Workshop in Digital Humanities

Mathematical Concepts and Programming in Machine Learning

Thomas Schnake & David Lassner

19. July 2021

Roadmap

1. Introduction to Statistical Learning Theory
2. What is a Learning Algorithm?
3. How Does an Algorithm Learn?
4. Examples for Applications (Programming)
5. Questions & Discussion

1. Introduction to Statistical Learning Theory

1. Introduction to Statistical Learning Theory

Background

- Originated in Russia in the 1960s.
- Gained wide popularity in 1990s with *Support Vector Machines.*

How can a machine, a computer, “learn” specific tasks by following specified learning algorithms?

[1] V. Vapnik, *Statistical Learning Theory,* Wiley, New York, 1998.

[2] V. Vapnik, The Nature of *Statistical Learning Theory,* Springer, New York, 1995.

[3] B. Schölkopf and A. Smola, *Learning with Kernels, Section 5, MIT Press*, Cambridge, MA, 2007.

Motivation

Problems that an be tackled with statistical learning theory.

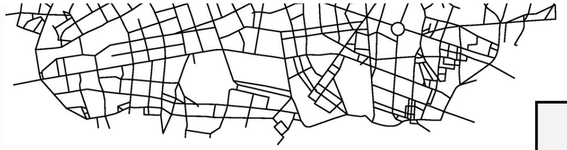
How long does it take to the airport?

Is this a cat or a dog?

What's the weather like tomorrow?

Different Concepts for Solutions

How long does it take to the airport?



Deterministic Approach

a) Explore all possible paths

b) Take the shortest path

\+ Finds the shortest path

\- Very expensive

Statistical Approach

a) Explore random paths.

b) Prefer lanes which lead to short paths.

c) Stop any time.

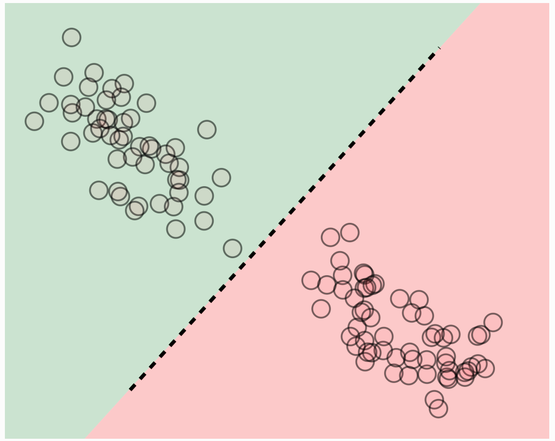
\- May not find shortest path.

\+ Cheap computation

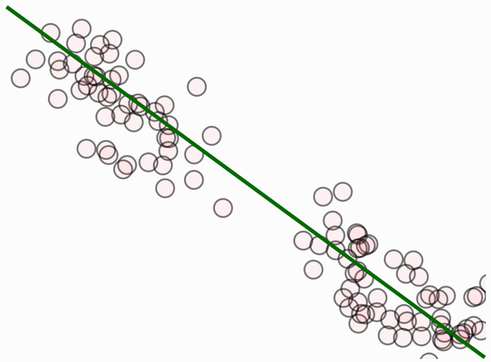
Typical Scenarios

Supervised Problems

Classification



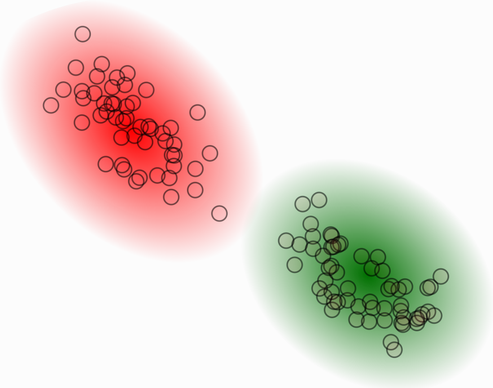
Regression



Exampels?

Unsupervised Problems

Clustering



Reinforcement learning

etc.

What is a Statistical Learning Problem?

