

Supervised Machine Learning Methods

Konrad U. Förstner

ZB MED – Information Centre for Life Science & TH Köln

Workshop *Systems Biology: From large datasets to biological insight*

2021-05-21



1 Introduction

2 Concepts and terminology

3 Selected supervised learning methods

4 Summary

1 Introduction

2 Concepts and terminology

3 Selected supervised learning methods

4 Summary



After the lecture you should have a basic understanding of supervised machine learning approaches and potential applications in research.

After the practical part you should be able to implement them with Python and the package scikit-learn.

We will not cover the mathematical background. This is not needed at this level but recommended later.

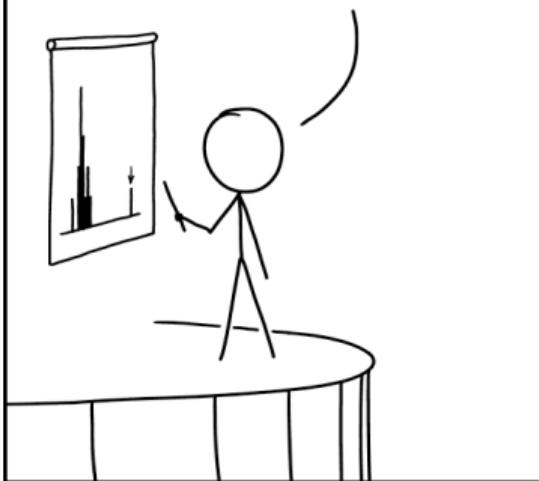


<https://www.pexels.com/photo/photo-of-code-projected-over-woman-3861969/> – CC0



[https://commons.wikimedia.org/wiki/File:Tools_and_tools_\(Unsplash\).jpg](https://commons.wikimedia.org/wiki/File:Tools_and_tools_(Unsplash).jpg) - CC0

DESPITE OUR GREAT RESEARCH
RESULTS, SOME HAVE QUESTIONED
OUR AI-BASED METHODOLOGY.
BUT WE TRAINED A CLASSIFIER
ON A COLLECTION OF GOOD AND
BAD METHODOLOGY SECTIONS,
AND IT SAYS OURS IS FINE.



Artificial Intelligence

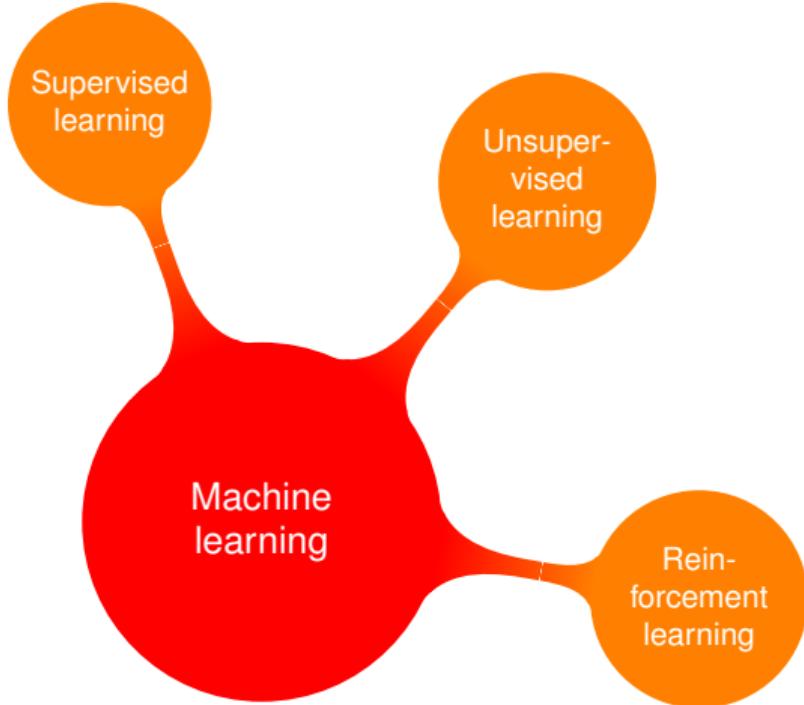
Methods that let machines immitated human behaviour.

Machine Learning

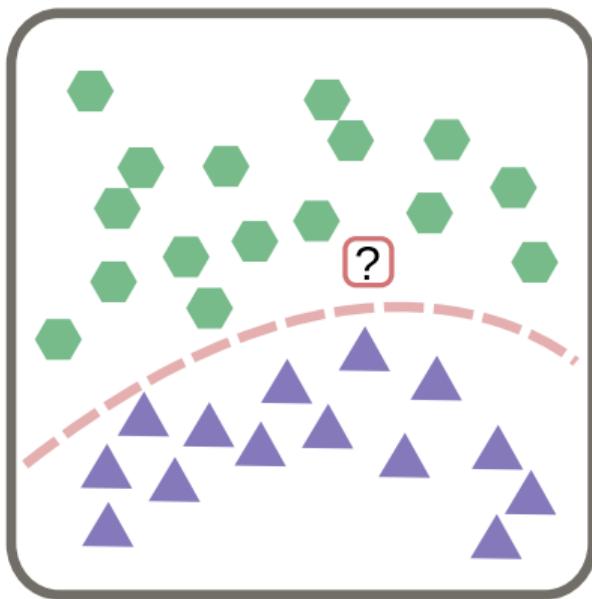
Programs learn based on data without being explicitly programmed.

Deep Learning

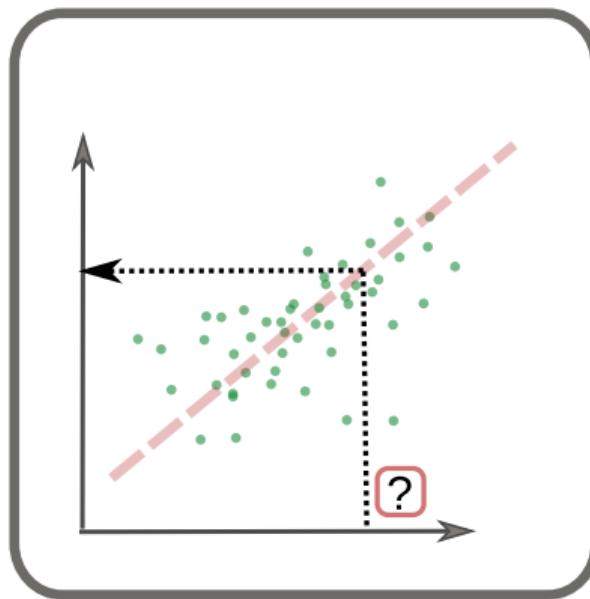
Machine learning based on artificial neural networks with many layers.



Two types of tasks that can be solved with supervised learning

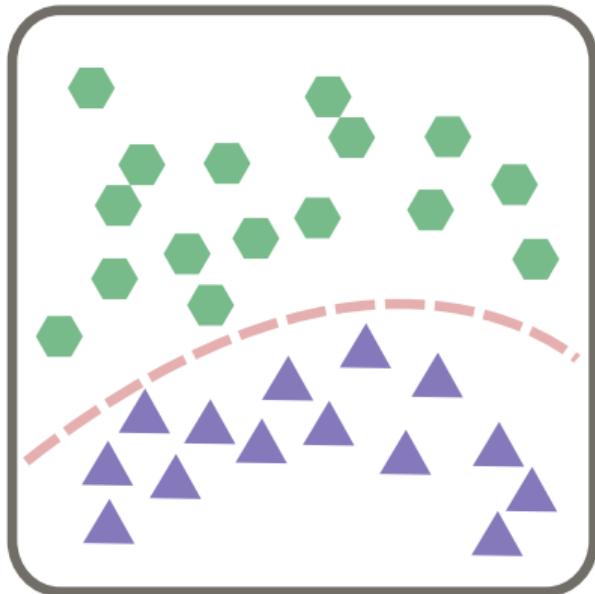


Classification

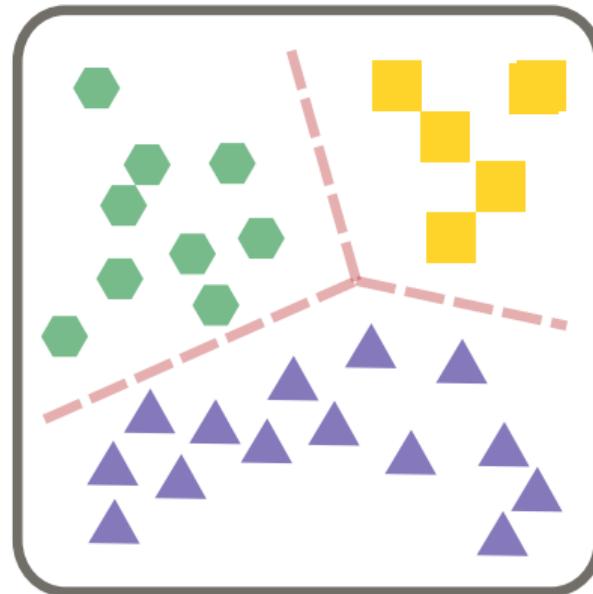


Regression

Classification types



Binary classification



Multi-class classification



Supervised learning means to generate models that generalize from given examples.

