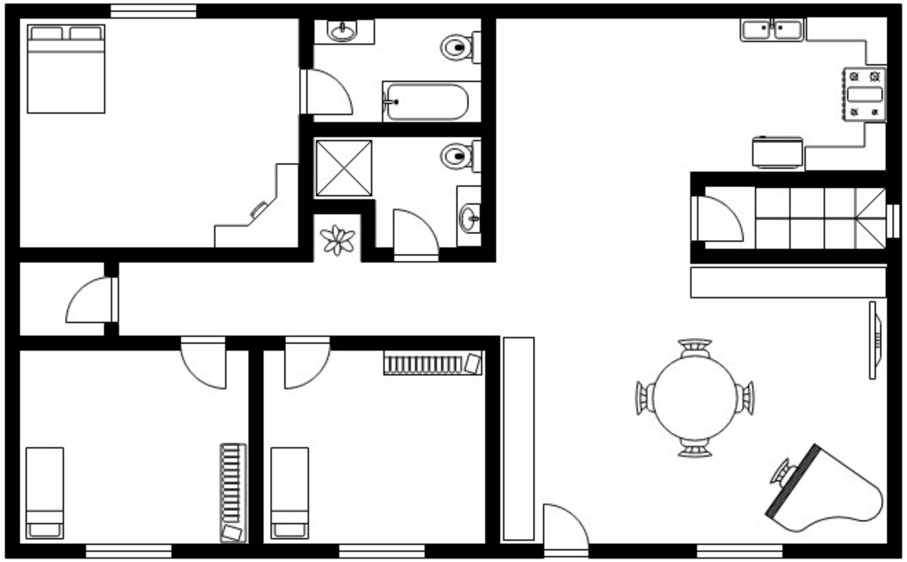
Machine Learning

- Learning from Data
- Models
- Features and Targets
- Dimensions of Machine Learning

Regression

Donnerstag, 13. Januar 2022 14:25

Estimating Appartment Prices



Features? -----------------------> Target: Price

Regression: Predicting a real number value

Classification

Donnerstag, 13. Januar 2022 14:37

Kind of Iris Plant

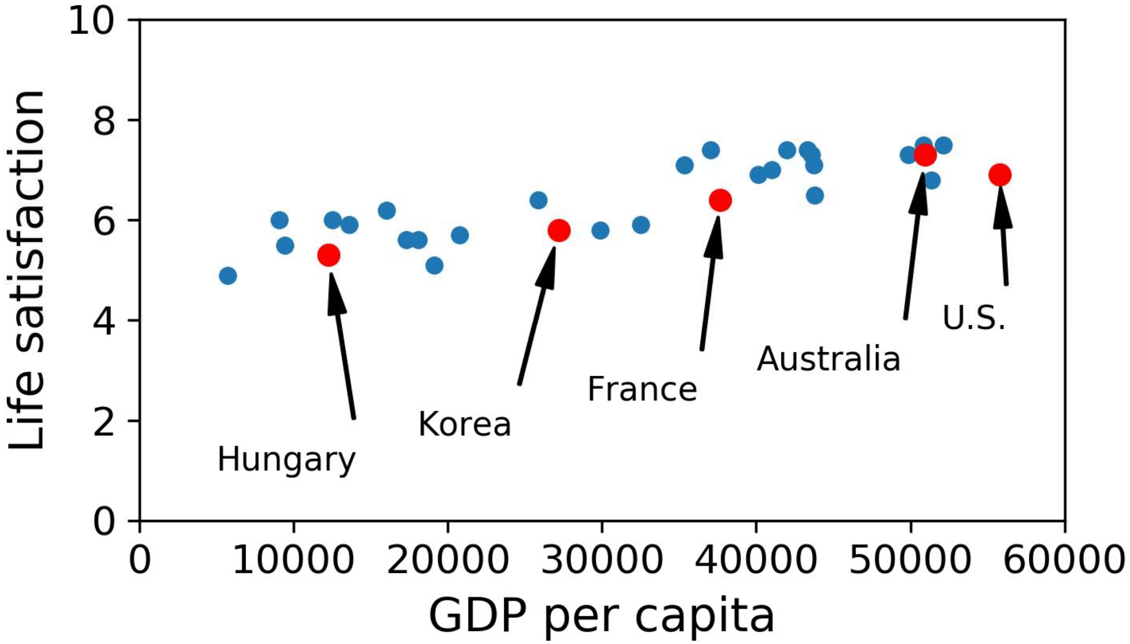
Features? -----------------------> Target: Kind

Classification: Predicting a label

Life Satisfaction

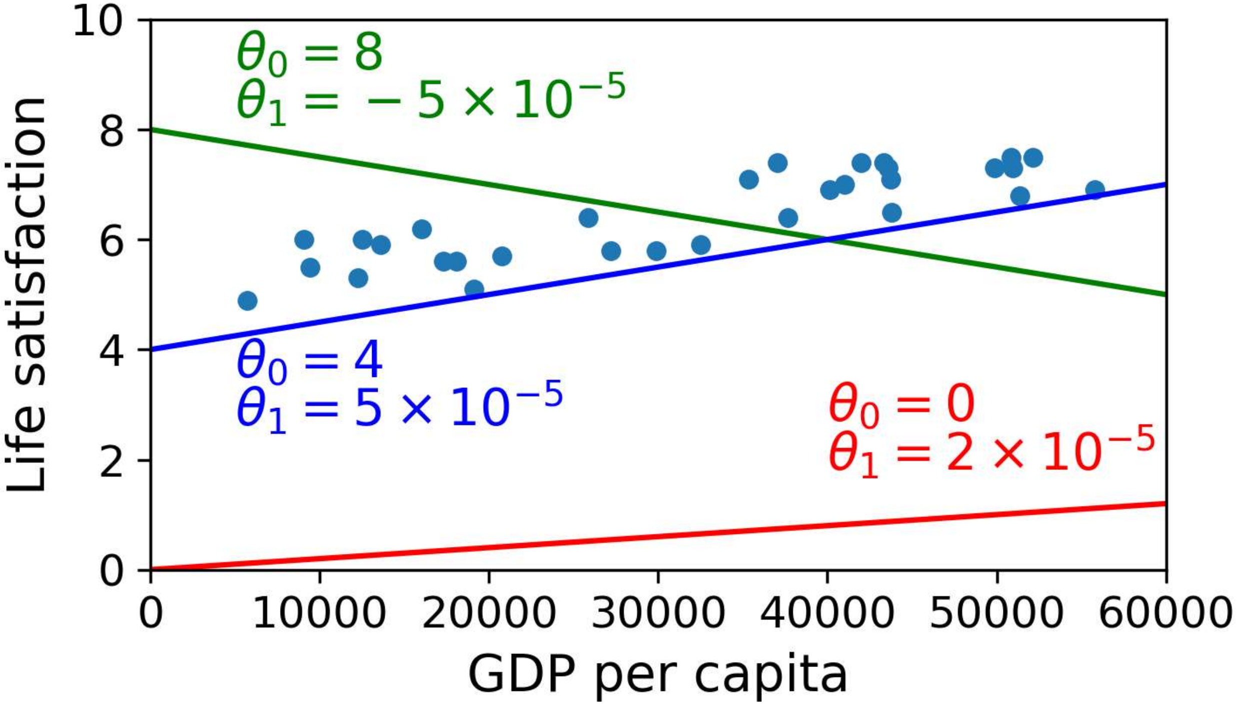
Donnerstag, 6. Mai 2021 09:58

Is there a trend?



