Data-driven Modeling - Machine Learning

Instructor: Heinz Koeppl

Department of Electrical Engineering and Information Technology
Technische Universität Darmstadt

COURSE INFORMATION

TECHNISCHE UNIVERSITÄT DARMSTADT

- Lecture will be uploaded on Moodle
- 5 non-mandatory theory exercises will accompany the lecture. We highly encourage you to solve these exercises.
- Solutions of theory exercises will be presented in the exercise sessions on Wednesdays. Dates are announced on Moodle beforehand.
- There will be 5 non-mandatory programming exercises. Solving these, you can achieve a bonus of 0.3/0.4 points for the exam. However, it can not convert a failing grade to a pass.
- For content-related questions, please use the Discussions-forum on Moodle
- Please contact Sebastian Wirth (sebastian.wirth@tu-darmstadt.de), Yannick Eich (yannick.eich@tu-darmstadt.de) or Matthias Schultheis (matthias.schultheis@tu-darmstadt.de) if you have questions regarding the organization.

RECOMMENDED TEXTBOOKS

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Kevin P. Murphy. *Machine learning: a probabilistic perspective*, The MIT Press, 2012.
2. Christopher M. Bishop. *Pattern recognition and machine learning*, Springer, 2006.
3. Daphne Koller and Nir Friedman, *Probabilistic graphical models: principles and techniques*, The MIT Press, 2009.
4. Ian Goodfellow, Yoshua Bengio and Arthur Courville. *Deep Learning*, MIT Press, 2016.
5. Kevin P. Murphy. *Probabilistic machine learning: advanced topics*, The MIT Press, 2023.

COURSE OUTLINE

TECHNISCHE UNIVERSITÄT DARMSTADT

- Introduction
- Recap of probability theory
- Probabilistic graphical models (PGMs)
- Inference in PGMs
- Specific PGMs
- Mixture models and clustering
- Regression and classification
- Variational inference
- Deep learning and generative (AI) models

Introduction, motivation, types of machine learning

WHAT IS MACHINE LEARNING?

TECHNISCHE UNIVERSITÄT DARMSTADT

“Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed” – *1959, Arthur L. Samuel*

“We define machine learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty” – *Kevin P. Murphy*

“Machine learning is programming computers to optimize a performance criterion using example data or past experience” – *Ethem Alpaydin*

“The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience” – *Tom M. Mitchell*

WHY AUTOMATED DATA ANALYSIS?

TECHNISCHE UNIVERSITÄT DARMSTADT

The era of big data

"We are drowning in information and starving for knowledge" – *John Naisbitt*

Examples of huge data sets

- more than 40 billion indexed web pages
- genomes of thousands of people have been sequenced
- 100 hours of video uploaded to YouTube every minute

$\rightarrow$ **Need for automated methods of data analysis**

PHYSICS-BASED VS DATA-DRIVEN MODELING

TECHNISCHE UNIVERSITÄT DARMSTADT

Increasing amount of data required for accurate prediction →

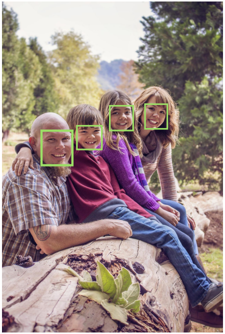
Governing equation (e.g. Maxwell equations) Symmetries Conservation relations Idealised geometries ….	Governing equation (e.g. Maxwell equations) Symmetries Conservation relations ….	Some partial governing equation Symmetries Conservation relations ….	General function representation (e.g. neural networks, universal approximations)
	Data/measurements: Complicated (nonlinear) material characteristics ….	Data/measurements: E.g. Time / spectral data to learn model parameters, parts of the equations	Data/measurements: Massive amount of data to adapt general structure to the specific system.
First principle (physics) model	Traditional engineering and physics in realistic domains	Physics-informed machine learning	General machine learning, artificial intelligence
Prediction without any prior measurements	Prediction require measurements-based model adaptation	Some model structure / topology is from physics - inductive bias	Features and functional dependency are inferred / parametrised from data

MACHINE LEARNING APPLICATIONS

TECHNISCHE UNIVERSITÄT DARMSTADT

Examples:

Face Detection

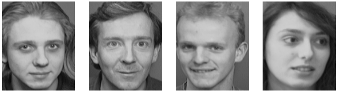


source:
https://pixabay.com/en/family-together-parents-people-838239/

Face Recognition

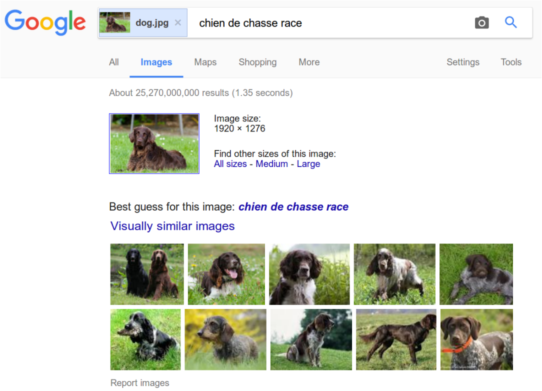
Training images

Test images



source: ORL dataset, AT&T laboratories, Cambridge UK,
http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

Example: Find similar images given an image



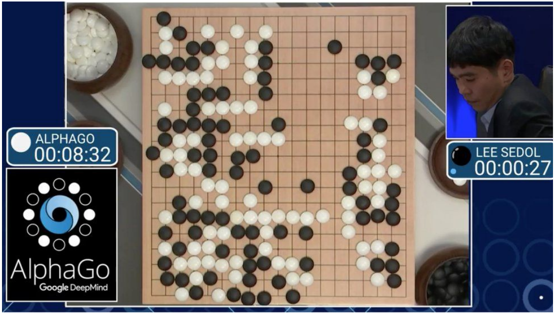
sources: Google Search by Image, `https://images.google.com/`
`https://pixabay.com/en/dog-animal-pet-782498/`

Example: Product Recommendations

source: www.amazon.de

MACHINE LEARNING APPLICATIONS

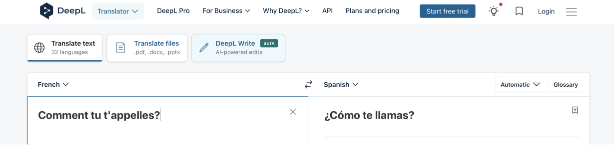
Example: The game of Go

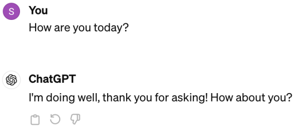


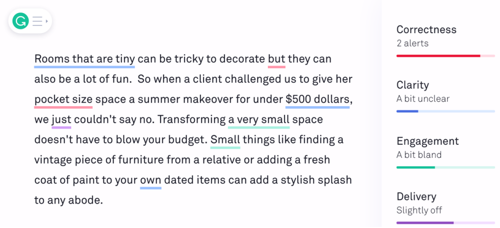
source: https://www.bbc.com/news/technology-35785875

Google DeepMind's AlphaGo (Learning a policy for the game GO)

Example: Language Tasks



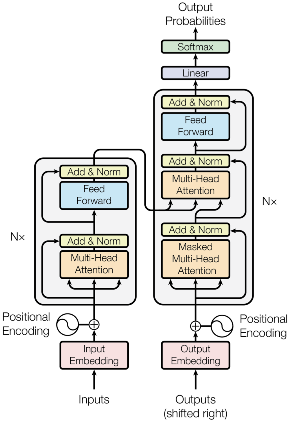




sources: DeepL, `https://www.deepl.com/translator`,
Grammarly, `https://www.ahead.ie/Grammarly`,
ChatGPT, `https://chat.openai.com`

Large Language Models

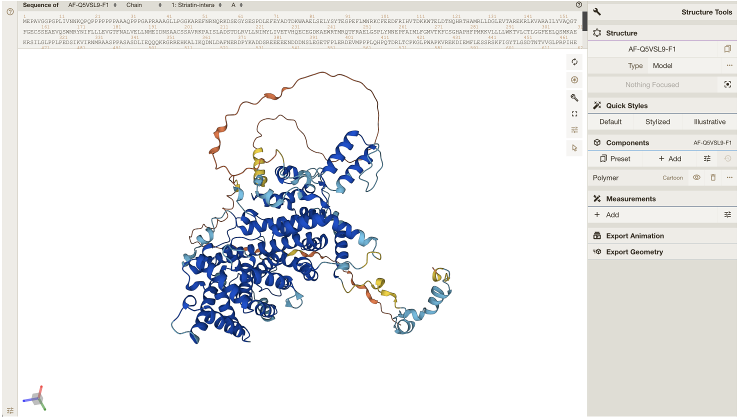
- Models with billions of parameters
- Trained on large amounts of text data
- Used for text classification and generation
- Applications: chatbots, translation, summarization, and more
- Examples are: ChatGPT, Cohere, Llama, Claude



source: Vaswani et al. (2017), "Attention Is All You Need"

