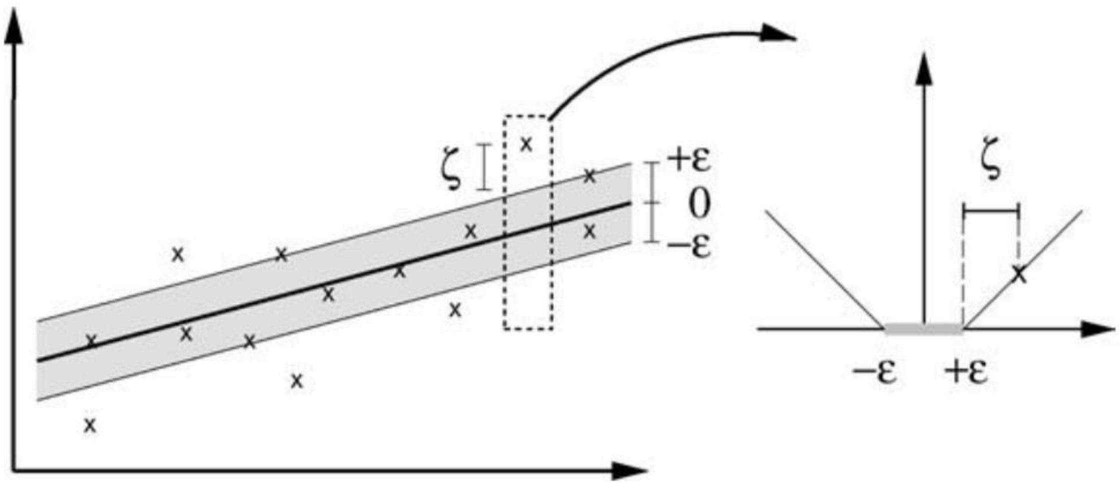
Adrian Loy
Antje Schubotz

Support Vector Regression

- SVR estimates a function, so that most points lie inside a tube of size ε around that function
- The function shall be as flat as possible and minimize the points outside the tube
- Optimization problem for SVR:
- $\text{argmin} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{t} (\xi_i + \xi_i^\star)$

  with subject to:

  - $y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i,$
  - $\langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i,$
  - $\xi_i, \xi_i^\star \geq 0$



Schölkopf and Smola, 2002

SVR Parameters

- $\operatorname{argmin} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{t} (\xi_i + \xi_i^\star)$

- Parameter C:
  - Controls the trade-off between the training error and the complexity of the model
  - Too small: Risk of underfitting: More points are outside
  - Too big: Risk of overfitting: More points inside
  - Rule of thumb: Choose C as the input range
  - *Rapidminer* heuristic: $C = \frac{n}{\sum k(i,i)}$

http://www.svms.org/parameters/

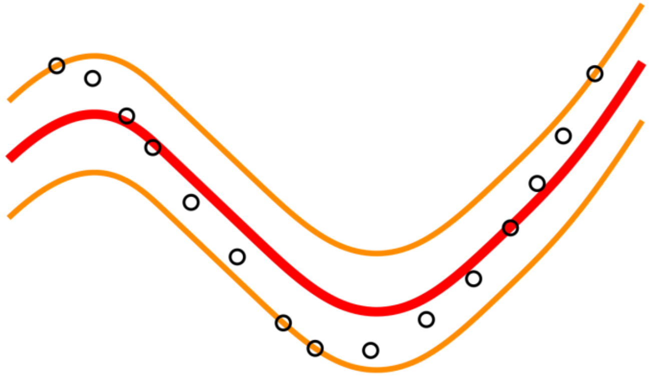
SVR Parameters

- $\text{argmin} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{t} (\xi_i + \xi_i^{\star})$

