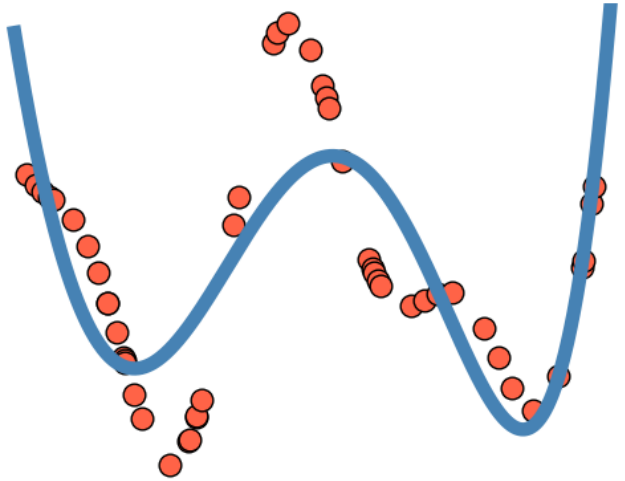
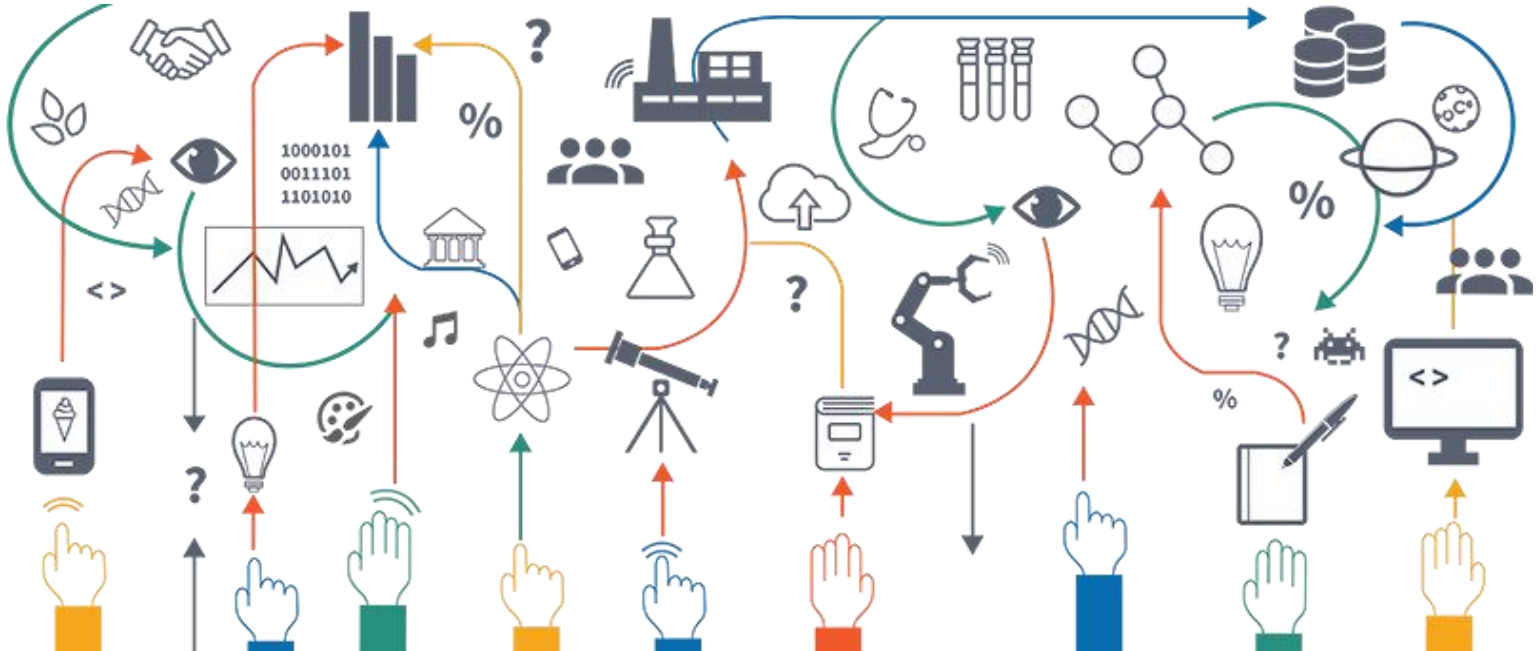
4 – Function Approximation and Supervised Learning

Mathematics of Data Science



Lecturer: Dominik Dold



University of Vienna, WiSe 2025

Master's programme in Data Science

Motivation

So far, we did not discuss “labelled” data!

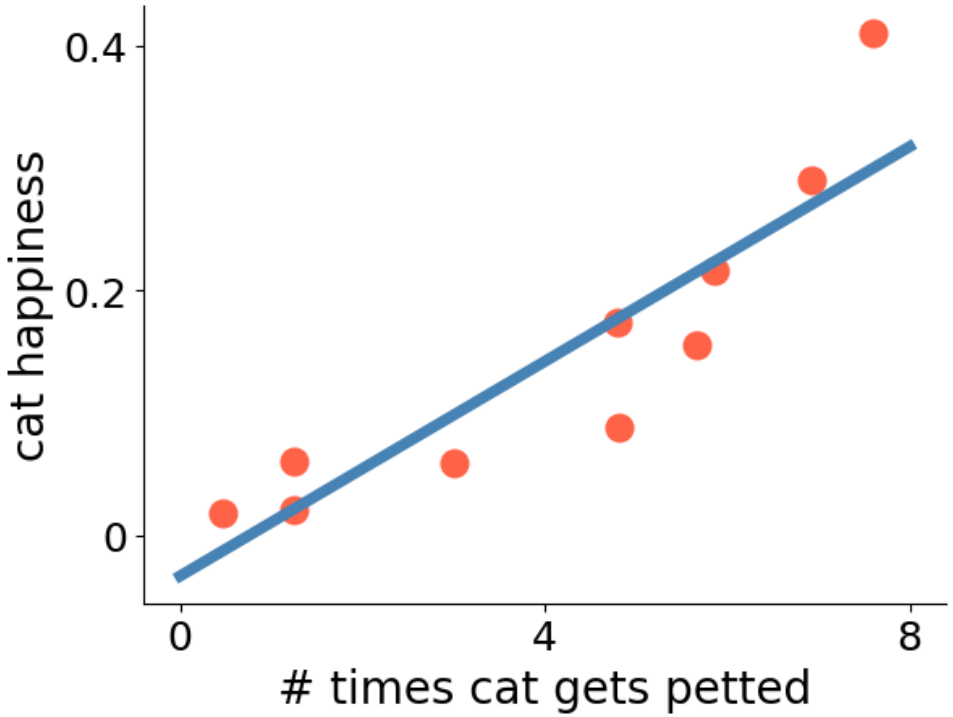
Meaning: we looked at $p(\boldsymbol{x})$, but not $p(\boldsymbol{y} \mid \boldsymbol{x})$

Classification

$\boldsymbol{x}$

$\boldsymbol{y}$ Cat Dog

Regression



Content

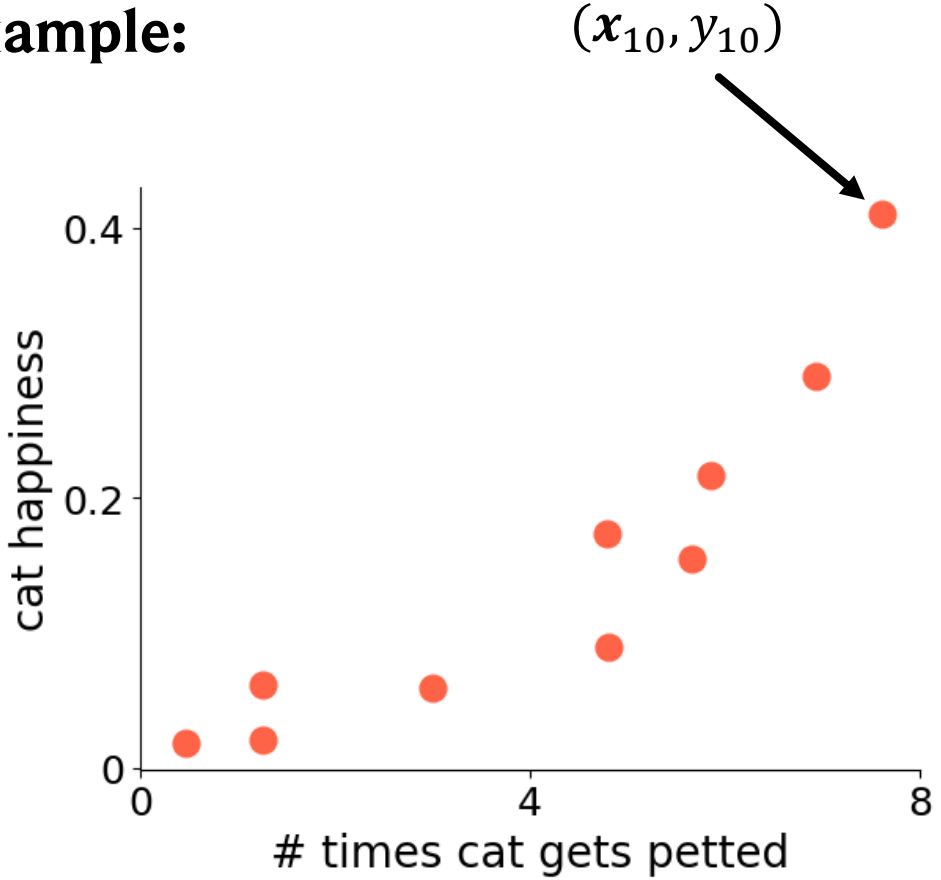
- Hypothesis set and risk
- Linear Regression
- Polynomials
- Universal function approximators
- Weierstrass approximation theorem
- Stone-Weierstrass theorem

