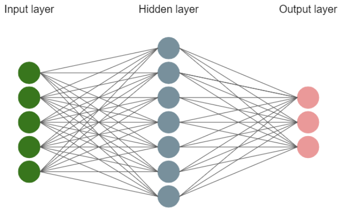
Introduction to Machine Learning

ML-Basics: Models & Parameters



Learning goals

- tbd

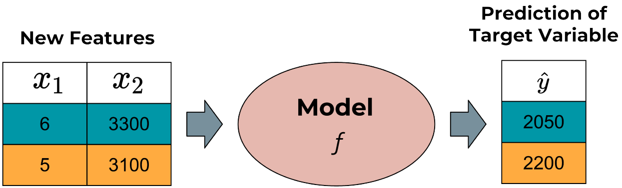
WHAT IS A MODEL?

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

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

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

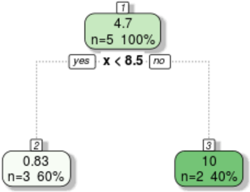
- Loosely speaking: if f is fed a set of features, it will output the target corresponding to these feature values under our hypothesis.

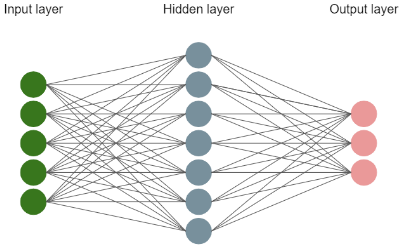


In conventional regression we will have $g = 1$; for classification g equals the number of classes, and output vectors are scores or class probabilities (details later).

WHAT IS A MODEL?

- 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., tree stumps) to reasonably complex (e.g., variational autoencoders), and how there is an infinite number of them.





- In fact, machine learning 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.
- The set of functions defining a specific model class is called a **hypothesis space** $\mathcal{H}$:

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

- BB: drop "family of curves" somewhere

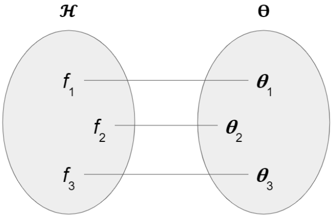
PARAMETERS OF A MODEL

- All models within one hypothesis space share a common functional structure.
- In fact, the only aspect in which they differ is the values of **parameters**.
- We usually subsume all these parameters in a **parameter vector** $\boldsymbol{\theta} = (\theta_1, \theta_2, ...)$ from a **parameter space** Θ.
- They are our means of configuration: once set, our model is fully determined.
- Therefore, we can re-write $\mathcal{H}$ as:

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

PARAMETERS OF A MODEL

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

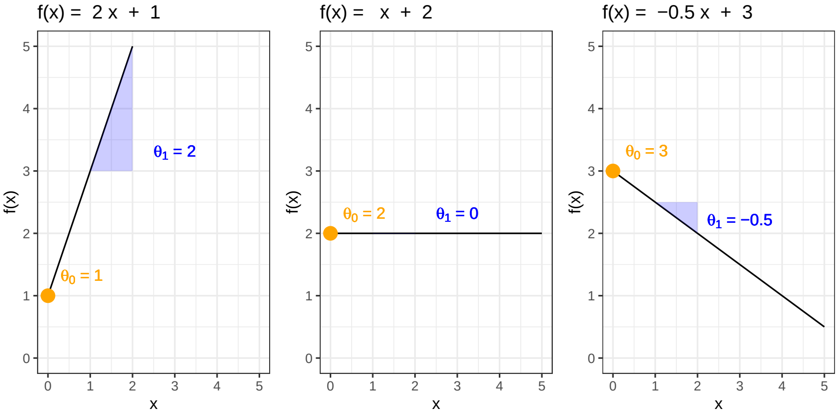


- $\boldsymbol{\theta}$ might be scalar or comprise thousands of parameters, depending on the complexity of our model.

EXAMPLES FOR HYPOTHESIS SPACES

Example 1: Hypothesis space of univariate linear functions

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



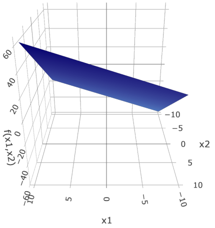
EXAMPLES FOR HYPOTHESIS SPACES

Example 2: Hypothesis space of bivariate quadratic functions

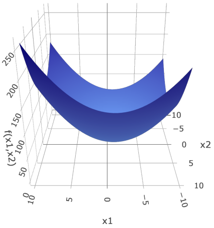
$$\mathcal{H} = \{f : f(\mathbf{x}) = \theta_0 + p^T\mathbf{x} + \mathbf{x}Q\mathbf{x}^T =$$
$$= \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2, \boldsymbol{\theta} \in \mathbb{R}^6\},$$

where $p = \begin{pmatrix} \theta_1 & \theta_2 \end{pmatrix}$ and $Q = \begin{pmatrix} \theta_3 & \frac{1}{2}\theta_5 \\ \frac{1}{2}\theta_5 & \theta_4 \end{pmatrix}$.

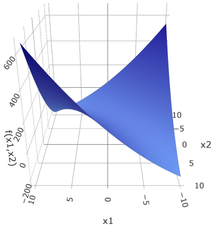
$f(x) = 3 + 2x_1 + 4x_2$



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



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



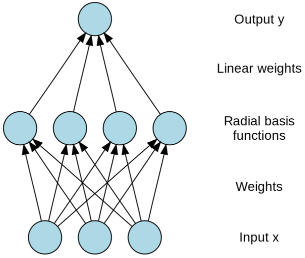
EXAMPLES FOR HYPOTHESIS SPACES

Example 3: Hypothesis space of radial basis function networks with Gaussian basis functions

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

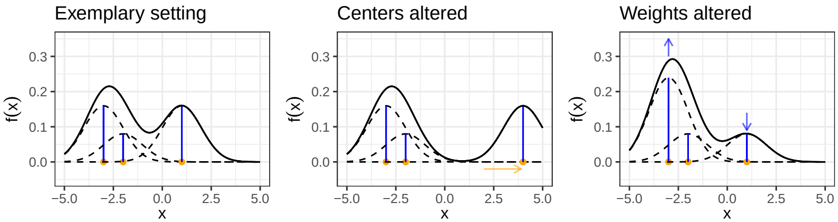
where

- 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, n, and the bandwidth β need to be set in advance (so-called *hyperparameters*).

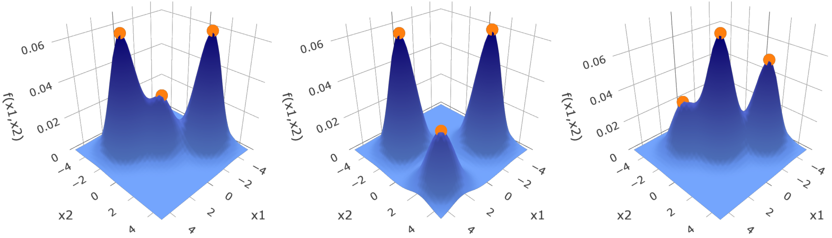
EXAMPLES FOR HYPOTHESIS SPACES



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

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

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



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

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

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