Supervised Learning

Input: $(x_1, x_2, \ldots x_n)$

Target: $(y_1, y_2, \ldots y_n)$

1) Find model f

2) Find parameter θ

Such that $f_\theta(x_i) = y_i$ $\quad i = 1, 2, \ldots, n$

What is a Statistical Learning Problem? Example

MNIST Dataset

70 000 handwritten digits from 0-9

Background

- Published 1998 (Y. LeCun et al.)
- Widely used for vision tasks
- Original error rate 0.8 %
- Todays (2020) error rate 0.17 % [5]

[4] Y. LeCun et al., *Gradient-Based Learning Applied to Document Recognition, In Proceedings of IEEE, 86 (11), 2278-2324*,1998.

[5] https://github.com/Matuzas77/MNIST-0.17

MNIST as a Supervised learning problem

- Pictures are grayscale
- Pictures are of dimension 28x28

$$x_i \in [0,1]^{28\times 28}$$

Input:

Target:

$$(0, 1, \ldots, 6)$$

1) Find model f 2) Find parameter θ

$$f_\theta(\ \ ) = 0$$

2. What is a Learning Algorithm?

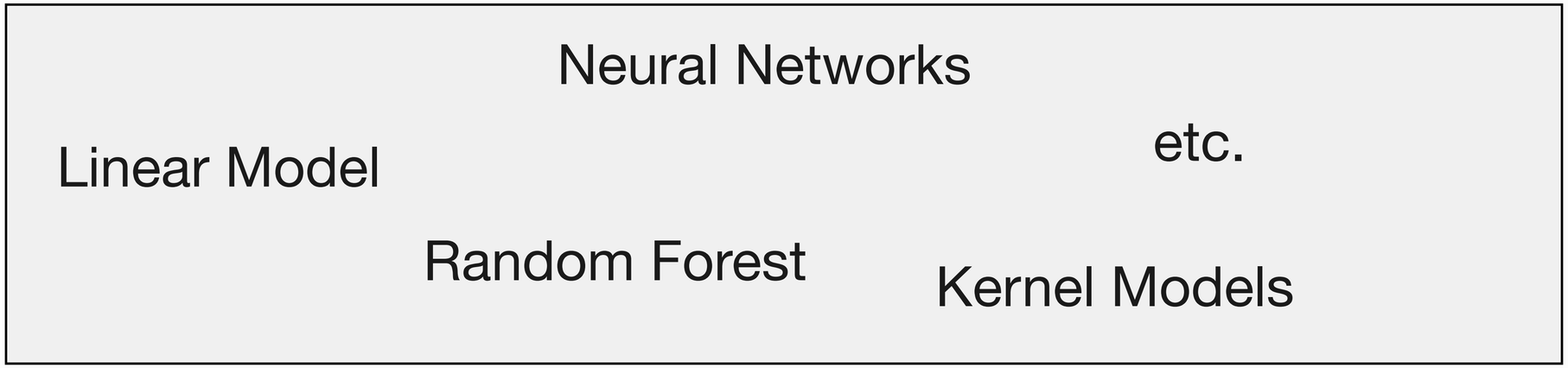
Choose f

2. What is a Learning Algorithm?

General Introduction

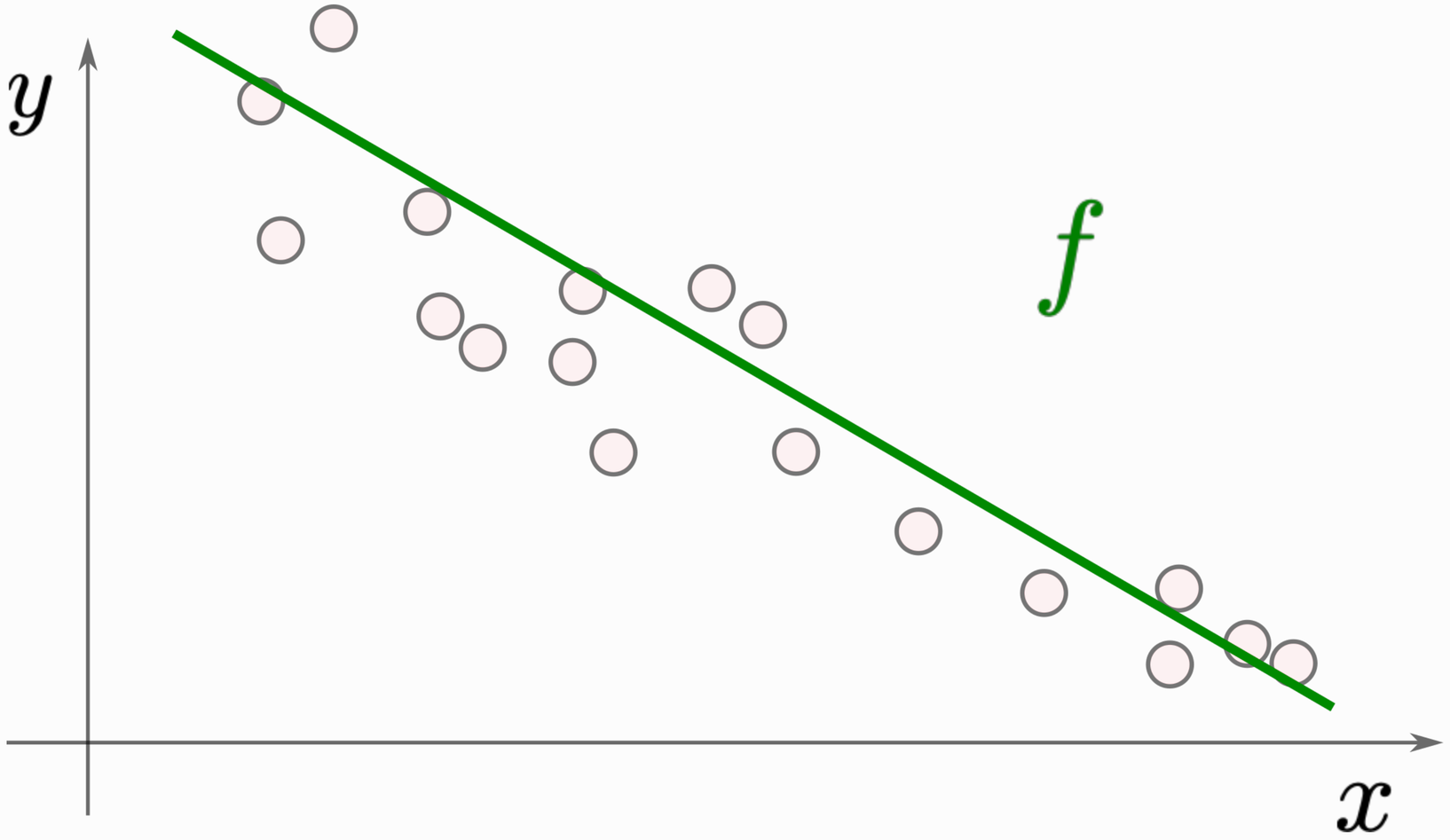
- Problem specific
- Inductive Bias
- Complexity vs. Simplicity