Fitting data: the hypothesis set

What is our usual setup?

We have n data samples $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, where $\boldsymbol{x}_i$ is a feature vector and $\boldsymbol{y}_i$ whatever we want to predict!

Example:



Task: For new, unlabelled data samples, we want to predict $\boldsymbol{y}$!

We cannot try to fit all possible functions, so we restrict ourselves to a **hypothesis set**:

$\mathcal{H} =$ set of functions to explain the data

For example: $\mathcal{H} = \{x, \ \ 2 \cdot x^2\}$

We can also have a whole span of functions:

$$\mathcal{H} = \{\alpha + \beta x + \gamma x^2 \mid \alpha, \beta, \gamma \in \mathbb{R}\}$$

Empirical risk minimization

Callback to the random number lectures:
we assume that our data is sampled from an underlying distribution $D = p(\boldsymbol{y}, \boldsymbol{x})$.
Let's use a loss function $\mathcal{L}$ to measure how well our model performs (e.g., $\mathcal{L}(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|^2$).

Risk: For any function $h \in \mathcal{H}$, its risk $R(h)$ tells us how well the function fits the data

$$R(h) = \mathrm{E}_{(\boldsymbol{x},\boldsymbol{y})\sim D}[\mathcal{L}(h(\boldsymbol{x}), \boldsymbol{y})]$$

Empirical risk: Assume we have m data samples. Then the empirical risk $\hat{R}(h)$ is

$$\hat{R}(h) = \sum_{i=1}^{m} \mathcal{L}(h(\boldsymbol{x}_i), \boldsymbol{y}_i)$$

Fitting the data: Find $\hat{h} \in \mathcal{H}$ **that minimizes** $\hat{R}$: $\hat{h} = \mathrm{argmin}_{h\in\mathcal{H}}\left(\hat{R}(h)\right)$

The risk of our empirical risk minimizer is bounded by two terms:

$$R(\hat{h}) \leq \sup_{h\in\mathcal{H}} \left|R(h) - \hat{R}(h)\right| + \inf_{g\in\mathcal{H}} \hat{R}(g)$$

How well does our fit work on the full data distribution?

Generalization error

Interpolation error

How well can we fit the training data?

Why the supremum?

$$R(\hat{h}) \leq \sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}(h) \right| + \inf_{g \in \mathcal{H}} \hat{R}(g)$$

What is our aim?

1. Given a finite sample from a dataset, we want to minimize the risk (= solving our task).
2. Depending on the dataset (and training samples), **any of our functions** in $\mathcal{H}$ could be the empirical risk minimizer!
3. We do **not** know which function in $\mathcal{H}$ this is in advance!
4. Thus, the function $\hat{h}$ minimizing the empirical risk **might be one with high generalization error** $\left| R(\hat{h}) - \hat{R}(\hat{h}) \right|$.

To account for this, we use the supremum to **bound** the generalization error.
In other words: this is the **worst generalization error** we can get using a hypothesis in $\mathcal{H}$ (again, empirical risk minimization might select exactly such a "bad" function, as any function in $\mathcal{H}$ is allowed as a solution).

Note: this is a property of the whole set $\mathcal{H}$, i.e., shared by **all** functions in $\mathcal{H}$.

Also note: usually, there is a trade-off between generalization and interpolation error. However, both can be low if, for instance, our data is native to $\mathcal{H}$. E.g., if the underlying data relationship is linear, then linear functions will fit the training data well and generalize well to new data!

Example: linear regression

Let's look at a concrete example: linear regression (your baseline for everything!)

In this case, we have $\mathcal{H} = \{\boldsymbol{w}^{\mathbf{T}}\boldsymbol{x} \mid \boldsymbol{w} \in \mathbb{R}^N\}$, the set of all affine functions.

Empirical risk minimization gives us:

$$\inf_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}) = \inf_{\boldsymbol{w}} \left\| \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{w}^T \boldsymbol{x}_i - y_i) \right\|^2$$

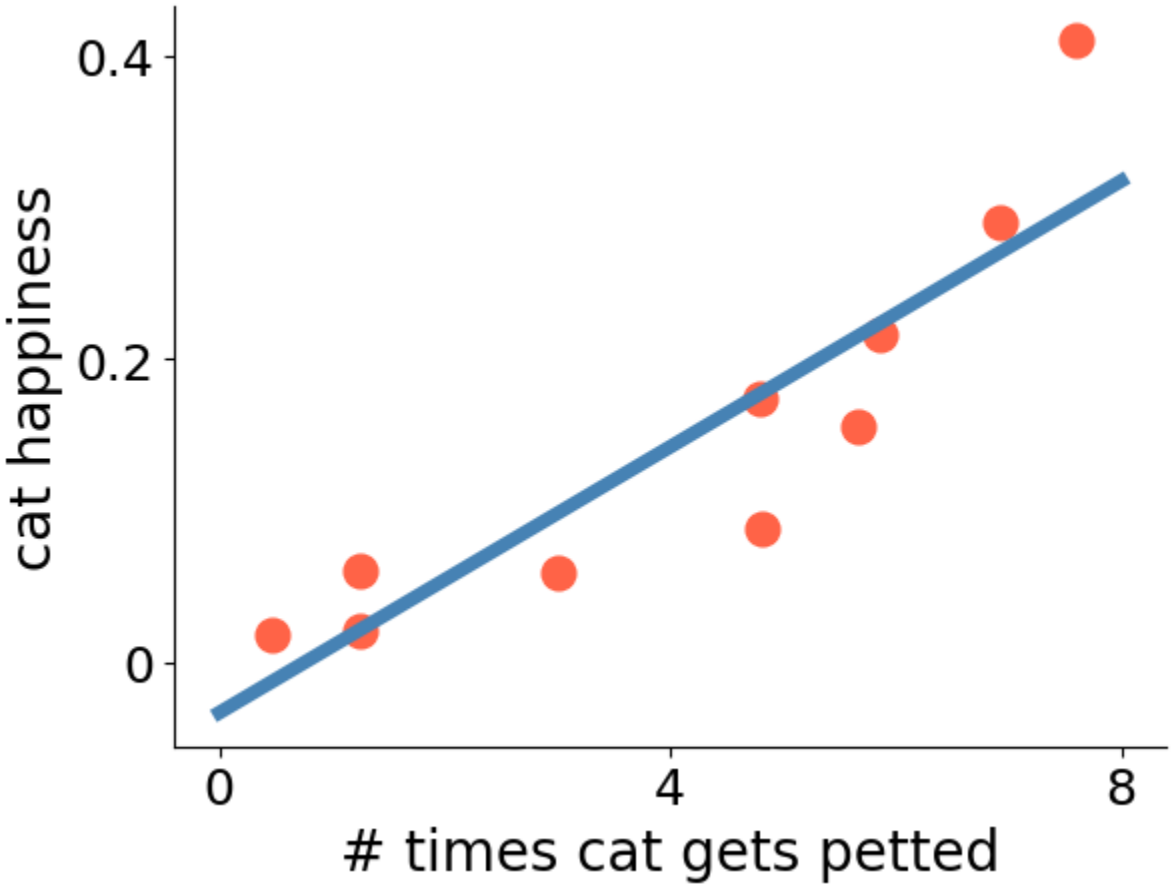
$$= \inf_{\boldsymbol{w}} \|(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})\|^2$$

