Lecture 2: (Supervised) Machine Learning Concepts

COMP 343, Spring 2022
Victoria Manfredi

WESLEYAN UNIVERSITY

Acknowledgements: These slides are based primarily on content from the book "Machine Learning" by Tom Mitchell, slides created by Vivek Srikumar (U of Utah), Dan Roth (UPenn), and lectures by Balaraman Ravindran (IIT Madras), as well as from courses at Stanford (http://cs229.stanford.edu/) and UC Berkeley (http://ai.berkeley.edu/)

Today's Topics

Homework 1 out

- Due Wed., February 9 by 5p

Recap, Badges data

Using supervised learning

1. What is our instance space?
2. What is our label space?
3. What is our hypothesis space?

Recap

Learning as generalization

Given unseen photo, classify it as dog or not!

Learning has to go beyond just memorizing what has been seen so far: this is generalization

From slides of Vivek Srikumar (U of Utah)

Learning as generalization

Tom Mitchell, Machine Learning book (1997)

- "A computer program is said to learn from **experience E** with respect to some **class of tasks T** and **performance measure P**, if its performance at tasks in T, as measured by P, improves with experience E."

Performance measure P

- To determine whether learning is happening
- E.g.,
  - Spam emails correctly identified
  - Winning rate: whether you won or lost game
  - # of patients that were accurately diagnosed

Different learning paradigms

Supervised learning

- Learn with a teacher

Our focus in this class will be supervised learning!

Unsupervised learning

- Learn without a teacher

Semi-supervised learning

- Learn with and without a teacher

Active learning

- Learner and teacher interact with each other

Reinforcement learning

- Learn by interacting in environment

Key issues in machine learning

Modeling

- How to formulate your problem as a machine learning problem?
- How to represent data? Do you have enough data?
- Is the data of sufficient quality (e.g., errors in data, missing values)
- Which algorithms to use?

Representation

- What functions should we learn (hypothesis spaces) ?
- How to map raw input to an instance space?
- Any rigorous way to find these? Any general approach?

Algorithms

- What is a good learning algorithm?
- What is success? How confident can I be of results?
- Generalization vs. overfitting

Badges data

Badges data

What is the function used to label badges with + or -?

Data:

```
+ Naoki Abe
+ Kamal M. Ali
- Chidanand Apte
+ Javed Aslam
+ Timothy P. Barber
+ Peter Bartlett
- Shai Ben-David
+ Malini Bhandaru
- Avrim Blum
+ Carla E. Brodley
- Andrey Burago
- Claire Cardie
+ Jason Catlett
+ Mark Changizi
+ Wan P. Chiang
+ William Cohen
- Antoine Cornuejols
+ Lindley Darden
- Brian D. Davidson
- Scott E. Decatur
- Thomas G. Dietterich
+ Harris Drucker
- Thomas Ellman
+ Bob Evans
- Usama Fayyad
+ David Finton
+ Seth Flanders
+ Judy A. Franklin
+ Merrick L. Furst
+ Ricard Gavalda
+ David Gillman
+ Paul W. Goldberg
- Myriam Abramson
- Eric Allender
+ Minoru Asada
+ Haralabos Athanassiou
+ Michael W. Barley
- Eric Baum
+ George Berg
+ Bir Bhanu
- Anselm Blumer
+ Nader Bshouty
+ Tom Bylander
+ Richard A. Caruana
+ Nicolo Cesa-Bianchi
+ Pang-Chieh Chen
- Steve A. Chien
+ David Cohn
+ Mark W. Craven
- Chris Darken
+ Michael de la Maza
+ Gerald F. DeJong
+ Michael J. Donahue
- Chris Drummond
+ Tapio Elomaa
- Claudio Facchinetti
+ Aaron Feigelson
+ John Fischer
+ Lance Fortnow
+ Yoav Freund
+ Jean Gabriel Ganascia
+ Melinda T. Gervasio
- Attilio Giordana
+ Sally Goldman
+ David W. Aha
+ Dana Angluin
+ Lars Asker
+ Jose L. Balcazar
- Cristina Baroglio
+ Welton Becket
+ Neil Berkman
+ Reinhard Blasig
+ Justin Boyan
- Wray Buntine
+ Bill Byrne
+ John Case
- Philip Chan
- Zhixiang Chen
+ Jeffery Clouse
- Clare Bates Congdon
+ Robert P. Daley
- Bhaskar Dasgupta
- Olivier De Vel
+ Kan Deng
+ George A. Drastal
+ Hal Duncan
+ Susan L. Epstein
+ Tom Fawcett
+ Nicolas Fiechter
+ Paul Fischer
- Ameur Foued
+ Johannes Furnkranz
+ William Gasarch
+ Yolanda Gil
+ Kate Goelz
+ Diana Gordon
```

3min: talk with your neighbors

1. What are good features for badges data? Why/why not?
2. What is the labeling function you guessed? How did you arrive at it?