Some Models



General Introduction to Learning Algorithm

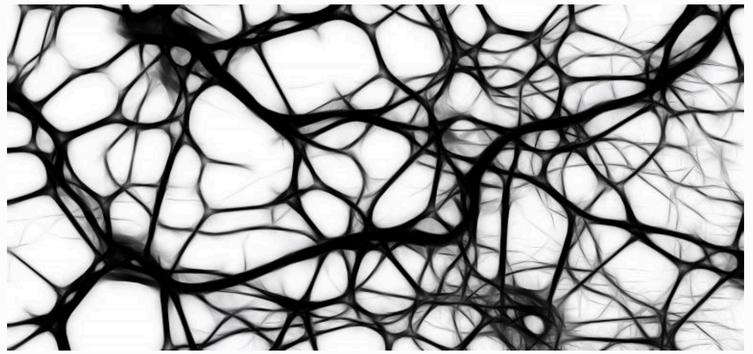
Linear Model

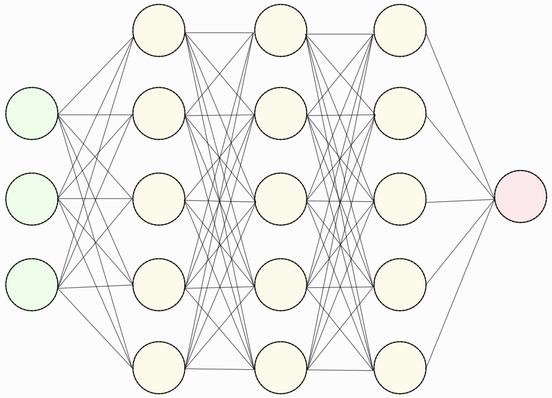


2. What is a Learning Algorithm?

Neural Networks: Background and History

- First attends on imitating *neural plasticity* 1940s [6]
- Creation of *perceptron* in 1958 [7]
- Introduction of *backpropagation* in 1975 [8]
- Introduce LSTM, CNNs, RNN … (Deep Learning)
- Running NNs on GPUs since 2015 [9]





[6] D. Hebb, *The Organization of Behavior,* New York: Wiley, 1949.

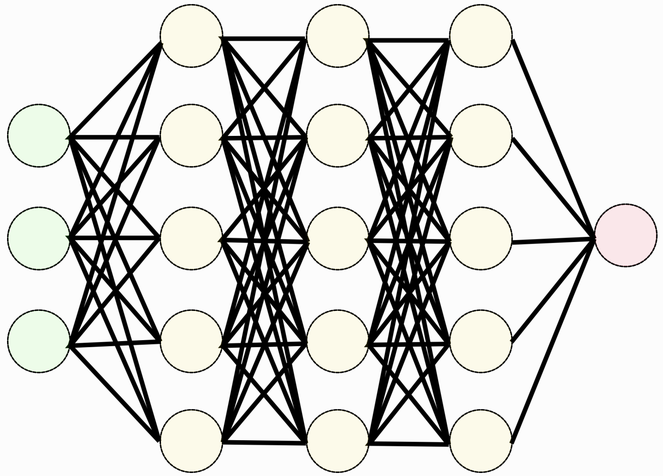
[7] F. Rosenblatt, The Perceptron: *A Probabilistic Model For Information Storage An Organization In The Brain, Psychological Reviews, p.* 386-408, 1949.

[8] P.J. Werbos, Beyond Regression: *New Tools for Prediction and Analysis in Behavioral Sciences,* Harvard University, 1975.

[9]A. Krizhevsky et al., *ImageNet classification with deep convolutional neural networks,* Communications of ACM, p. 84-90, 2017.