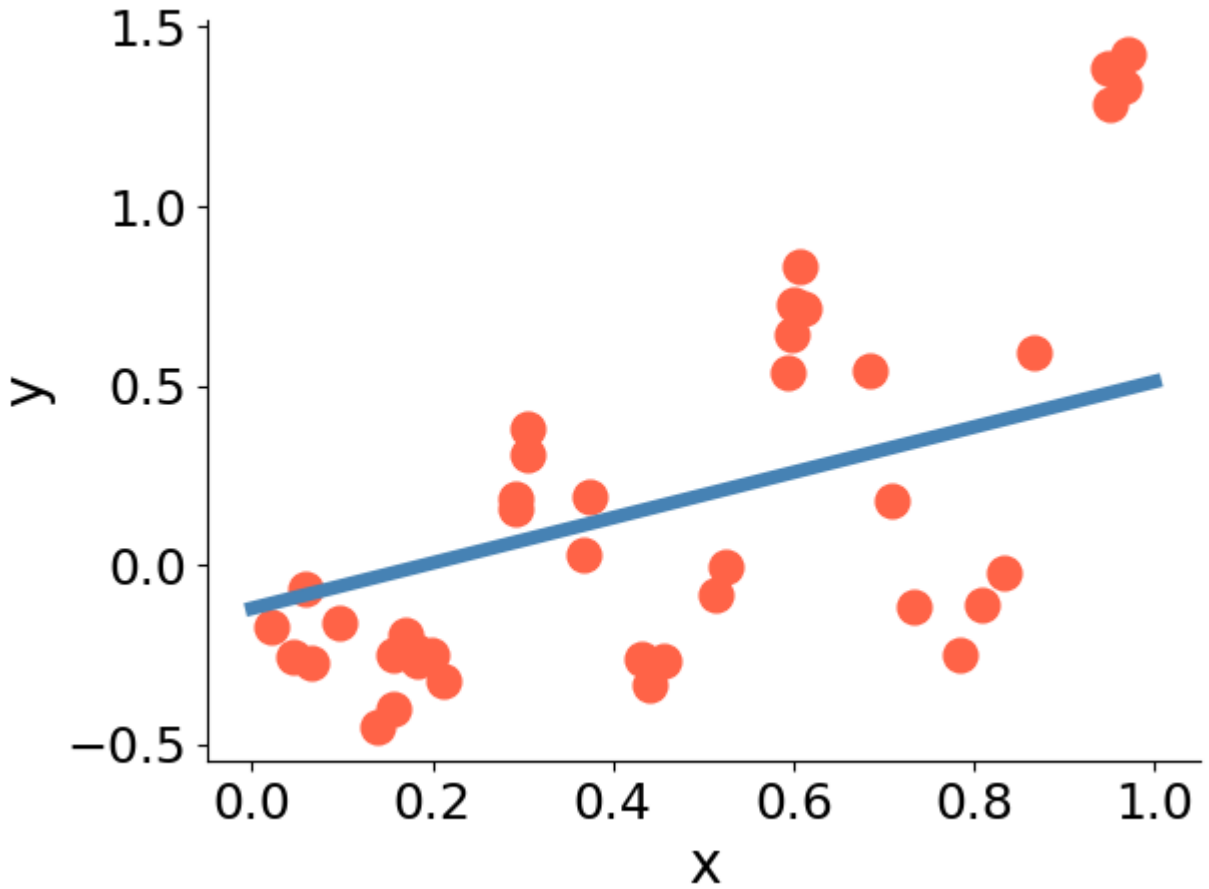
rows of $\boldsymbol{X} \rightarrow \boldsymbol{x}_i$
elements of $\boldsymbol{y} \rightarrow y_i$

This is solved using the gradients:

$$\nabla_{\boldsymbol{w}} \mathcal{L} = \boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) = 0 \quad \leftrightarrow \quad \boldsymbol{w} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

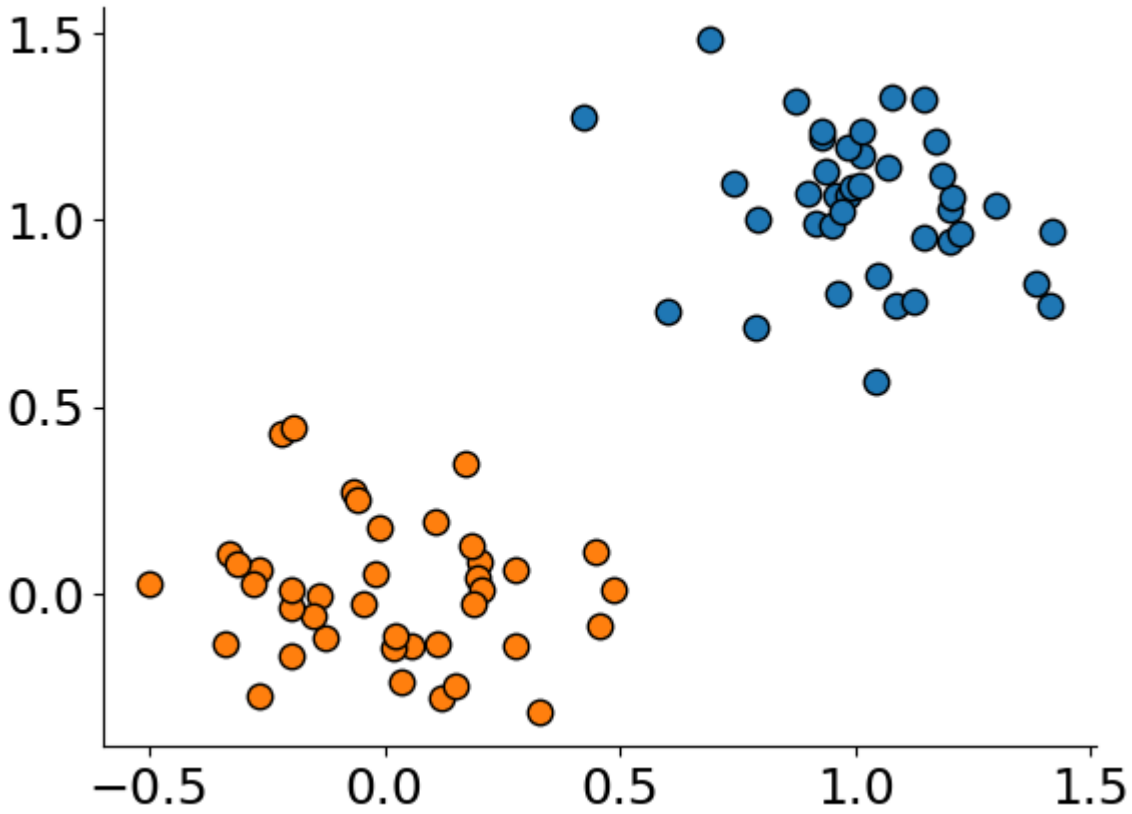
Example: linear regression





What about classification?