Questions to think about

What is the labeling function you guessed?

Questions to think about

What is the labeling function you guessed?

How can you be certain that you got the right function?

- How did you arrive at it?

vumanfredi@wesleyan.edu

Questions to think about

What is the labeling function you guessed?

How can you be certain that you got the right function?

- How did you arrive at it?

Remember, function maps domain to range

- What is our domain?

Questions to think about

What is the labeling function you guessed?

How can you be certain that you got the right function?

- How did you arrive at it?

Remember, function maps domain to range

- What is our domain? all possible names, essentially infinite

Questions to think about

What is the labeling function you guessed?

How can you be certain that you got the right function?

- How did you arrive at it?

Remember, function maps domain to range

- What is our domain? all possible names, essentially infinite
- But we have only 200 names! There may be other names that our function doesn't correctly map to badge label?

Questions to think about

Learning issues

- Is this predicting things about new names or just modeling existing data? Is there a difference?
- How did you know that you should look at the letters?
- What background knowledge about letters did you use? How did you know that is relevant?
- What “learning algorithm” did you use?

Using Supervised Learning

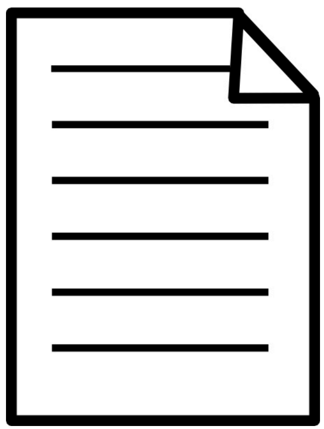
OVERVIEW

Instances and labels

Running example: automatically tag news articles

Instances and labels

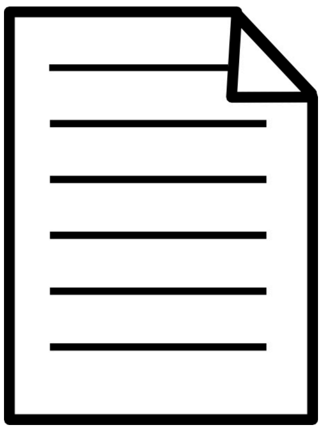
Running example: automatically tag news articles



An instance of a news article that needs to be classified

Instances and labels

Running example: automatically tag news articles

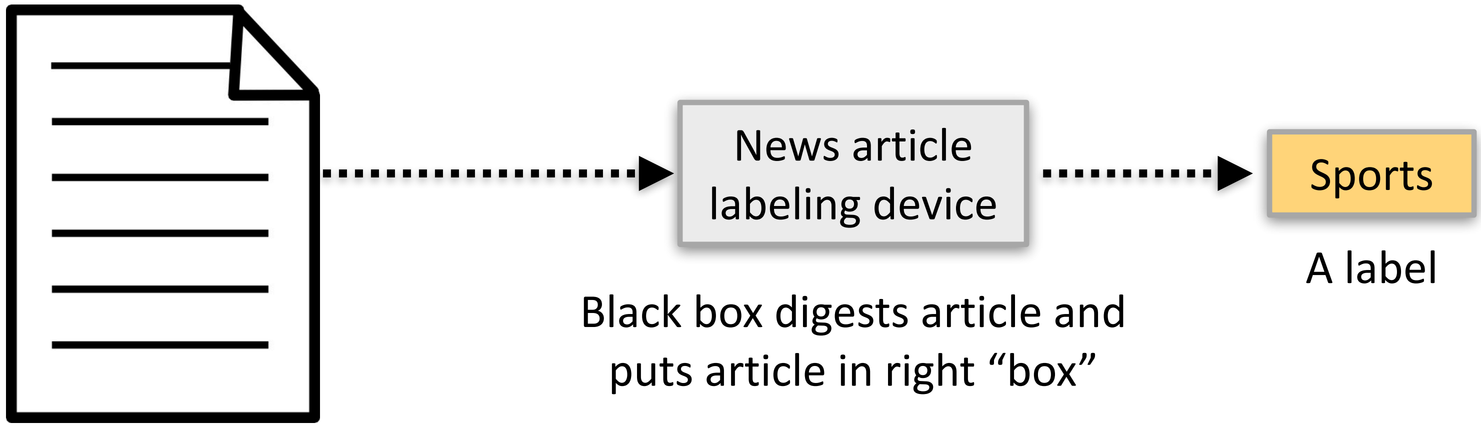


An instance of a news article that needs to be classified

What should this article be tagged as?

