Lecture 1

Introduction and Overview

Sujay Sanghavi

What is this course about ?

This is a course on machine learning: using computation to make sense of data.

Our focus: fundamental principles, algorithms.

Sister class 460J: Data Science Lab

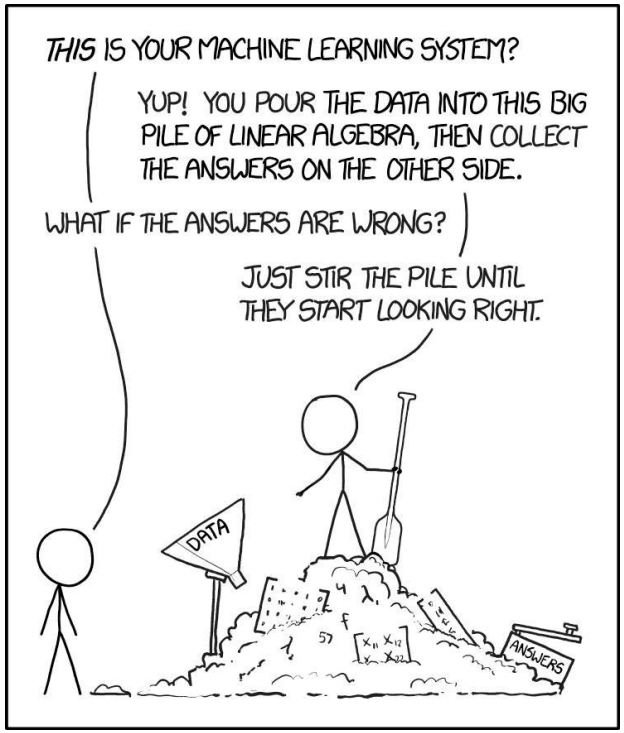


Figure: credit: xkcd

Supervised Learning

Task: Predict **target** y given **features** x, where $x = (x_1, \cdots, x_d)$

Set up:

- We are given **training data** with n samples $(y^{(1)}, x^{(1)}), (y^{(2)}, x^{(2)}), \ldots$
- Based on this, make a **model** (function/rule/algorithm) $f : \mathcal{X} \to \mathcal{Y}$.
  - e.g. by minimizing "training error" ...
- This f is then evaluated on **new** samples, called **test data**.

One Example of Supervised Learning

Dataset: IMAGENET

14 million images, 20k categories

Large Scale Visual Recognition Challenge

Classification task: 1000 categories, 1.2 million images ($\sim 10^5$ pixels per image)

Localization task: predict the position (x, y, h, w) of an object.

Supervised Learning in 461P

Methods

- Linear Regression, LASSO, Ridge
- Linear Classifiers, SVMs
- Nearest Neighbors
- Decision trees, random forests

Concepts

- Measuring accuracy for regression and classification
- Overfitting
- Bias-variance, Confidence, Bootstrap
- Boosting, Gradient Boosting

Unsupervised Learning

Task: Learn to represent data in a way that best suits downstream applications

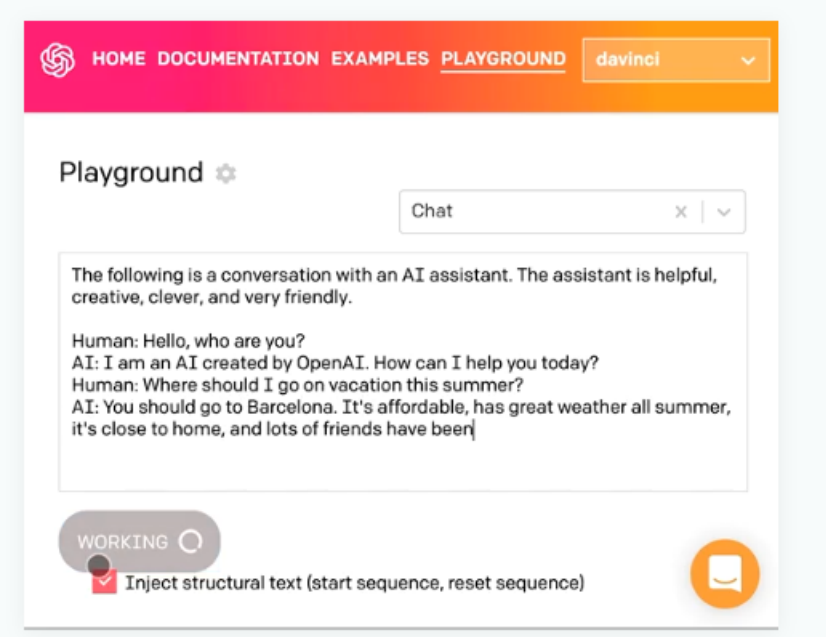
Set up:

- We are given **unlabeled data** with n samples $x^{(1)}, x^{(2)}, \ldots$
- Based on this, we learn a **repersentation** (function/rule/algorithm) $f : \mathcal{X} \to \mathcal{X}'$.
- This f is then evaluated on another separate **downstream task**.

Unsupervised Learning in 461P

- Dimensionality reduction - PCA and related
- Clustering, k-means, GMMs
- Topic modeling
- Matrix factorization, recommendation systems

One Example of Unsupervised Learning



Language Modelling: make a probability distribution on words/sentences/paragraphs in natural language

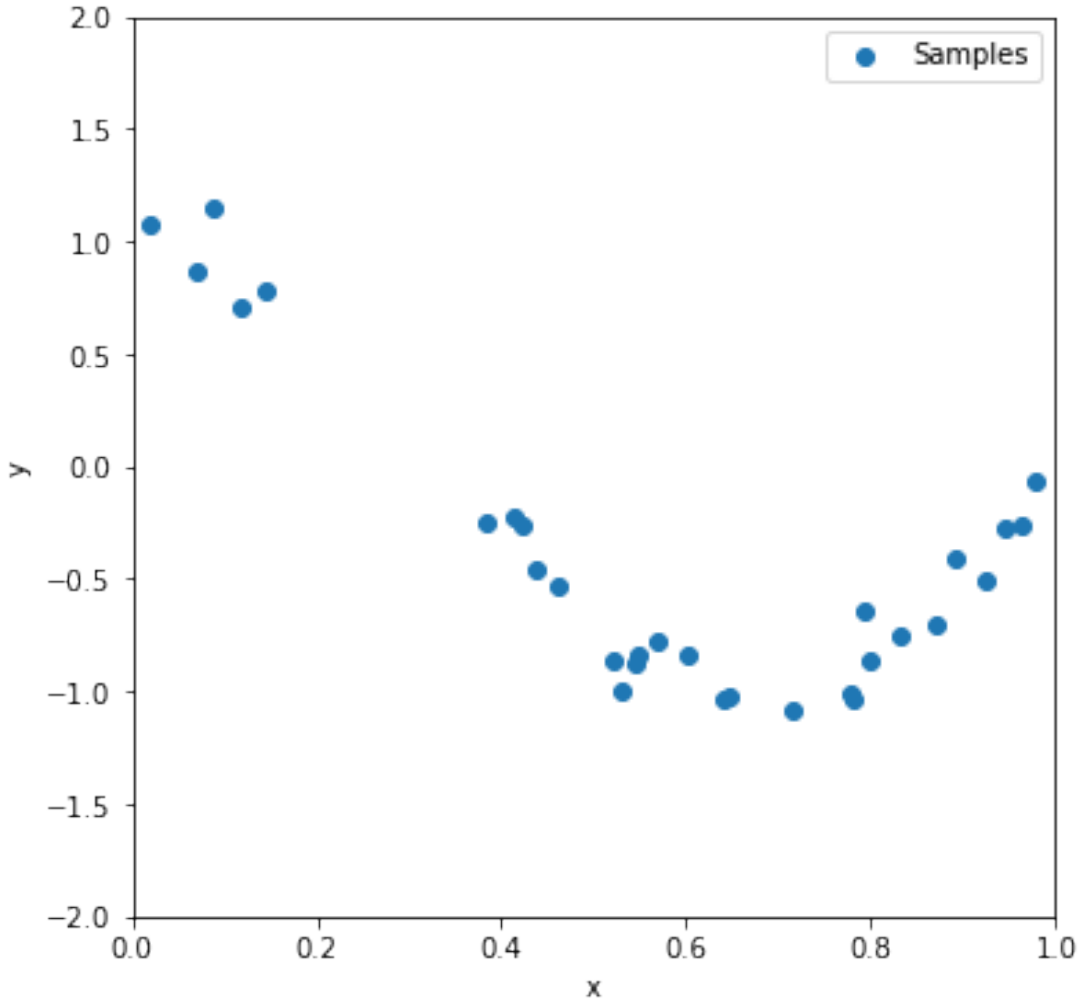
Task: Predict the next word.