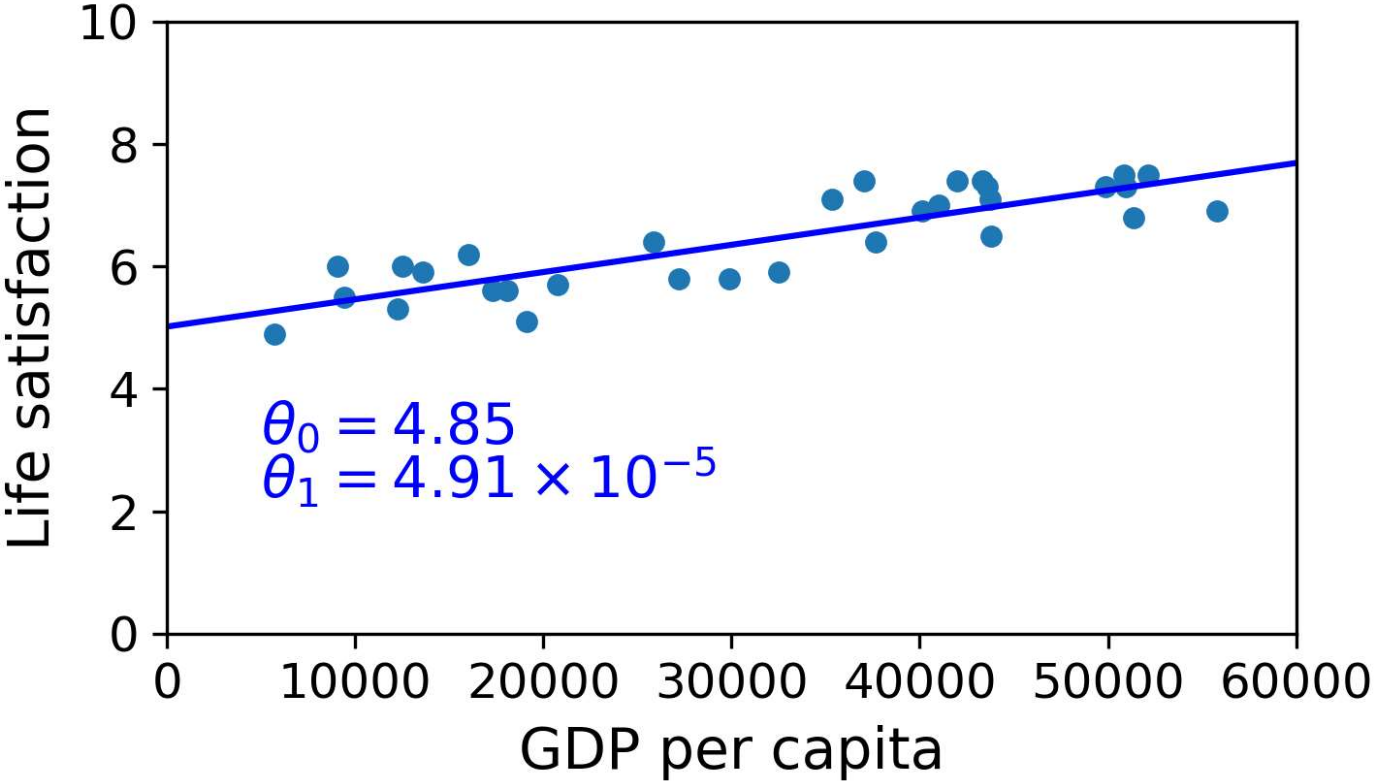
Linear model: $sat = \theta_0 + \theta_1 * gdp\_per\_capita$

