Introduction to Machine Learning

Basic Concepts and Programming in Machine Learning

Thomas Schnake

12.02.2024

Roadmap

Session 1

a. Historical Background of Machine Learning

b. What is a Learning Algorithm?

c. How Does an Algorithm Learn?

Session 2

d. Implementing a Machine Learning Algorithm

a. Historical Background of Machine Learning

Historical Background of ML (Highlights)

- 1958 - Frank Rosenblatt introduced the first *perceptron*
- 1960s - Origin of Statistical Learning Theory (SLT).
- 1990s - SLT gained wide popularity with *Support Vector Machines.*
- 1990s - Convolutional and Recurrent Neural Network where introduced
- 2012 - Usage of GPUs in Neural Network increased performance drastically

[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.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. *In Proceedings of the 25th International Conference on Neural Information Processing Systems*, 1097–1105.

General Motivation in ML

How can a computer “learn” specific tasks, only by considering data?

Problems that an potentially be tackled with machine Learning.

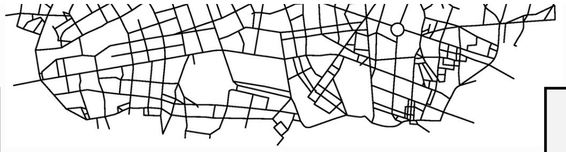
How long does it take to the airport?

Is this a cat or a dog?

What’s the weather like tomorrow?

Comparing deterministic and statistical approach to solve problems

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 may lead to shorter paths.

c) Stop any time.

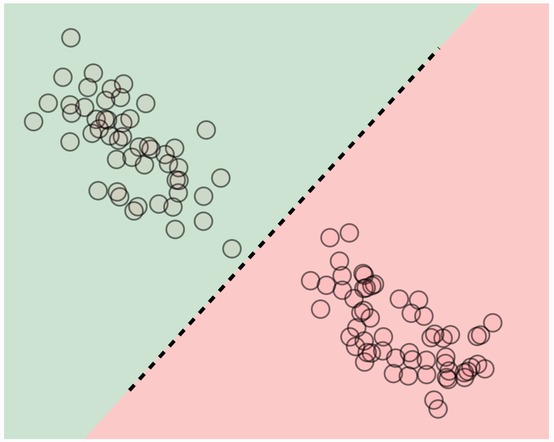
\- May not find shortest path.

\+ Cheap computation

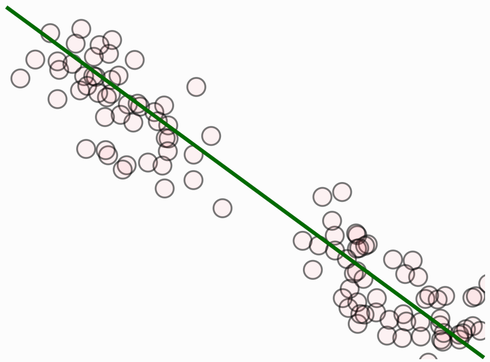
Typical scenarios in ML

Supervised Problems

Classification

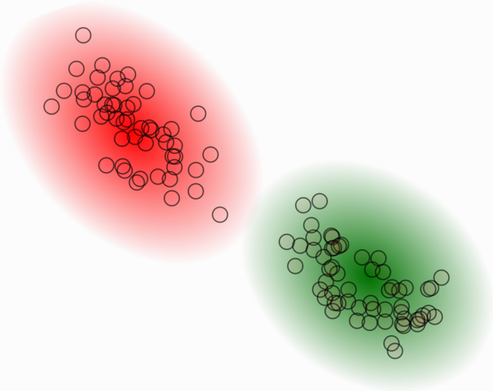


Regression



Unsupervised Problems

Clustering



Reinforcement learning

...

Common Machine Learning Problem - Mathematical Formulation

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

Common Machine Learning Problem - Example Dataset

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]

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

[6] 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

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

b. What is a Learning Algorithm?