Basic concept of supervised machine learning

The model / function maps from a given two-dimensional matrix X
to an output vector y with labels (classification)
or numerical values (regression).

$$X_1 \rightarrow y_1$$

$$X_2 \rightarrow y_2$$

$$X_3 \rightarrow y_3$$



In the actual training / learning process the parameters of the model / function are estimated. The model is then able to project the input variable X to the output variable y .

$$y = f(X)$$

Example of classification



Cancer classification based on single-cell
gene expression data.

Example of regression



Predicting the gene expression level of a gene based on the gene expression levels of several regulators.

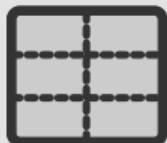
1 Introduction

2 Concepts and terminology

3 Selected supervised learning methods

4 Summary

Entities and their features

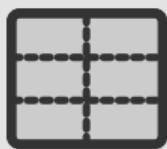


Entities (aka. samples, data points) are described by **features** (aka. covariates, attributes) that have **values**.

E.g. for different cell lines (entities) the relative expression (values) of several genes (features).

Entities and their features

Features can be



- categorical
 - Nominal (e.g. cell line, cancer type, eye color, gender)
 - Ordinal (e.g. very bad, bad, good, very good)
- numerical
 - Discrete (e.g. gene length in nucleotides, number cells)
 - Continuous (e.g. cell length, concentration, relative expression)

Feature selection

Choosing features with high variance.

Feature A	Feature B	Feature C	Feature D	...
10.00	5.01	102.01	120	...
20.91	5.01	102.00	200	...
80.03	5.01	102.09	980	...
90.19	5.00	103.00	700	...
50.99	5.02	102.31	703	...
80.63	5.01	102.30	443	...

Feature scaling

Normalizing the feature values to their ranges e.g. min/max normalization, mean normalisation, standard score / z-score normalization.

Feature A	Feature B
4.3	537
5.3	703
2.2	510
1.5	200
5.2	760



Scaled Feature A	Scaled Feature B
0.736	0.601
1.000	0.898
0.184	0.554
0.000	0.00
0.974	1.00

Features encoding

Translating categorical values into numerical values
(e.g. via one-hot encoding)

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

e.g. AATTGC becomes:

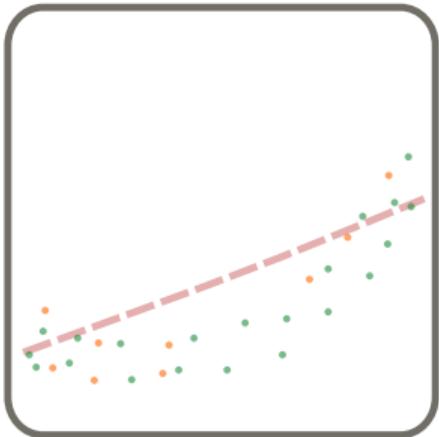
1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0

How well does the model fit?

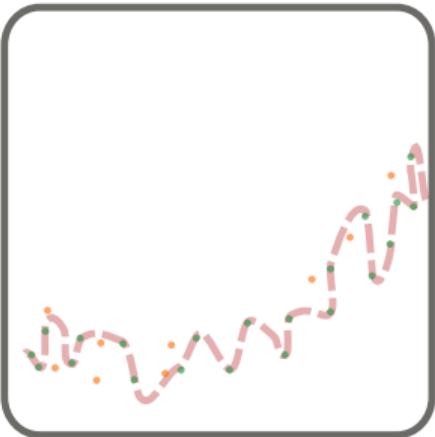
Overfitting: Good performance on the training data,
poor generalization to other data

Underfitting: Poor performance on the training data
and poor generalization to other data

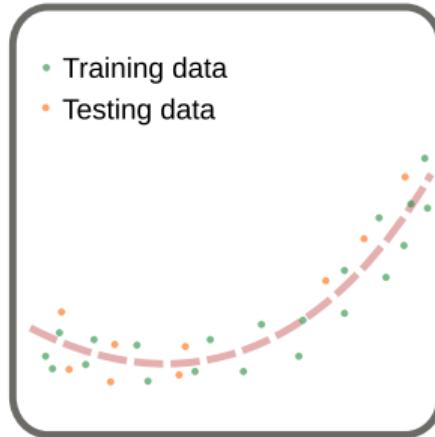
Regularization: Different methods to prevent overfitting



Underfitting



Overfitting



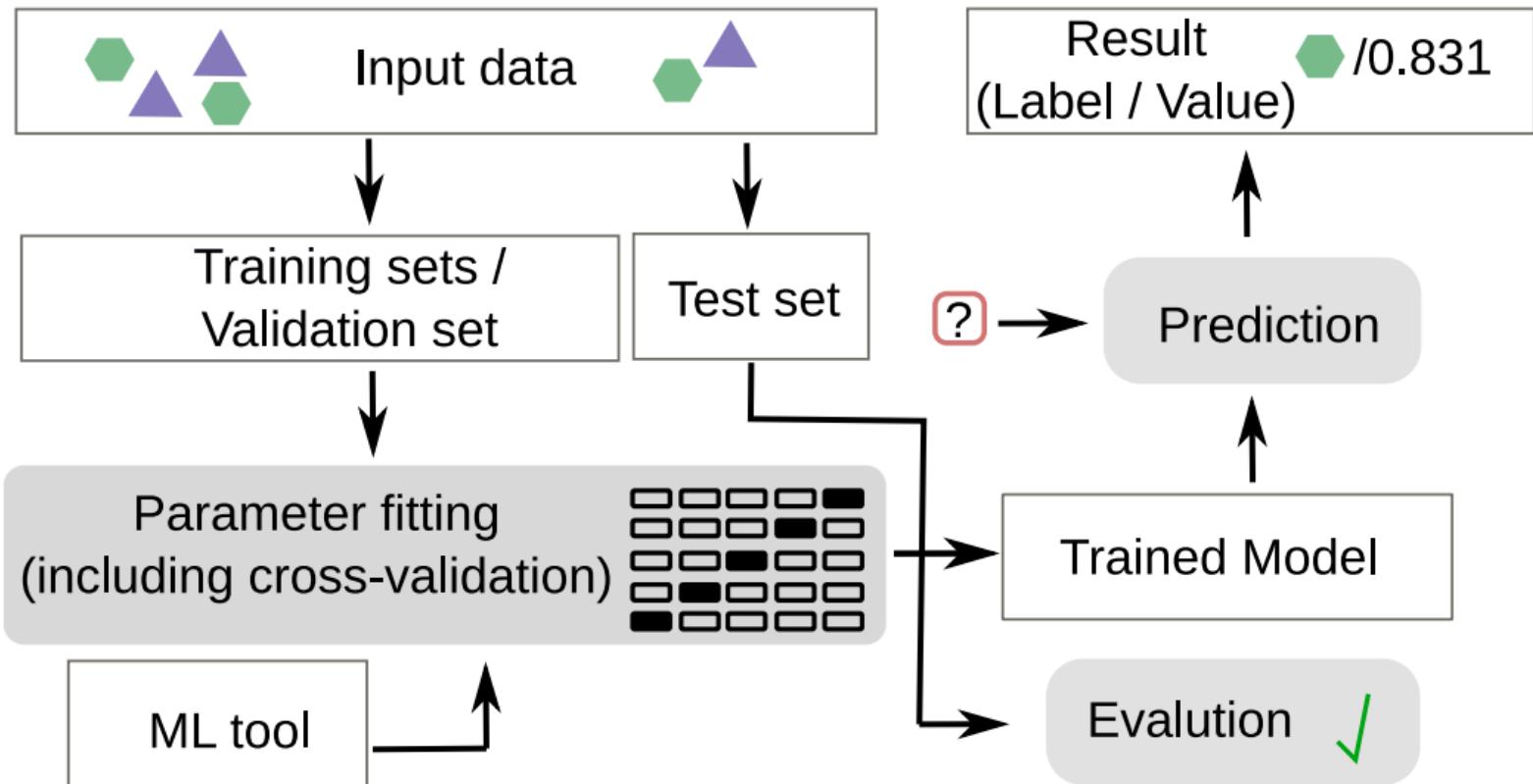
Good fit

- Training data
- Testing data

Workflow for parameter fitting and evaluation

- 1.) Split into training and test/validation set (e.g. 75%/25%)
- 2.) Train model by estimating the parameters with the training set
- 3.) Evaluate the performance by using the test/validation set
(e.g. scored as accuracy)