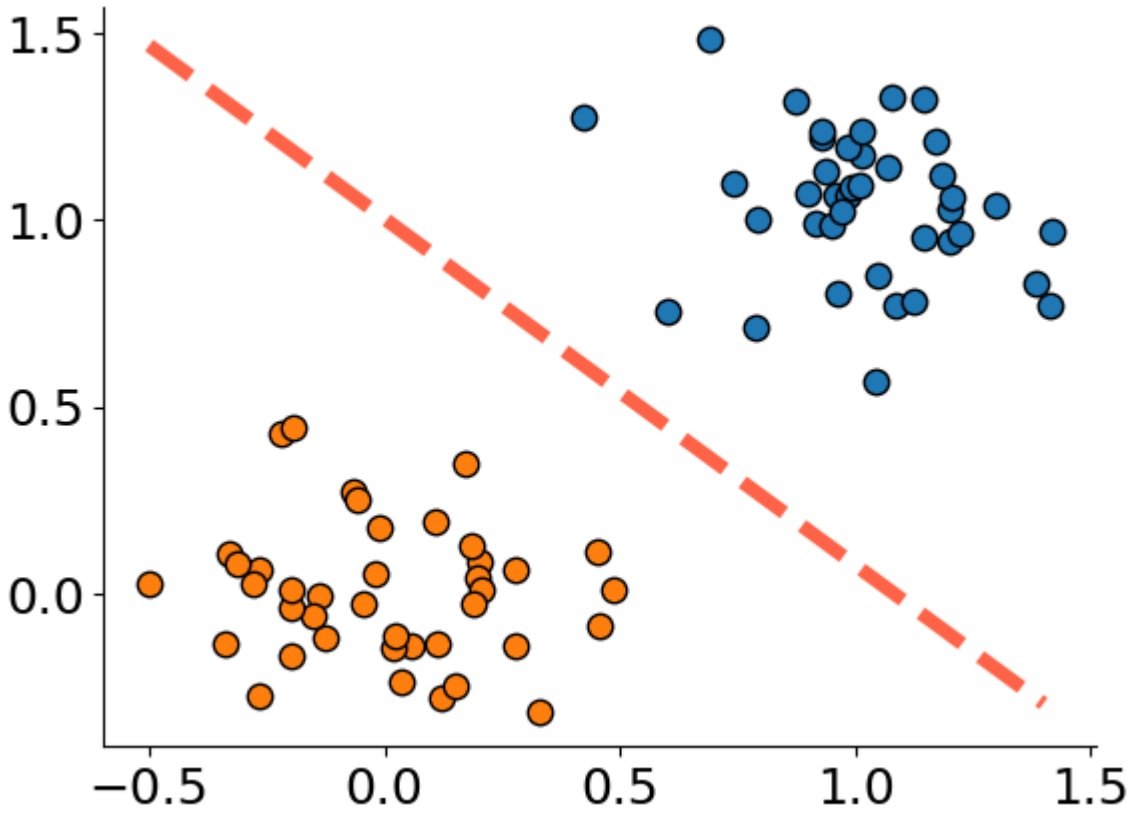
Very similar! In case of two classes, we model the probability of a data sample to belong to the first class by $p(\boldsymbol{x}) = \sigma\left(b + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}\right)$ and to the second class $1 - p(\boldsymbol{x})$.



A simple decision rule is to assign class 1 if $p \geq 1$, and assign class 0 otherwise.

What about classification?

Very similar! In case of two classes, we model the probability of a data sample to belong to the first class by $p(\boldsymbol{x}) = \sigma\left(b + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}\right)$ and to the second class $1 - p(\boldsymbol{x})$.



A simple decision rule is to assign class 1 if $p \geq 1$, and assign class 0 otherwise.

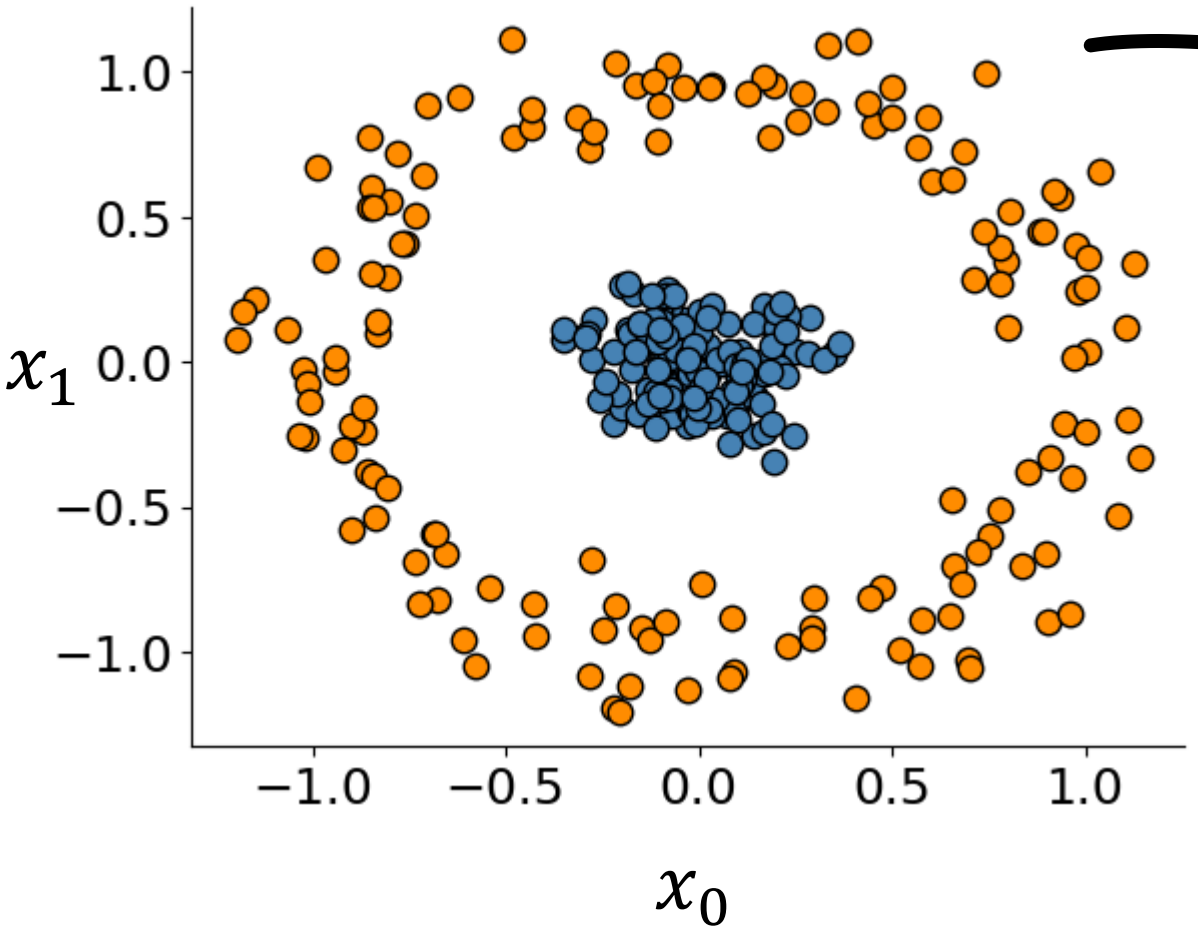
This is equivalent to:

$$\text{Class 1: } b + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} \geq 0$$

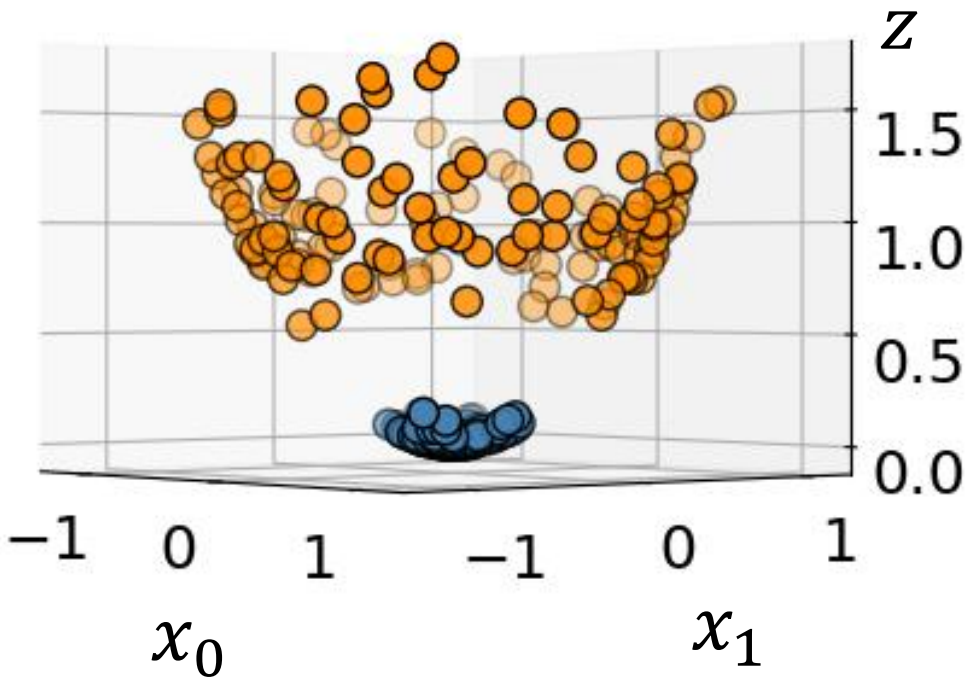
$$\text{Class 2: } b + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} < 0$$

The class is given depending on which side of the **hyperplane $\boldsymbol{w}$** the sample lies!

And… what now?!