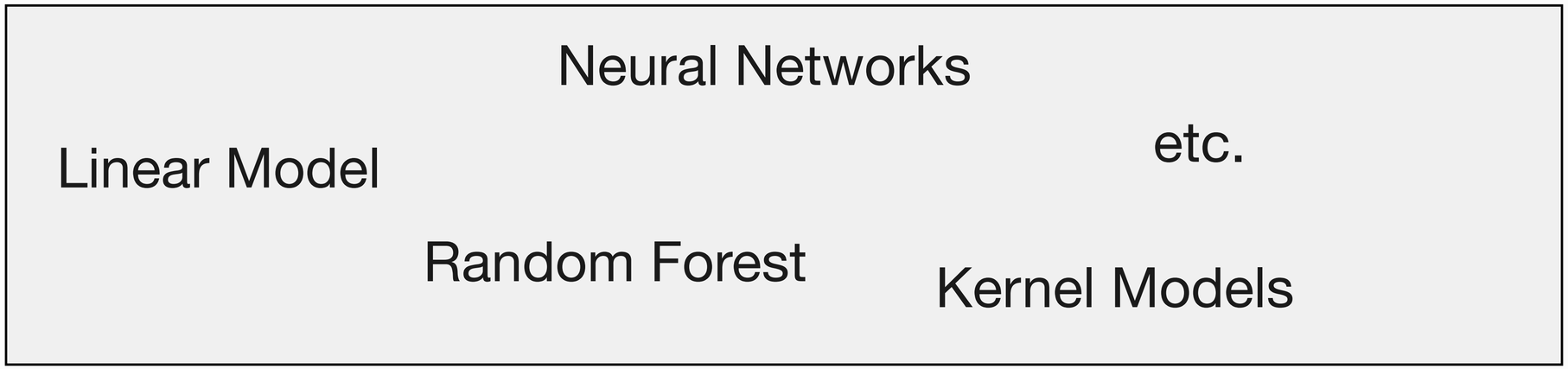
Choose f

2. What is a Learning Algorithm?

General Introduction

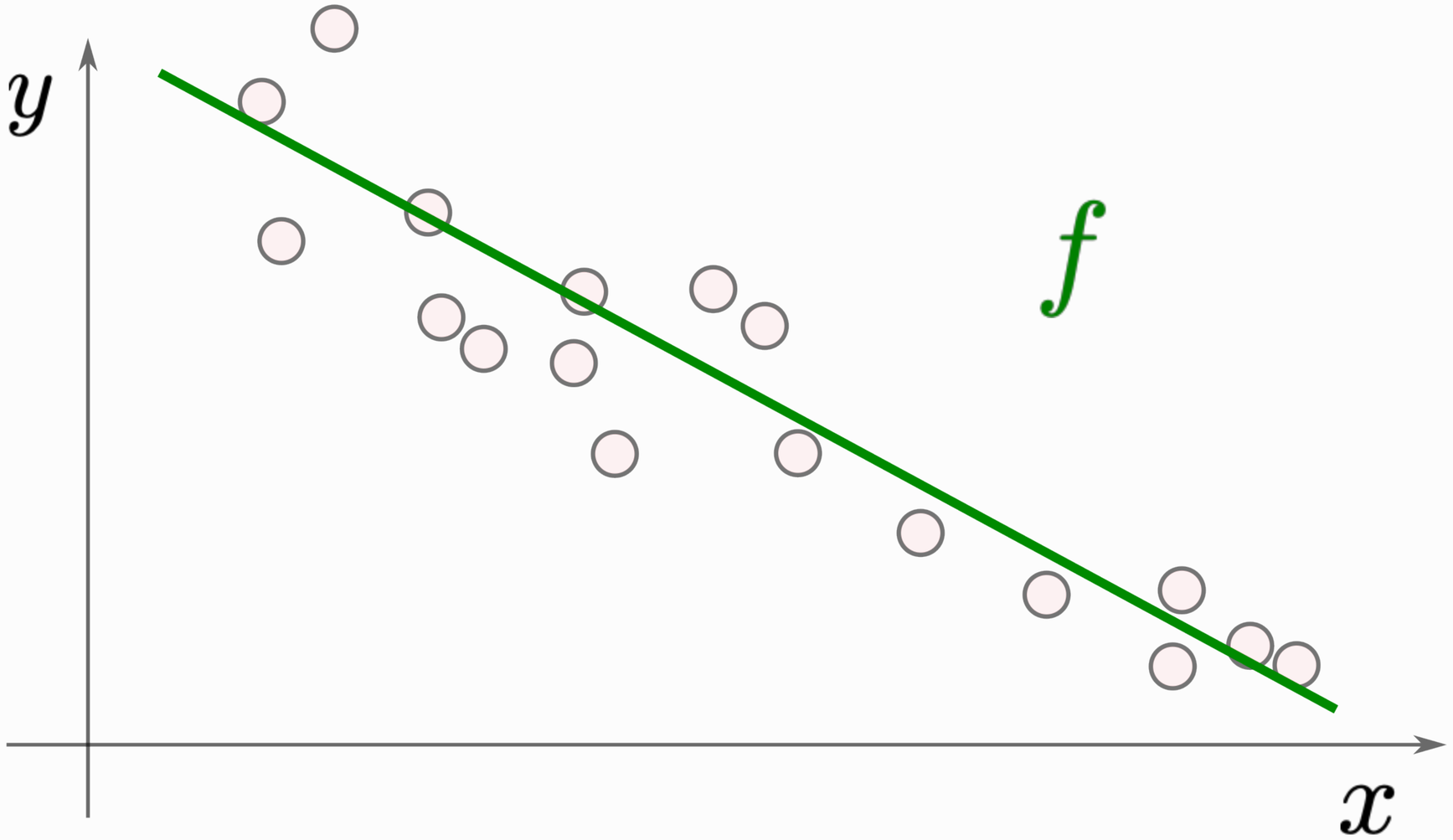
- Choice is problem specific
- *Inductive bias* plays a role
- Complexity vs. Simplicity