Workflow with cross-validation



1 Introduction

2 Concepts and terminology

3 Selected supervised learning methods

4 Summary

Overview of different methods

- K-Nearest neighbor
- Naive Bayes
- Linear Regression
- Logistic Regression
- Decision trees
- Artificial Neural Network (multilayer perceptron)
- Genetic Programming

1 Introduction

2 Concepts and terminology

3 Selected supervised learning methods

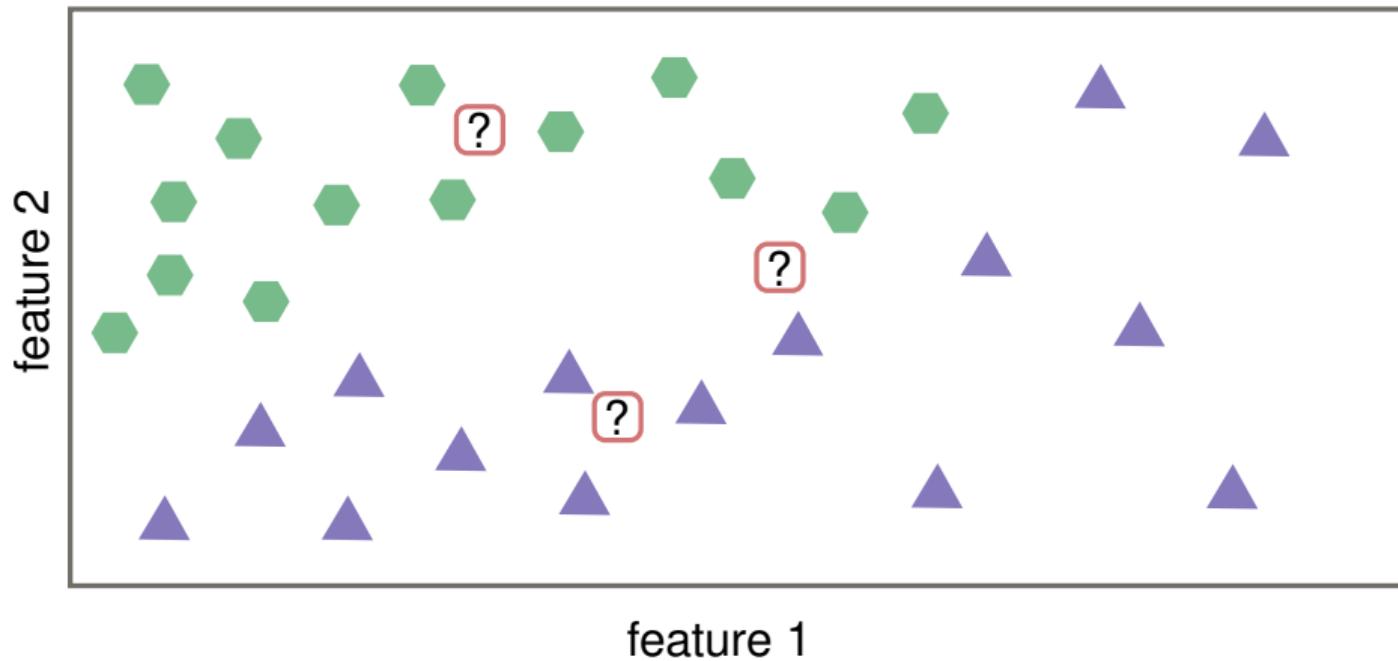
- k-Nearest Neighbors
- Linear models
- Support Vector Machines (SVMs)
- Decision Trees and Random Forest
- Artificial Neural Networks

4 Summary

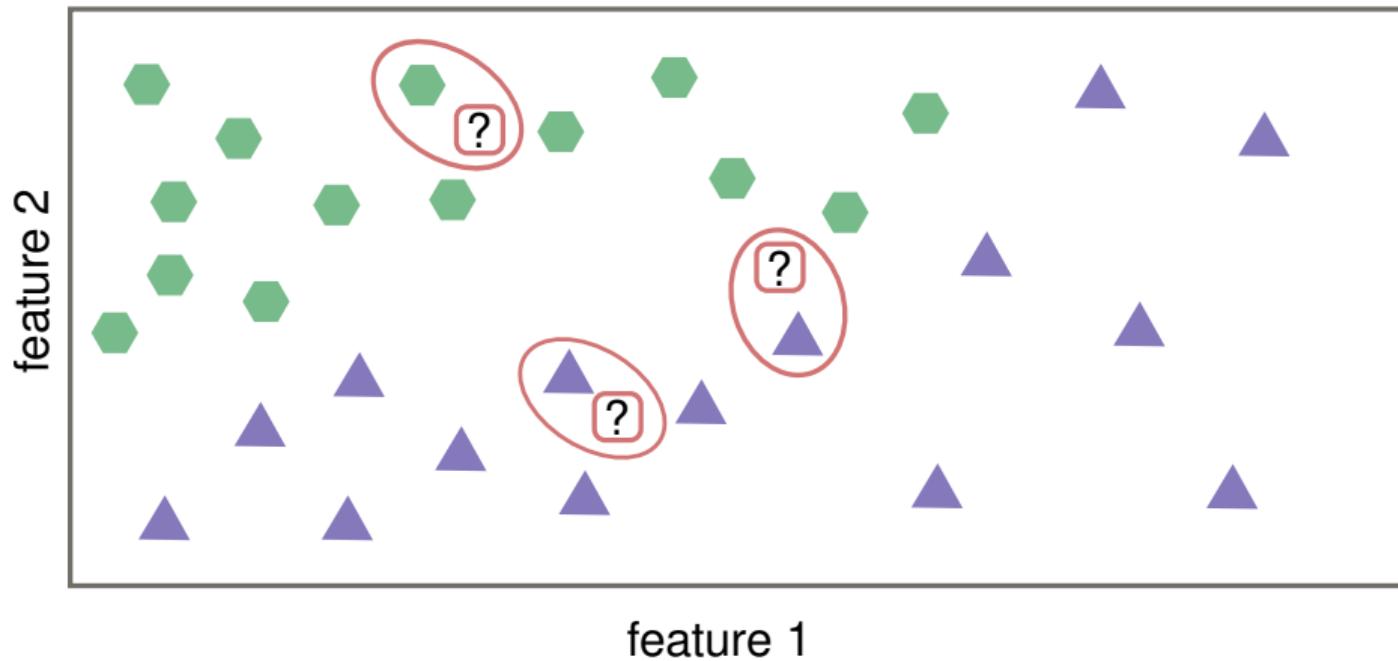
k-Nearest Neighbors

- For classification and regression
- Simplest case of supervised machine learning
- Can be easily applied to multi-class classification

