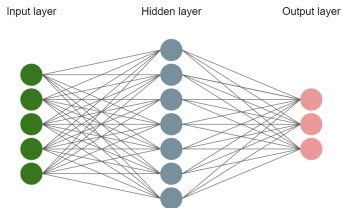


# Introduction to Machine Learning

## ML-Basics

## Models & Parameters



### Learning goals

- Understand that an ML model is simply a parametrized curve
- Understand that the hypothesis space lists all admissible models for a learner
- Understand the relationship between the hypothesis space and the parameter space

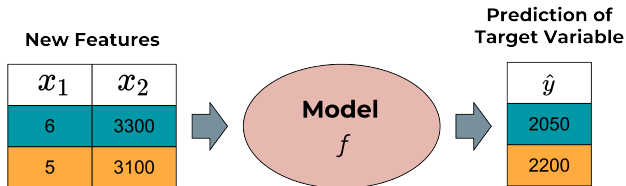
# WHAT IS A MODEL?

- A **model** (or **hypothesis**)

$$f : \mathcal{X} \rightarrow \mathbb{R}^g$$

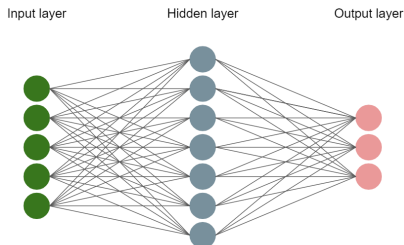
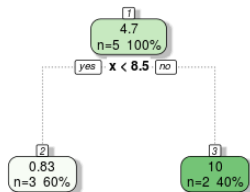
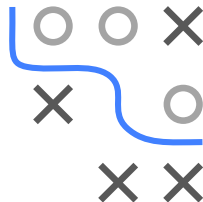
is a function that maps feature vectors to predicted target values.

- In conventional regression:  $g = 1$ ; for classification  $g$  is the number of classes, and output vectors are scores or class probabilities (details later).



# WHAT IS A MODEL? / 2

- $f$  is meant to capture intrinsic patterns of the data, the underlying assumption being that these hold true for *all* data drawn from  $\mathbb{P}_{xy}$ .
- It is easily conceivable how models can range from super simple (e.g., linear, tree stumps) to very complex (e.g., deep neural networks) and there are infinitely many choices how we can construct such functions.

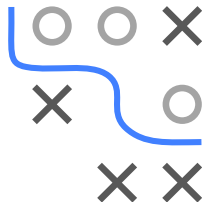


- In fact, ML requires **constraining**  $f$  to a certain type of functions.

# HYPOTHESIS SPACES

- Without restrictions on the functional family, the task of finding a “good” model among all the available ones is impossible to solve.
- This means: we have to determine the class of our model *a priori*, thereby narrowing down our options considerably. We could call that a **structural prior**.
- The set of functions defining a specific model class is called a **hypothesis space**  $\mathcal{H}$ :

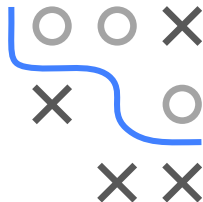
$$\mathcal{H} = \{f : f \text{ belongs to a certain functional family}\}$$



# PARAMETRIZATION

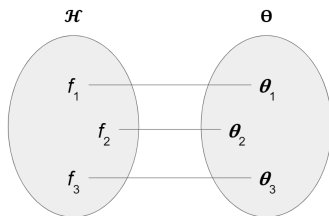
- All models within one hypothesis space share a common functional structure. We usually construct the space as **parametrized family of curves**.
- We collect all parameters in a **parameter vector**  $\theta = (\theta_1, \theta_2, \dots, \theta_d)$  from **parameter space**  $\Theta$ .
- They are our means of fixing a specific function from the family. Once set, our model is fully determined.
- Therefore, we can re-write  $\mathcal{H}$  as:

$$\mathcal{H} = \{f_{\theta} : f_{\theta} \text{ belongs to a certain functional family} \\ \text{parameterized by } \theta\}$$

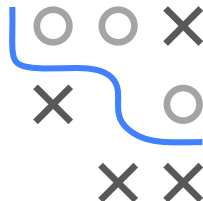


# PARAMETRIZATION / 2

- This means: finding the optimal model is perfectly equivalent to finding the optimal set of parameter values.
- The relation between optimization over  $f \in \mathcal{H}$  and optimization over  $\theta \in \Theta$  allows us to operationalize our search for the best model via the search for the optimal value on a  $d$ -dimensional parameter surface.

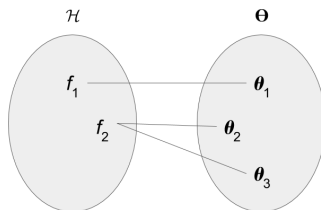
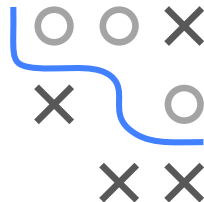


- $\theta$  might be scalar or comprise thousands of parameters, depending on the complexity of our model.