Some Models



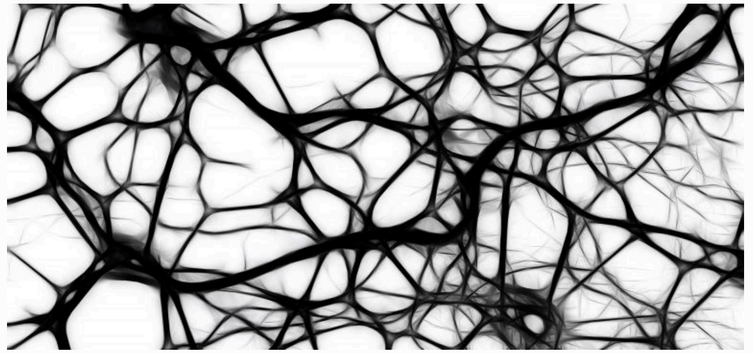
General Introduction to Learning Algorithm

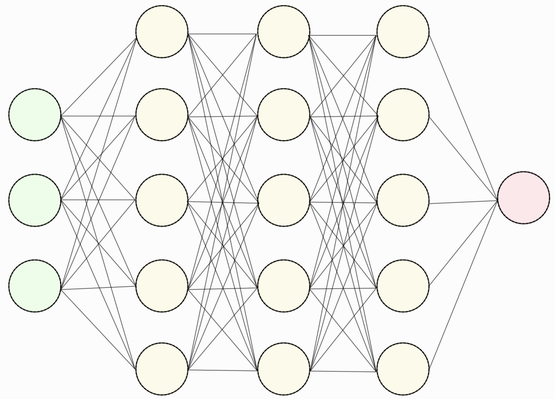
Linear Model



Neural Networks: Background and History

- First attends on imitating *neural plasticity* 1940s [7]
- Creation of *perceptron* in 1958 [8]
- Introduction of *backpropagation* in 1975 [9]
- Introduce LSTM, CNNs, RNN … (Deep Learning)
- Running NNs on GPUs since 2012 [4]