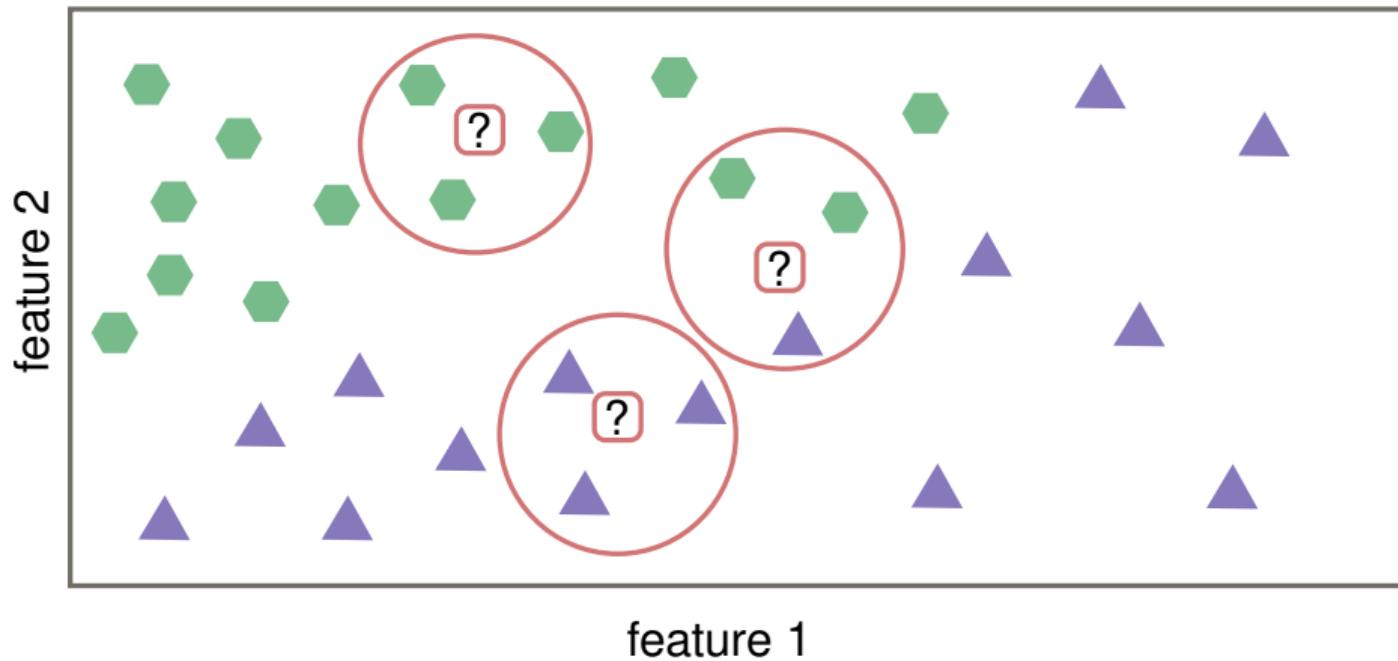
k-Nearest Neighbors



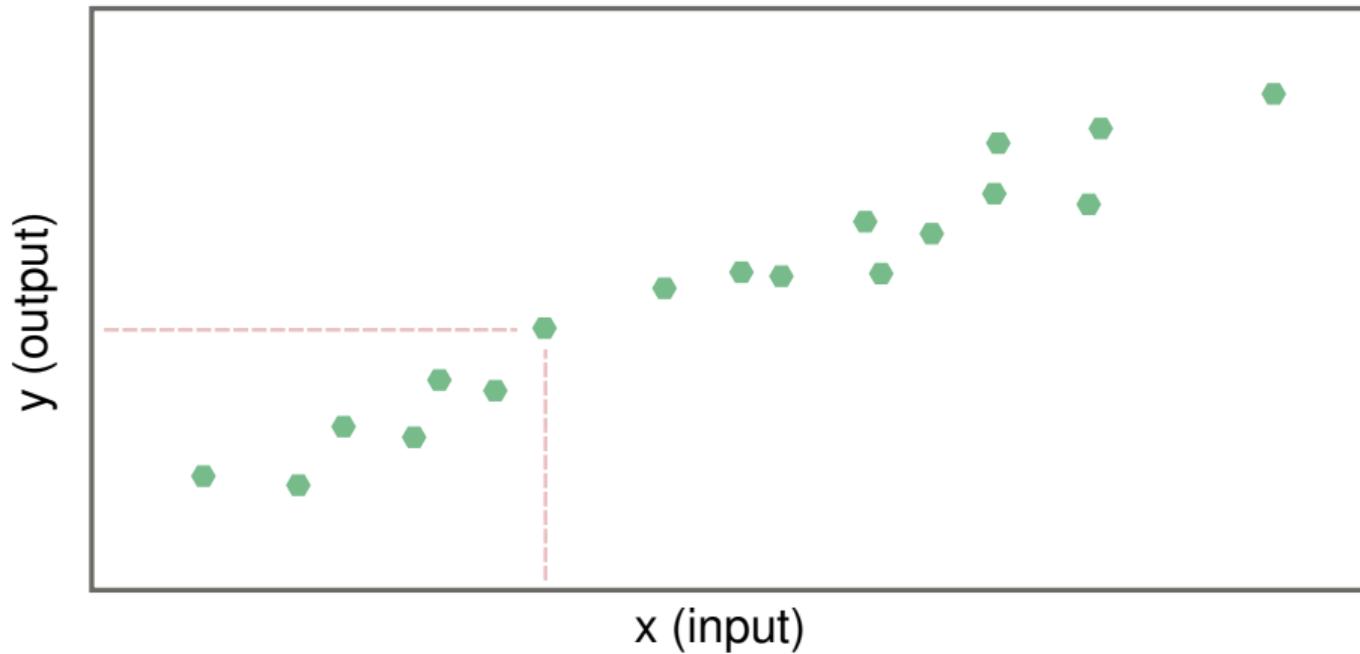
k-Nearest Neighbors



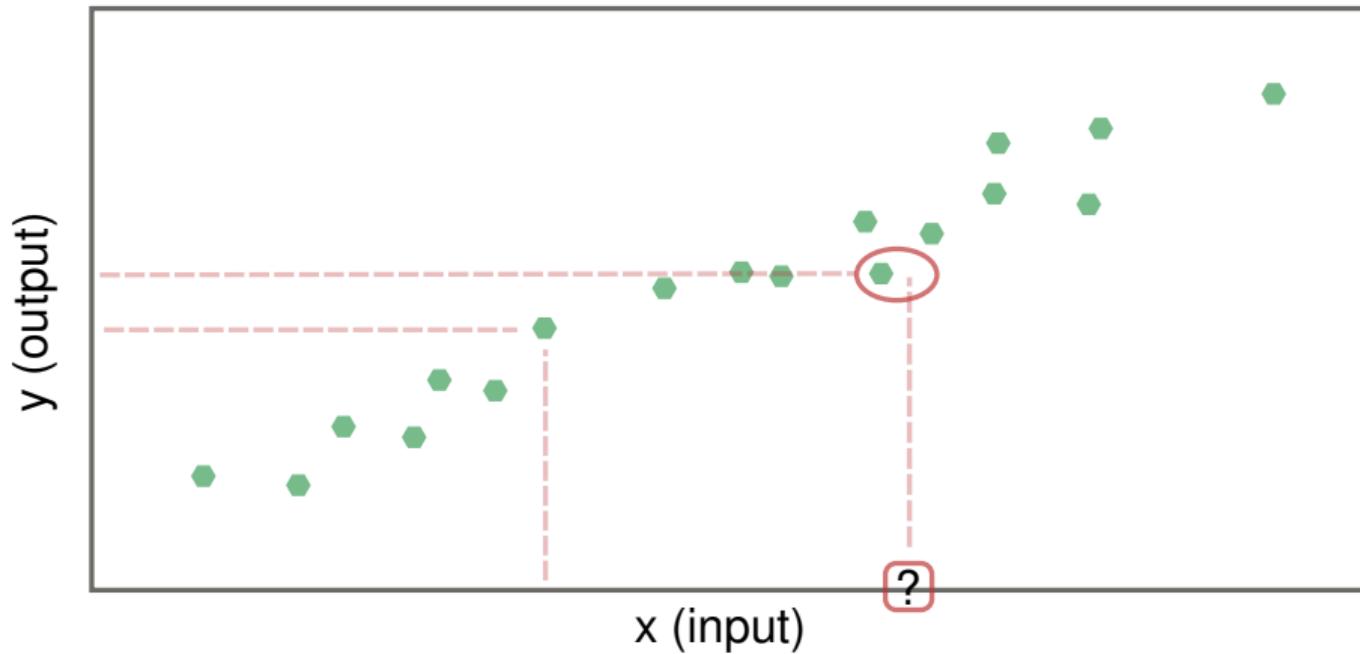
k-Nearest Neighbors



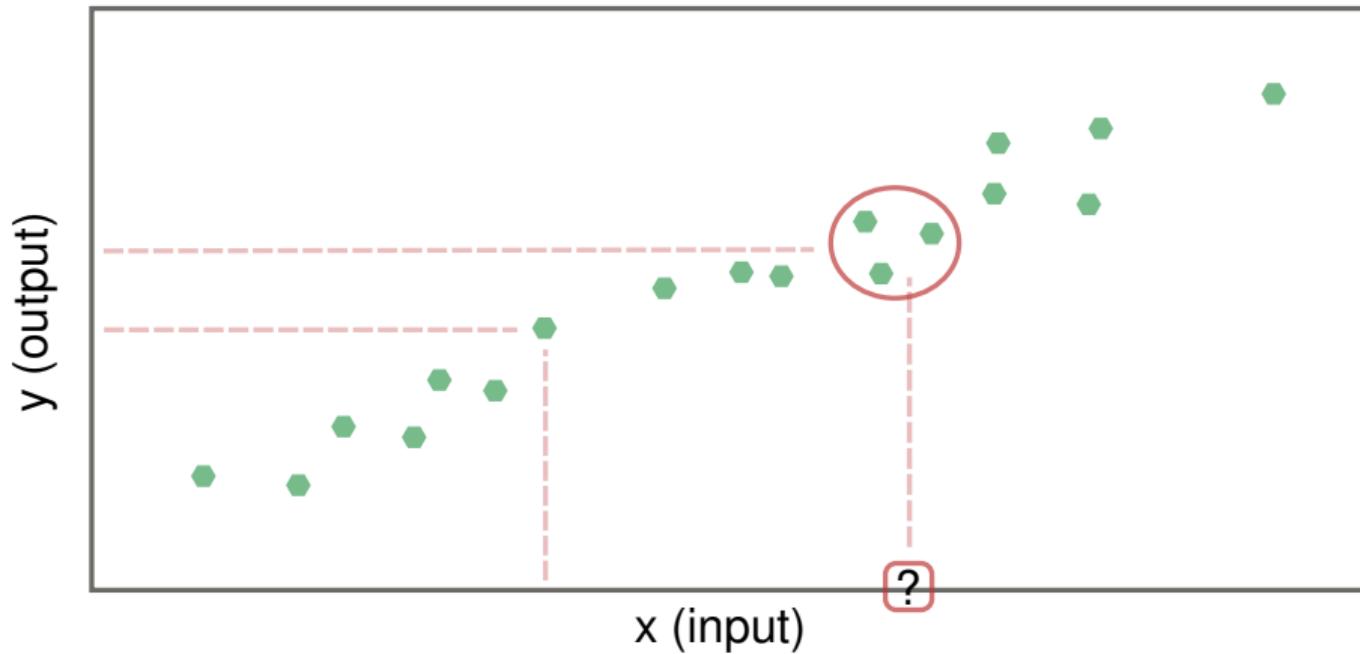
k-Nearest Neighbors



k-Nearest Neighbors



k-Nearest Neighbors



1 Introduction

2 Concepts and terminology

3 Selected supervised learning methods

- k-Nearest Neighbors
- Linear models
- Support Vector Machines (SVMs)
- Decision Trees and Random Forest
- Artificial Neural Networks

4 Summary

Linear models

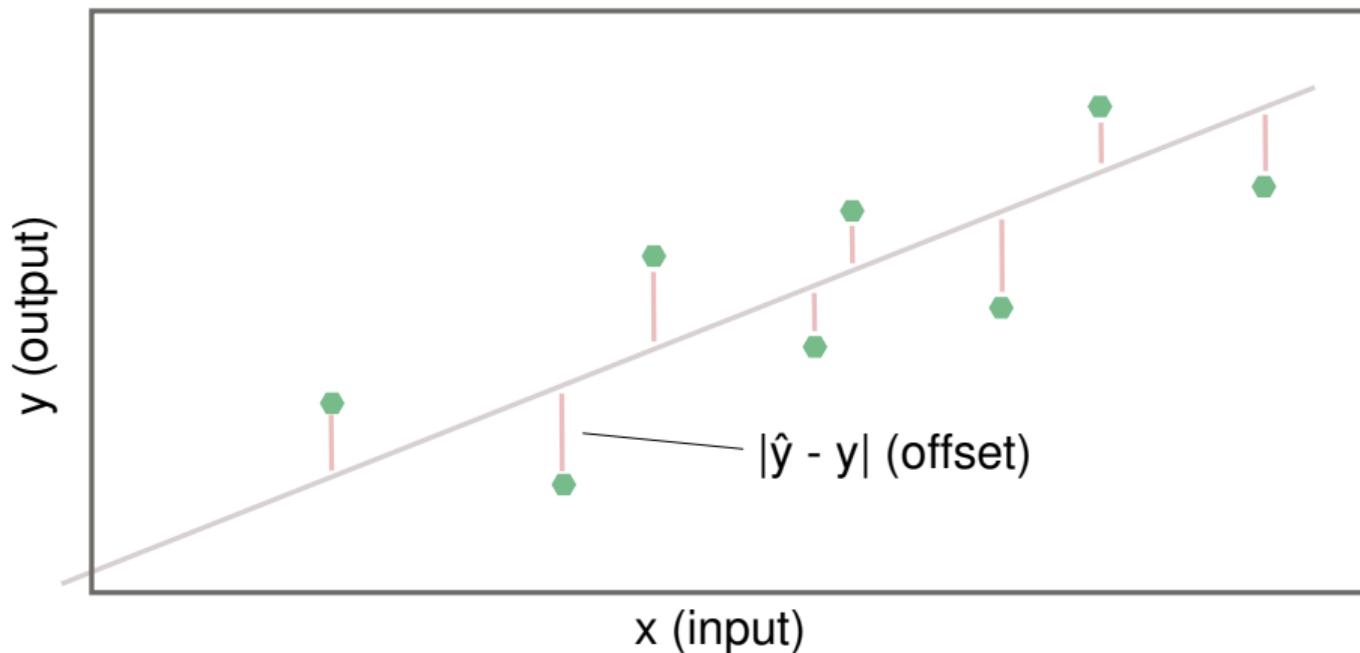
$$\hat{y} = w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_nx_n + b$$

with n as the number of features
 w are the different weights/coefficients
 b the intercept

Different ways to estimate the parameters

- Ordinary Least Squares
 - no parameters - easy to use but no possibility to adapt
- Ridge
 - coefficients should be close to zero
 - more resistant against overfitting
- Least Absolute Shrinkage and Selection Operator (LASSO)

Ordinary least squares (OLS)



Minimize the offset between \hat{y} and y the mean squared error (MSE) or sum of squared errors (SSE).

Once the parameters (b and the weights w) of

$$\hat{y} = w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_nx_n + b$$

are estimated the prediction can be performed by putting the x values of the data points into the equation to predict the y value.