Finding good models?
How is "good" defined



Best Model via Linear Regression

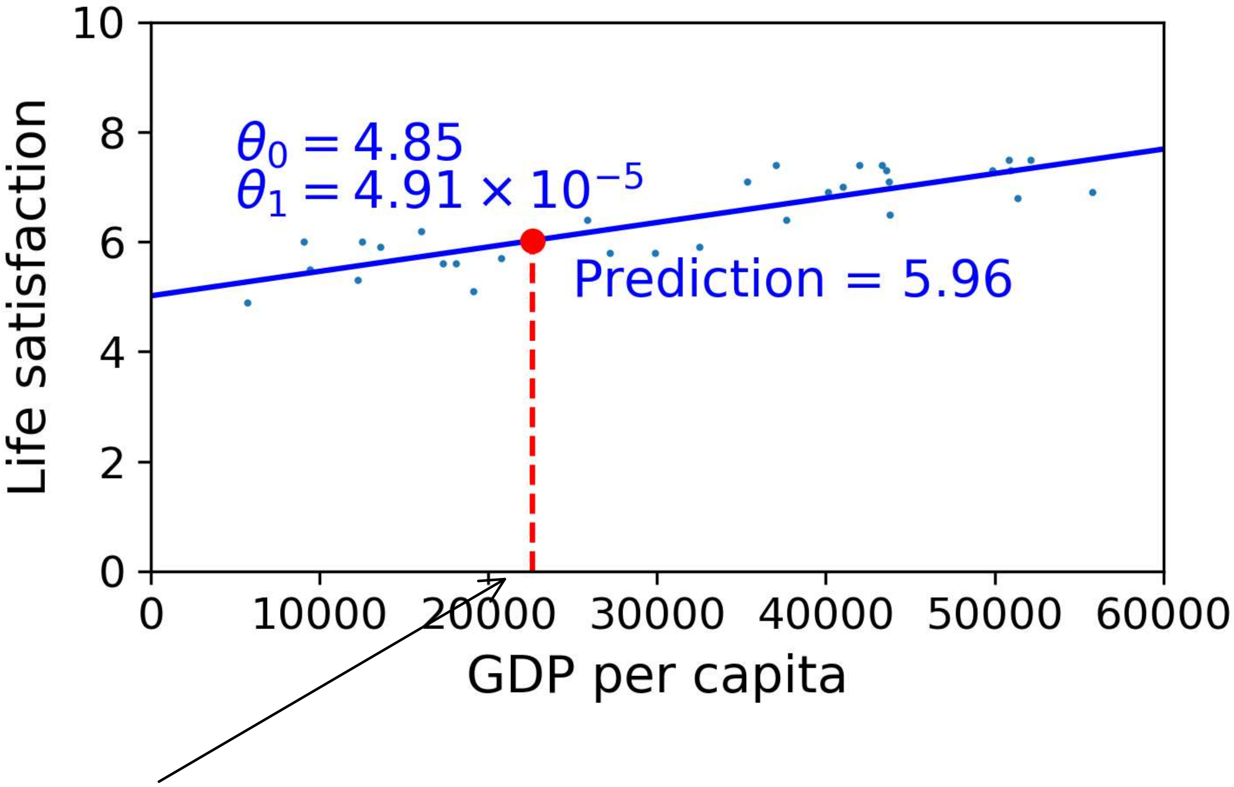
Donnerstag, 6. Mai 2021 10:19



Prediction

Donnerstag, 6. Mai 2021 10:30

Prof. Dr. Ingo Claßen



Take GDB per capita and predict sat

Learning from Data - Problems

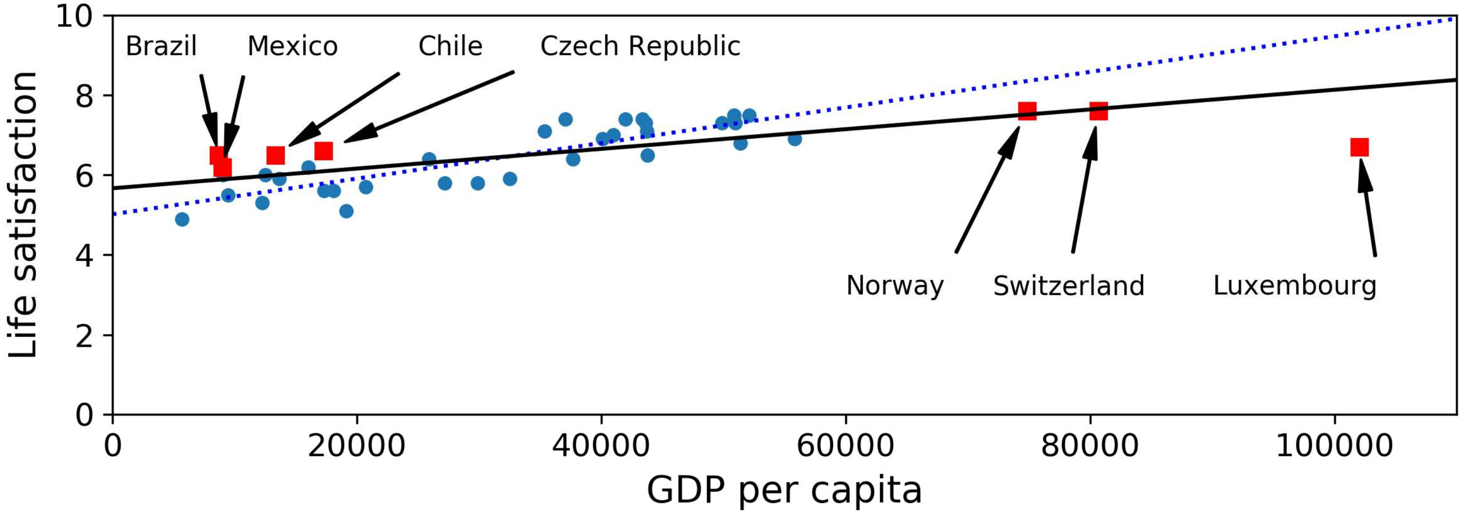
Donnerstag, 6. Mai 2021 10:21

- Not enough data
- Data is not representative
- Bad data quality
- Irrelevant features
- Overfitting
- Underfitting

Data not representative

Donnerstag, 6. Mai 2021 10:23

Adding Brasil, Mexico, …, Norway, … changes model
Linear model does not really fit

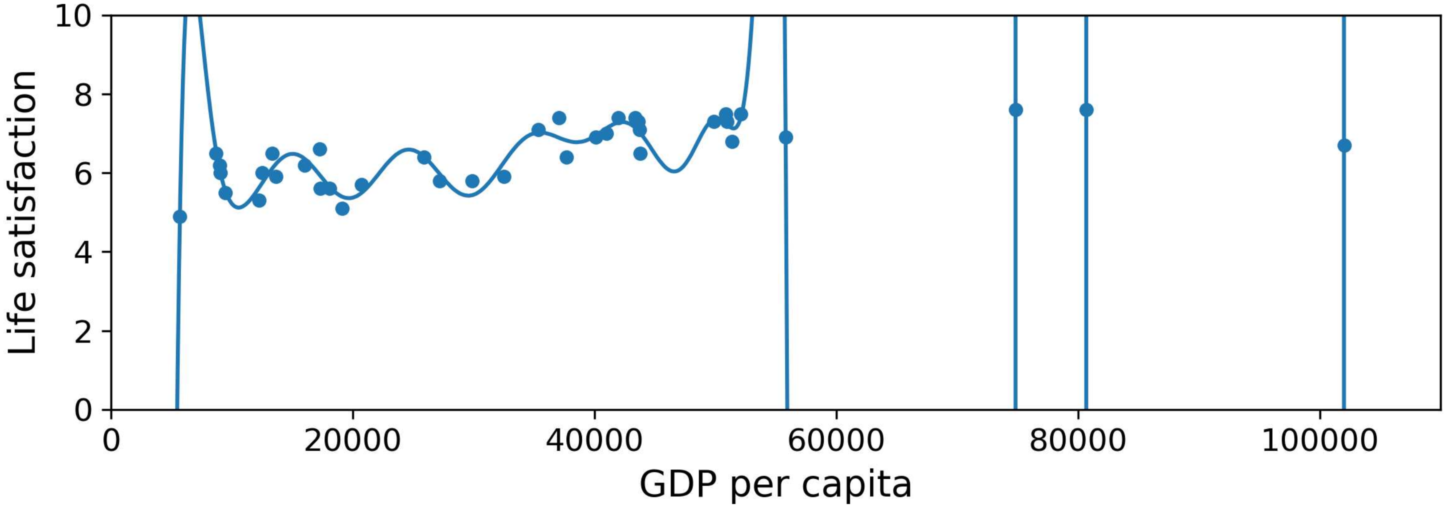


Countries with high GDP per capita but lower sat
Countries with low GDP per capita but higher sat

Overfitting

Donnerstag, 6. Mai 2021 10:27

Komplex model
Polynomial with high degree



Goal:

- Find relationship between features and targets
- Discover patterns
- Find correlations

Common terms:

- Machine Learning
- Data Mining
- Statistical Learning

Reasons why learning from data

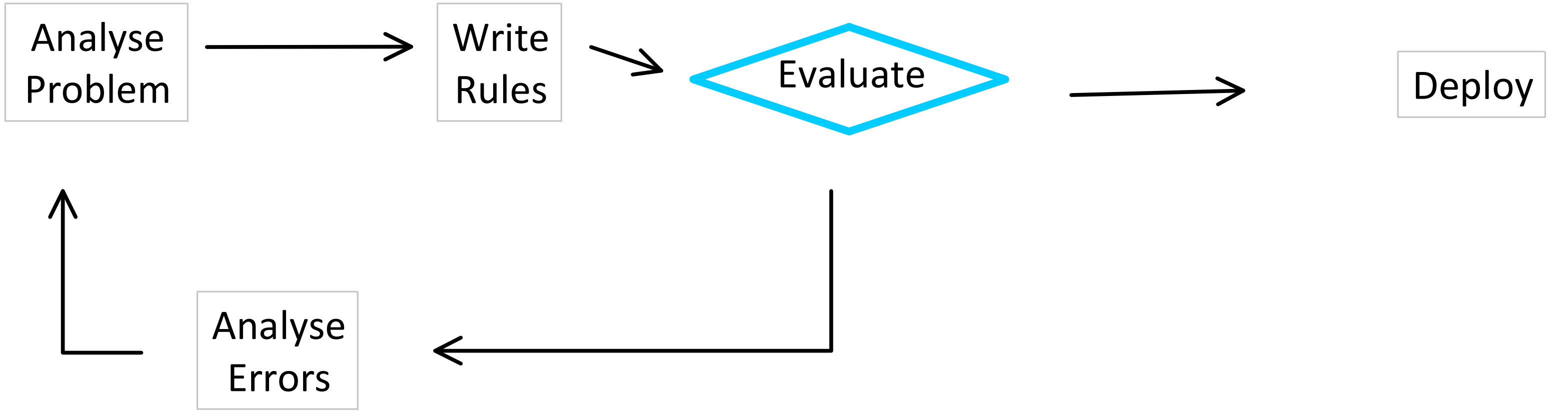
- Constructing an abstraction for the relationship between features and targets (model)
- Complex problems where algorithmic formulation is hard
- Environments where data changes continuously
- Problems that require a long list of complicated rules