E.g. GPT-3 from OpenAI

Trained on 499 Billion tokens, sourced from Commow Crawl, WebText2 etc.

Few-shot learner, does not need task specific fine-tuning.

Taster: Fitting a Model to Data

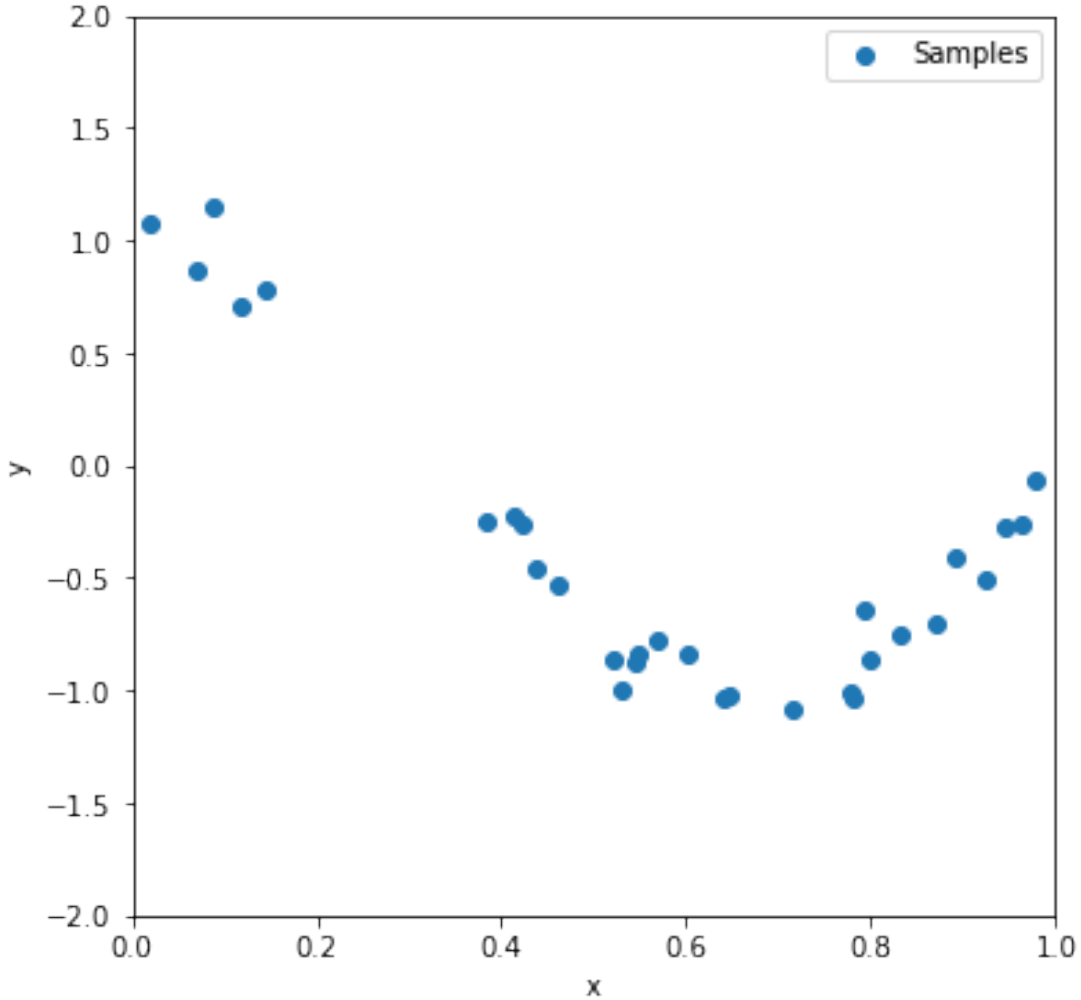