Linearly separable (by a plane!) in a **new non-linear feature** $z = x_0^2 + x_1^2$



Polynomial regression

Use **non-linear transformations** of the features: $F(x) = \boldsymbol{w}^{\mathrm{T}}(1, x, x^2, x^3) = \underbrace{\sum_{i=0}^{3} w_i x^i}_{\text{Third-degree polynomial}}$

For fitting data, using the basis functions $\{x^i \mid i \in \mathbb{N}\}$ is generally not recommended.
Instead, use an orthogonal set of basis polynomials!

E.g., **Legendre polynomials**: $P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2 - 1)^n$

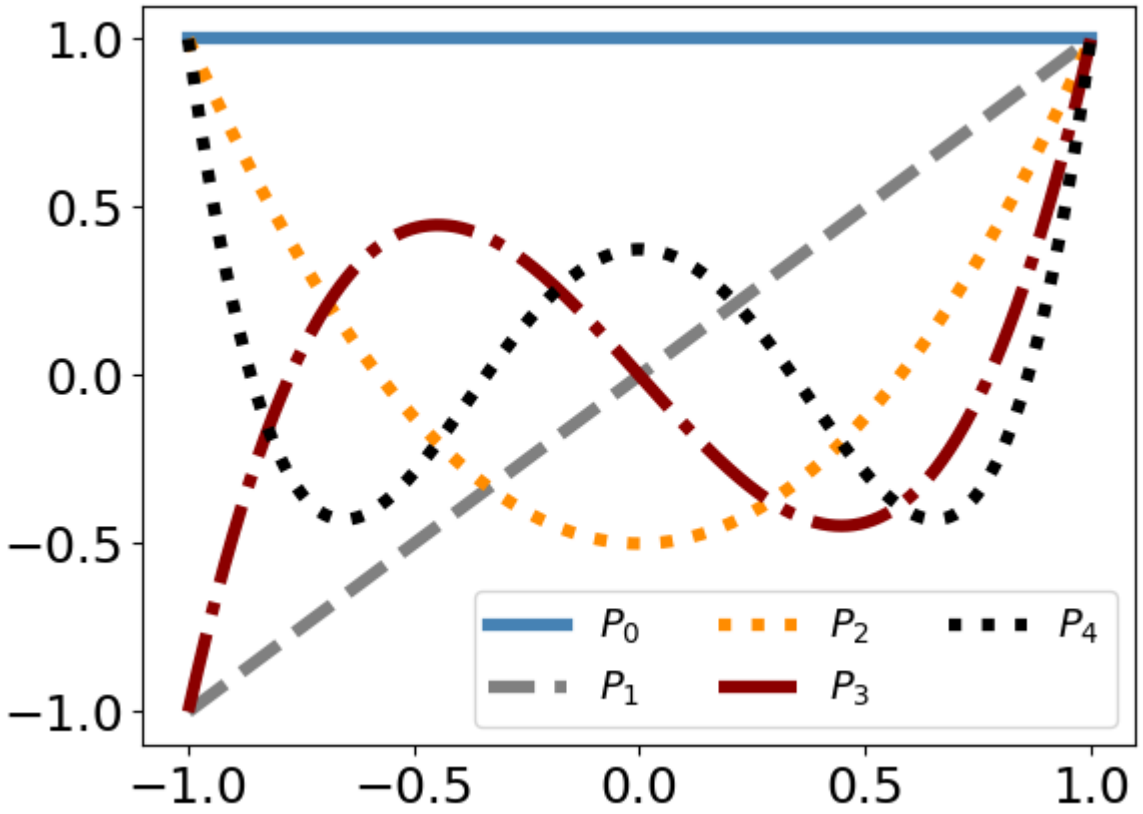
$P_0 = 1$ $\quad$ $P_2 = \frac{1}{2}(3x^2 - 1)$

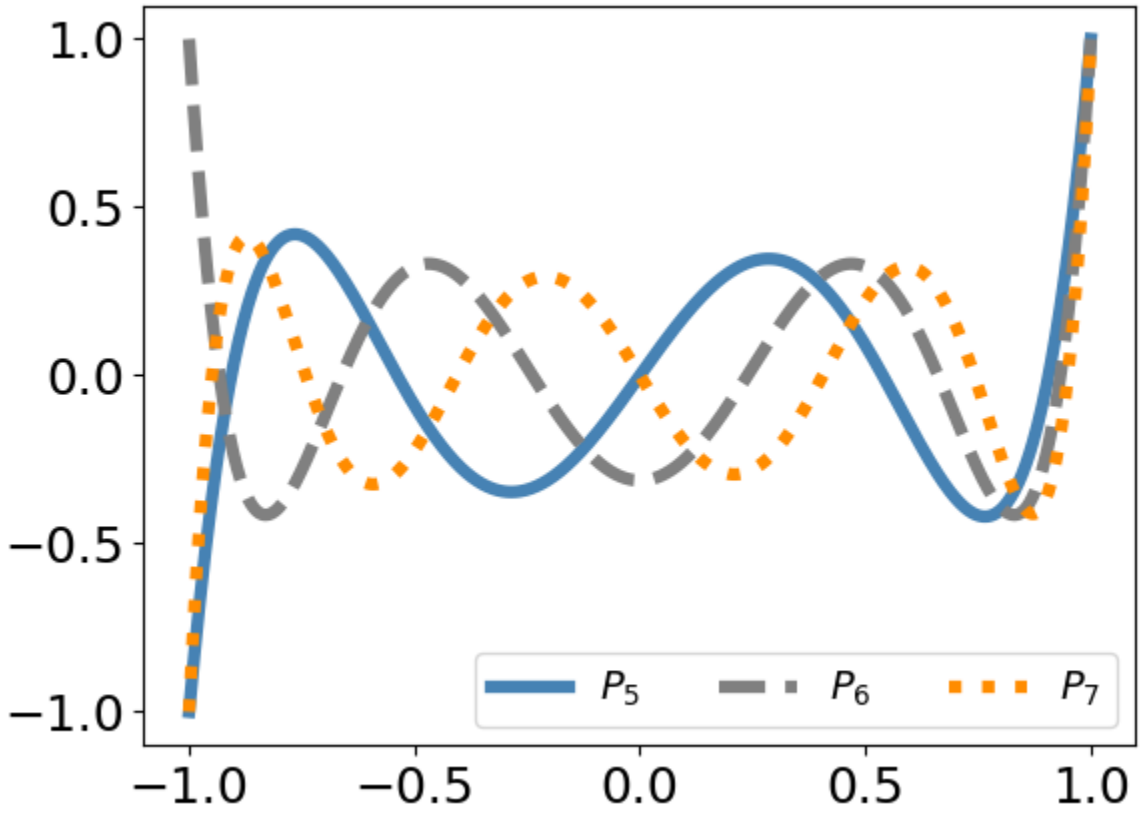
$P_1 = x$ $\quad$ $P_3 = \frac{1}{2}(5x^3 - 3x)$

These basis polynomials satisfy:

$$\int_{-1}^{1} P_m(x) P_n(x) \mathrm{d}x = 0 \text{ if } n \neq m$$

Some example Legendre polynomials





Fitting data using polynomials

This is exactly like linear regression, just with different "data features".

Assume the case of N one-dimensional data features $x \in [-1, 1]$. Then our regression model is:

$$\begin{pmatrix} P_0(x_1) & \cdots & P_d(x_1) \\ \vdots & \ddots & \vdots \\ P_0(x_N) & \cdots & P_d(x_N) \end{pmatrix} \cdot \begin{pmatrix} w_0 \\ \vdots \\ w_d \end{pmatrix} = \boldsymbol{X}\boldsymbol{w} \qquad \text{or} \qquad (\boldsymbol{X}\boldsymbol{w})_k = \sum_{i=0}^{d} w_i P_i(x_k)$$