1 Introduction

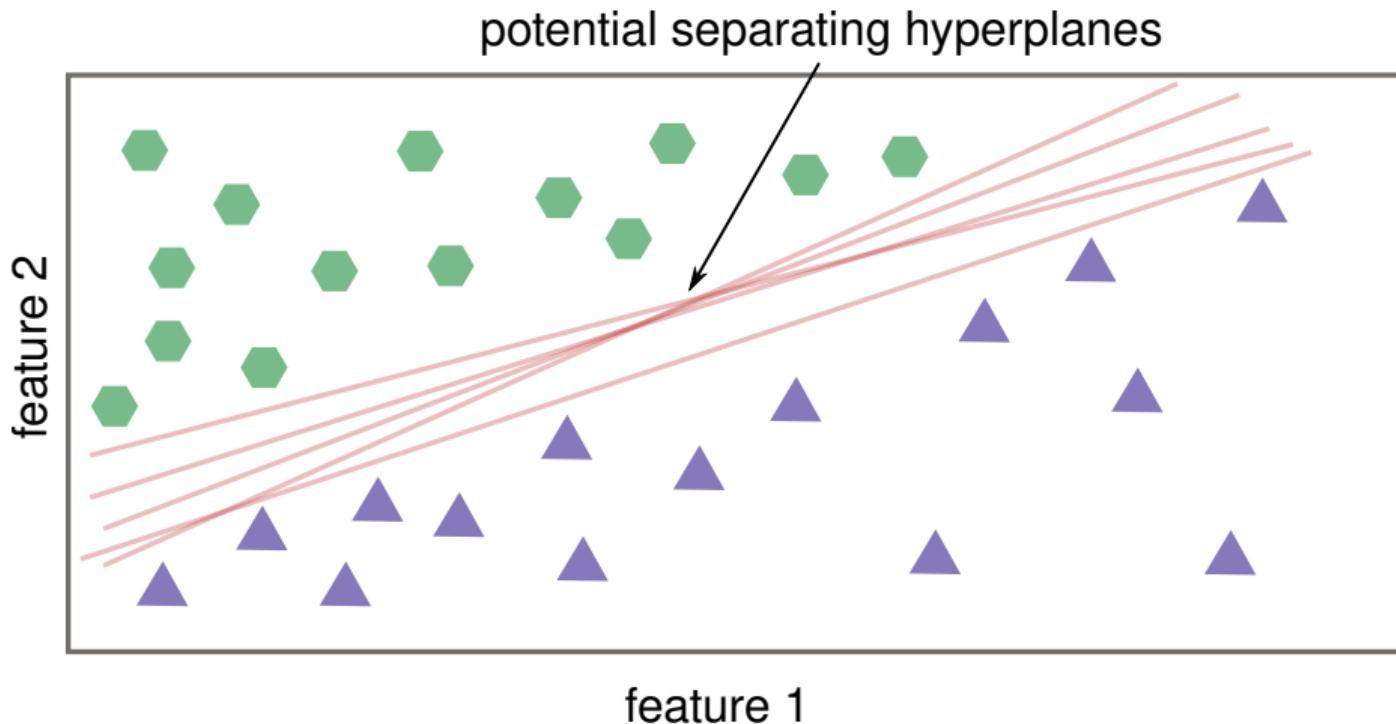
2 Concepts and terminology

3 Selected supervised learning methods

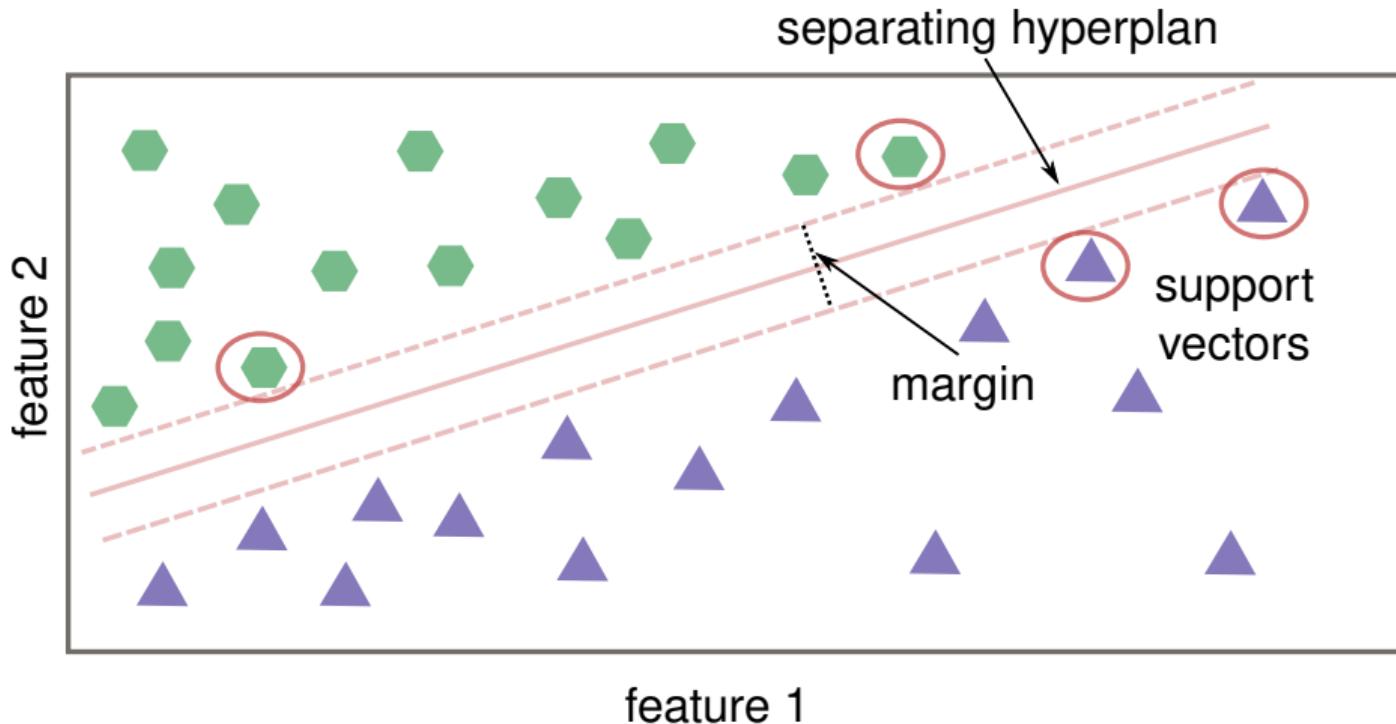
- k-Nearest Neighbors
- Linear models
- Support Vector Machines (SVMs)**
- Decision Trees and Random Forest
- Artificial Neural Networks

4 Summary

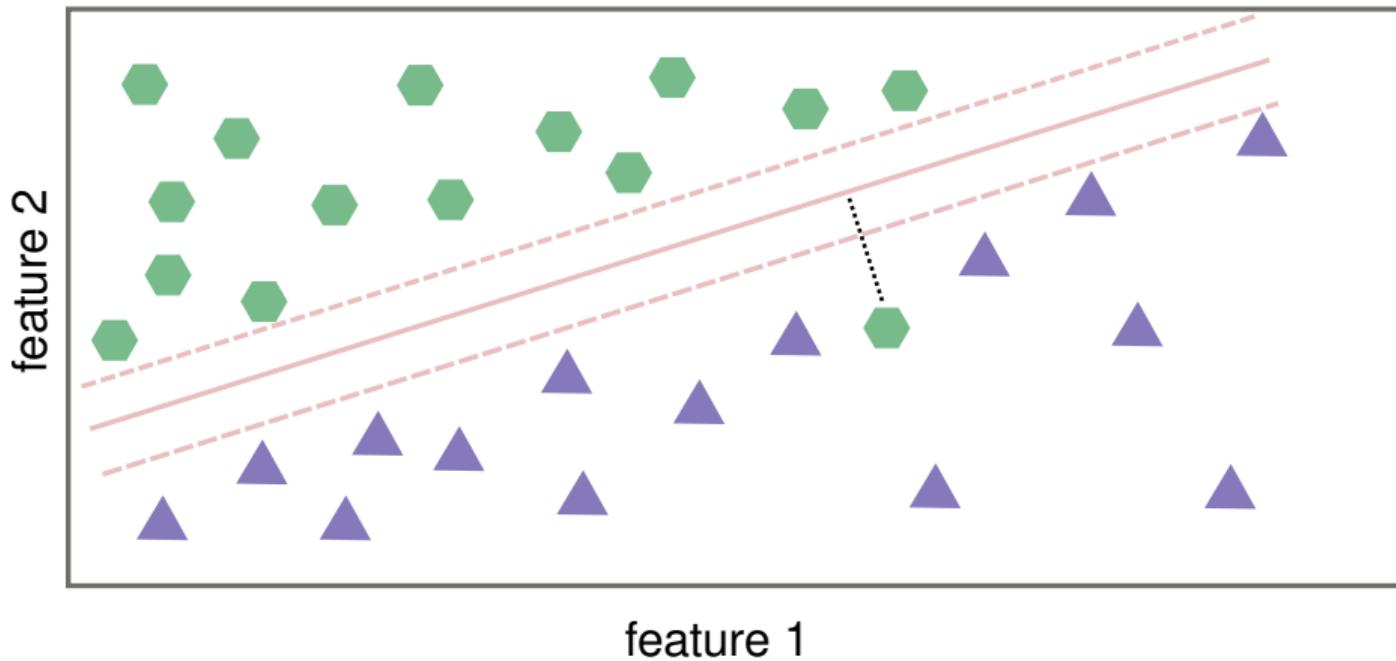
Support Vector Machines (SVMs) – Separating hyperplane



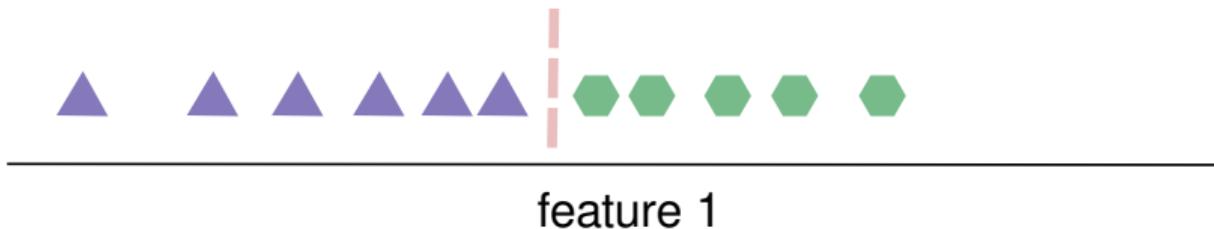
Support Vector Machines (SVMs) – Margin



Support Vector Machines (SVMs) – Soft Margin



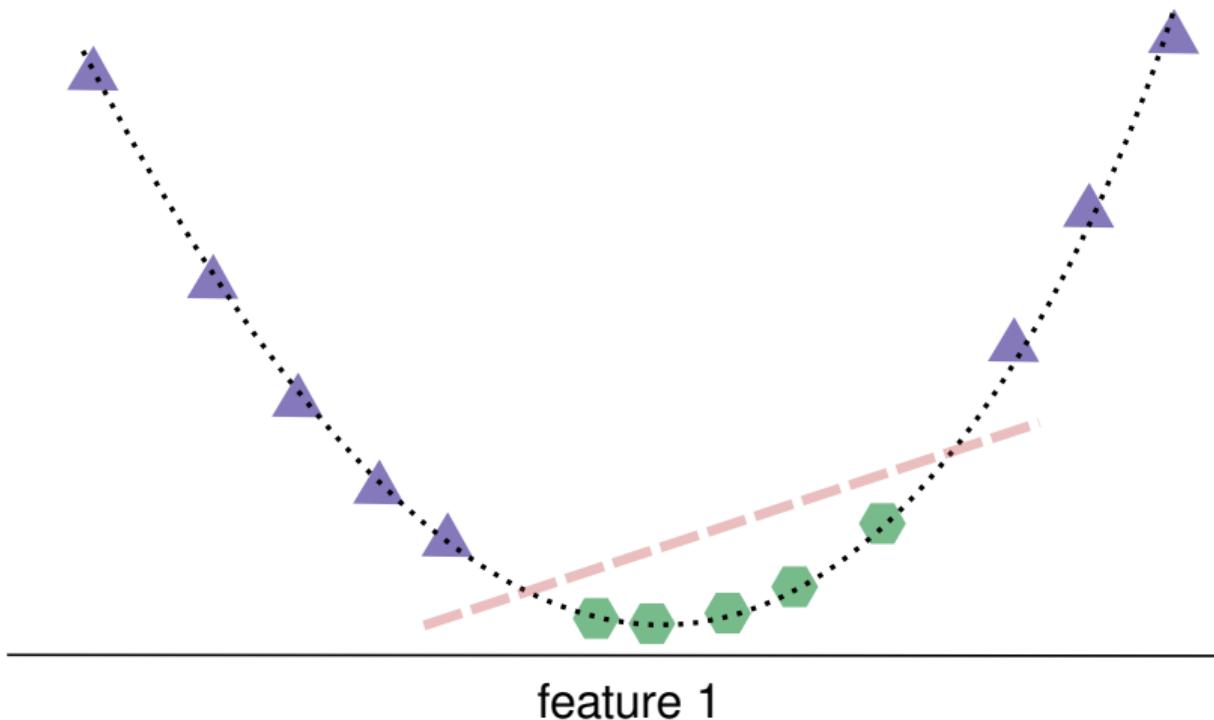
Support Vector Machines (SVMs) – Kernel trick



SVM – Kernel trick



Support Vector Machines (SVMs) – Kernel trick



1 Introduction

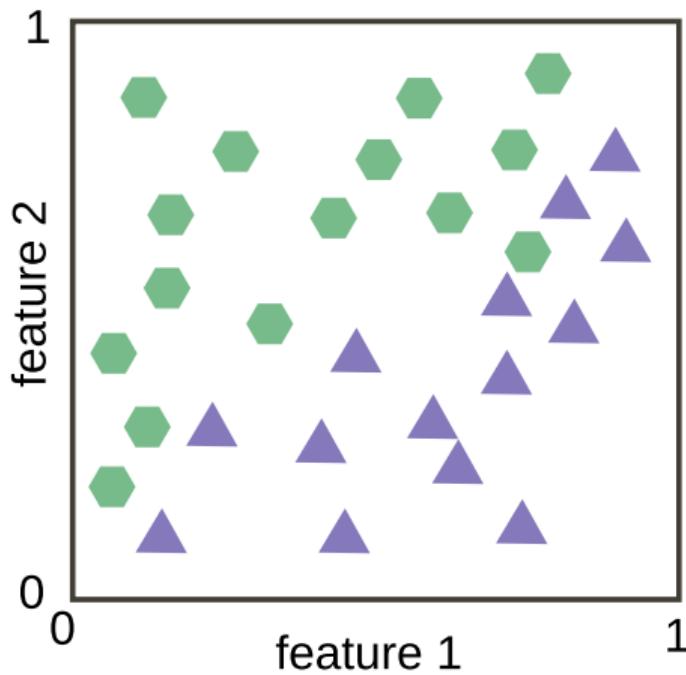
2 Concepts and terminology

3 Selected supervised learning methods

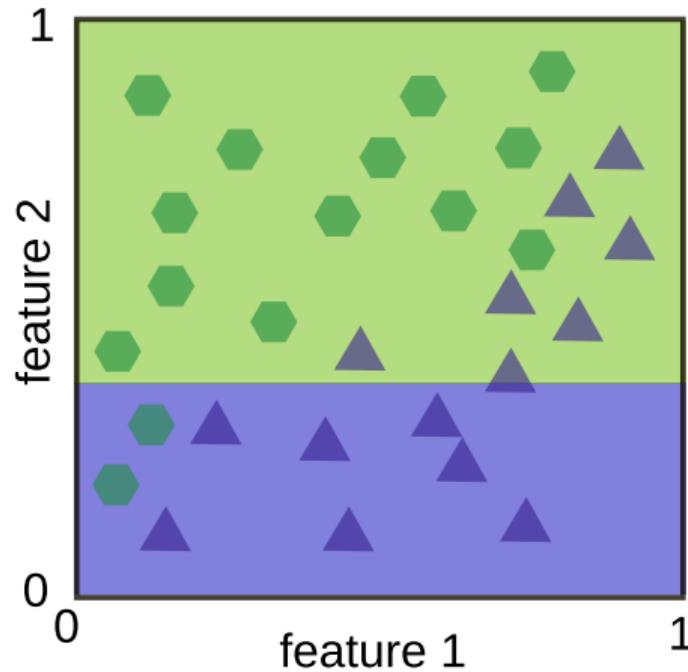
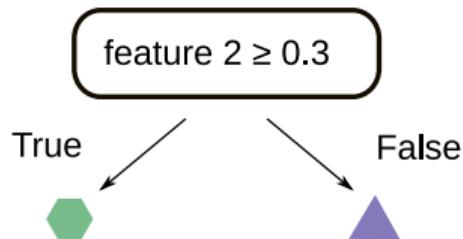
- k-Nearest Neighbors
- Linear models
- Support Vector Machines (SVMs)
- Decision Trees and Random Forest**
- Artificial Neural Networks