Example Spam Filter - Rule-based

Prof. Dr. Ingo Claßen

Mittwoch, 5. Mai 2021 10:14

Analysis of spam emails, identifiction of offensive words and phrases

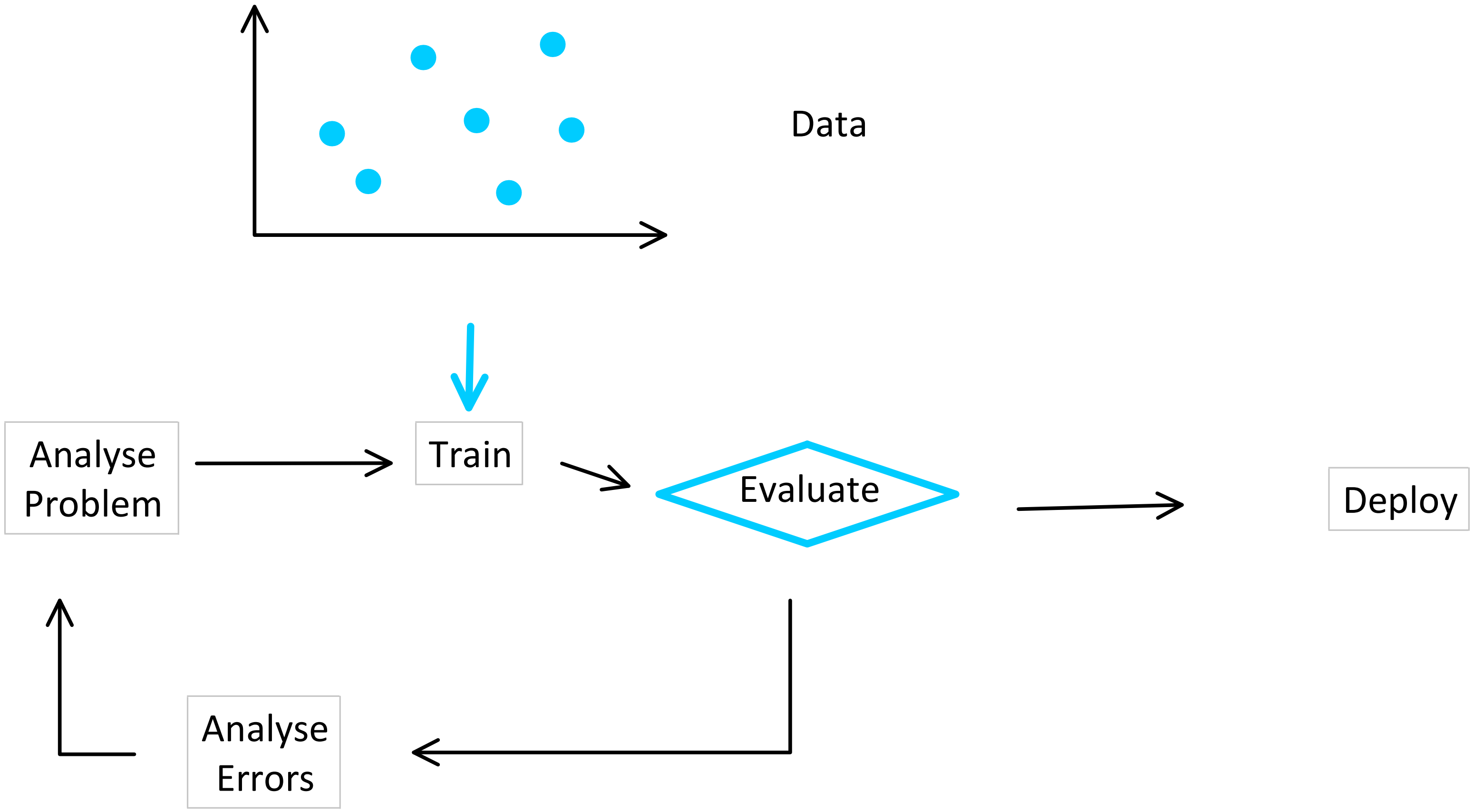


Long list of rules - not easily maintainable
Adaption to changes are hard to do

Example Spam Filter - Learned from Data

Mittwoch, 5. Mai 2021 10:27

Identification auf words and phrases by learning from data using suitable algorithms.



Data

Analyse Problem

Train

Evaluate

Deploy

Analyse Errors

Adaption to changes much easier: just take new data and learn again

Features and Targets

Mittwoch, 30. Juni 2021 08:27

Features

- Attributes that characterize a particular instance/object
- Also called predictors