- Parameter ε:
  - Controls the size of the tube and therefore the accuracy
  - Effects the "flatness" (generalization) and the amount of support vectors
  - Rule of thumb: Choose ε so that 50% are support vectors

http://www.svms.org/parameters/

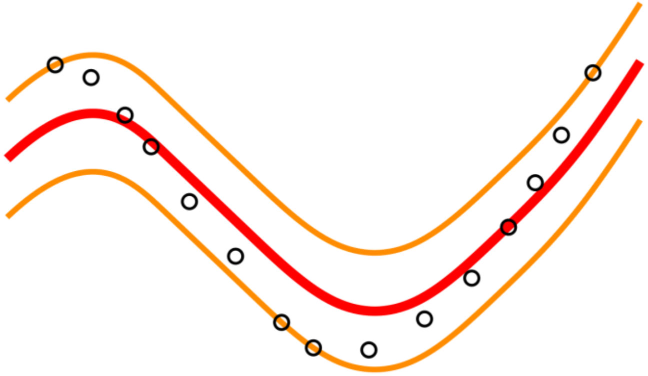
Given this diagram with data points, function and tube. How can we improve the result?

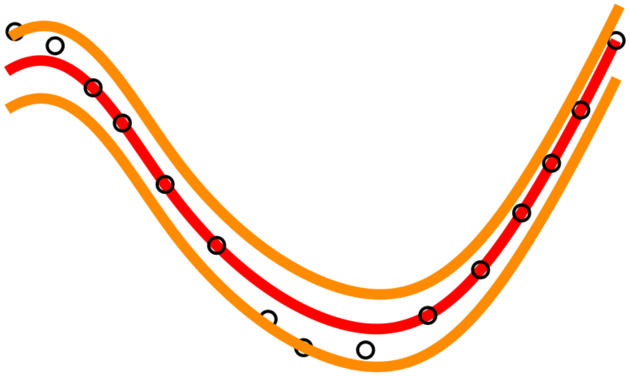


A. Bigger C
B. Smaller C
C. Bigger ε
D. Smaller ε

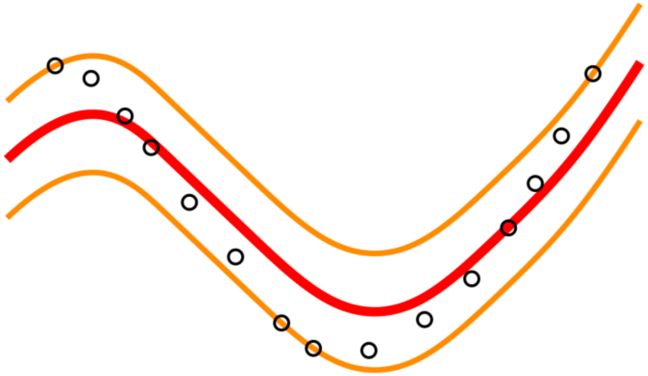
Given this diagram with data points, function and tube. How can we improve the result?

Answer: D. Smaller ε





What would happen if C is set to 0?



A. SVR produces a complex, overfitting model
B. SVR produces a flat line
C. Nothing changes
D. The model has more support vectors

Support Vector Regression (SVR)

+

- Memory efficient due to SVs
- Flexible with kernels
- Type of function can be controlled
- No requirements to the distribution of amount of data
- Can deal with outliers

-

- Runtime can be huge for some kernels
- Often domain specific knowledge is needed
- Difficult to evaluate and share
- Choosing and optimizing the parameters is really hard!

