Lecture
Machine Learning Basics

Learning from data

- A machine learning algorithm is an algorithm that is able to learn from seen data (training data), and predict on new, previously unseen data (test data).
- Machine learning allows us to tackle tasks that are too difficult to solve with fixed programs (rules) written and designed by human beings.
- Roughly, data-driven approaches are good at perception (感知), rather than cognition (认知).

Category by training data

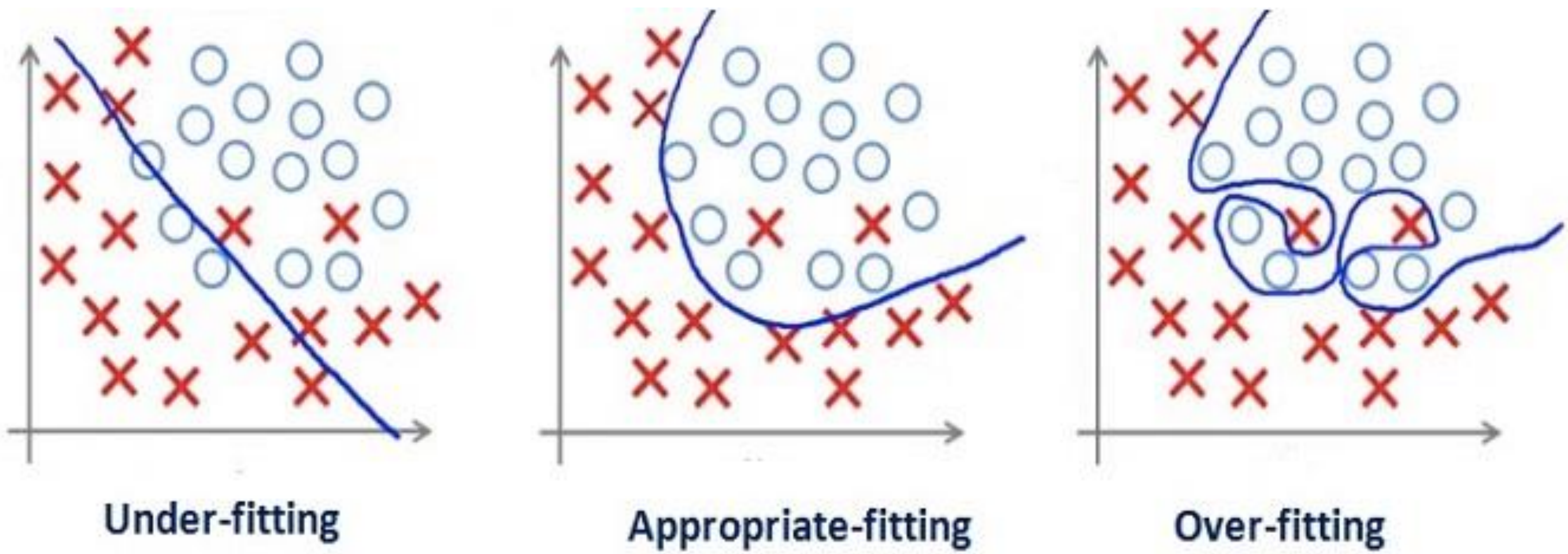
- Supervised learning: data with label
- Unsupervised learning: data without label
- Semi-supervised learning: a mix of labeled and unlabeled data
- Multi-instance learning: a number of bags of examples, which are labeled as containing or not containing at least one example of a class

Category by outputs

- Classification $f : \mathbb{R}^n \to \{1, \ldots, k\}$
- Regression $f : \mathbb{R}^n \to \mathbb{R}$
- Structured prediction $f : \mathbb{R}^n \to \{1, \ldots, k\}^m$
- Density estimation $f : \mathbb{R}^n \to \mathbb{R}\ (\mathbf{x} \to p(\mathbf{x}))$

Underfitting and Overfitting

- The objectives of an ML algorithm
  1) Make the training error small. If fails → underfitting (欠拟合)
  2) Make the gap between training and test error small. If fails → overfitting (过拟合)



Capacity

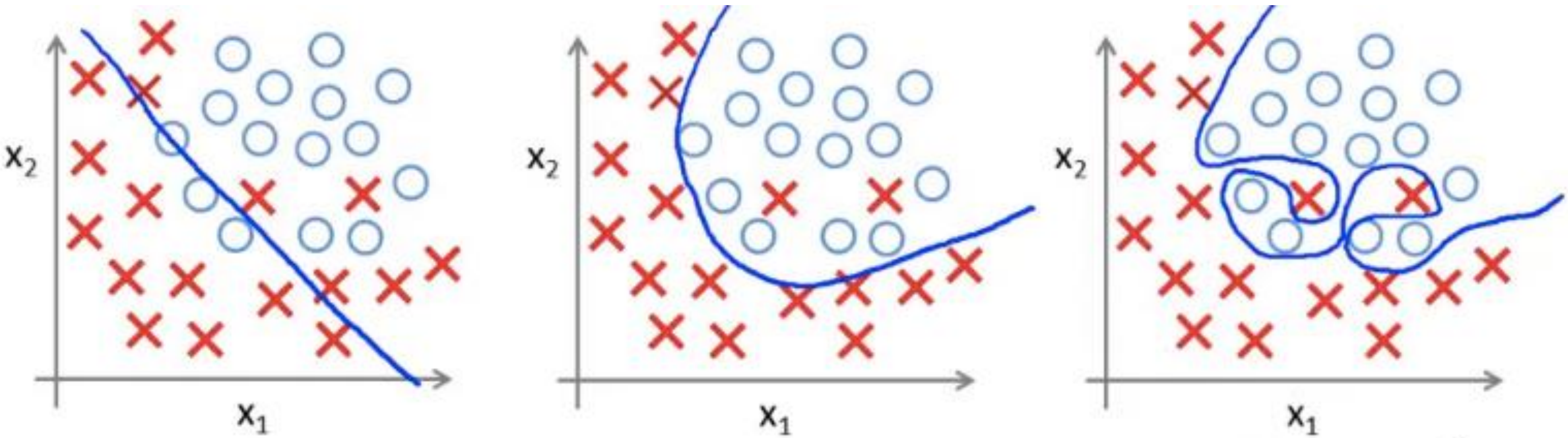
- Definition: the ability to fit a wide variety of functions
- One way to control the capacity of a learning algorithm is by choosing its hypothesis space, the set of functions that the learning algorithm is allowed to select as being the solution.
- Example: a linear regression model

$$\hat{y} = b + wx$$

has a lower capacity than a quadratic one

$$\hat{y} = b + w_1 x + w_2 x^2$$

Capacity



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

(g = sigmoid function)

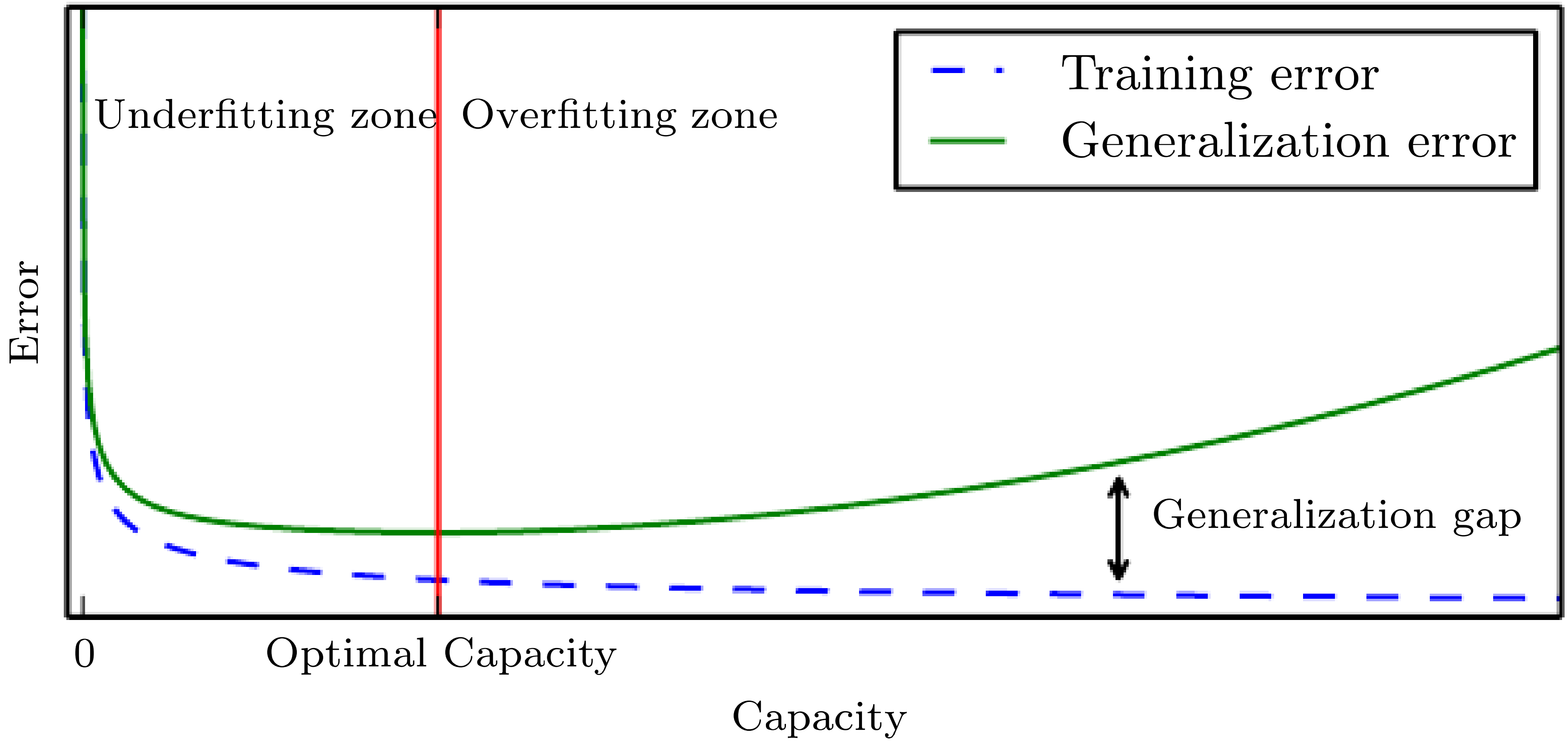
$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 \\ +\theta_3 x_1^2 + \theta_4 x_2^2 \\ +\theta_5 x_1 x_2)$$

$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 \\ +\theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 \\ +\theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots$$

Underfitting
Low capacity
High bias

Overfitting
High capacity
High variance

Capacity vs. Error



Hyperparameter and Validation Sets

- Hyperparameters are settings that we can use to control the behavior of the learning algorithm. The values of hyperparameters are not adapted by the learning algorithm itself.
- We do not learn the hyperparameter because it is not appropriate to learn that hyperparameter on the training set (e.g., all hyperparameters that control model capacity).
- We split the training data into two disjoint subsets. One of these subsets is used to learn the parameters. The other subset is our validation set, used to estimate the generalization error during or after training, allowing for the hyperparameters to be updated accordingly.

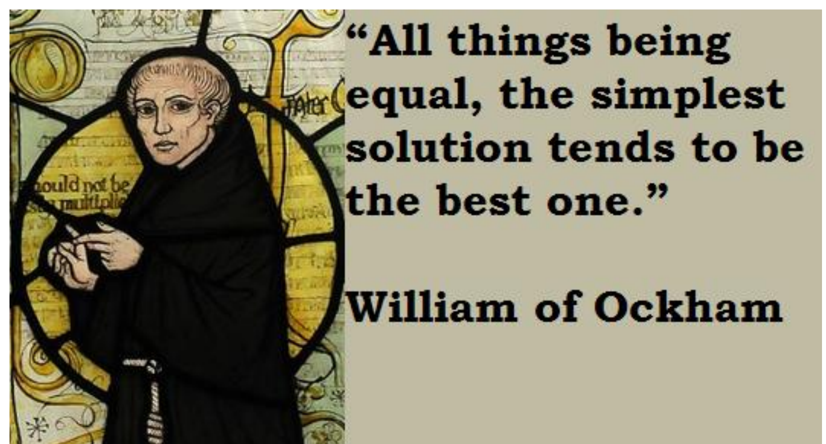
Regularization in a narrow sense

$$L(W) = \frac{1}{N}\sum_{i=1}^{N} L_i(f(x_i, W), y_i) + \lambda R(W)$$

Data loss: Model predictions should match training data

Regularization: Model should be “simple”, so it works on test data

Motivation:
Occam’s Razor



Regularization in a broad sense

- Regularization is any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error.
- Regularization controls a model's capacity by expressing preferences for different solutions, both implicitly and explicitly
- Neural networks are usually regularized implicitly in many ways (e.g. dropout, data augmentation, early stopping)

Building a machine learning algorithm

- A dataset
- A model
- A loss function
- An optimization procedure

Loss functions