Instances and labels

Running example: automatically tag news articles



An instance of a news article that needs to be classified

Instances and labels

Running example: automatically tag news articles

Instance space: set of all possible news articles

Instance space includes articles we have as well as articles we don't have

Instances we have are our examples

Instances and labels

Running example: automatically tag news articles

Mapped by the learned function to

Black box is a function mapping instances to labels

Goal is to place article in 1 of 4 boxes

Instance space: set of all possible news articles

Business

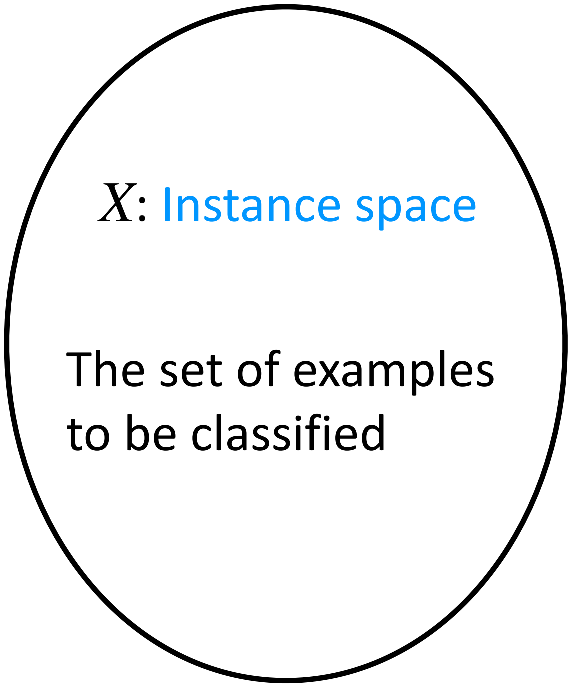
Entertainment

Politics

Sports

Label space: set of all possible labels

Instances and labels



E.g., the set of all possible names, documents, sentences, images, emails, …

X: instance space

$\mathbf{x}$: individual example in X

$$\mathbf{x} \in X$$

vumanfredi@wesleyan.edu

Instances and labels

X: Instance space

The set of examples to be classified

E.g., the set of all possible names, documents, sentences, images, emails, …

Instances are what we need to categorize or label

Input to our classifier

Instances and labels

X: Instance space

The set of examples to be classified

Y: Label space

The set of all possible labels

E.g., the set of all possible names, documents, sentences, images, emails, …

E.g., {Spam, Not-Spam}, {+, -}, …

vumanfredi@wesleyan.edu

Instances and labels

X: Instance space

The set of examples to be classified

E.g., the set of all possible names, documents, sentences, images, emails, …

Y: Label space

The set of all possible labels

E.g., {Spam, Not-Spam}, {+, -}, …

Y: instance label

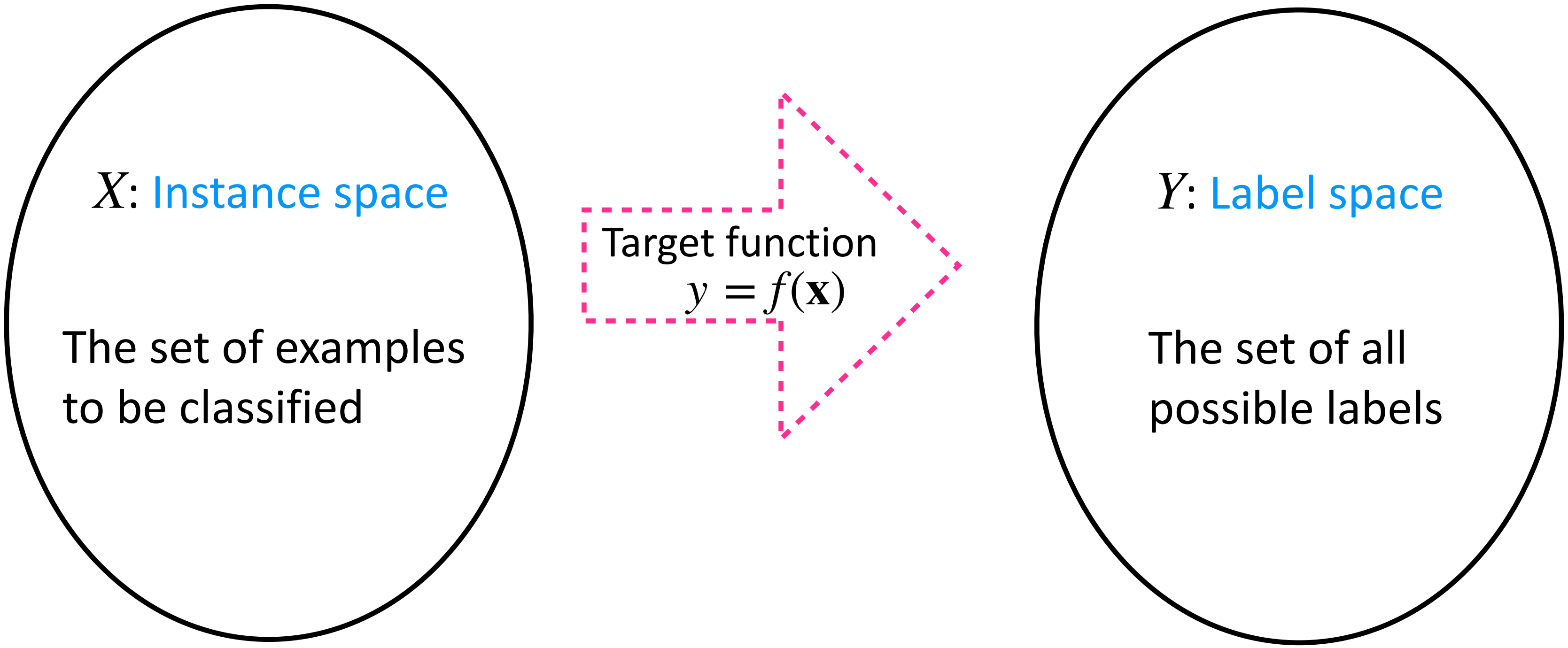
y: individual label in Y

$$y \in Y$$

Instances and labels

Note: we define instance and label space in context of a particular task: e.g., classifying emails, labeling badges, locating dogs in photos …