Local Polynomial Regression (LPR)

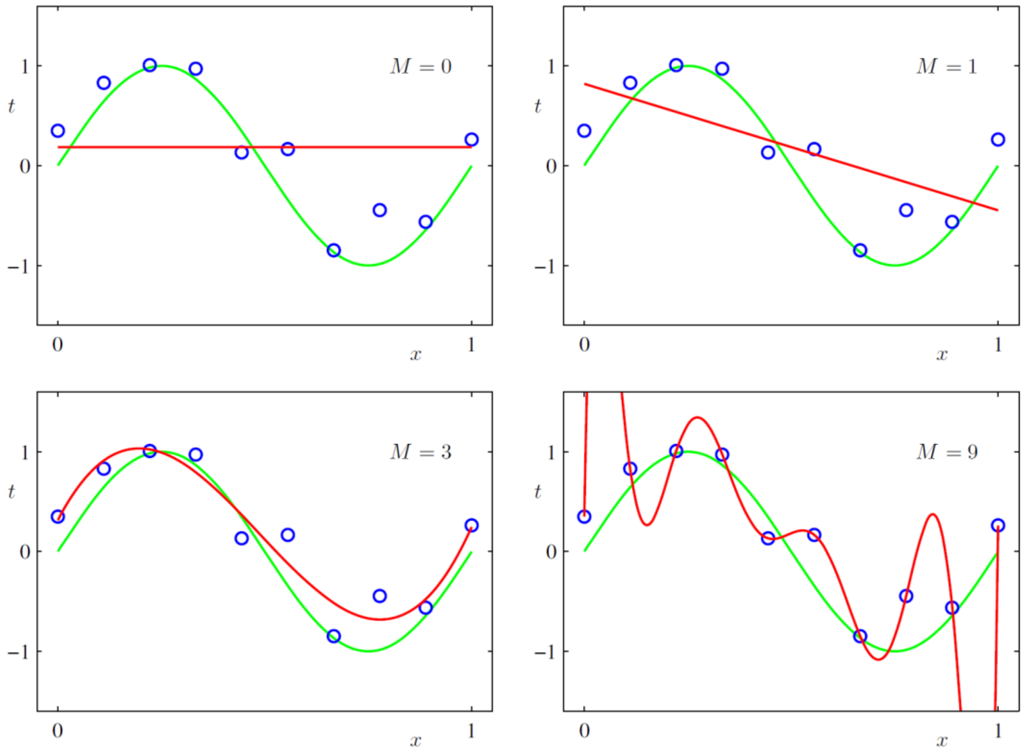
Polynomial (degree p)

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_p x^p$$

bivariate version:

$$f(x, y) = a_0 + a_1 x + a_2 y + a_3 xy + a_4 x^2 + a_5 y^2 + \cdots + a_p x^p y^p$$

Polynomial Regression



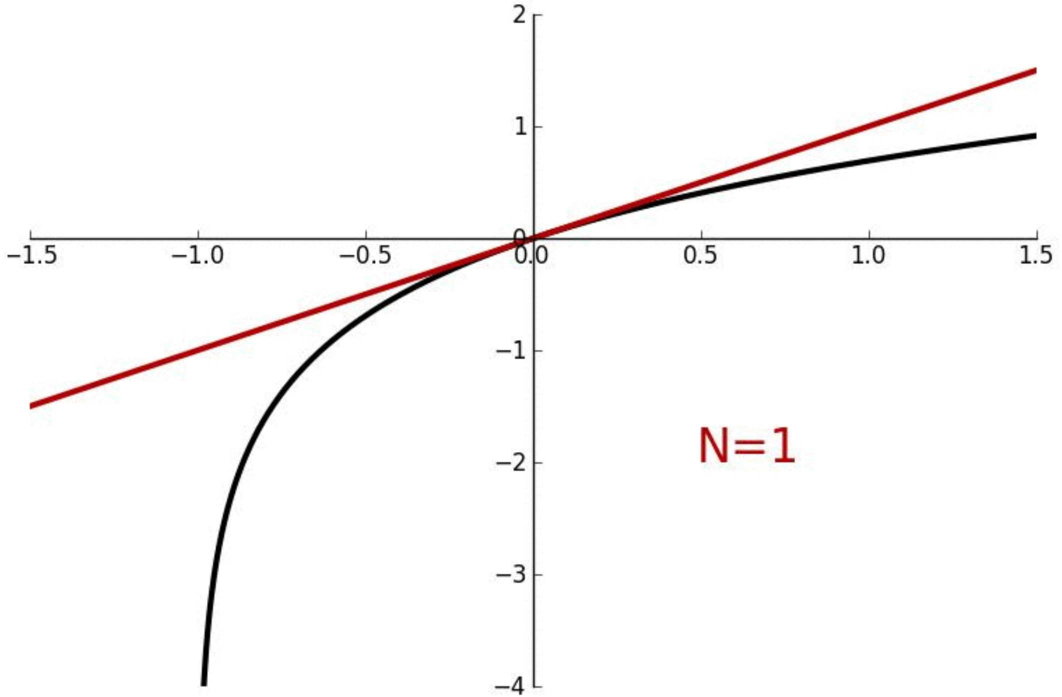
Bishop, "Pattern Recognition and Machine Learning", 2006, page 7