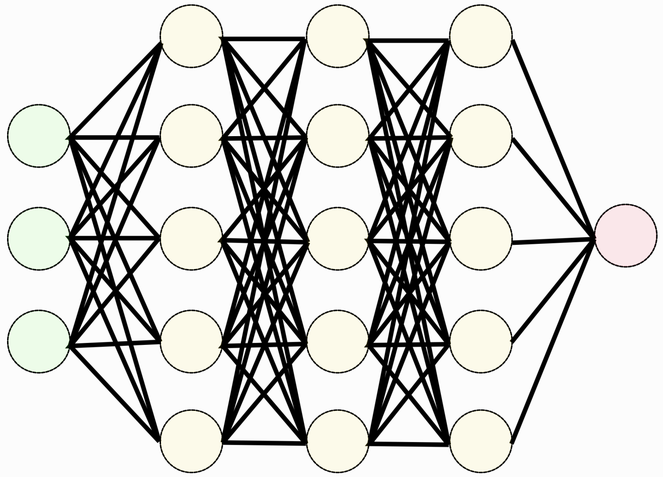
[7] D. Hebb, *The Organization of Behavior,* New York: Wiley, 1949.

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

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

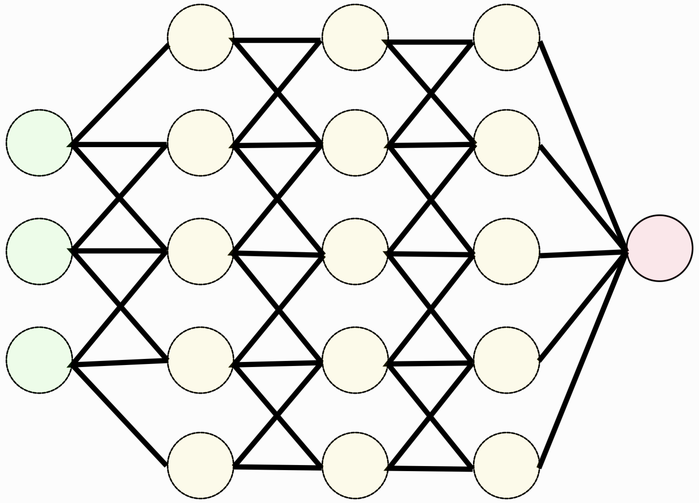
Neural Networks: Types

Vectorial and Transformer



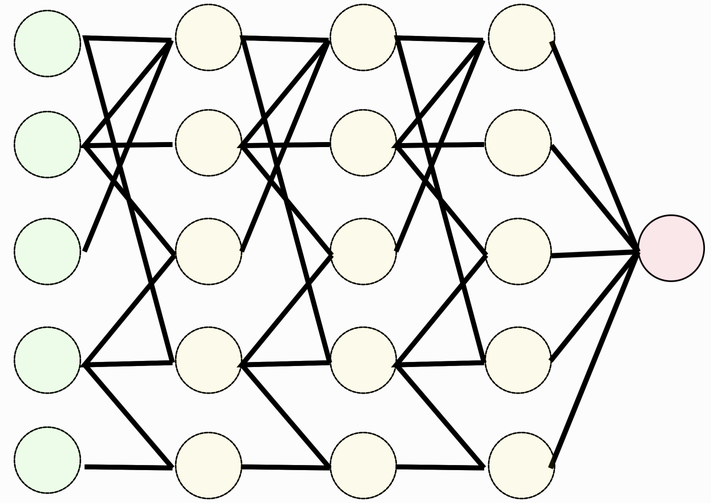
Fully connected

Convolutional



Locally connected

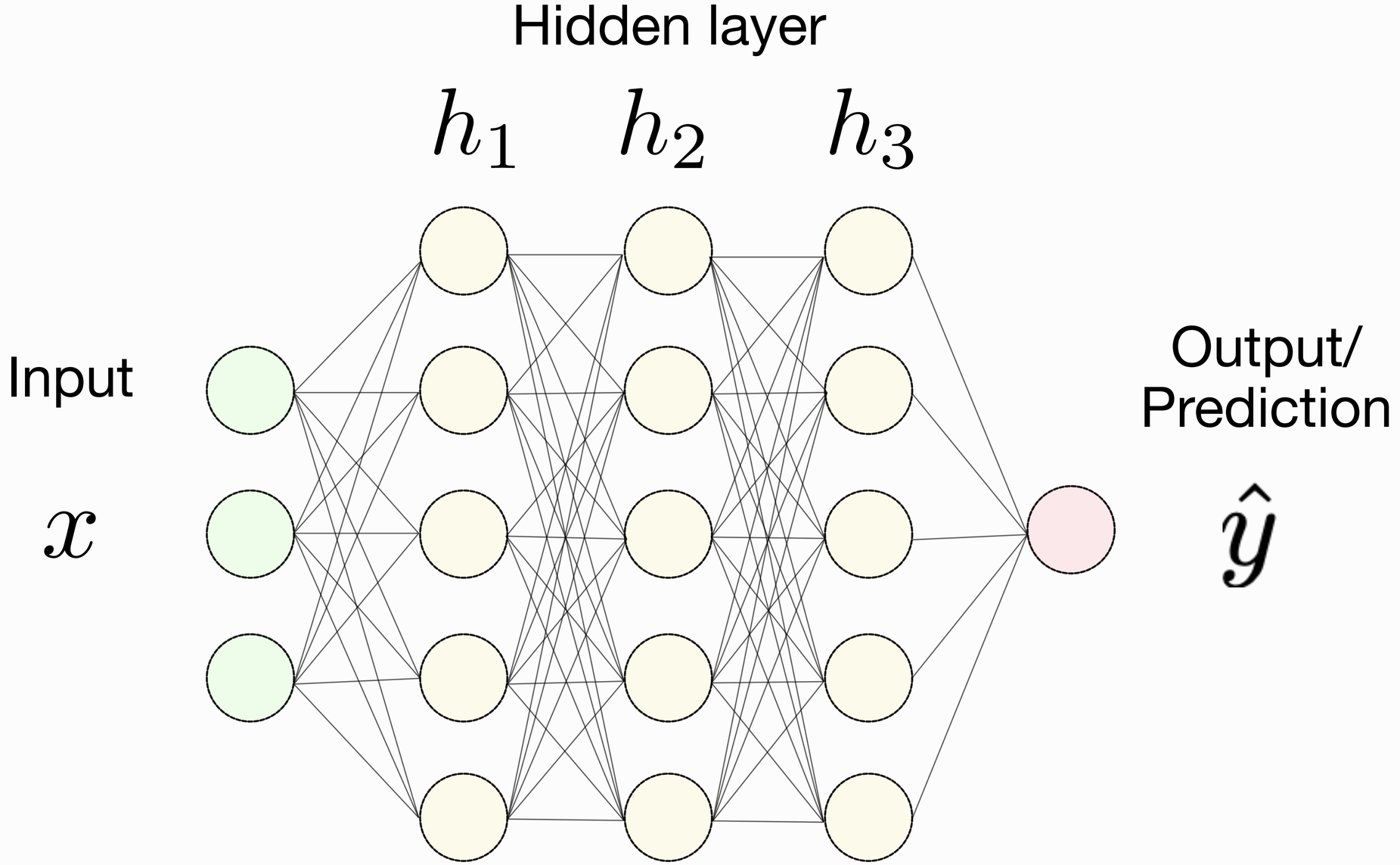
Graph



Custom connectivity

Recurrent: Sequential application

Layer-wise Visualization of a Vectorial NN

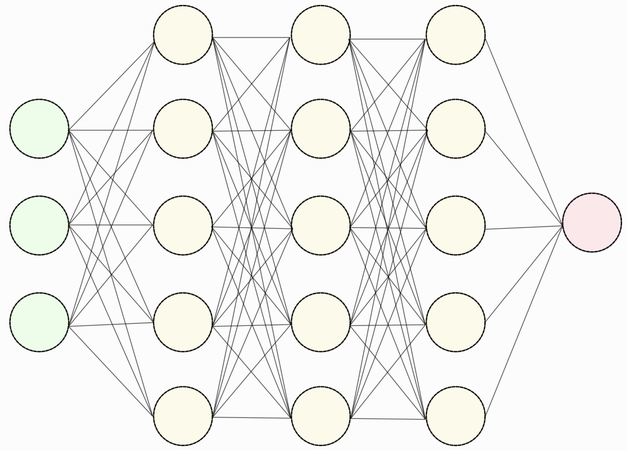


Propagation Rules in a Vectorial NN

Layer-wise 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

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

c. How Does an Algorithm Learn?

Choose θ

Introduction to optimization theory

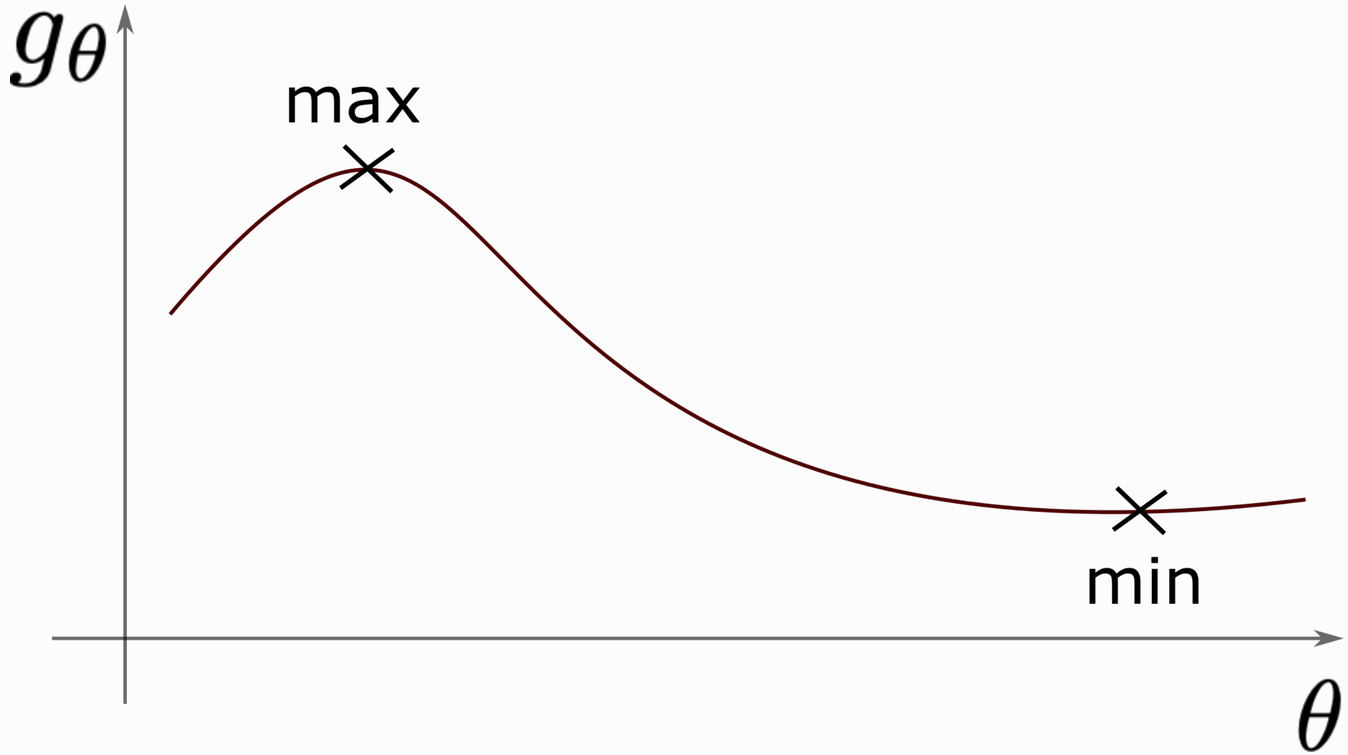
General problem:

Given a function g_θ, for what parameter θ is the function smallest

I.e. $\min_\theta g_\theta$

Good properties for g_θ

- Linearity
- Convexity



Optimization problem for Neural Networks

Our case

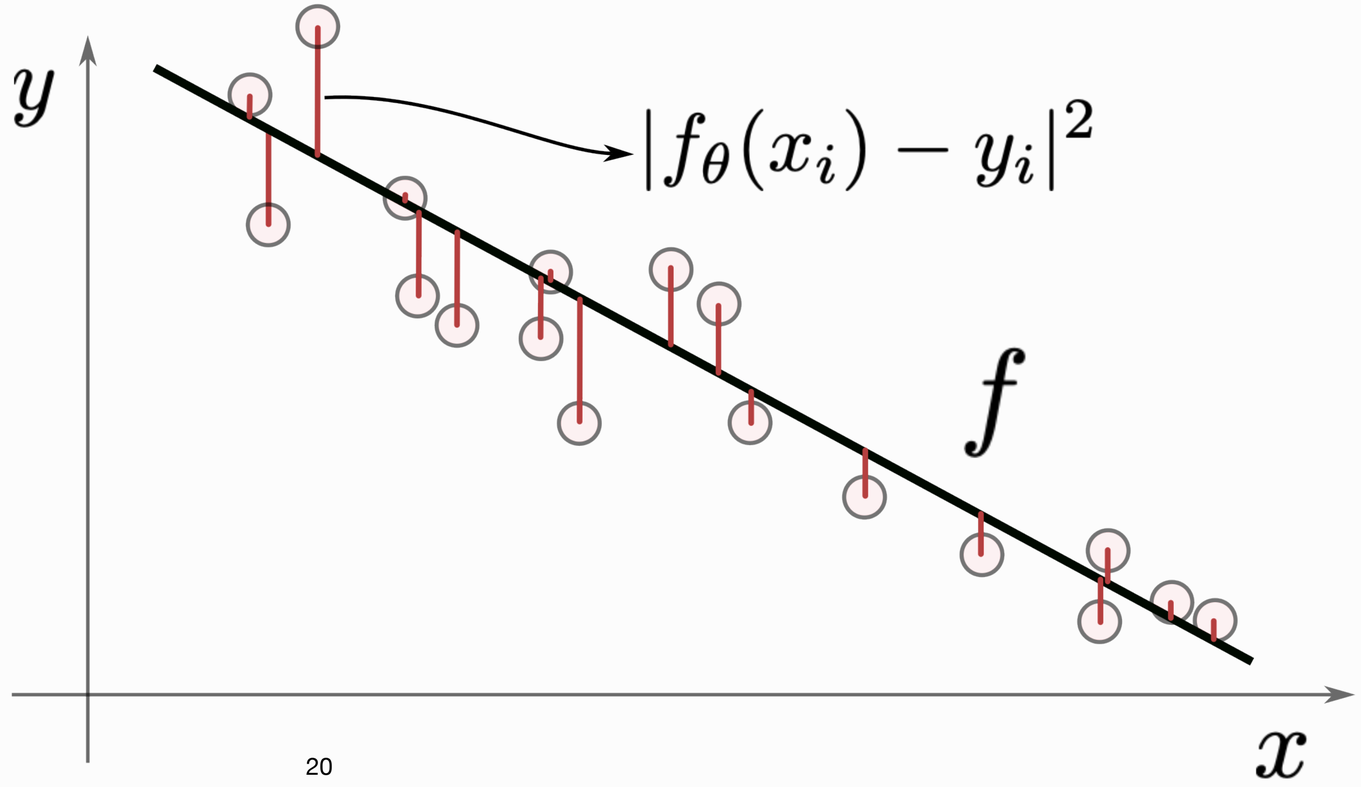
$$g_\theta = d(f_\theta, y)$$

d is e.g. l_2 distance

$$d(f_\theta, y) = \sum_i |f_\theta(x_i) - y_i|^2$$

Problem, g_θ is

- **Not** linear
- **Not** convex

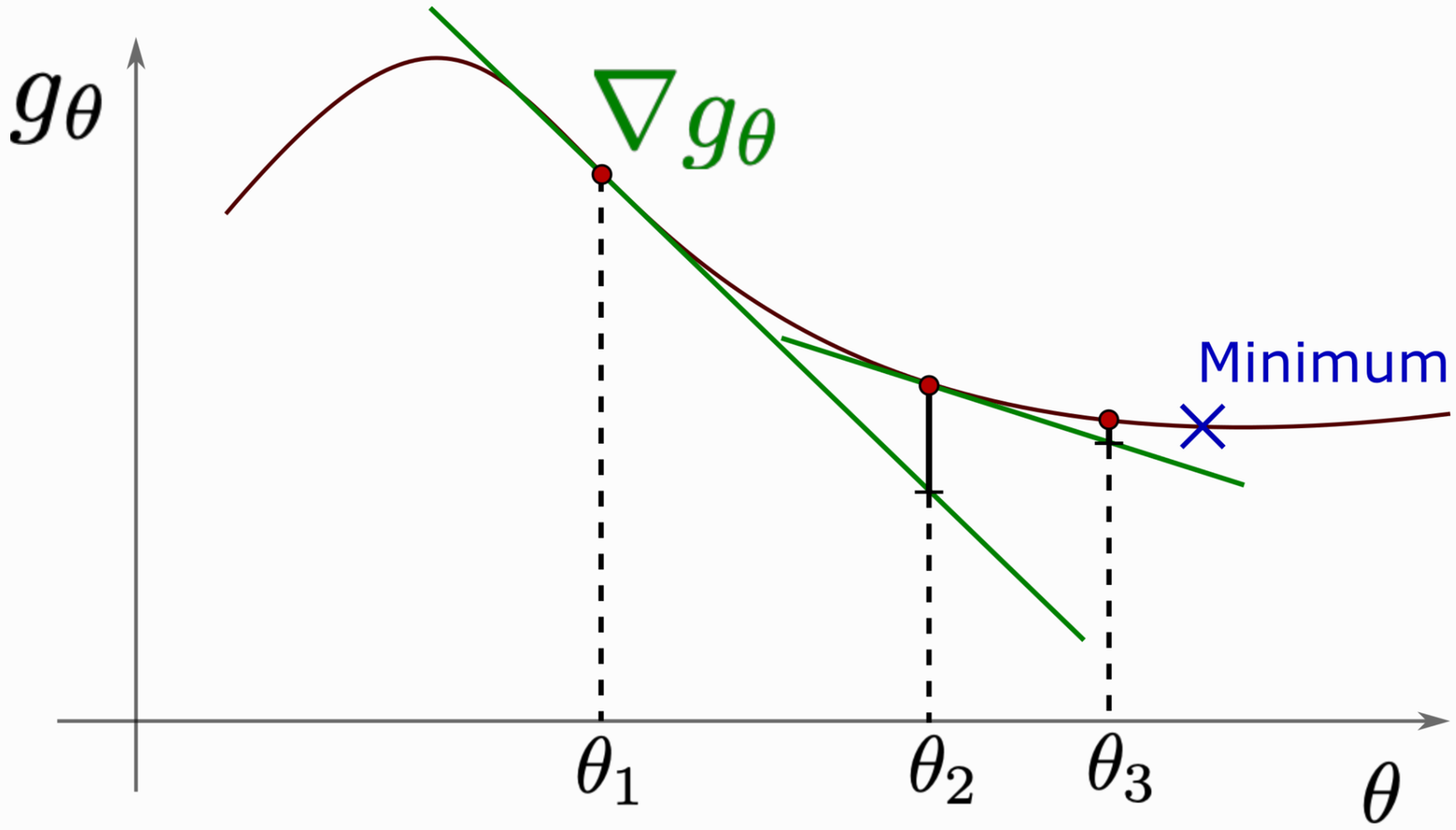


Finding NN parameters: Gradient Descent

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

+ Always applicable

- Greedy approach



Problems with gradient decent

- Local Minima
- Overfitting

Interpretation in the shortest path problem

- Stuck in a long path
- Only find good path in known cities

Additional Considerations

Datasplitting

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

Parameter Initialization

- The parameters needs to be initialized
- Typical is centered Gaussian distribution variance equal to the hidden dimension.

d. Implementing a Machine Learning Algorithm

Coding Session

Thank you for you attention