4 Summary

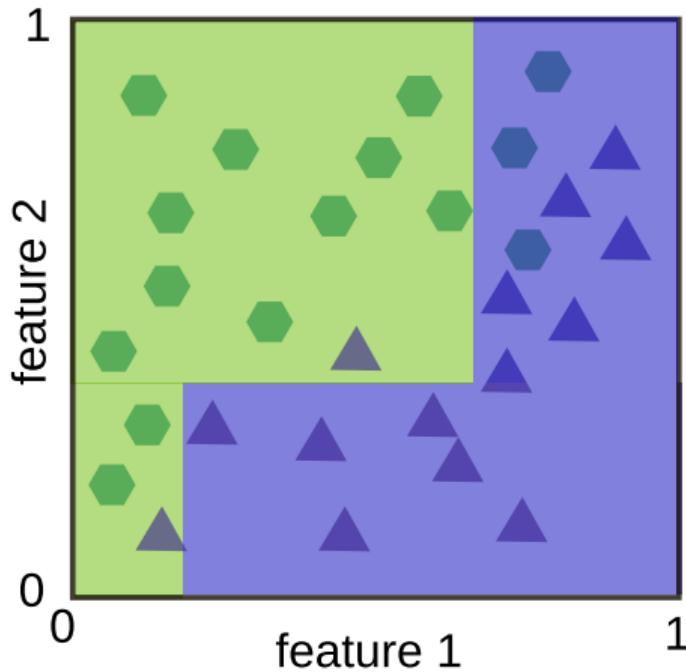
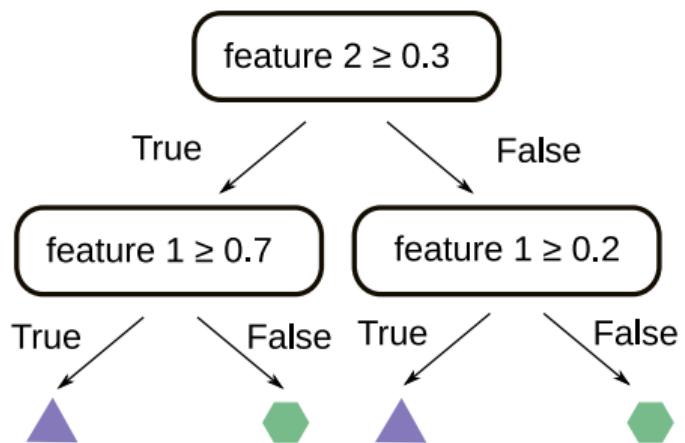
Decision Trees



Decision Trees



Decision Trees



Random forest

- In the random forests approach many different decision trees are generated by a randomized tree-building algorithm.
- The training set is sampled with replacement to produce a modified training set of equal size to the original but with some training items included more than once.
- In addition, when choosing the question at each node, only a small, random subset of the features is considered.
- Decision is happening by presenting the data to all tree and then do a voting.

1 Introduction

2 Concepts and terminology

3 Selected supervised learning methods

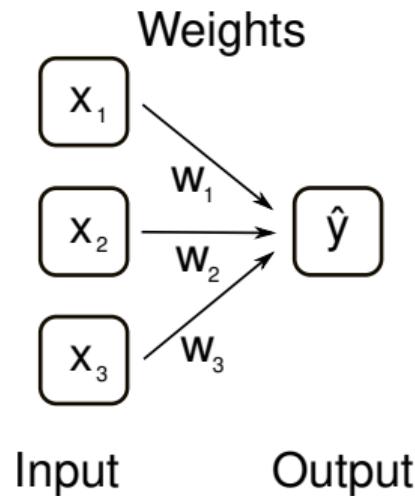
- k-Nearest Neighbors
- Linear models
- Support Vector Machines (SVMs)
- Decision Trees and Random Forest
- Artificial Neural Networks

4 Summary

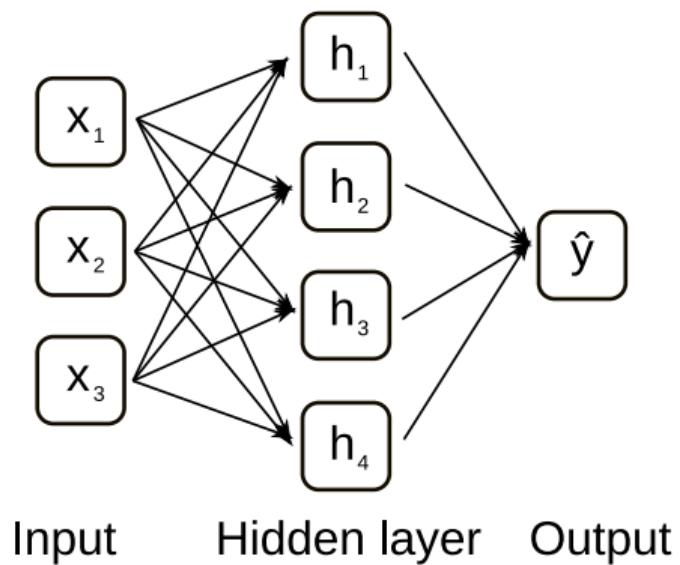
Artificial Neural Networks

- Aka. Multilayer perceptrons or Feed-forward neural networks
- Inspired by natural neural networks
- For classification or regression

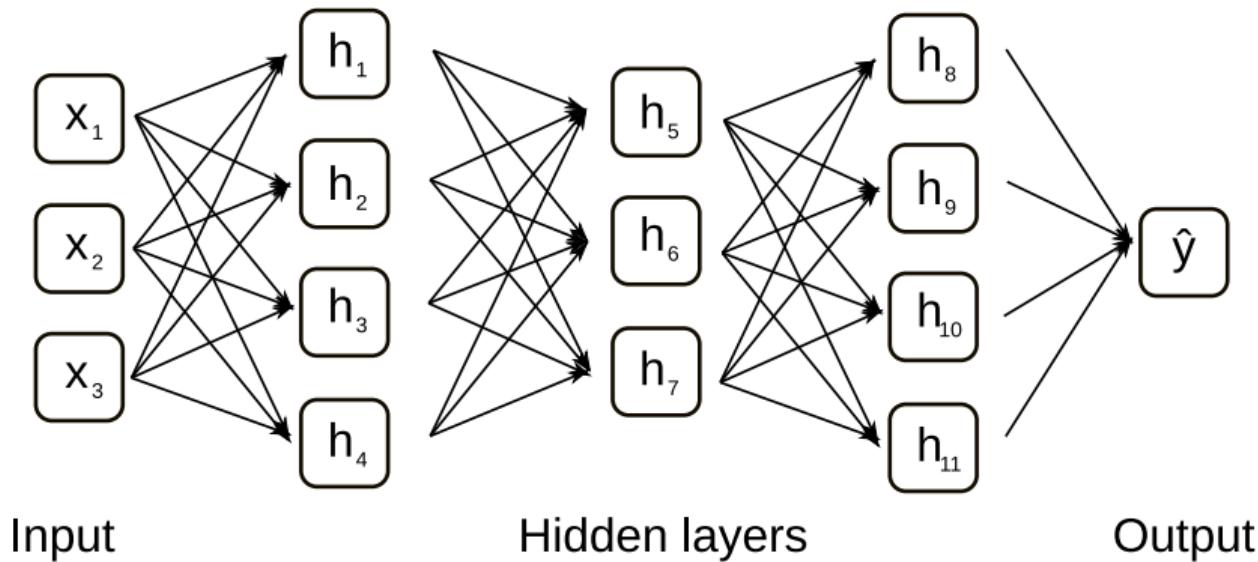
Artificial Neural Networks



Artificial Neural Networks



Artificial Neural Networks



1 Introduction

2 Concepts and terminology

3 Selected supervised learning methods

4 Summary

Summary

- Supervised machine learning can be used for classification and regression.
- The parameters of models are estimated based on training data.
- Features have to be selected and potentially encoded or scaled.
- There are numerous machine learning approaches with different strength and weaknesses available.

Thank you for your attention

konrad.foerstner.org / @konradfoerstner

zbmed.de / @ZB_MED

th-koeln.de / @th_koeln



**Technology
Arts Sciences
TH Köln**