Task: predict y given x

Given: data samples, i.e. (x, y) pairs

Any Ideas ?

Taster: Fitting a Model to Data



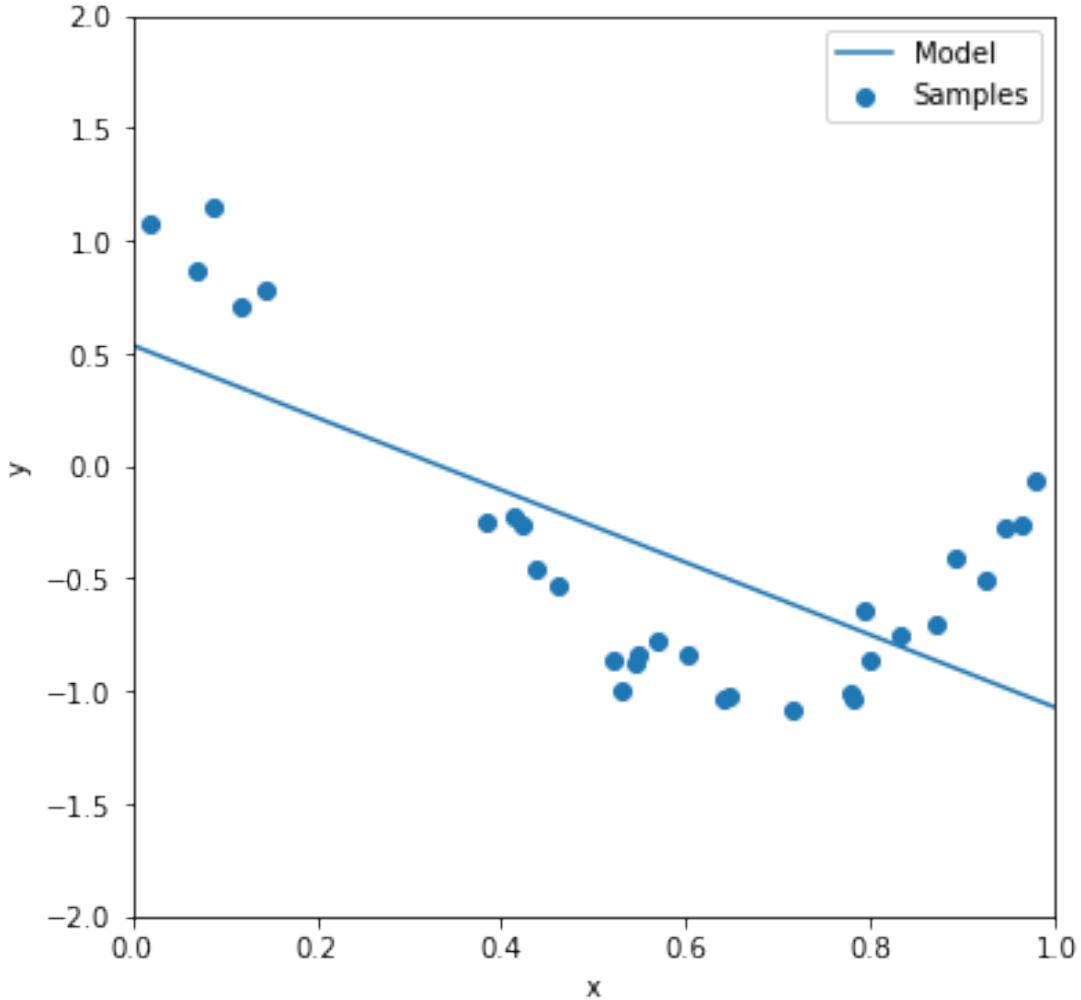
Idea 1: "fit" a **line**
$y = ax + b$

i.e.: "I believe the data (approximately) comes from a line."