Approximation with Taylor Series

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_p x^p$$

$$\approx f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2 + \cdots + \frac{f^{(p)}(x_0)}{p!}(x - x_0)^p$$

Approximation with Taylor Series



https://en.wikipedia.org/wiki/Taylor_series#/media/File:Logarithm_GIF.gif

Approximation with Taylor Series

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_p x^p$$

$$\approx f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2 + \cdots + \frac{f^{(p)}(x_0)}{p!}(x - x_0)^p$$

$$= \beta_0 + \beta_1 (x - x_0) + \beta_2 (x - x_0)^2 + \cdots + \beta_p (x - x_0)^p$$

$$= \sum_{k=0}^{p} \beta_k (x - x_0)^k$$

Local Polynomial Regression (LPR)

$$\underset{\beta_0,\beta_1,\beta_2,\ldots,\beta_p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} w_i(x) \cdot \left[y_i - \sum_{k=0}^{p} \beta_k (x - x_0)^k \right]^2 \right\}$$

Main Parameters of LPR

- Weighting function $w_i(x)$
  - Defines neighborhood
- Smoothing kernel
  - Used to calculate weights of distant examples
- Degree p
  - $p > 2$ is computationally costly

Local Polynomial Regression (LPR)

+

- No assumptions about target function
- Simple
- Flexible
- Good estimator
- Easy and fast training

-

- Evenly distributed data points necessary
- Outliers problematic
- Difficult to evaluate and share
- Expensive to apply

Preprocessing with *rapidminer*

„Our **visually-based software** accelerates the process of creating **predictive analytics models** and makes it easy to get the results embedded in business operations."