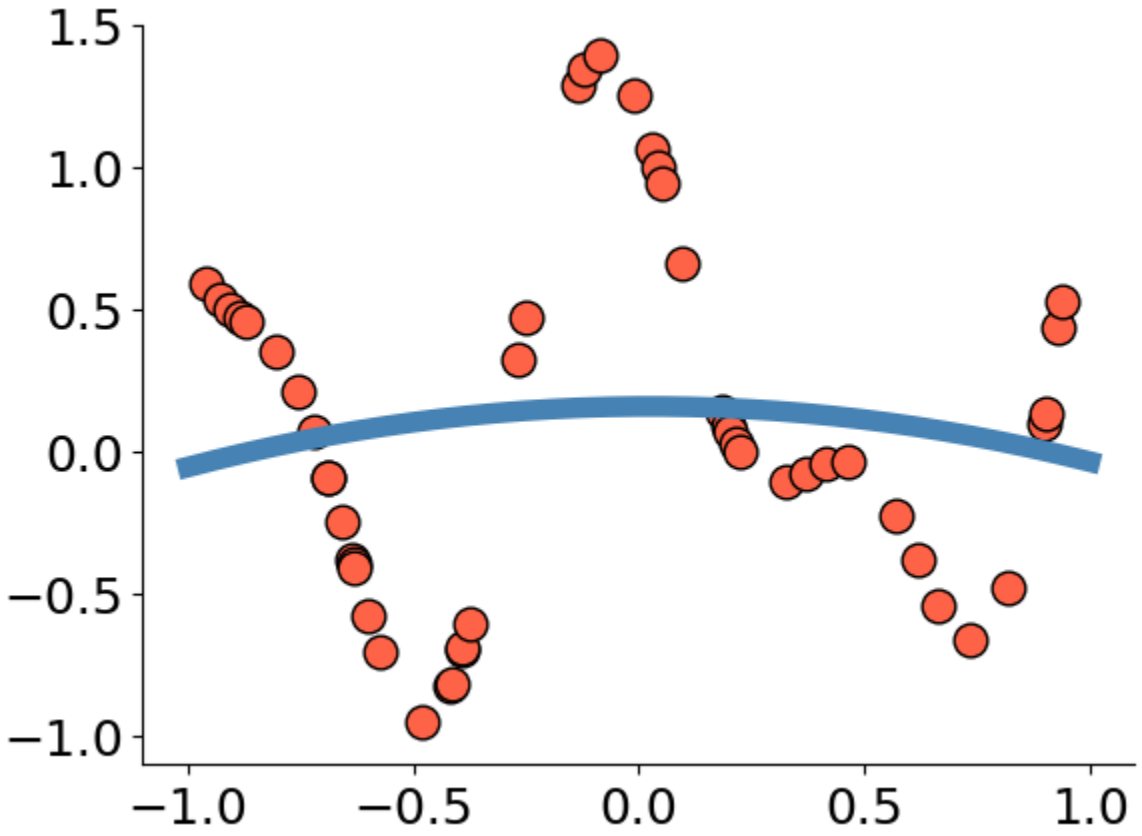
where d is the degree of the polynomial we use. We can find a solution again using:

$$\boldsymbol{w} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

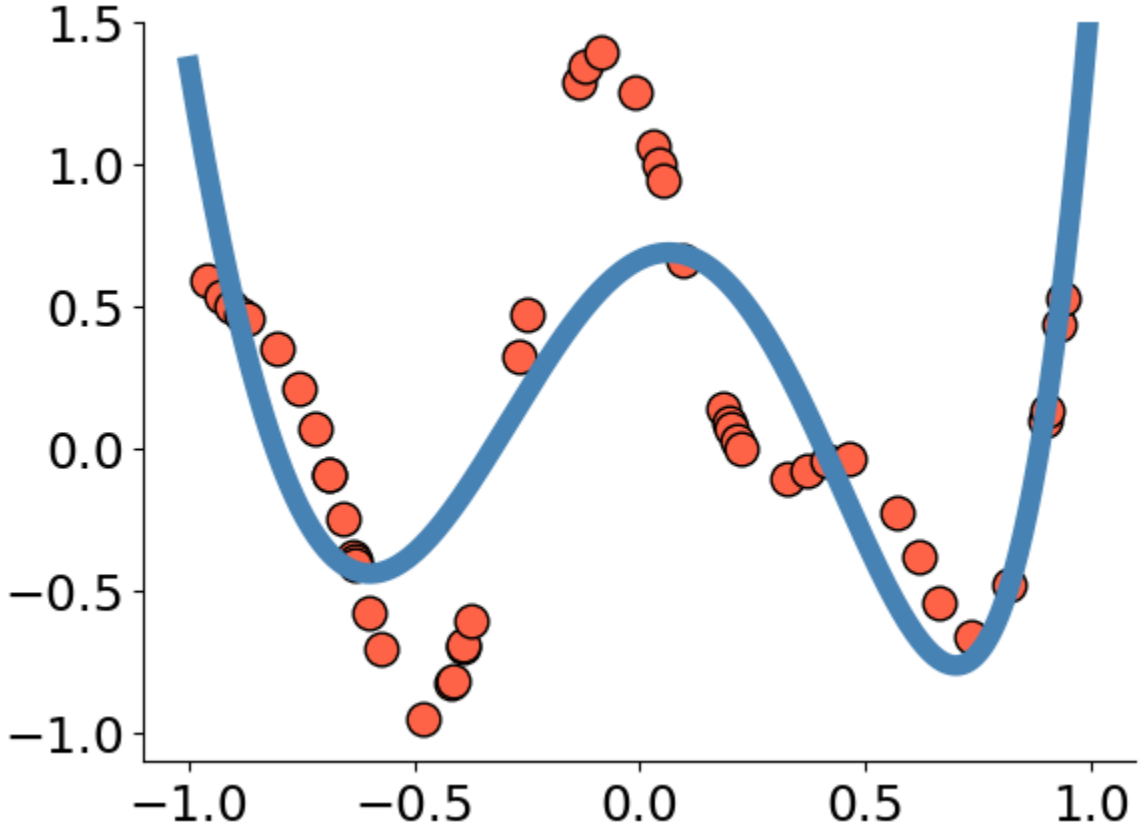
Note: we could also use any other function as basis functions here, not just polynomials!