fit == find the **best** a and b using the samples

But what does "best" mean ?

Taster: Fitting a Model to Data



Idea 1: "fit" a **line**
$y = ax + b$

fit == find the **best** a and b using the samples

Here best == minimize squared error

$$\min_{a,b} \sum_i (y_i - ax_i - b)^2$$

Taster: Fitting a Model to Data

Idea 2: fit a **polynomial**

$y = a_0 + a_1 x + a_2 x^2 + \ldots a_d x^d$

fit == find the **best** $a_0, \ldots, a_d$ using the samples

Here best == minimize squared error

$$\min_a \sum_i \left(y_i - \sum_{j=0}^{d} a_j x_i^j \right)^2$$

Taster: Fitting a Model to Data

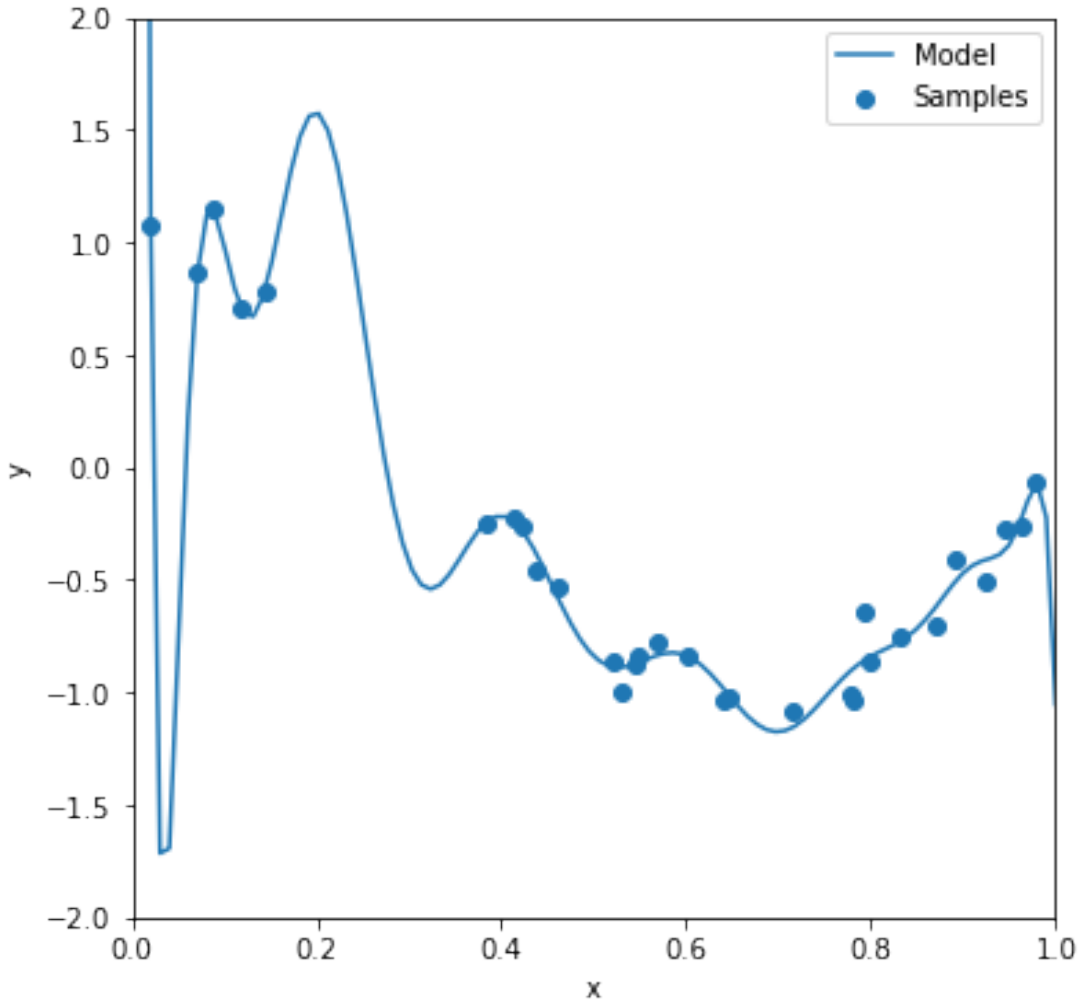


Figure: degree $d = 15$

Idea 2: fit a **polynomial**

$y = a_0 + a_1 x + a_2 x^2 + \ldots a_d x^d$

fit == find the **best** $a_0, \ldots, a_d$ using the samples

Here best == minimize squared error

$$\min_a \sum_i \left(y_i - \sum_{j=0}^{d} a_j x_i^j \right)^2$$