https://rapidminer.com/us/

Operators in *rapidminer*

Set Role Label

- Label is value we want to predict
- Here: dialysis length

Nominal to Numerical

- Algorithms cannot handle them
- Categories to quantitative data

Missing Values

- Algorithms cannot handle them
- Impute average value

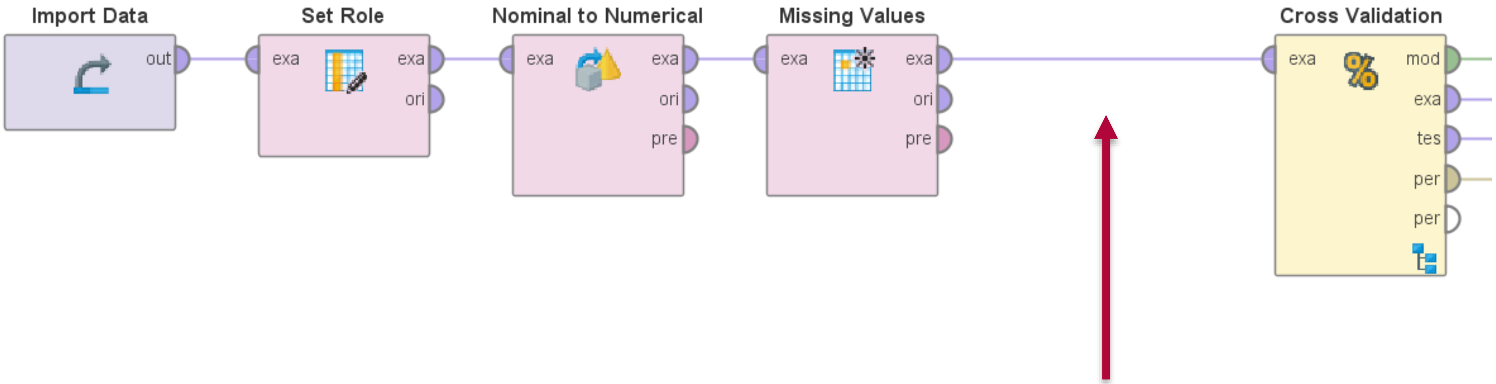
Cross Validation

- Split into training and testing
- Determines performance

LIVE DEMO

Optimization Challenges

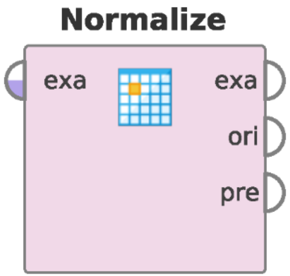
Current Process

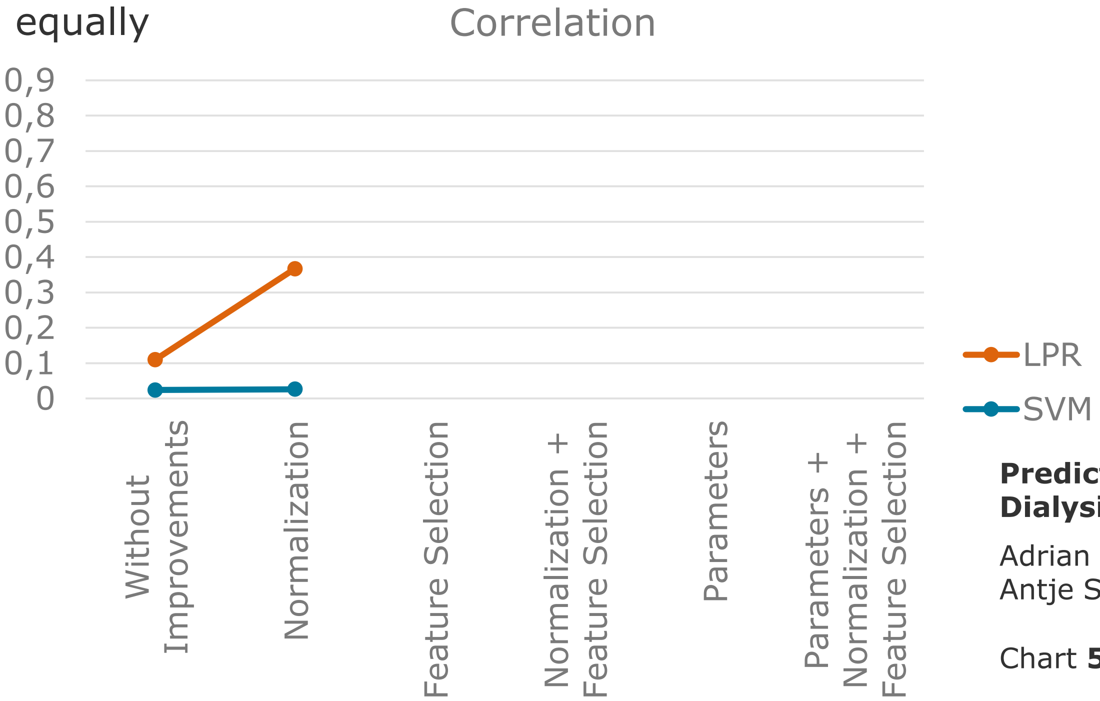


Room for improvement

Normalization (1/3)