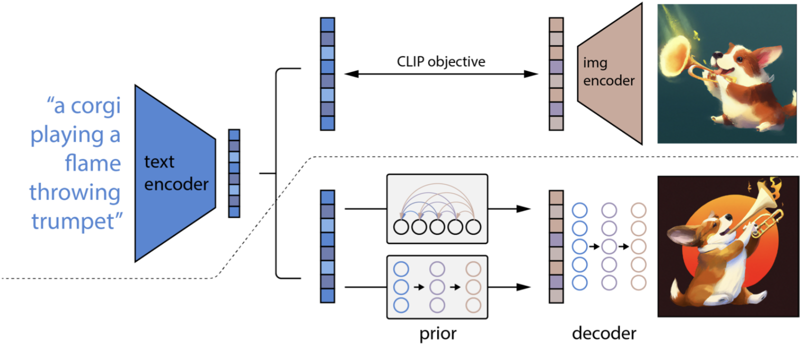
TECHNISCHE UNIVERSITÄT DARMSTADT

Example: Protein Structure Prediction



source: AlphaFold, `https://www.alphafold.ebi.ac.uk/entry/Q5VSL9`

Image Generation with Dall-E



sources: Ramesh et al. (2022), "Hierarchical Text-Conditional Image Generation with CLIP Latents"

Image Generation

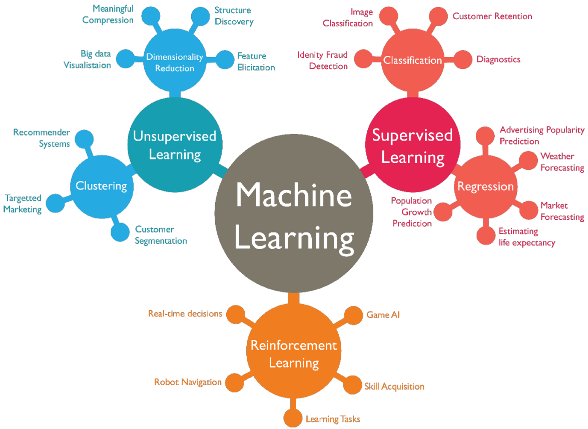
DALL-E

source: https://openai.com/dall-e-2

Midjourney

source: https://commons.wikimedia.org/wiki/File:Midjourney_as_it_imagines_itself.png

TYPES OF MACHINE LEARNING PROBLEMS



Source: `https://www.mactores.com/services/aws-big-data-machine-learning-cognitive-services` (27.03.2018)

TYPES OF MACHINE LEARNING PROBLEMS

Supervised Learning	Unsupervised Learning	Reinforcement Learning
- Classification - Regression - Self-supervised Learning - …	- Clustering - Dimensionality Reduction - …	Not part of this lecture

TYPES OF MACHINE LEARNING PROBLEMS

TECHNISCHE UNIVERSITÄT DARMSTADT

Supervised Learning	Unsupervised Learning	Reinforcement Learning
- Classification - Regression - Self-supervised Learning - ...	- Clustering - Dimensionality Reduction - ...	Not part of this lecture

SUPERVISED LEARNING

Supervised learning deals with:

- Inferring a mapping from an input to an output
- Given a set of labeled training data

SUPERVISED LEARNING

TECHNISCHE UNIVERSITÄT DARMSTADT

General Setting

Given: A labeled set of input-output pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \in (\mathcal{X} \times \mathcal{Y})^N$

Goal: Learn a mapping from inputs $\mathbf{x} \in \mathcal{X}$ to outputs $y \in \mathcal{Y}$