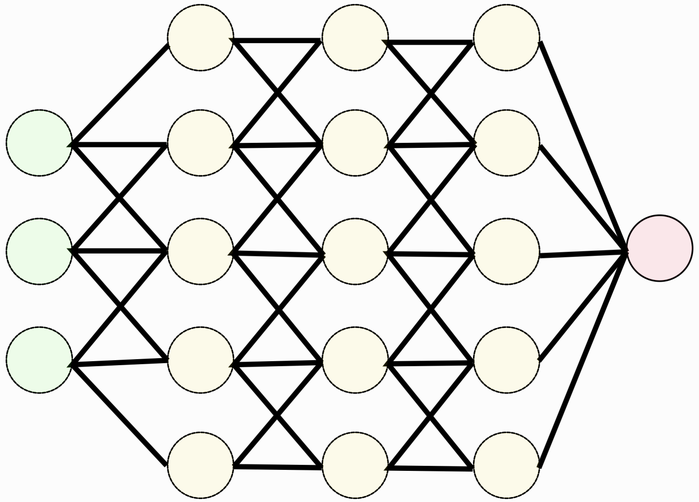
Neural Networks: Types

Vectorial



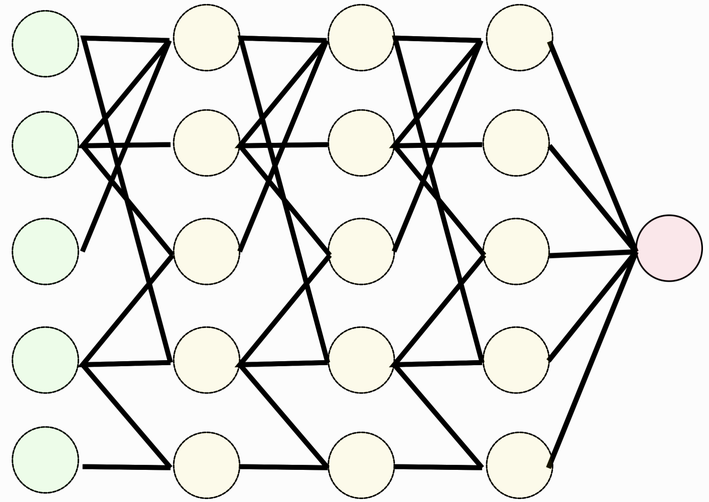
Fully connected

Convolutional



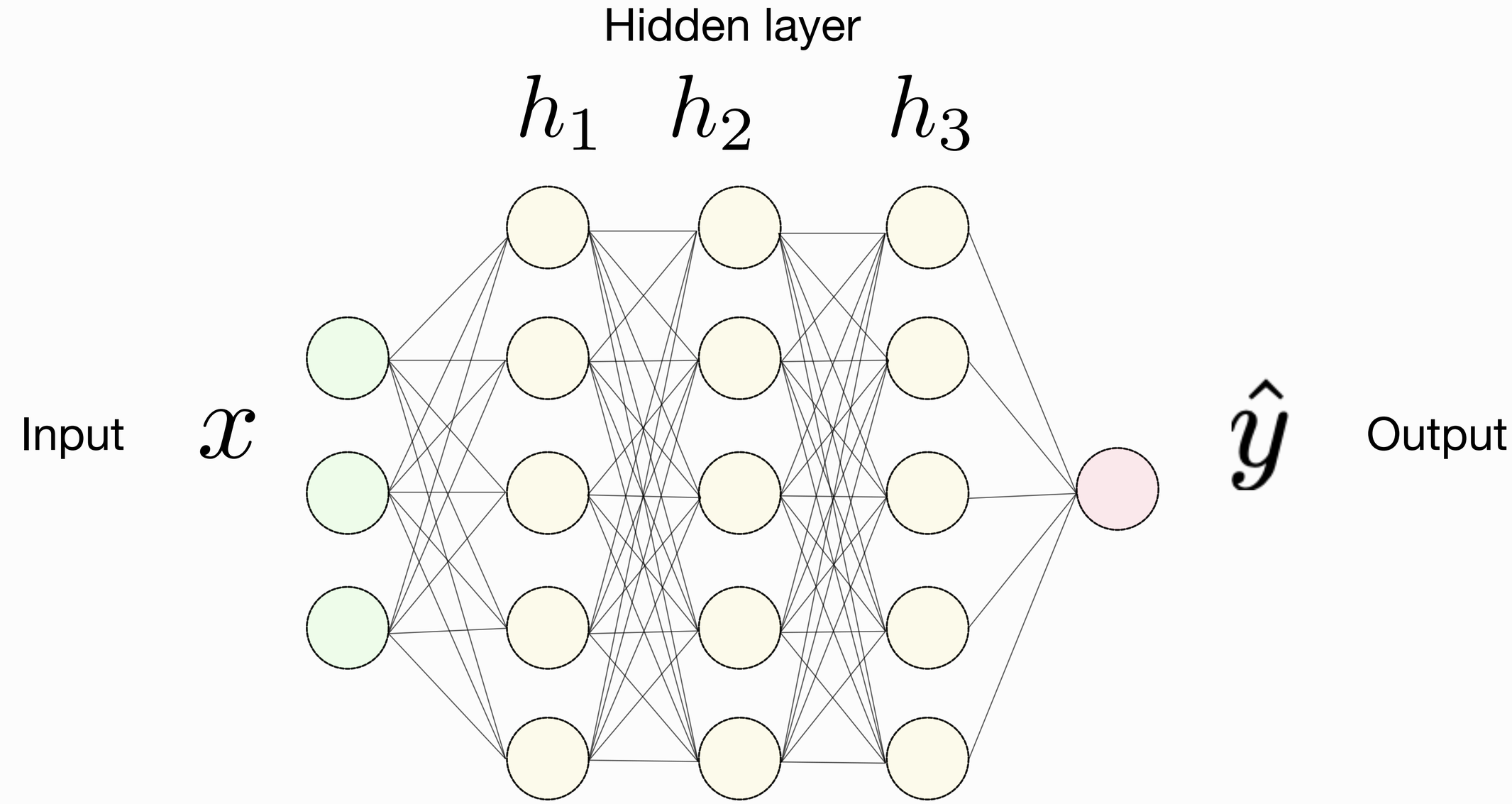
Locally connected

Graph



Custom connectivity

Neural Networks: Layerwise Visualisation



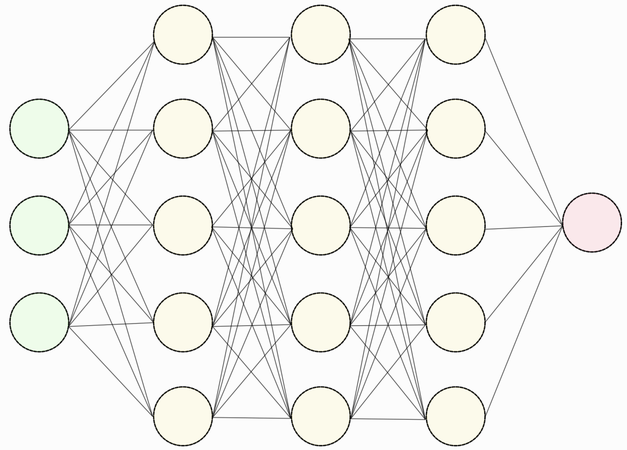
Neural Networks: Propagation Rules

Vectorial Neural Network