Targets

- Scalar value in case of regression
- Nominal value in case of classification

Features / Targets

Regression

Bedrooms	Bathrooms	Latitude	Longitude	Price
1	1	40.71	-73.94	3055
…	…	…	…	…

Classification

Income	Job	IsMarried	Age	Loan
…	…	…	…	Yes
…	…	…	…	No

Feature values, vectors, matrix

Mittwoch, 30. Juni 2021 08:34

x1	=
x2	=
…	
xm	=

x11	x12	x13	…	x1n
x21	x22	x23	…	x2n
…				
xm1	xm2	xm3	…	xmn

y1
y2
…
ym

X

y

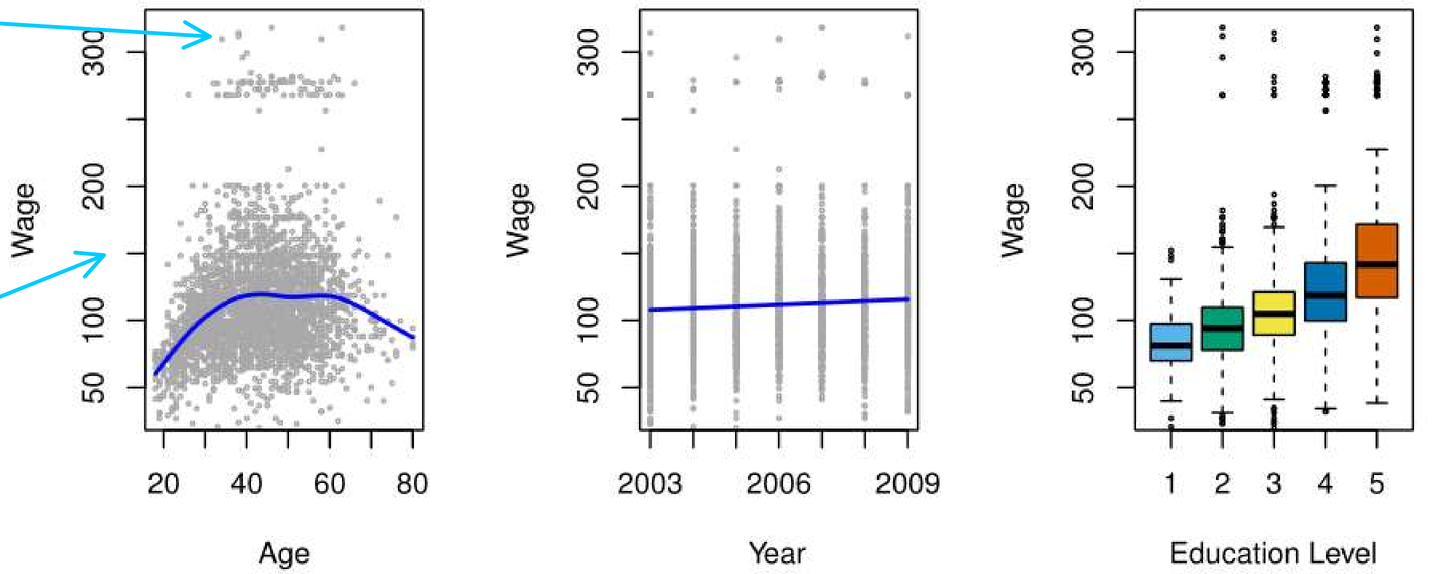
xi	Feature vector
xij	Feature value
yi	Target value
X	Feature matrix
y	Target vector

- All entries must be numbers, i.e. all feature and target values must be transformed into numbers
- Each feature vector is one instance/object

Visualization of Features and Targets

Mittwoch, 5. Mai 2021 10:14

each point is one person (object, instance)



Scatterplot

Boxplot

Learn from Training Data

Mittwoch, 30. Juni 2021 08:53

Learning should

- be as accurate as possible
- generalize to new data as good as possible

Conflicting goals!

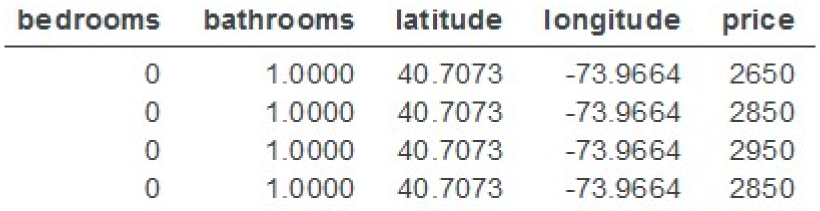
Learning Regimes

Prof. Dr. Ingo Claßen

Freitag, 18. Juni 2021 09:41

Just remember target value wrt. feature combination

- Good accuracy
- Bad generalization
- Combinations might not exist
- Combinations might not be unique



bedrooms	bathrooms	latitude	longitude	price
0	1.0000	40.7073	-73.9664	2650
0	1.0000	40.7073	-73.9664	2850
0	1.0000	40.7073	-73.9664	2950
0	1.0000	40.7073	-73.9664	2850

Build coarse grained groups, e.g. price per bedrooms and bathroom combinations

- Bad accuracy
- Good generalization

Bedrooms	Bathroom	Price
1	1	3000
2	1	3700
1	2	3010
…	…	…

Learn from neighbors
(k nearest neighbor)

- Powerful
- But must store complete training data set
- No real learning
- Finding nearest neighbors requires search, might be slow

Build a linear model:

Price = w0 +
w1*bedrooms +
w2*bathrooms +
w3*latitude +
w4*longitude

Might be not suitable: higher number of bathrooms doesn't mean higher price

Dimensions of Machine Learning

Mittwoch, 5. Mai 2021 10:34

Supervision

- Supervised
- Unsupervised
- Semisupervised
- Reinforcement Learning

Data Usage

- Batch
- Online

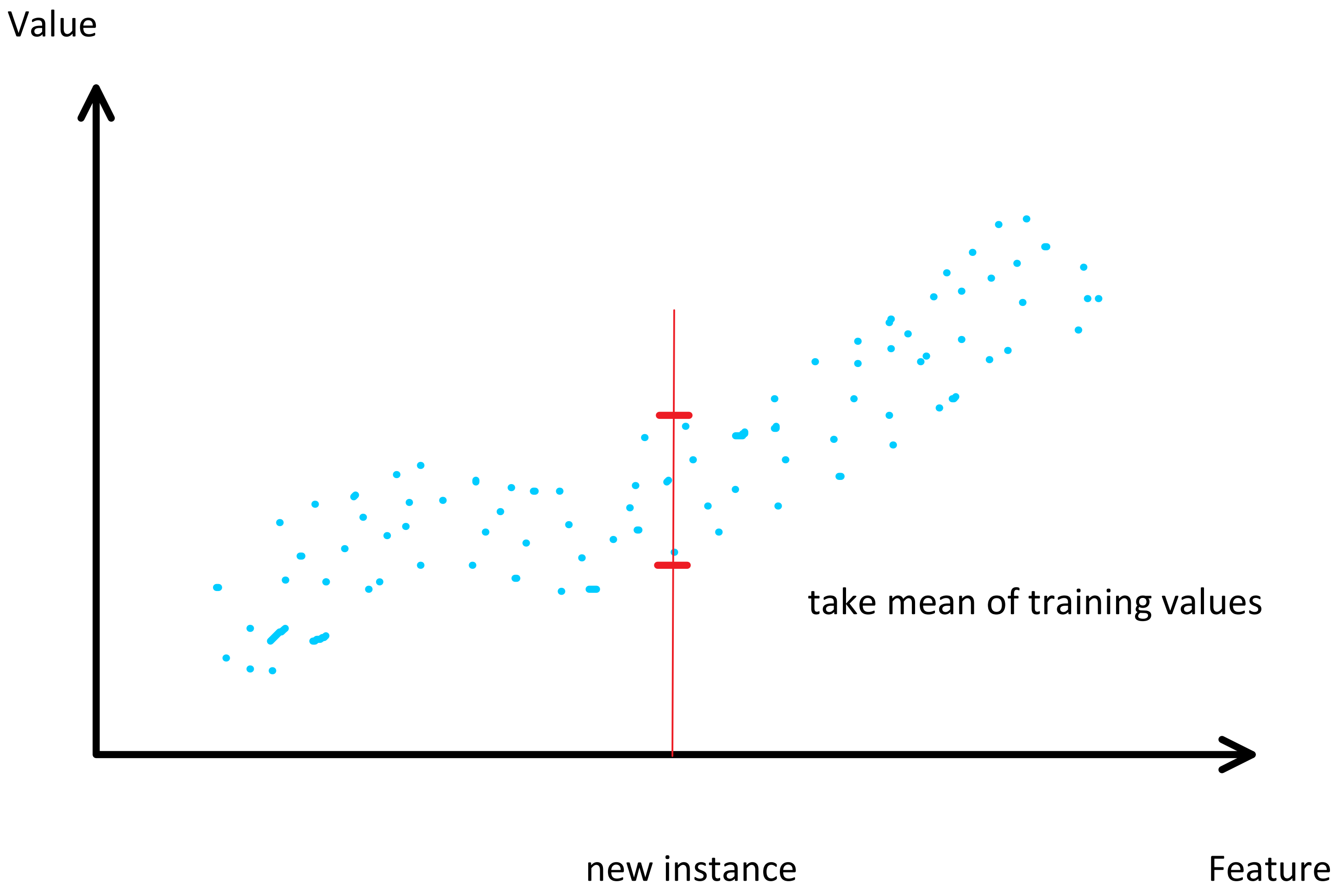
Method

- Instance-Based
- Model-Based

Supervised - Regression

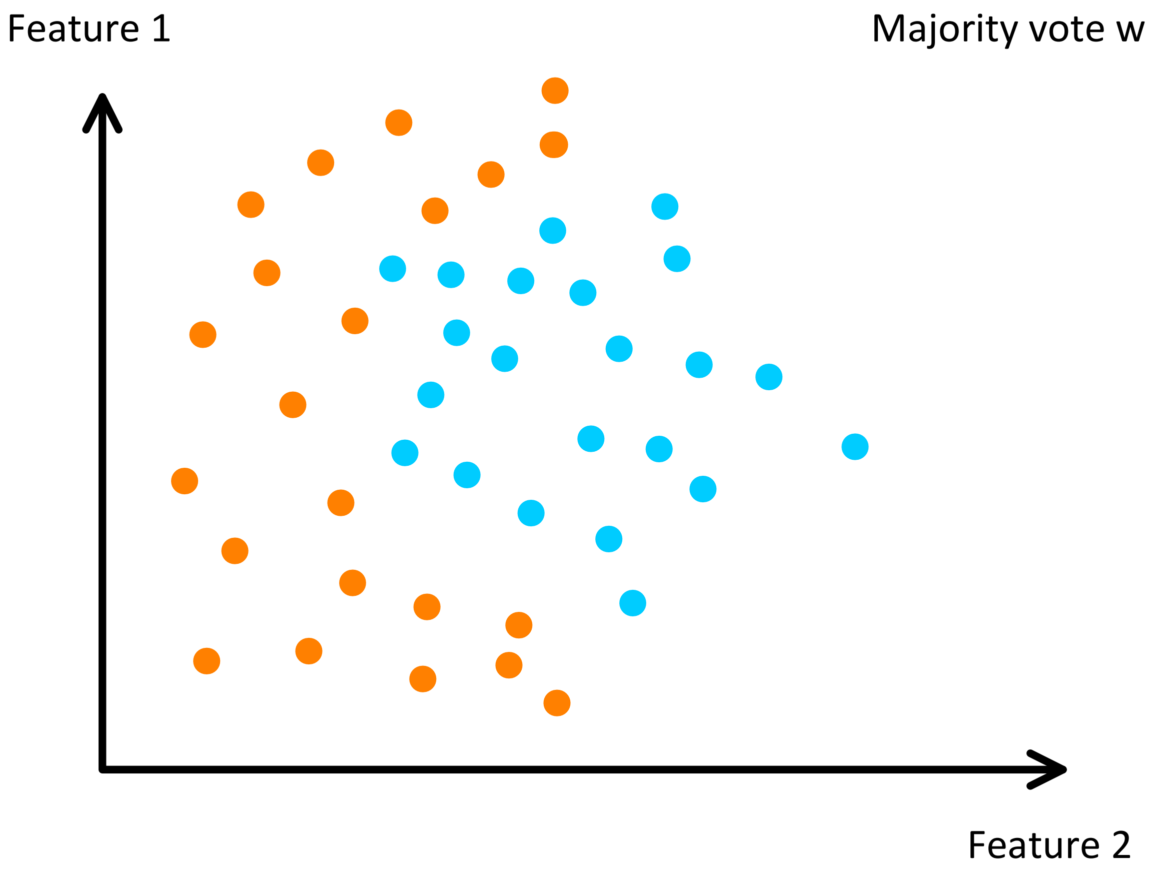
Mittwoch, 5. Mai 2021 10:42

Features of instances + target values



Supervised - Classification

Mittwoch, 5. Mai 2021 10:37



Majority vote wrt. neighbors

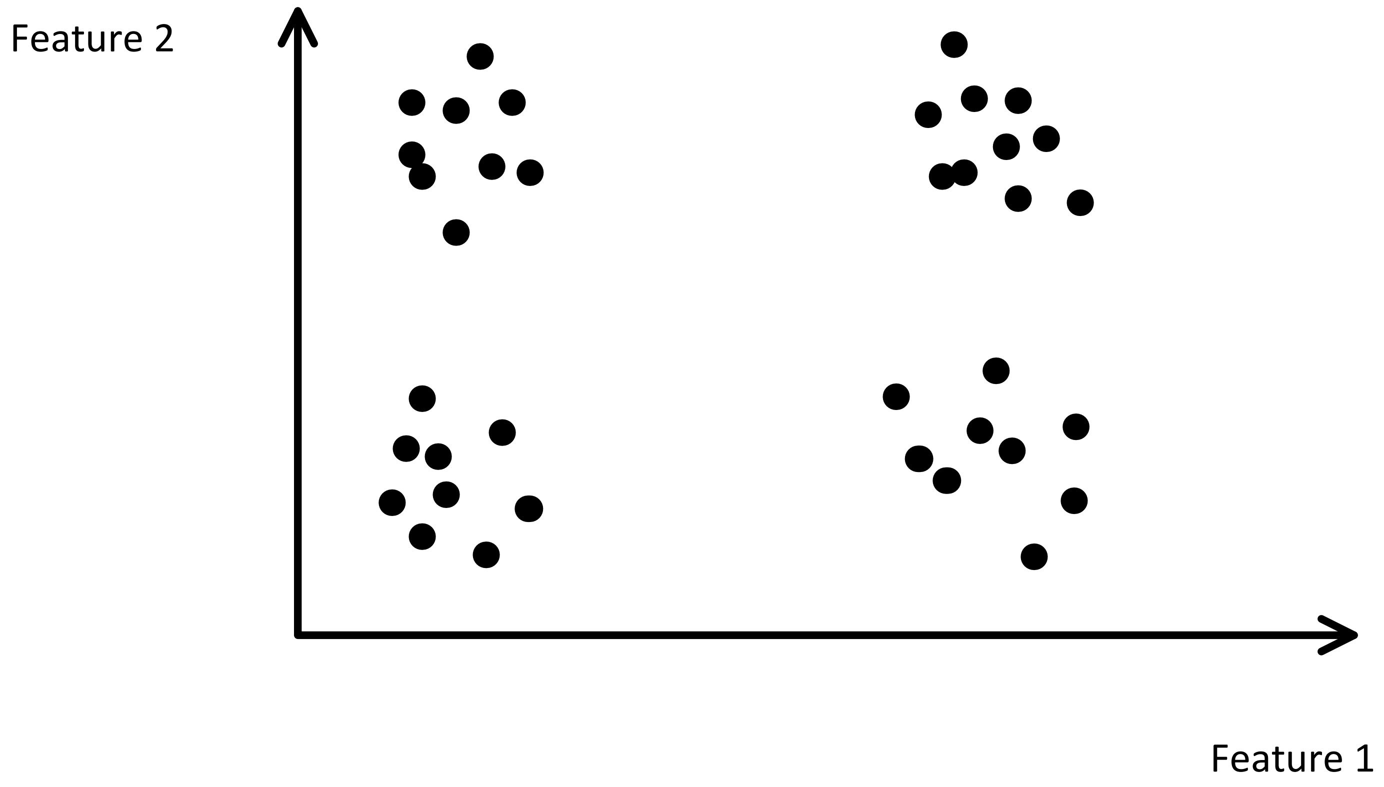
Class tags:

- Class1: orange
- Class2: blue

Unsupervised - Clustering

Mittwoch, 5. Mai 2021 10:47

Features of instances and no class tags



Find sets of data with strong coupling
and large distance between sets

Unsupervised - Anomaly Detection

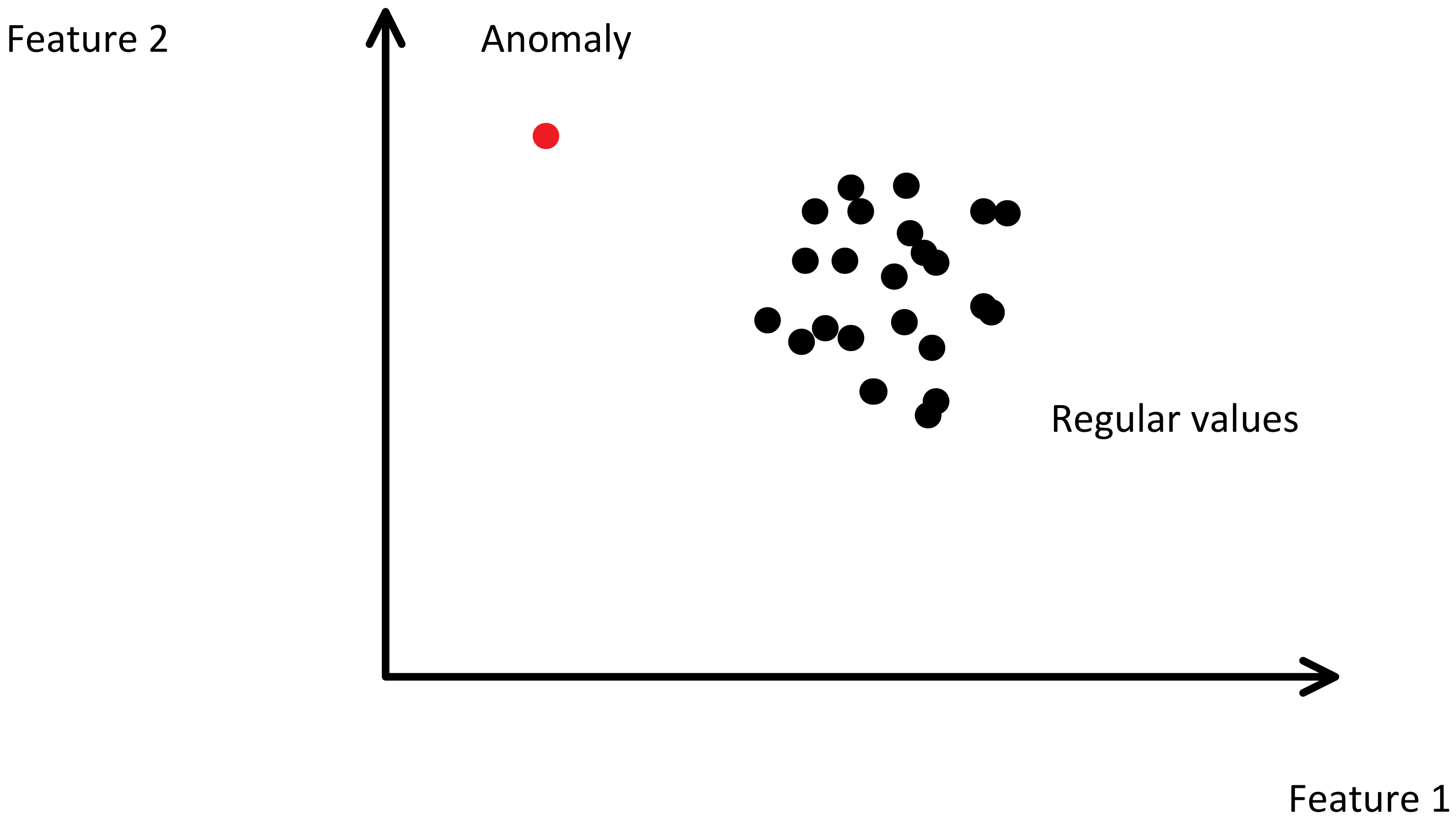
Prof. Dr. Ingo Claßen

Mittwoch, 5. Mai 2021 10:49

Features of instances

Compute cluster of regular instances

Calculate distance



Unsupervised - Association Learning

Mittwoch, 5. Mai 2021 10:53

Market Basket Analysis
Learning of rules

{ Milk, Bread } -> { Butter }

{ PC, Monitor, Graphics Card } -> { PC Game }

Friday evening
{ Beer } -> { Diapers }

Semi Supervised

Mittwoch, 5. Mai 2021 10:57

Example: Identification of persons on your private fotos

- Upload of fotos
- System detects same persons on different fotos (unsupervised)
- You tag with name (supervised)

Reinforcement Learning

Prof. Dr. Ingo Claßen

Mittwoch, 5. Mai 2021 11:00

- Agent observes environment
- Chooses action according to strategy
- Executes action
- Gets reward or punishment
- Learns from feedback - adopts strategy
- Repeat until optimal solution found (or at least a satisfying solution)

Data Usage

Mittwoch, 5. Mai 2021 11:04

Batch Learning

- Take all data at once and learn from it
- Complete new take if data changes
- Requires long execution time and ressources in case of large training data set

Online Learning

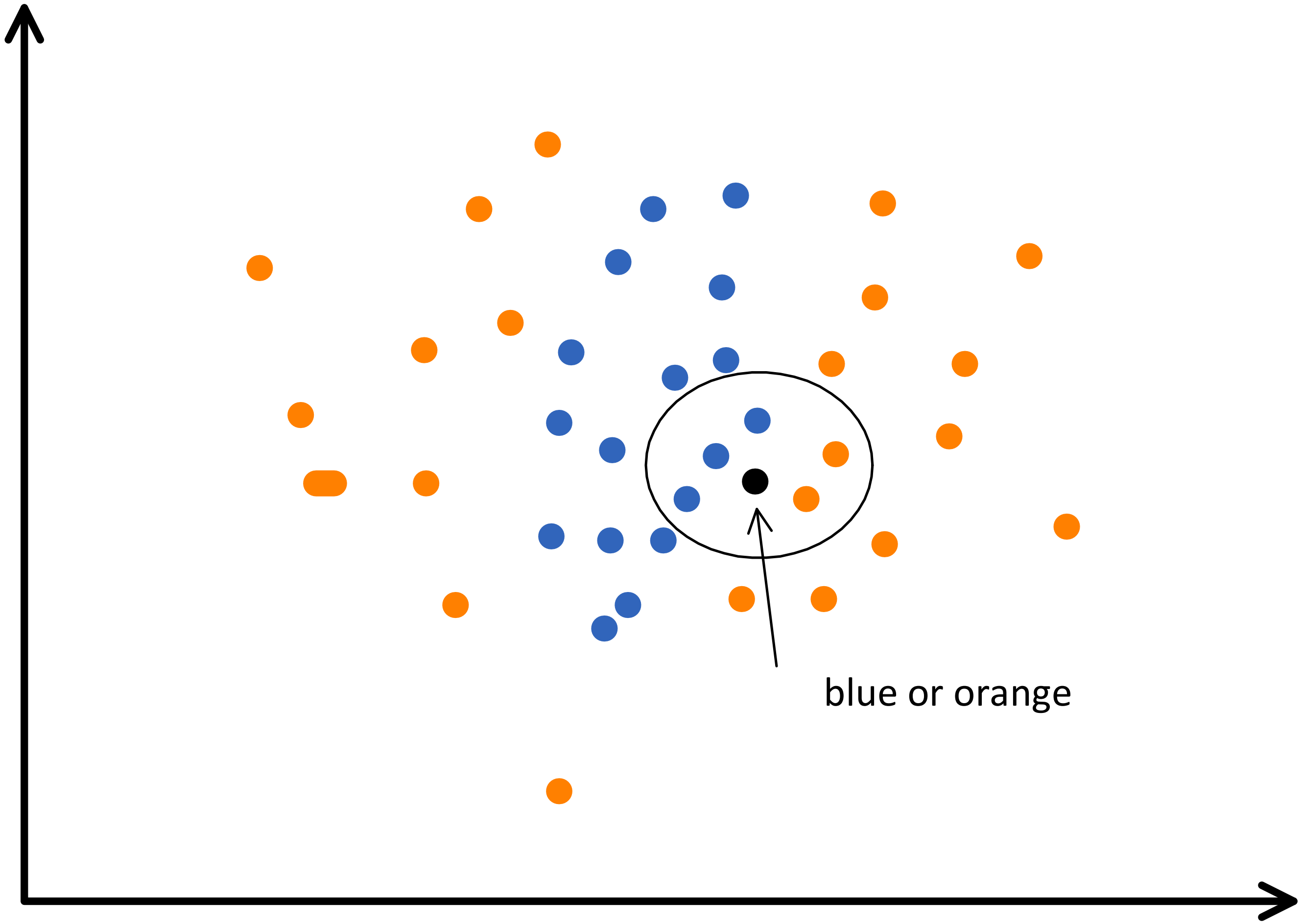
- Incremental usage of data
- Single data points or minibatches
- Faster adoption to new data
- More volatile models

Instance-Based Learning

Prof. Dr. Ingo Claßen

Mittwoch, 5. Mai 2021 11:09

Similarity to training data



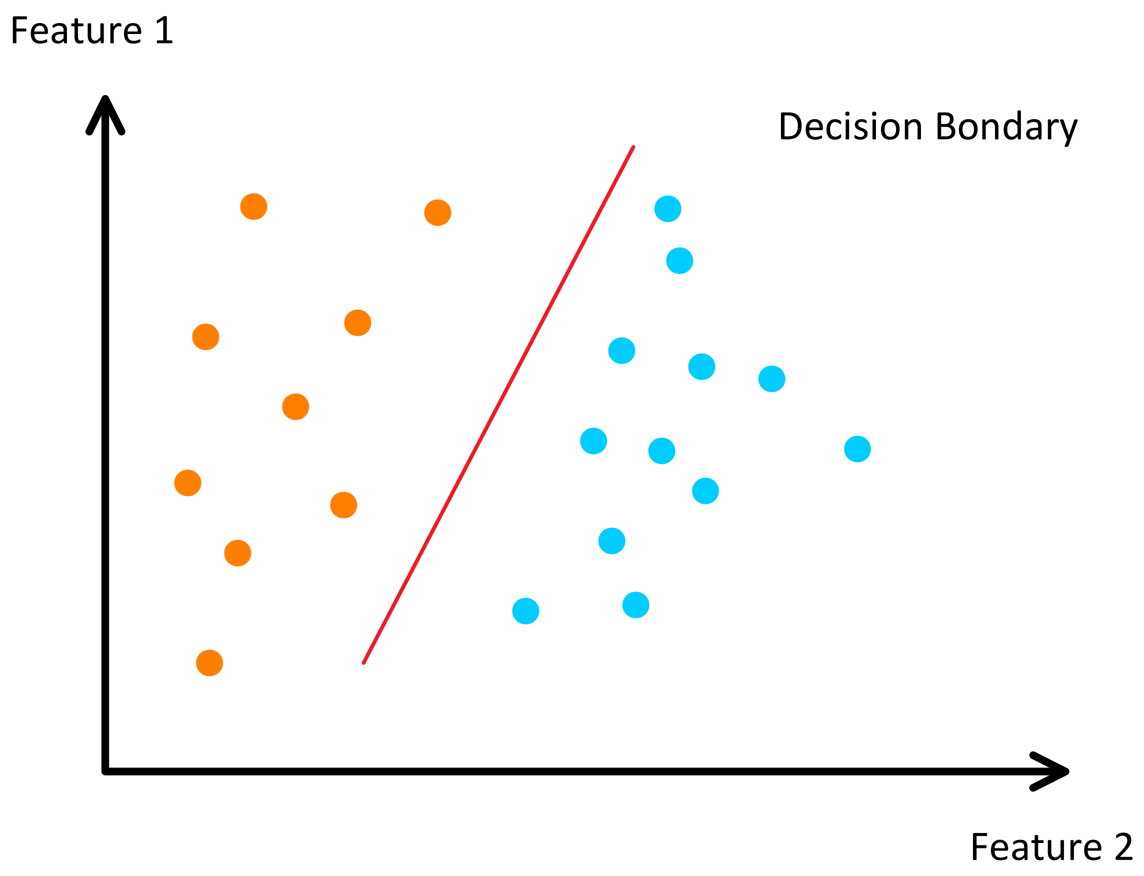
Keep all training data

Use distance metrics

Model-Based Learning

Prof. Dr. Ingo Claßen

Donnerstag, 6. Mai 2021 08:20



No storage of trainings data
Just keeping model parameters, e.g. slope and intercept of separating line

References

Donnerstag, 6. Mai 2021 10:20

Parr, Howard: The Mechanics of Machine Learning, https://mlbook.explained.ai/

James, Witten, Hastie, Tibshirani: An Introduction to Statistical Learning,
Springer, 2013, http://www.statlearning.com/

Géron: Hands-On Machine Learning with Scikit-Learn and TensorFlow,
O'Reilly, 2017, http://shop.oreilly.com/product/0636920052289.do