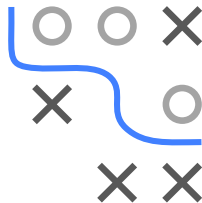
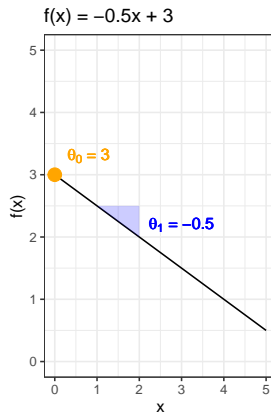
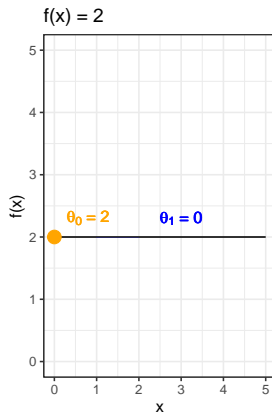
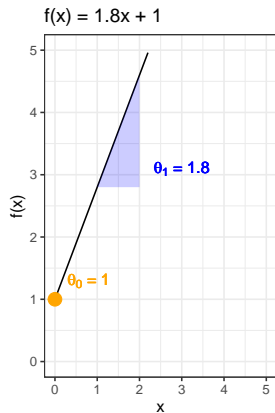
# PARAMETRIZATION / 3

- Short remark: In fact, some parameter vectors, for some model classes, might encode the same function. So the parameter-to-model mapping could be non-injective.
- We call this then a non-identifiable model.
- But this shall not concern us here.



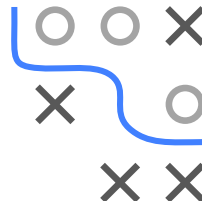
# EXAMPLE: UNIVARIATE LINEAR FUNCTIONS

$$\mathcal{H} = \{f : f(\mathbf{x}) = \theta_0 + \theta_1 x, \theta \in \mathbb{R}^2\}$$



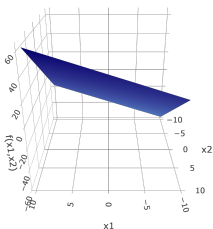


# EXAMPLE: BIVARIATE QUADRATIC FUNCTIONS

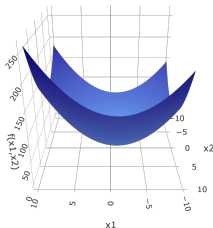


$$\mathcal{H} = \{f : f(\mathbf{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2, \theta \in \mathbb{R}^6\},$$

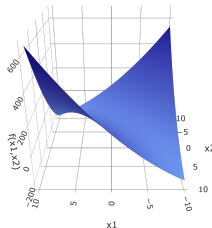
$$f(x) = 3 + 2x_1 + 4x_2$$



$$f(x) = 3 + 2x_1 + 4x_2 + 1x_1^2 + 1x_2^2$$



$$f(x) = 3 + 2x_1 + 4x_2 + 1x_1^2 + 1x_2^2 + 4x_1 x_2$$



# EXAMPLE: RBF NETWORK

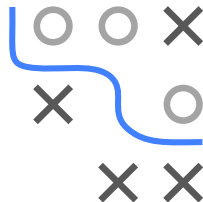
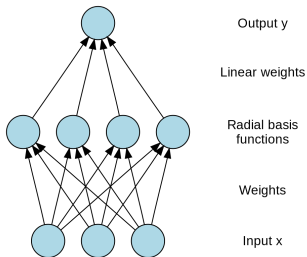
Radial basis function networks with Gaussian basis functions

$$\mathcal{H} = \left\{ f : f(\mathbf{x}) = \sum_{i=1}^k a_i \rho(\|\mathbf{x} - \mathbf{c}_i\|) \right\},$$

where

- $a_i$  is the weight of the  $i$ -th neuron,
- $\mathbf{c}_i$  its center vector, and
- $\rho(\|\mathbf{x} - \mathbf{c}_i\|) = \exp(-\beta\|\mathbf{x} - \mathbf{c}_i\|^2)$  is the  $i$ -th radial basis function with bandwidth  $\beta \in \mathbb{R}$ .


Usually, the number of centers  $k$  and the bandwidth  $\beta$  need to be set in advance (so-called *hyperparameters*).



## ©

A 3D surface plot of the function  $f(x_1, x_2)$ . The surface is blue and has two prominent peaks and a saddle point. The peaks are located at  $(x_1, x_2) = (-2, 2)$  and  $(x_1, x_2) = (2, 2)$ , both reaching a value of 0.06. The saddle point is located at  $(x_1, x_2) = (0, 0)$  with a value of 0.03. The axes are labeled  $x_1$  and  $x_2$  on the horizontal plane, and  $f(x_1, x_2)$  on the vertical axis. The vertical axis has tick marks at 0.02, 0.04, and 0.06. The horizontal axes have tick marks at -4, -2, 0, 2, and 4.

A 3D surface plot of the function  $f(x_1, x_2)$ . The surface is blue and has three distinct peaks. The central peak is at  $(0, 0)$  with a height of approximately 0.01. The two side peaks are at  $(2, 2)$  and  $(2, -2)$ , both reaching a height of approximately 0.06. The axes are labeled  $x_1$  and  $x_2$  on the horizontal plane, and  $f(x_1, x_2)$  on the vertical axis. The vertical axis has tick marks at 0.02, 0.04, and 0.06. The horizontal axes have tick marks at -4, -2, 0, 2, and 4.



A 3D surface plot of a function  $f(x_1, x_2)$ . The surface is blue and has three distinct peaks. Each peak is marked with an orange dot. The vertical axis is labeled  $f(x_1, x_2)$  and ranges from 0 to 0.06. The horizontal axes are labeled  $x_1$  and  $x_2$ , both ranging from -4 to 4.

$$a_1 = 0.2, a_2 = 0.45, a_3 = 0.35$$

$$c_1 = (2, -2), c_2 = (0, 0),$$

$$c_3 = (-3, 2)$$