Taster: Fitting a Model to Data

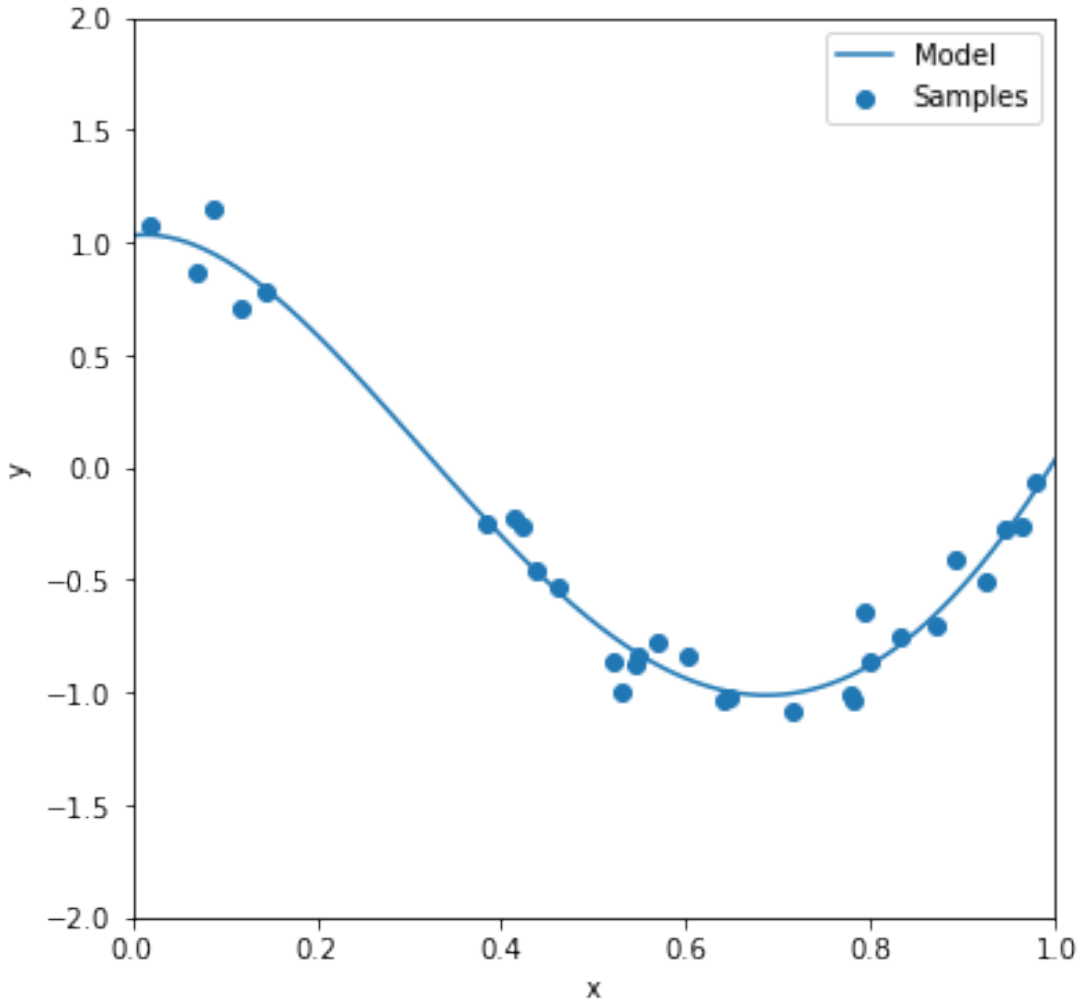


Figure: degree $d = 4$

Idea 2: fit a **polynomial**

$y = a_0 + a_1 x + a_2 x^2 + \ldots a_d x^d$

fit == find the **best** $a_0, \ldots, a_d$ using the samples

Here best == minimize squared error

$$\min_a \sum_i \left(y_i - \sum_{j=0}^{d} a_j x_i^j \right)^2$$

Taster: Fitting a Model to Data

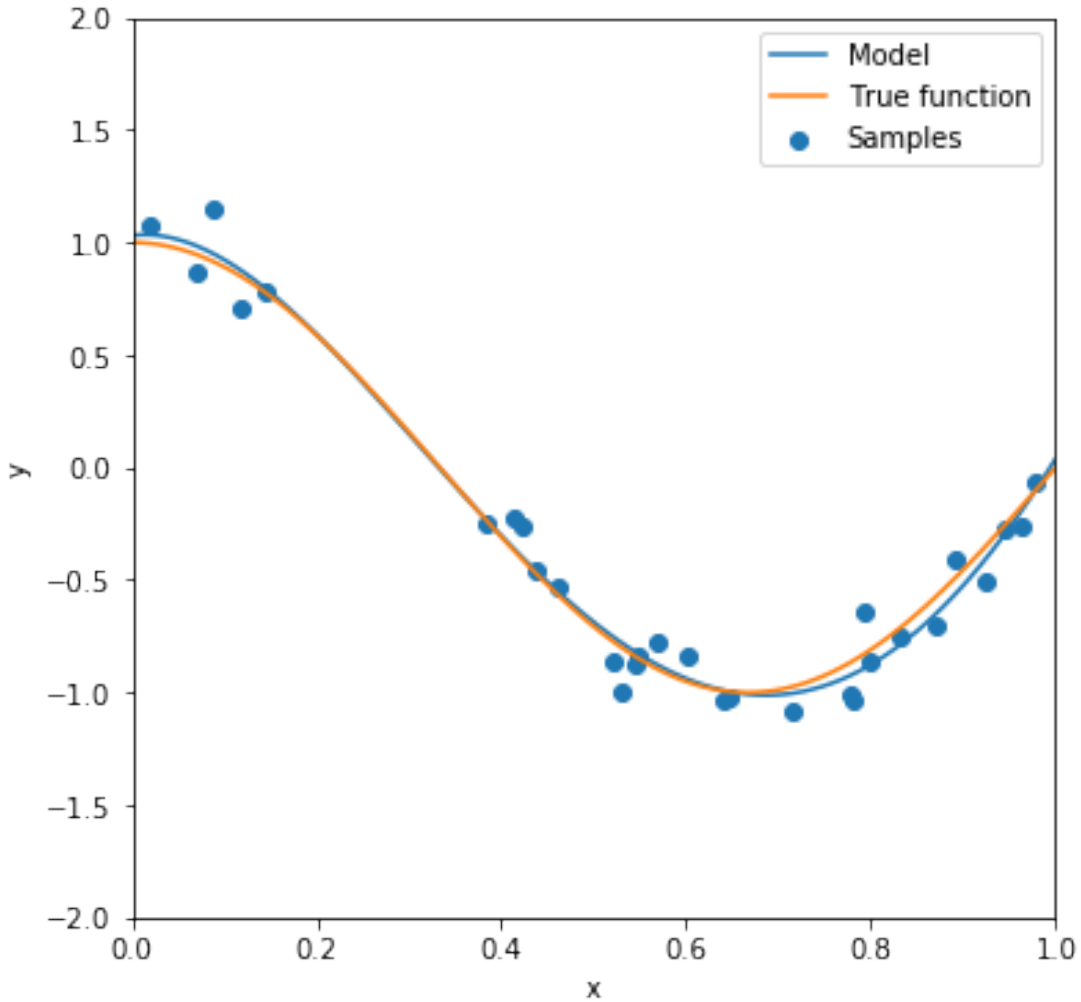


Figure: degree $d = 4$ was the "truth"