- $\mathcal{D}$ – training set
- N – number of training examples
- $\mathbf{x}_i$ – training input, e.g., a D-dimensional vector of numbers. The components are called the features/attributes
- $\mathbf{x}$ – test input
- y_i (y) – output/ response variable for training input $\mathbf{x}_i$ (test input $\mathbf{x}$)
- $\mathcal{X}$ – input space
- $\mathcal{Y}$ – output space

Classification

Given: A labeled set of input-output pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N} \in (\mathcal{X} \times \mathcal{Y})^N$ with categorical outputs $y_i \in \mathcal{Y} = \{1, ..., C\}$, where C is the number of classes

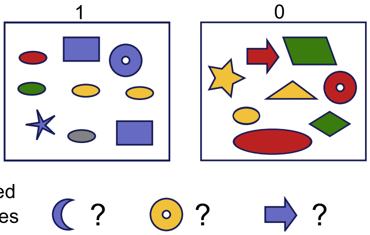
Goal: Learn a mapping from inputs $\mathbf{x} \in \mathcal{X}$ to outputs $y \in \mathcal{Y} = \{1, ..., C\}$

- $\mathcal{X}$ – input space
- $\mathcal{Y} = \{1, ..., C\}$ – a finite set of discrete classes / labels
- $C = 2$: “binary” classification
- $C > 2$: “multiclass” classification

Applications: e.g., e-mail spam filtering, face detection and recognition

An Example of Binary Classification

labeled training data



unlabeled test cases

D features (attributes)

N cases

Color	Shape	Size (cm^2)	Label
Blue	Square	10	1
Red	Ellipse	2.4	1
Red	Ellipse	20.7	0
...	...	...	...

training data represented in N x D design matrix

Based on an example from Kevin P. Murphy, *Machine learning – A probabilistic perspective*, The MIT Press, 2012.

Regression

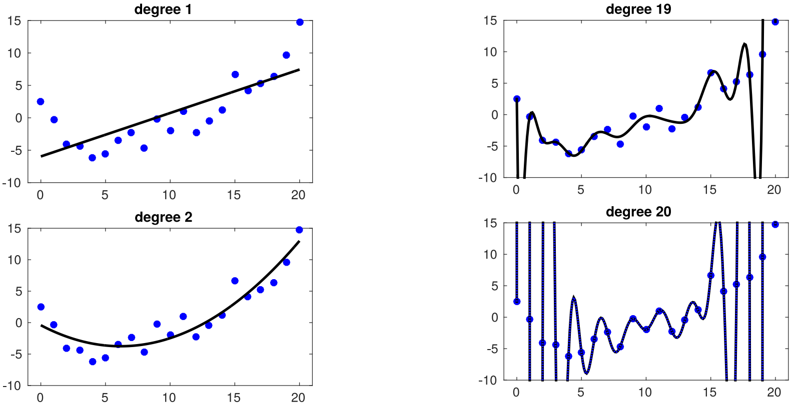
Given: A labeled set of input-output pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \in (\mathcal{X} \times \mathcal{Y})^N$ with continuous outputs $y_i \in \mathcal{Y} \subseteq \mathbb{R}$

Goal: Learn a mapping from inputs $\mathbf{x} \in \mathcal{X}$ to outputs $y \in \mathcal{Y}$

- $\mathcal{X}$ – input space
- $\mathcal{Y} \subseteq \mathbb{R}$ – continuous output space

Applications: e.g., stock market price prediction, temperature prediction

An Example of Regression



Polynomial regression applied to 1D data using different polynomial degrees.

Figures generated by `linregPolyVsDegree` using the freely available MATLAB package PMTK

SELF-SUPERVISED LEARNING

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Self-supervised learning deals with:

- Uses data itself to generate supervisory signals
- No need for labeled training data

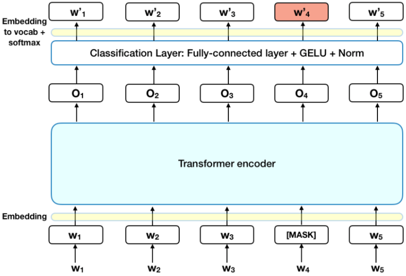
Examples:

- Bidirectional Encoder Representations from Transformers (BERT) Model,
- General Pretrained Transformer (GPT)

TECHNISCHE UNIVERSITÄT DARMSTADT

Example: BERT Model

Data: Sentences: {The quick brown fox jumps over the lazy dog., ...}
Model Input: Masked Sentences:
{The quick brown fox jumps [MASK] the lazy dog., ...}
Goal: Predict masked word, depending on the context:



source: `https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270`

TYPES OF MACHINE LEARNING PROBLEMS

Supervised Learning	Unsupervised Learning	Reinforcement Learning
- Classification - Regression - Self-supervised Learning - ...	- Clustering - Dimensionality Reduction - ...	Not part of this lecture

UNSUPERVISED LEARNING

Unsupervised learning deals with:

- Finding regularities in the data

General Setting

Given: A set of data $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N}$

Goal: Find “interesting” patterns in the data

- $\mathcal{D}$ – data set
- N – number of data examples
- $\mathbf{x}_i$ – data example

Clustering

Given: A set of data $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N}$

Goal: Divide the data into clusters

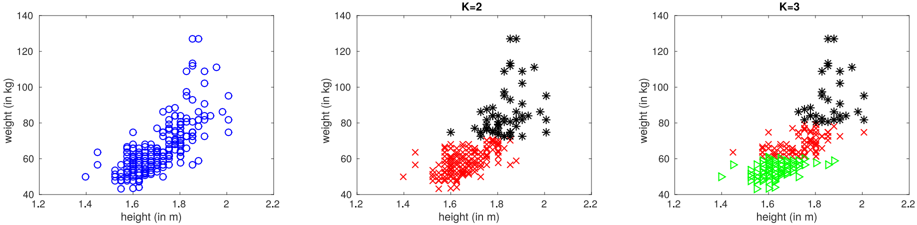
1. Estimate the distribution $p(k|\mathcal{D})$ over the number k of clusters and decide on a number of clusters
2. Estimate to which cluster each point belongs

Applications: e.g., customer segmentation in customer relationship management, clustering of astrophysical or biological data

UNSUPERVISED LEARNING

TECHNISCHE UNIVERSITÄT DARMSTADT

An Example of Clustering



Clustering of some 2D data (height and weight of a group of 210 people) using k-means algorithm.

Figures generated by `kmeansHeightWeight` using the freely available MATLAB package PMTK

Dimensionality Reduction

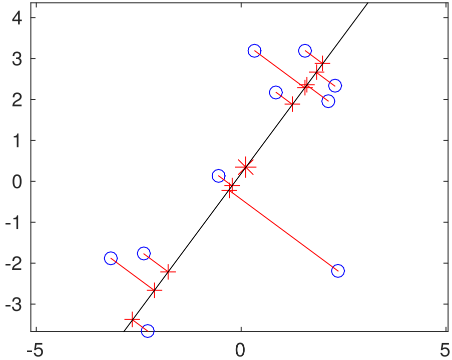
Given: A set of data $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$

Goal: Find a low dimensional representation of the data

- project the data to a lower dimensional subspace which captures the "essence" of the data
- discover "latent factors"

Applications: e.g., interpretation of biological data, document retrieval in natural language processing, signal separation in signal processing

An Example of Dimensionality Reduction (PCA)

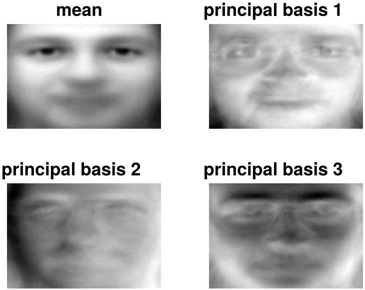


A projection of a 2D data set to a 1D line

Figure generated by `pcaDemo2d` using the freely available MATLAB package PMTK

Another Example of Dimensionality Reduction

Some images from the Olivetti face database



The mean and the first 3 principal component basis vectors

Figures generated by `pcaImageDemo` using the freely available MATLAB package PMTK

TYPES OF MACHINE LEARNING PROBLEMS

TECHNISCHE UNIVERSITÄT DARMSTADT

Supervised Learning	Unsupervised Learning	Reinforcement Learning
- Classification - Regression - Self-supervised Learning - …	- Clustering - Dimensionality Reduction - …	*Not part of this lecture*

REINFORCEMENT LEARNING

Reinforcement learning deals with:

- Learning how to act
- Given occasional reward or punishment signals

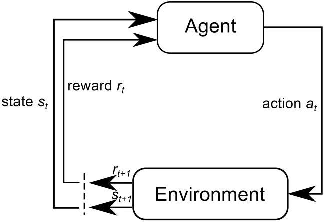
$\rightarrow$ goal-directed learning from interaction

General Setting

- An agent and its environment interact in discrete time steps $t = 1, 2, 3, ...$
- At each time step t, the agent
  - observes its environment's state $s_t \in \mathcal{S}$
  - makes a decision, by selecting an action $a_t \in \mathcal{A}$
  - receives a numerical reward $r_{t+1} \in \mathbb{R}$

Goal: Find the policy (=which action to select in which state), which maximizes the total reward over the long run

- $\mathcal{A}$ – action space
- $\mathcal{S}$ – state space



Based on a figure from R.S. Sutton and A.G. Barto, *Reinforcement Learning*, The MIT Press, 2012.

An Application Example: A Bin-Picking Robot

FANUC's bin-packing robot

Video available at: `https://www.youtube.com/watch?v=ydh_AdWZflA`

Task: Learn how to pick bins from a box

- *States*: The visual representation of the bins in the box (as sensed by the robot's camera)
- *Actions*: The position of its arm to grab
- *Reward*: Success or failure when grabbing

TYPES OF MACHINE LEARNING MODELS

TECHNISCHE UNIVERSITÄT DARMSTADT

Parametric vs. Non-Parametric Models

	Parametric	Non-Parametric
Properties	fixed number of parameters	number of parameters grows with amount of training data
Pros	faster to use	more flexible to use
Cons	stronger assumptions on the data distributions	often computationally intractable for large data sets
Example	classification using logistic regression	classification using K-nearest neighbor classifier (KNN)

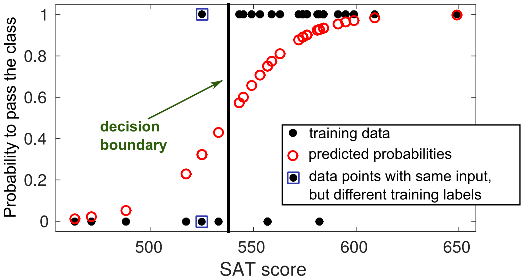
TECHNISCHE UNIVERSITÄT DARMSTADT

A Parametric Model for Classification: Classification via Logistic Regression[1]

parametric model, estimate w_0 and w_1:

$$y = \frac{\exp(w_0 + w_1 x)}{\exp(w_0 + w_1 x) + 1}$$

Logistic regression on SAT (Scholastic Assessment Test) score, threshold = 537.9



decision rule: if SAT > 537.9, predict that class will be passed

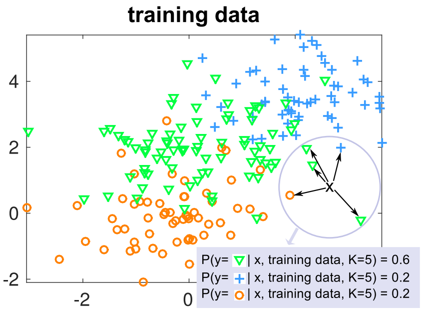
Logistic regression for SAT scores

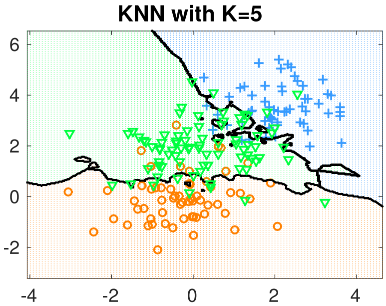
Figure generated by `logregSATdemo` using the freely available MATLAB package PMTK

[1] Even though called logistic “regression”, this is a form of classification, not regression!

TECHNISCHE UNIVERSITÄT DARMSTADT

A Non-Parametric Model for Classification: The K-nearest neighbor classifier





Classification of some 2D data using K-nearest neighbor (KNN) algorithm.

Figures generated by `knnClassifyDemo` using the freely available MATLAB package PMTK

Discriminative vs. Generative Models: An Analogy

Task: Determine the language that someone is speaking

Generative Approach: Learn each language and determine to which language the speech belongs

Discriminative Approach: Determine the linguistic differences without learning any language

$\rightarrow$ The discriminative approach is much easier, but the generative approach may afterwards be used to "generate" speech samples of that exact language