Results for different degrees

$d = 2$

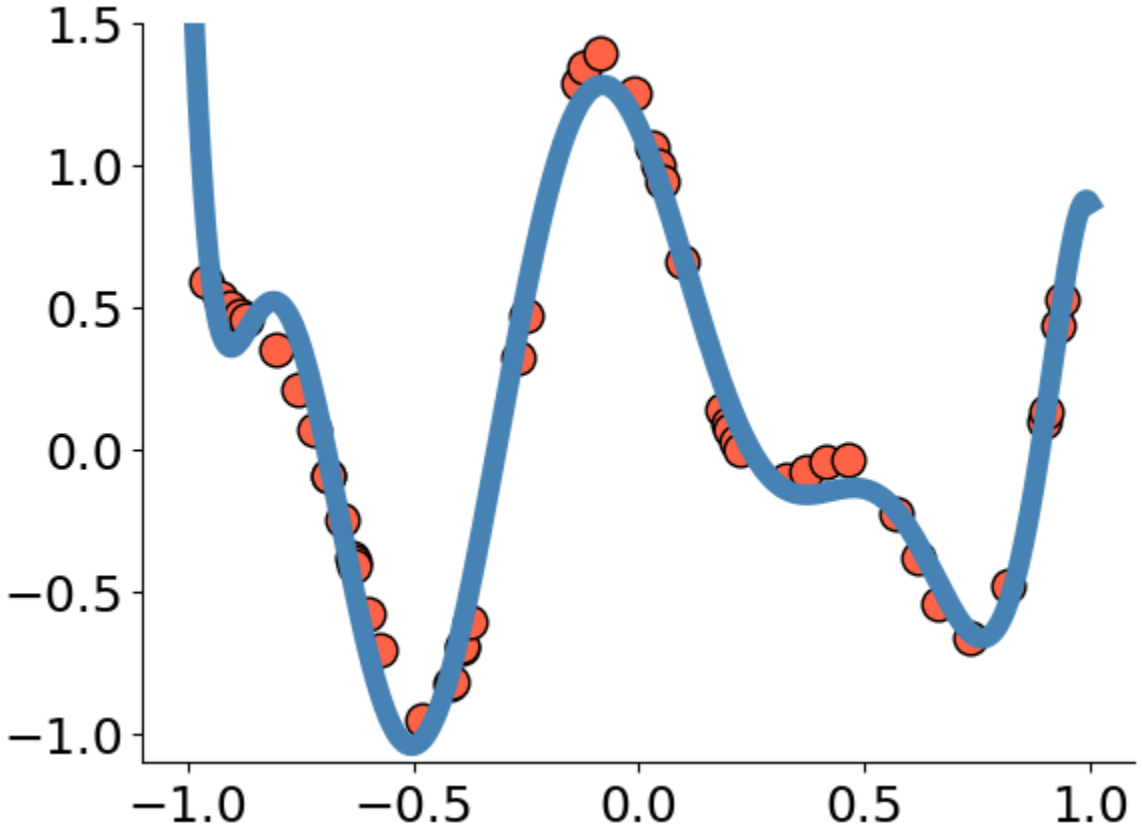


$d = 5$

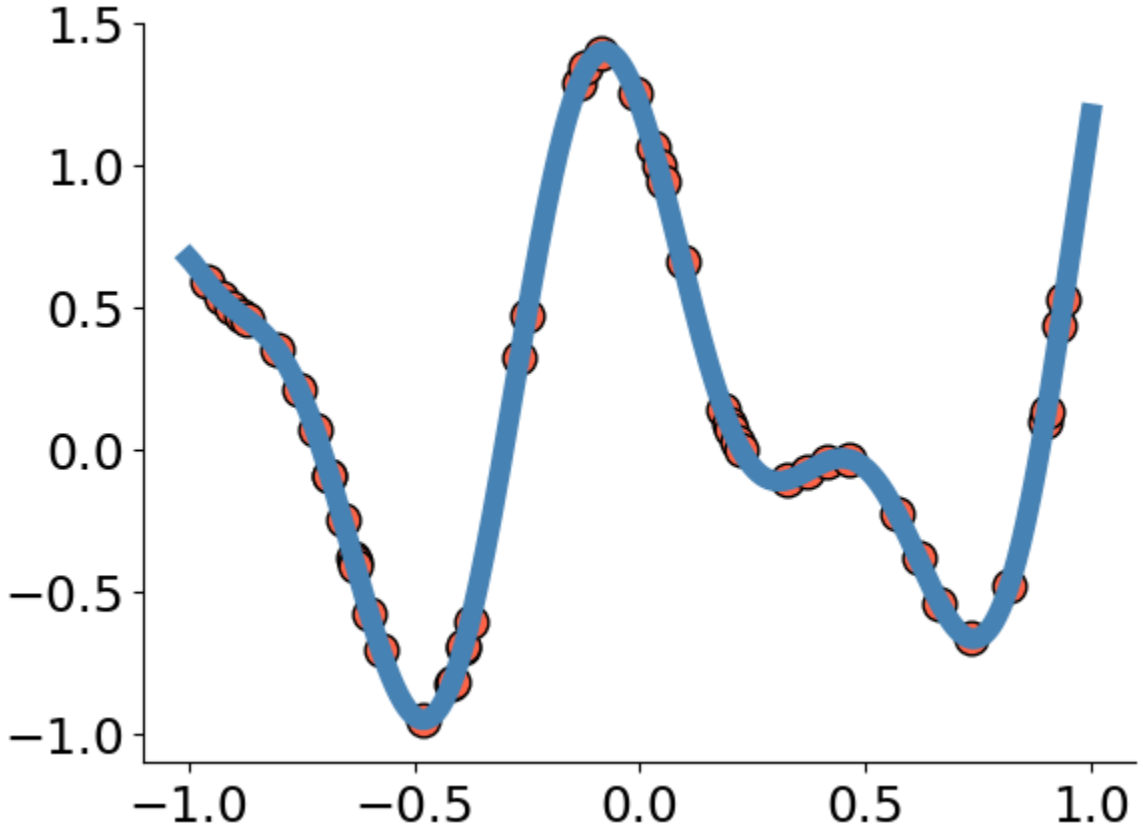


Results for different degrees

$d = 9$



$d = 15$



Why polynomials? Universal function approximation

Let $d \in \mathbb{N}$. A set of functions $\mathcal{H}$ from $\mathbb{R}^d$ to $\mathbb{R}$ is a **universal approximator** if

- for every $\epsilon > 0$, *"any error"*
- every compact $K \subset \mathbb{R}^d$, *"no matter which inputs"*
- and every continuous function $f\colon \mathbb{R}^d \to \mathbb{R}$, *"no matter which function"*

there exists $g \in \mathcal{H}$ such that:

$$\sup_{x \in K} |f(x) - g(x)| < \epsilon$$

Thus, a universal approximator can approximate a continuous function **everywhere** up to **any error**!

Weierstrass approximation theorem

Karl Weierstrass
(1815-1897)

The set of all **polynomials** (of all degrees) is a **universal approximator**.

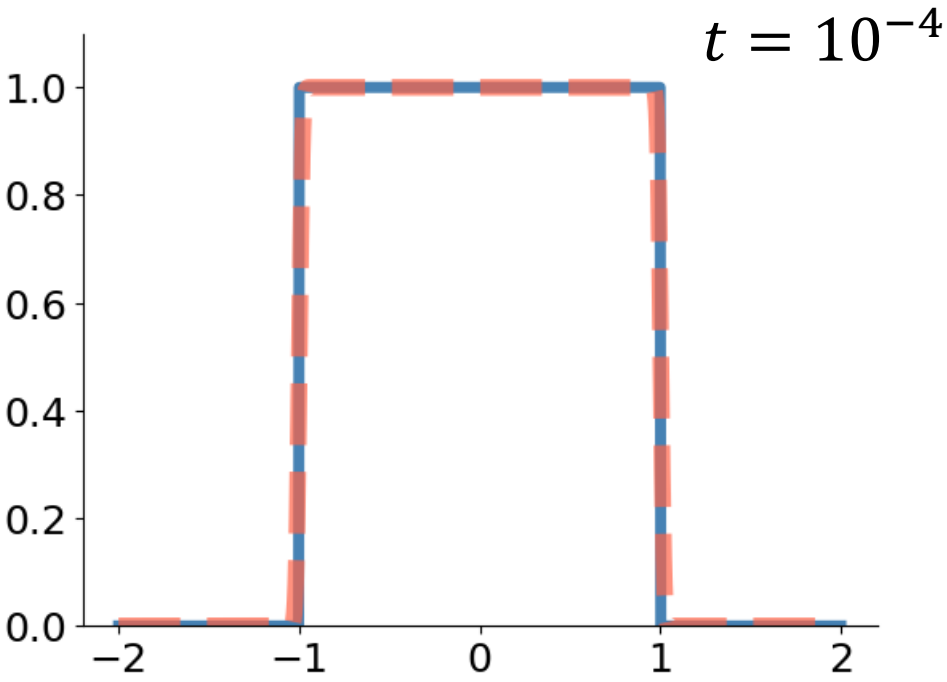
We will sketch the proof here, which is quite short and uses only two major ideas!

First, we "smear" out the actual function f we would like to approximate using:

$$\phi(x,t) = \frac{1}{\sqrt{2\pi t}} \int f(y) e^{-\frac{(x-y)^2}{2t}} \, \mathrm{d}y$$

For $t \to 0$, $\phi(x,t)$ uniformly approximates f.
Thus, given an $\epsilon > 0$, we can find a t_0 such that

$$|\phi(x,t_0) - f(x)| < \frac{\epsilon}{2}$$



Weierstrass approximation theorem

Karl Weierstrass
(1815-1897)

The set of all **polynomials** (of all degrees) is a **universal approximator**.

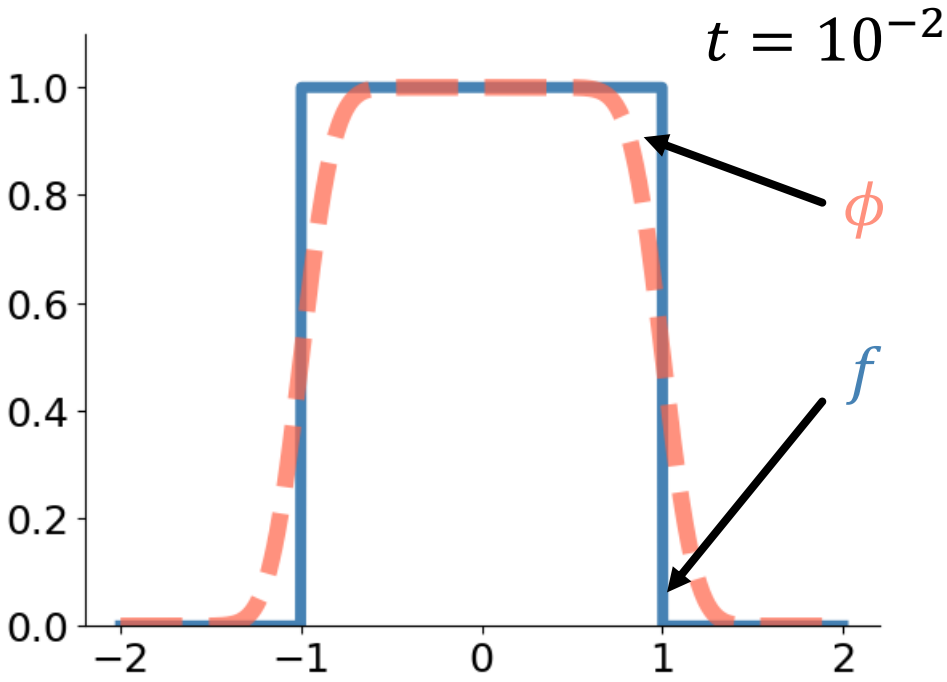
We will sketch the proof here, which is quite short and uses only two major ideas!

