Introduction to Machine Learning

ML-Basics
Models & Parameters

Learning goals

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

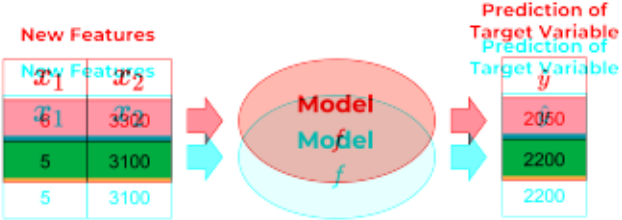
WHAT IS A MODEL?

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

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

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

- In regression: $g = 1$; in classification, g is the number of classes, and output vectors are scores or class probabilities.

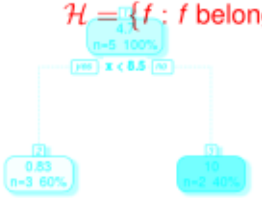


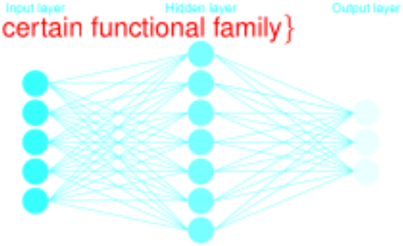
HYPOTHESIS SPACES

- Without restrictions on the functional family, the task of finding a "good" model among all available models is impossible to solve.
- We have to determine the class of our model *a priori*, thereby narrowing down the search space. We call this a **structural prior**.
- The set of functions defining a specific model class is called a **hypothesis space** $\mathcal{H}$:

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

- 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.

PARAMETRIZATION

- All models within a hypothesis space share a common functional structure. We usually construct the space as **parametrized family of functions**.
- We collect all parameters in a **parameter vector** $\theta = (\theta_1, \theta_2, \dots, \theta_d)$ from **parameter space** Θ.
- 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:

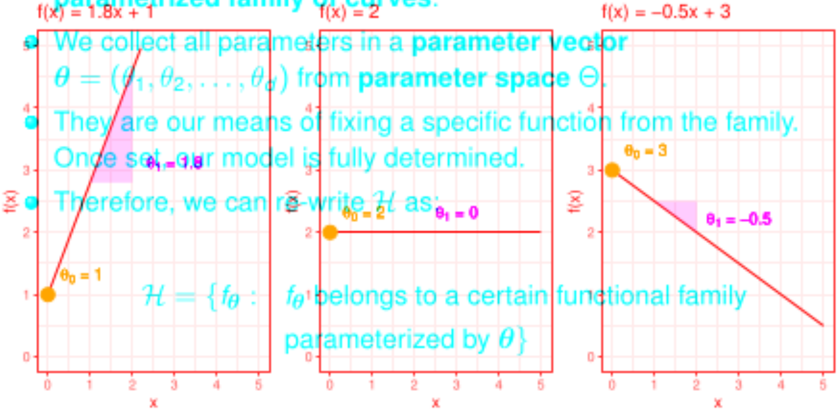
$$\mathcal{H} = \{f_\theta : \; f_\theta \text{ belongs to a certain functional family parameterized by } \theta\}$$

EXAMPLE: UNIVARIATE LINEAR FUNCTIONS

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

- 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** $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ from **parameter space** Θ.
- 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:

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

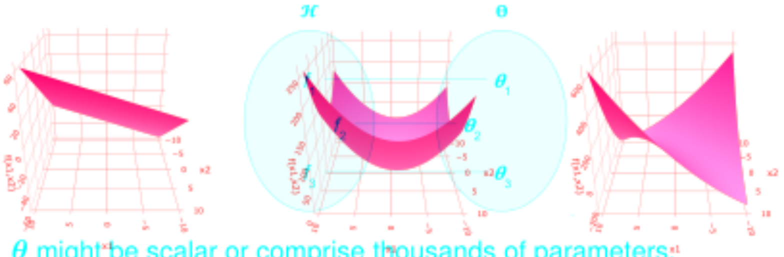


EXAMPLE: BIVARIATE QUADRATIC FUNCTIONS

- This means: finding the optimal model is perfectly equivalent to finding the optimal set of parameter values.

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

- 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.



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

EXAMPLE: RBF NETWORK

Radial basis function networks with Gaussian basis functions

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

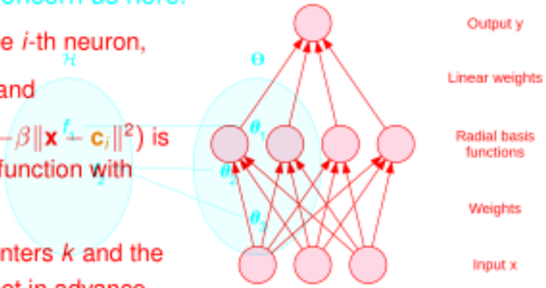
where

- 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 β need to be set in advance (so-called *hyperparameters*).