Idea 2: fit a **polynomial**

$y = a_0 + a_1 x + a_2 x^2 + \ldots a_d x^d$

fit == find the **best** $a_1, \ldots, a_d$ using the samples

Here best == minimize squared error

$$\min_a \sum_i (y_i - \sum_j a_j x^j)^2$$

Taster: Fitting a Model to Data

Even with the correct d, big errors if we do not have enough samples ...

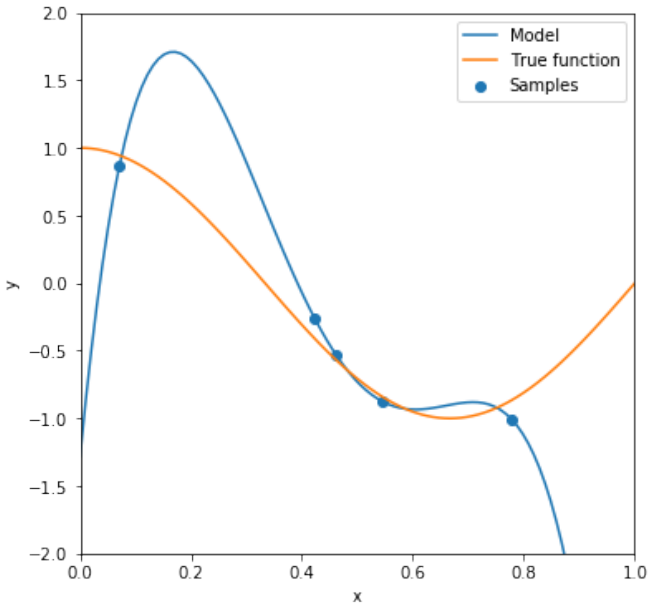


Figure: only 5 samples

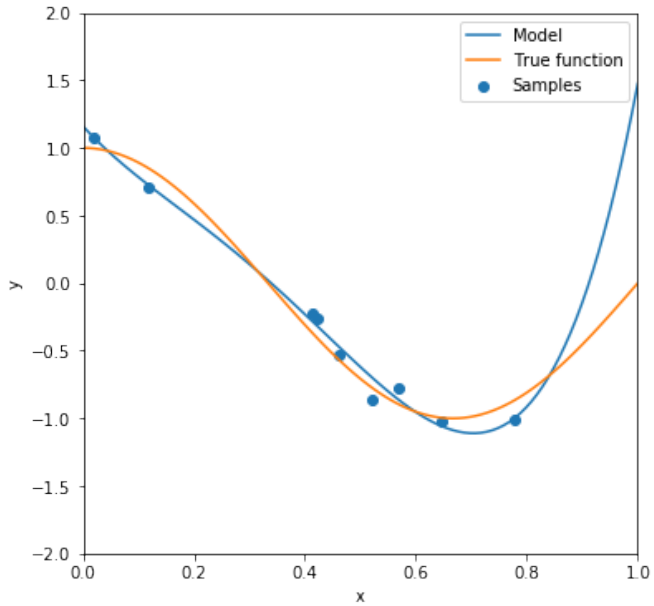


Figure: only 11 samples