Layerwise Parameter

$$A_l, b_l \sim \text{random} \qquad l = 1, \ldots, L$$

$$\begin{aligned} h_0 &= x \\ h_l &= \sigma(A_l h_{l-1} + b_l) \\ h_L &= \hat{y} \end{aligned}$$



ReLU activation

$$\sigma(x) = \max(0, x)$$

Global Parameter

$$\theta = (A_1, b_1, A_2, b_2, \ldots, A_L, b_L)$$

3. How Does an Algorithm Learn?

Choose θ

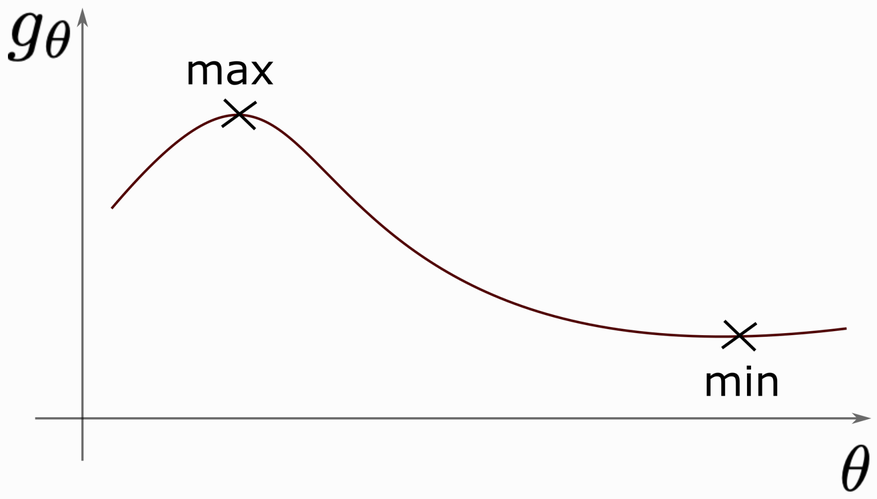
3. How does an Algorithm Learn?

Introduction

Optimization Theory

For what properties of g can we find

$$\min_{\theta} g_\theta$$



Good properties for g

- Linearity
- Convexity

Our case

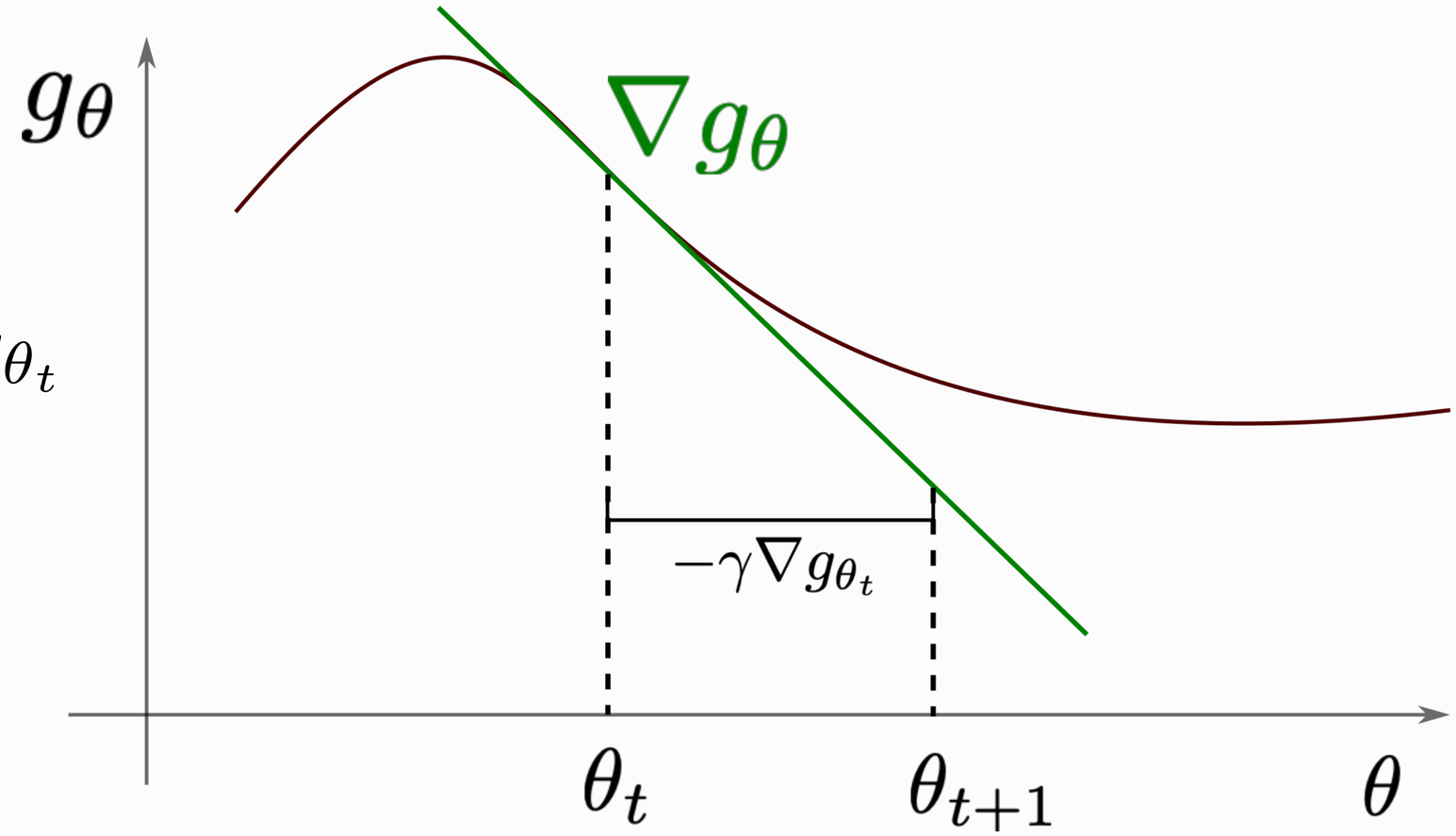
$$g_\theta = d(f_\theta, y) \quad d \text{ Some distance}$$

Gradient Descent

$$\theta_{t+1} \leftarrow \theta_t - \gamma \nabla g_{\theta_t}$$

+ Always applicable

- Greedy approach



Generalization

Problems with gradient decent

1. Local Minima
2. Overfitting

Applied to the shortest path problem

1. Stuck in a long path
2. Only know good path unprepared for construction sites

Datasplitting

- The given data is split into *train* and *validation* data.
- Typical amount:
  val-train <-> 20%-80%

Parameter Initialization

- Usual Gaussian distributed with mean 0 and variance equal to the hidden dimension.

4. Example for Applications

Coding Session

Thank you for you attention