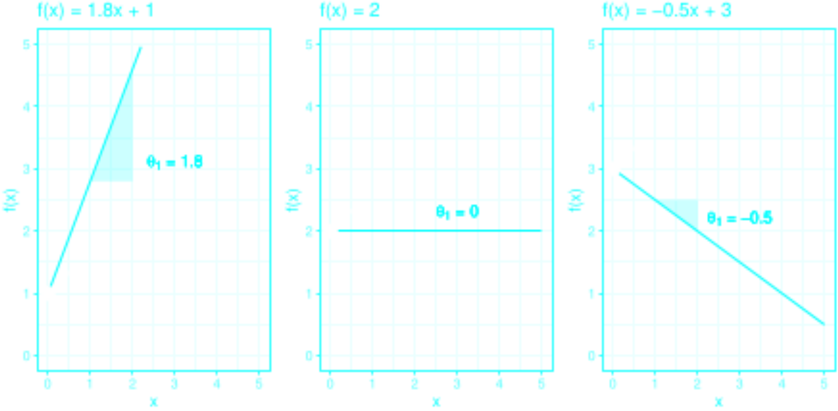
- ... model classes, might encode the same function. So the parameter-to-model mapping could be non-injective.
- We call this then a non-identifiable model.

But this shall not concern us here.



EXAMPLE: UNIVARIATE LINEAR FUNCTIONS

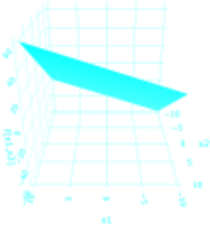
$$\mathcal{H} = \{f : f(\mathbf{x}) = \quad\quad + \theta_1 x, \boldsymbol{\theta} \in \mathbb{R}^2\}$$



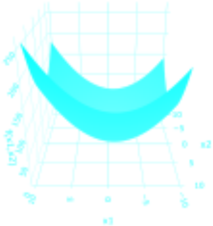
EXAMPLE: BIVARIATE QUADRATIC FUNCTIONS

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

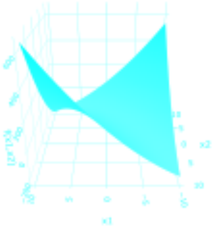
$f(x) = \quad + 2x_1 + 4x_2$



$f(x) = \quad + 2x_1 + 4x_2 +$
$\quad + \quad x_1^2 + \quad x_2^2$



$f(x) = \quad + 2x_1 + 4x_2 +$
$\quad + \quad x_1^2 + \quad x_2^2 + \quad x_1 x_2$



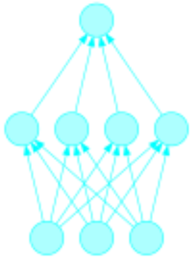
EXAMPLE: RBF NETWORK

Radial basis function networks with Gaussian basis functions

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

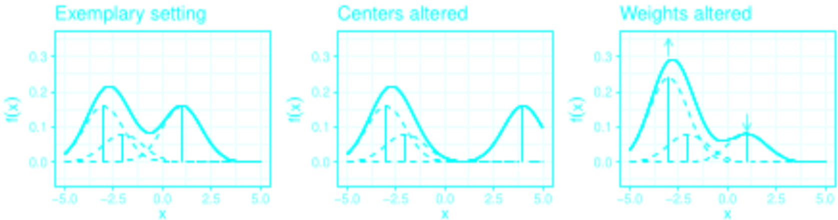
where

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



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

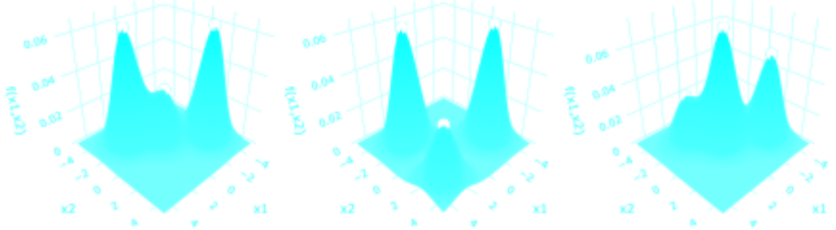
EXAMPLE: RBF NETWORK



$a_1 = 0.4, a_2 = 0.2, a_3 = 0.4$
$c_1 = -3, c_2 = -2, c_3 = 1$

$a_1 = 0.4, a_2 = 0.2, a_3 = 0.4$
$c_1 = -3, c_2 = -2,$

$a_1 = 0.6, a_2 = 0.2, a_3 = 0.2$
$c_1 = -3, c_2 = -2, c_3 = 1$



$a_1 = 0.4, a_2 = 0.2, a_3 = 0.4$
$c_1 = (2, -2), c_2 = (0, 0),$
$c_3 = (-3, 2)$

$a_1 = 0.4, a_2 = 0.2, a_3 = 0.4$
$c_1 = (2, -2),$,
$c_3 = (-3, 2)$

$a_1 = 0.2, a_2 = 0.45, a_3 = 0.35$
$c_1 = (2, -2), c_2 = (0, 0),$
$c_3 = (-3, 2)$