- Rescaling of features to the same scale
- All features are weighted equally

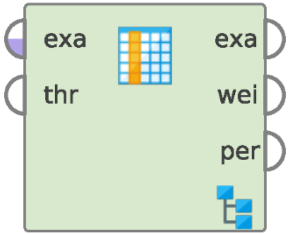




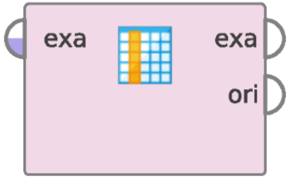
Feature Selection (2/3)

- Only subset of most important features
- Reduces noise and is faster

Optimize Selection



Select Attributes



LPR:

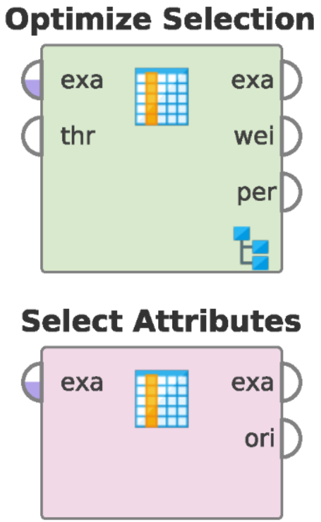
- CALCIUM_AVG_BEFORE
- CREATININE_AVG_BEFORE
- FREE_CALCIUM_AVG_BEFORE
- PH_AVG_BEFORE
- POTASSIUM_AVG_BEFORE

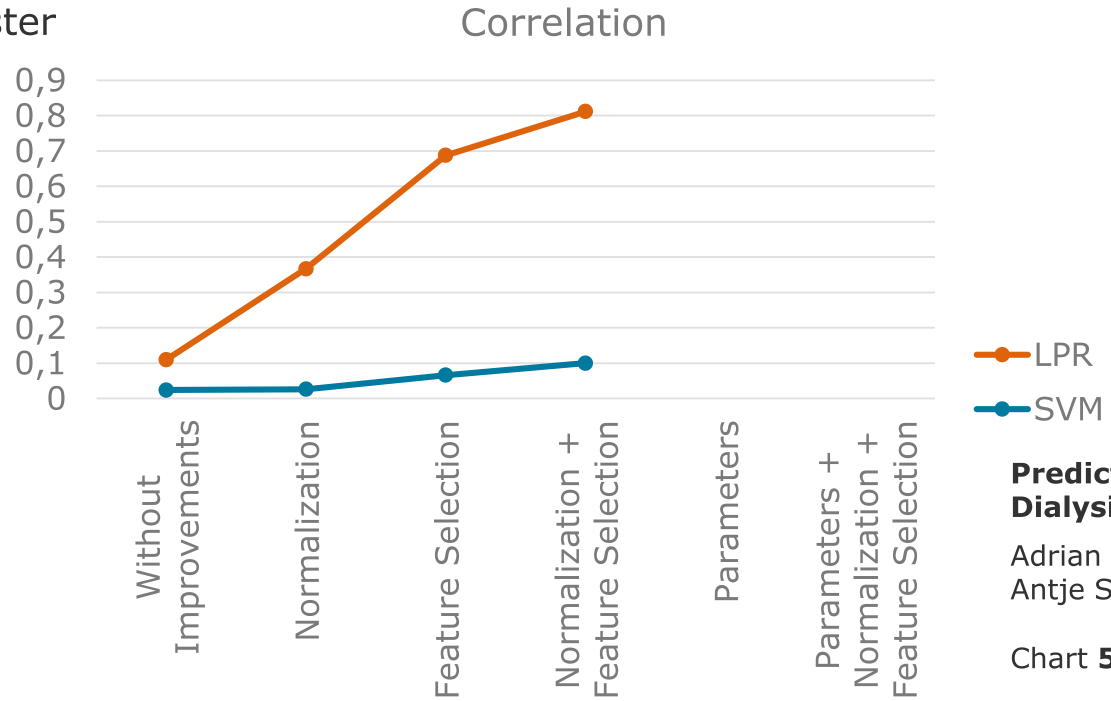
SVM:

- AGE
- HEIGHT
- ELIXHAUSER_VANWALRAVEN
- POTASSIUM_AVG_BEFORE
- SODIUM_AVG_BEFORE

Feature Selection (2/3)

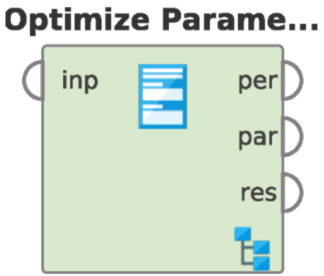
- Only subset of most important features
- Reduces noise and is faster

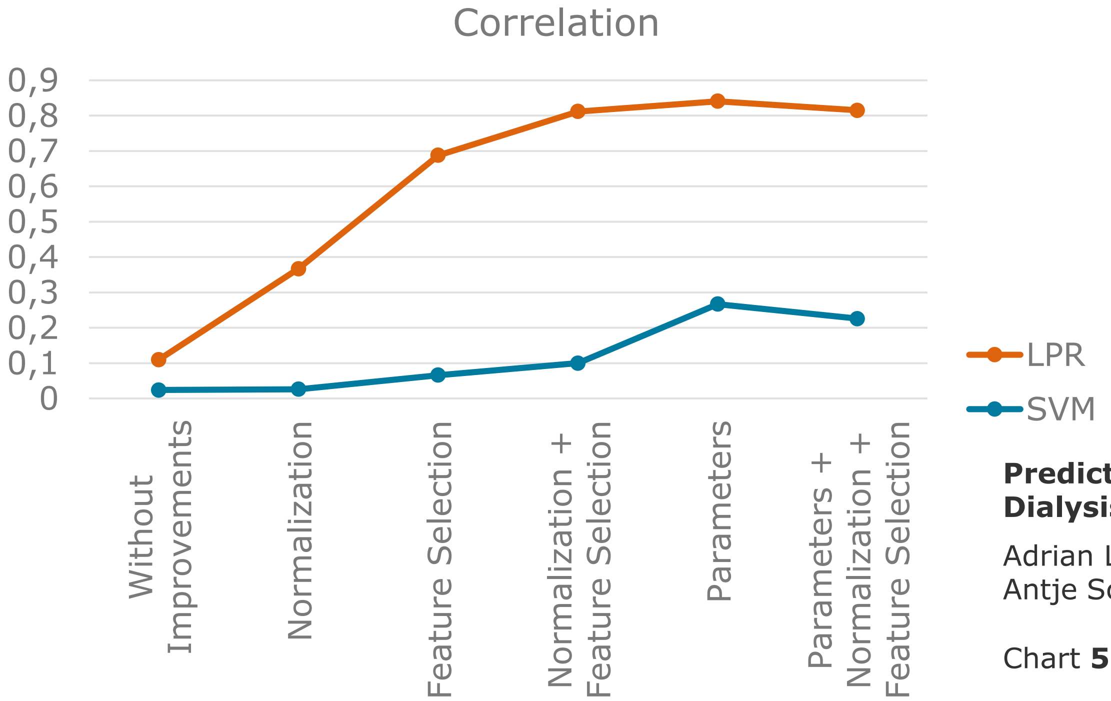




Parameters of Algorithms (3/3)

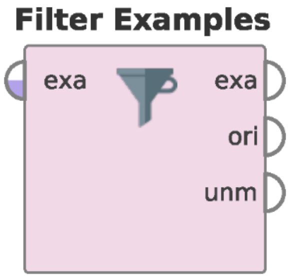
- LPR: weighting function, smoothing kernel, degree, …
- SVM: kernel type, C, …