First, we "smear" out the actual function f we would like to approximate using:

$$\phi(x,t) = \frac{1}{\sqrt{2\pi t}} \int f(y) e^{-\frac{(x-y)^2}{2t}} \, \mathrm{d}y$$

For $t \to 0$, $\phi(x,t)$ uniformly approximates f.
Thus, given an $\epsilon > 0$, we can find a t_0 such that

$$|\phi(x,t_0) - f(x)| < \frac{\epsilon}{2}$$



Weierstrass approximation theorem

Karl Weierstrass
(1815-1897)

The set of all **polynomials** (of all degrees) is a **universal approximator**.

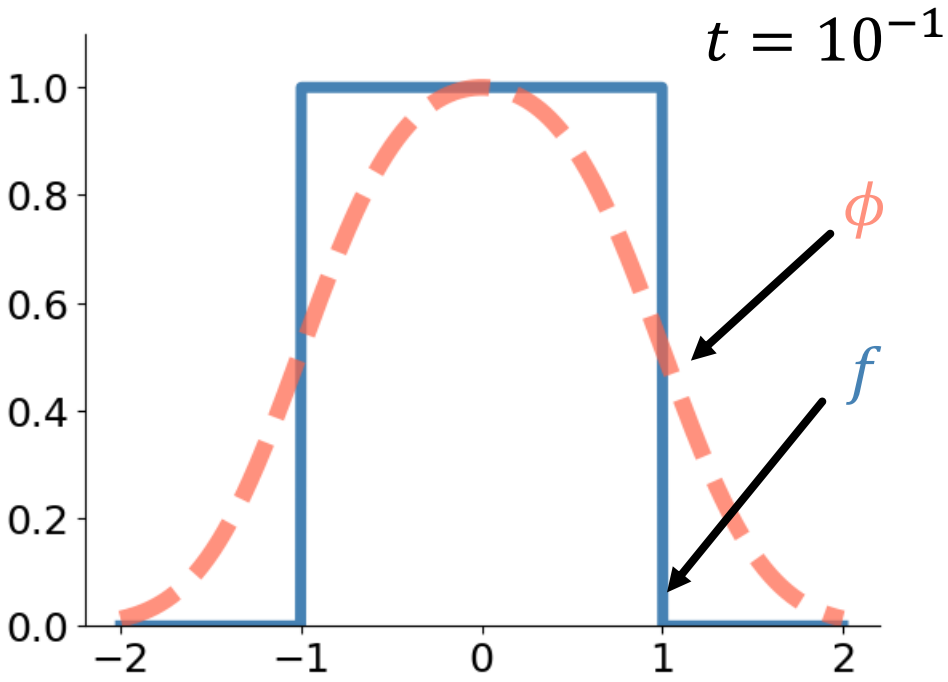
We will sketch the proof here, which is quite short and uses only two major ideas!

First, we "smear" out the actual function f we would like to approximate using:

$$\phi(x,t) = \frac{1}{\sqrt{2\pi t}} \int f(y) e^{-\frac{(x-y)^2}{2t}} \, \mathrm{d}y$$

For $t \to 0$, $\phi(x,t)$ uniformly approximates f.
Thus, given an $\epsilon > 0$, we can find a t_0 such that

$$|\phi(x,t_0) - f(x)| < \frac{\epsilon}{2}$$



Weierstrass approximation theorem

Karl Weierstrass
(1815-1897)

The set of all **polynomials** (of all degrees) is a **universal approximator**.

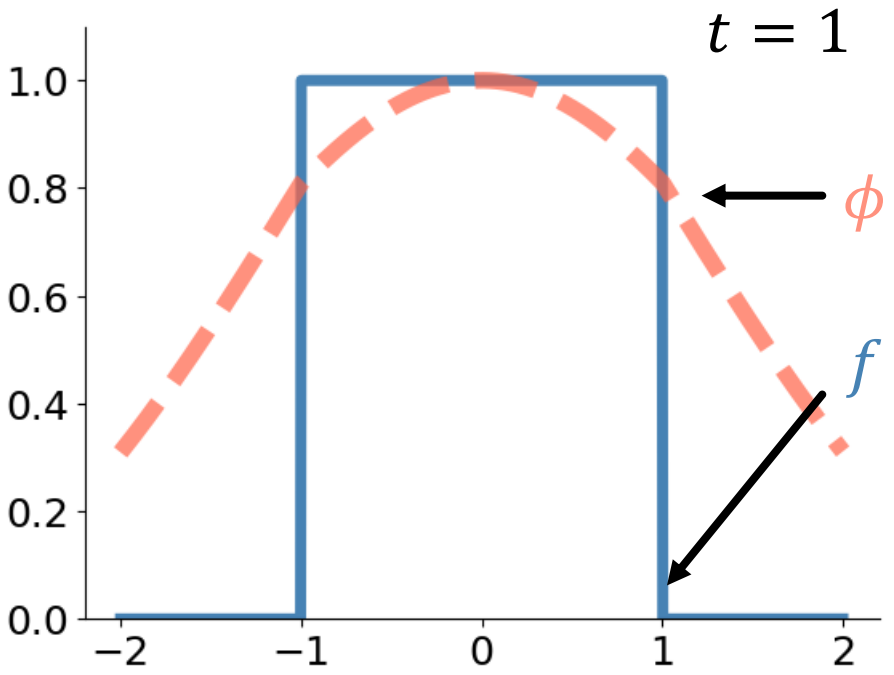
We will sketch the proof here, which is quite short and uses only two major ideas!

First, we "smear" out the actual function f we would like to approximate using:

$$\phi(x,t) = \frac{1}{\sqrt{2\pi t}} \int f(y) e^{-\frac{(x-y)^2}{2t}} \, \mathrm{d}y$$

For $t \to 0$, $\phi(x,t)$ uniformly approximates f.
Thus, given an $\epsilon > 0$, we can find a t_0 such that

$$|\phi(x,t_0) - f(x)| < \frac{\epsilon}{2}$$



Weierstrass approximation theorem

Karl Weierstrass
(1815-1897)

The set of all **polynomials** (of all degrees) is a **universal approximator**.

We will sketch the proof here, which is quite short and uses only two major ideas!

Second, by Taylor's theorem, $e^{-\frac{(x-y)^2}{2t}}$ can be approximated everywhere via a power series!

If we use a power series of high enough order and plug it into our definition of $\phi(x, t_0)$, we obtain a polynomial $p(x)$ such that:

$$|\phi(x, t_0) - p(x)| \leq \frac{\epsilon}{2}$$

And thus: $|f(x) - p(x)| < |f(x) - \phi(x, t_0)| + |\phi(x, t_0) - p(x)| < \epsilon$

□

There is an alternative proof by Bernstein using Bernstein polynomials which is really cool as well!

Stone-Weierstrass theorem

Marshall Harvey Stone
(1903 - 1989)

Generalization of the previous theorem!

Let $d \in \mathbb{N}$, let $K \subset \mathbb{R}^d$ be compact, and let our hypothesis set of continuous functions $\mathcal{H}$ from $\mathbb{R}^d$ to $\mathbb{R}$ satisfy that

- for all $\boldsymbol{x} \in K$, there exists $f \in \mathcal{H}$ such that $f(x) \neq 0$,
- for all $\boldsymbol{x} \neq \boldsymbol{y} \in K$, there exists $f \in \mathcal{H}$ *such that* $f(\boldsymbol{x}) \neq f(\boldsymbol{y})$,
- $\mathcal{H}$ is an algebra of functions, i.e., closed under addition, multiplication, and multiplication with scalars*.

Then $\mathcal{H}$ is a universal approximator.

This is a great tool: just check the criteria if you wanna know if $\mathcal{H}$ is a universal approximator!

**By closed, we mean: if* $f, g \in \mathcal{H}$ *and* $\alpha \in \mathbb{R}$, *then* $f + g \in \mathcal{H}$, $f \cdot g \in \mathcal{H}$, *and* $\alpha \cdot f, \alpha \cdot g \in \mathcal{H}$

Exercise

Using Stone-Weierstrass, explain which of the following hypothesis sets are universal approximators:

- Linear regression *(linear in the features)*
- Polynomials
- Decision Trees

Helpful illustration of a decision tree: