ECE595 / STAT598: Machine Learning I

Lecture 22.1: What Constitutes a Learning Problem?

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

Learning Theory

- Welcome to Part 3 of ECE595 / STAT598!
- Here is what we have learned:
  - Part 1: The machine learning pipeline
  - Part 2: Classification methods
- What are we going to do in Part 3?
  - Now that we have a method, then what?
  - Will it do well? How well?
  - Will it fail? When?
  - Complex model = better?
  - More sample = better?
  - Can every problem be solved by learning?
  - When do you overfit?
  - How to avoid overfit?

Outline

Today's Lecture:

- What constitutes a learning problem?
  - Training and testing samples
  - Target and Hypothesis function
  - Learning Model
- Is learning feasible?
  - An example
  - The power of probability
- Training versus Testing
  - In-sample error
  - Out-sample error
  - Probability bound

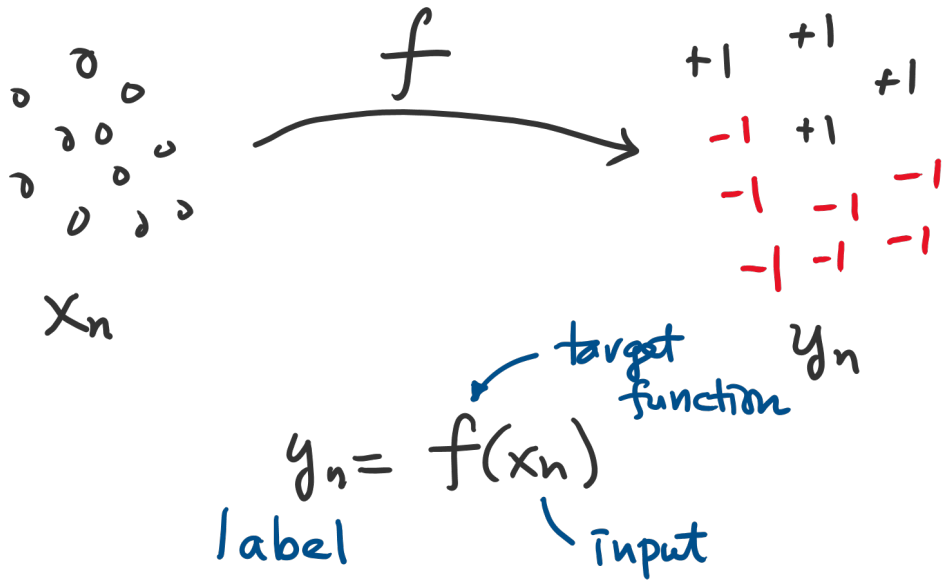
Reference:

- Learning from Data, chapter 1.3

Dataset

Let us first talk about a dataset:

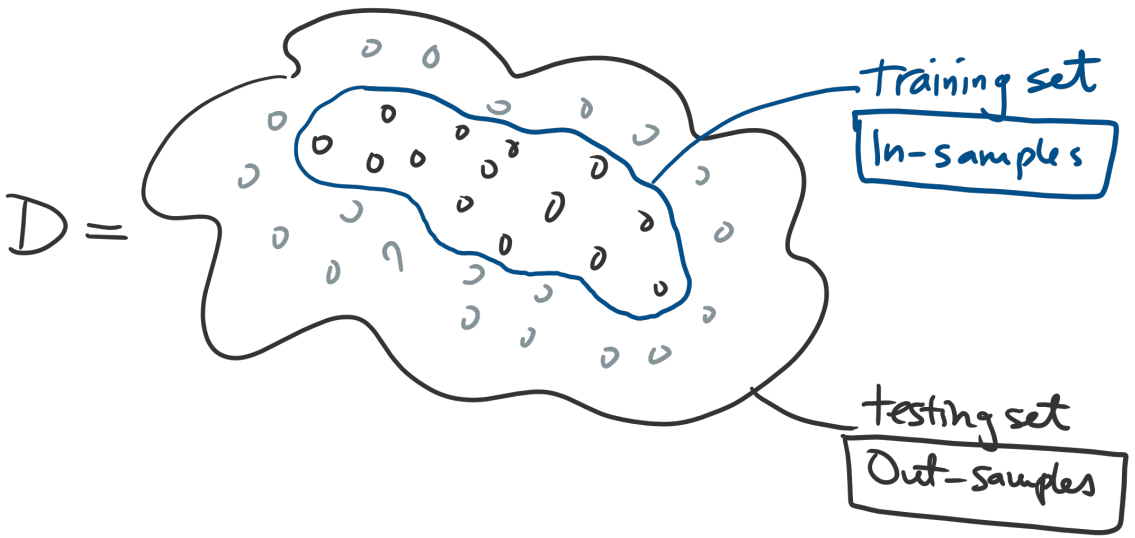
- Input vectors: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$
- Labels: $y_1, \ldots, y_N$
- Training set: $\mathcal{D}$
- Target function f: Maps $\boldsymbol{x}_n$ to y_n
- Target function is always **unknown** to you



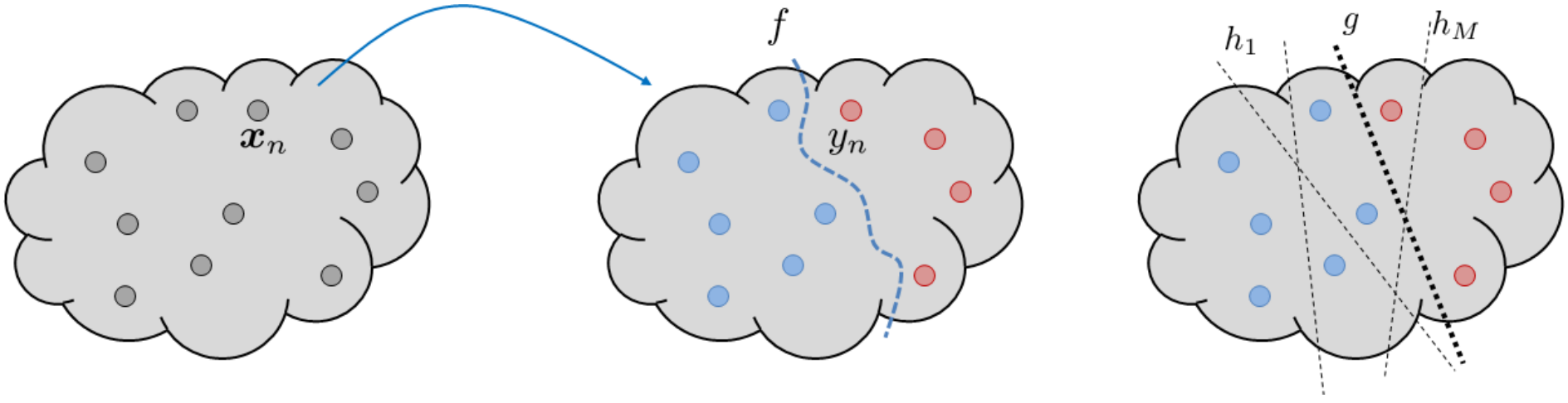
Training and Testing Set

Let us first talk about a dataset:

- **In-sample**: Samples that are inside the training set
- **Out-sample**: Samples that are outside the training set



Hypothesis Fucntion



- Hypothesis set: $\mathcal{H} = \{h_1, \ldots, h_M\}$: Possible decision boundaries
- Algorithm: Picks h_m from $\mathcal{H}$
- Final hypothesis: g: The one you found

Learning Model