Next Step: Datasets

- Split data (e.g., men/women, outliers, age)
- But: do not reduce number of data points too much

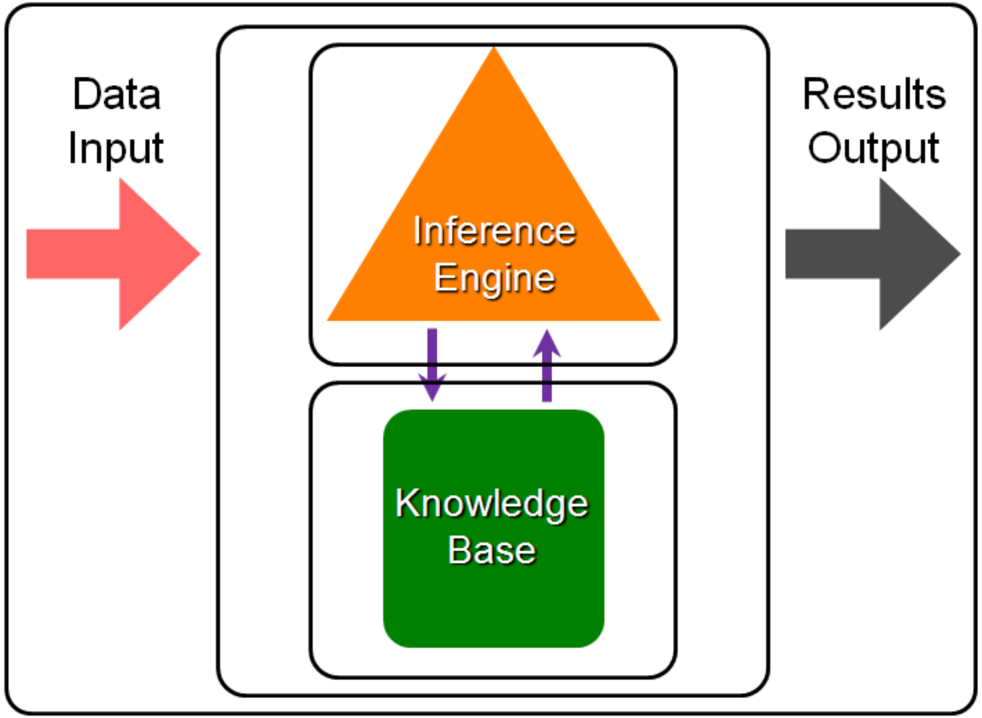


Agenda

1. Introduction
   - Dialysis
   - Research Questions and Objectives
2. Methodology
   - MIMIC-III
   - Algorithms SVR and LPR
   - Preprocessing with *rapidminer*
   - Optimization Challenges
3. **Preliminary Results**
4. Discussion

Preliminary Results

Preliminary Results



Bonney (2011)

Preliminary Results

- Predicting Dialysis length is important for quality of care
- But: it is not an easy task
- LPR works better than SVM so far
- Best correlations achieved:
  - SVM:
    - 0.682 on small subset (n=79)
    - Polynomial Kernel deg 2, C=50, ε=20
  - LPR:
    - 0.877 on whole dataset (n=2047)
    - 11/15 features selected, no normalization, weighting: fixed distance

Roadmap

Roadmap

- improve SVR performance:
  - Different parameters
  - Different datasets
- Gold standard:
  - Train model on subset of „good patients“
  - Apply model to all patients
  - Compare the results

Thank you for your attention!

Adrian Loy, Antje Schubotz

2 February 2017

Further Information

Further Information

- MIMIC-III: https://mimic.mit.edu/help/
- SVM: Chih-Wei Hsu, e. a.: A practical guide to SVMs
- SVR: Schölkopf, A tutorial on Support Vector Regression
- LPR: Avery, "Literature Review for Local Polynomial Regression"
- *rapidminer*: https://rapidminer.com/