

# **Introduction to Machine Learning**

## **Introduction: Models & Parameters**

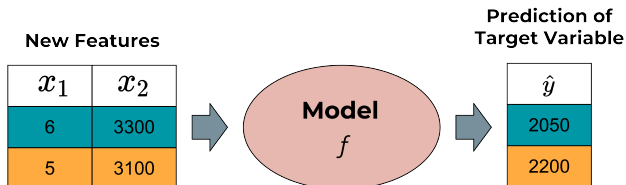
# WHAT IS A MODEL?

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

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

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

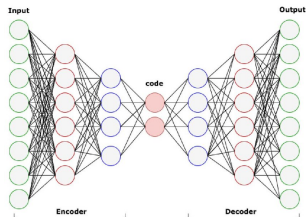
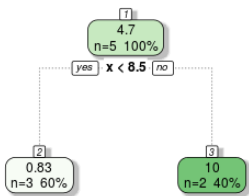
- Loosely speaking: if  $f$  is fed a set of features, it will output the target corresponding to these feature values under our hypothesis.



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**  $\mathcal{H}$ :

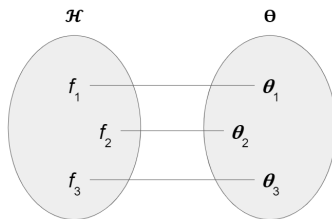
$$\mathcal{H} = \{f : f \text{ belongs to a certain functional family}\}$$

# 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**  $\theta = (\theta_1, \theta_2, \dots)$  from a **parameter space**  $\Theta$ .
- They are our means of configuration: once set, our model is fully determined.

# 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  $\theta \in \Theta$  allows us to operationalize our search for the best model via the search for the optimal value on a  $p$ -dimensional parameter surface.

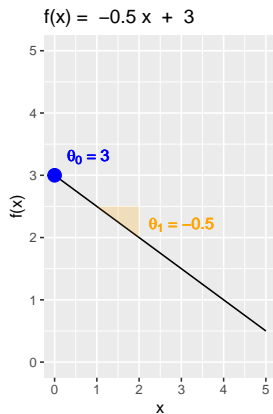
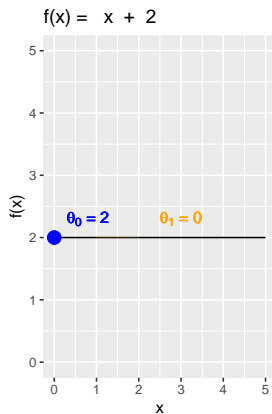
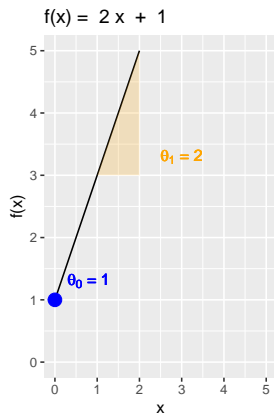


- $\theta$  might be scalar or comprise thousands of parameters, depending on the complexity of our model.

# 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

$$\begin{aligned}\mathcal{H} &= \{f : f(\mathbf{x}) = \theta_0 + P\mathbf{x}^T + \mathbf{x}Q\mathbf{x}^T = \\ &= \theta_0 + \theta_1x_1 + \theta_2x_2 + \theta_3x_1^2 + \theta_4x_2^2 + \theta_5x_1x_2, \boldsymbol{\theta} \in \mathbb{R}^6\}\end{aligned}$$



# EXAMPLES FOR HYPOTHESIS SPACES

## Example 3: Hypothesis space of radial basis function networks

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

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  $\beta$  need to be set in advance (so-called *hyperparameters*).