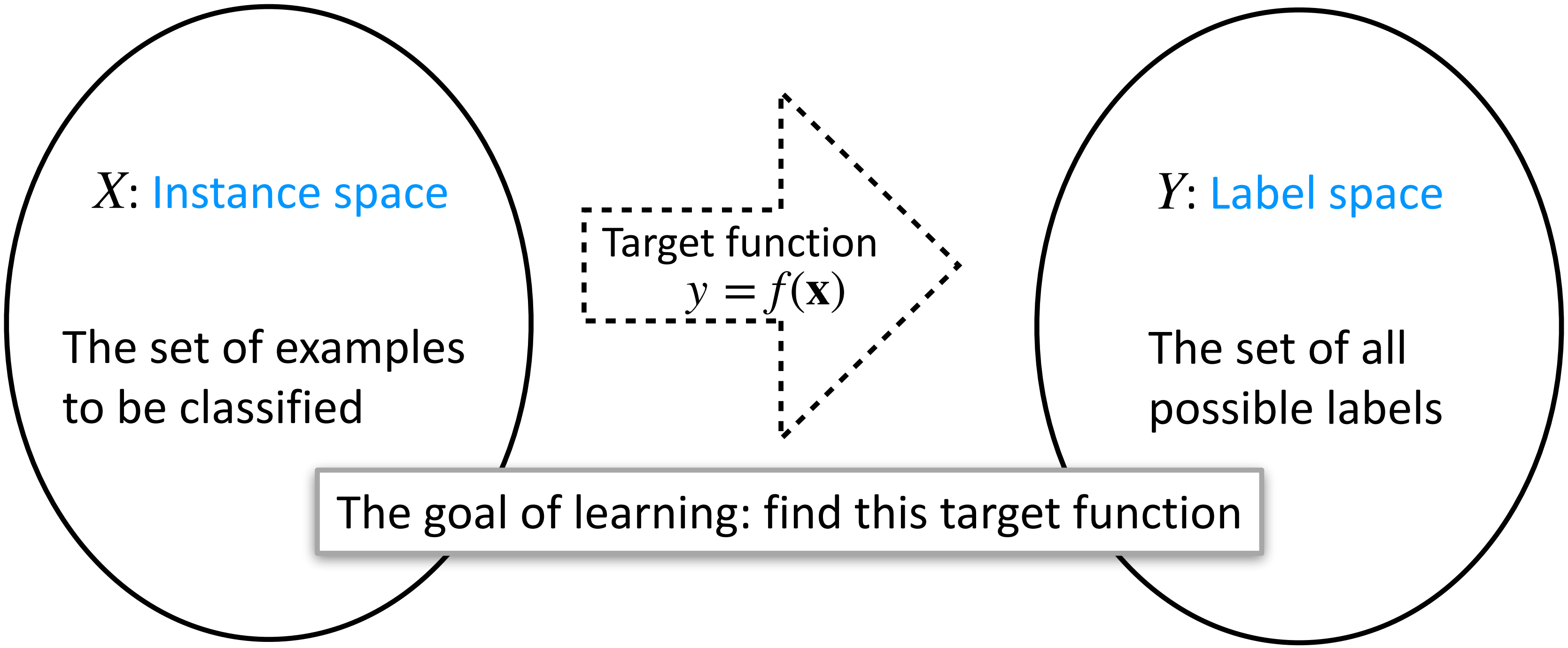
Target function



Imagine a perfect classifier that always gives you the right answer: this the target function, what we hope to find

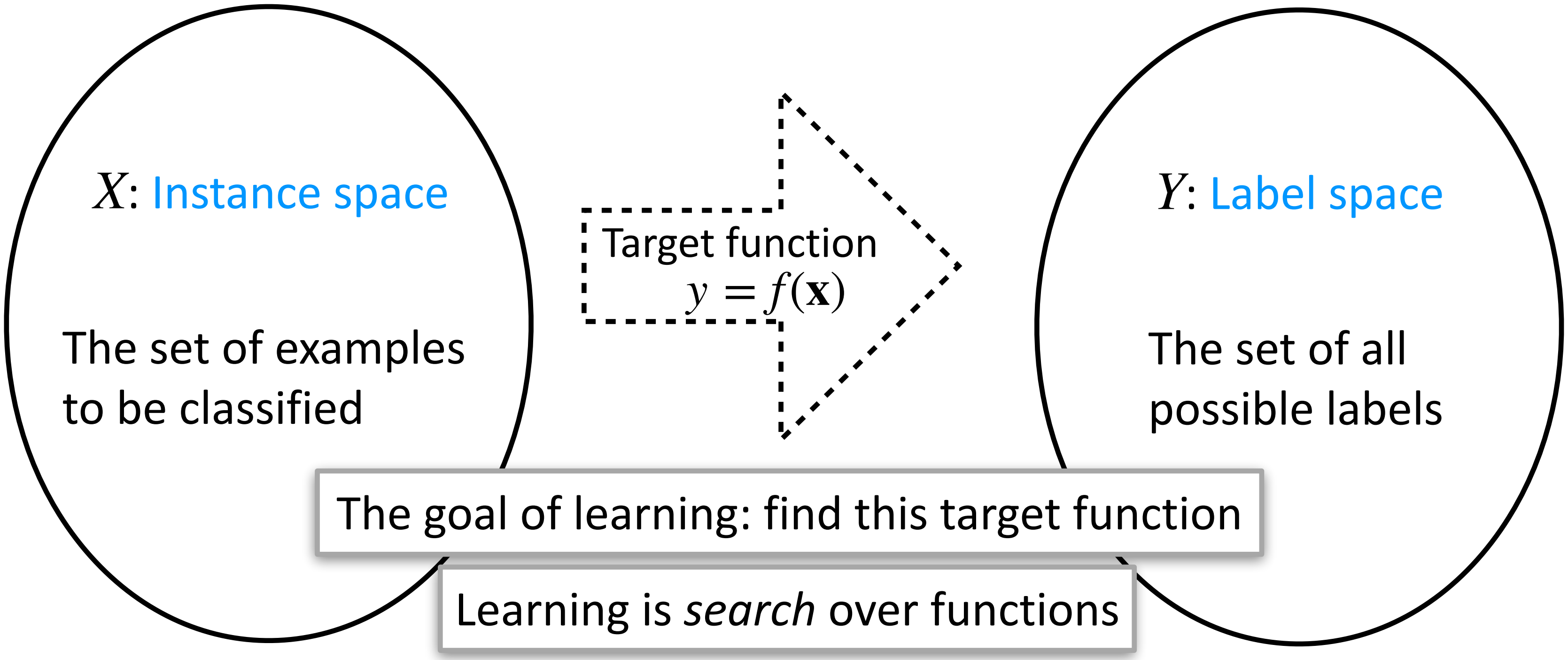
vumanfredi@wesleyan.edu

Target function



vumanfredi@wesleyan.edu

Target function



We need to search over the set of possible functions that exist to find the one function that maps instances to labels in the way we want

Supervised learning

X: Instance space

The set of examples to be classified

Target function
$y = f(\mathbf{x})$

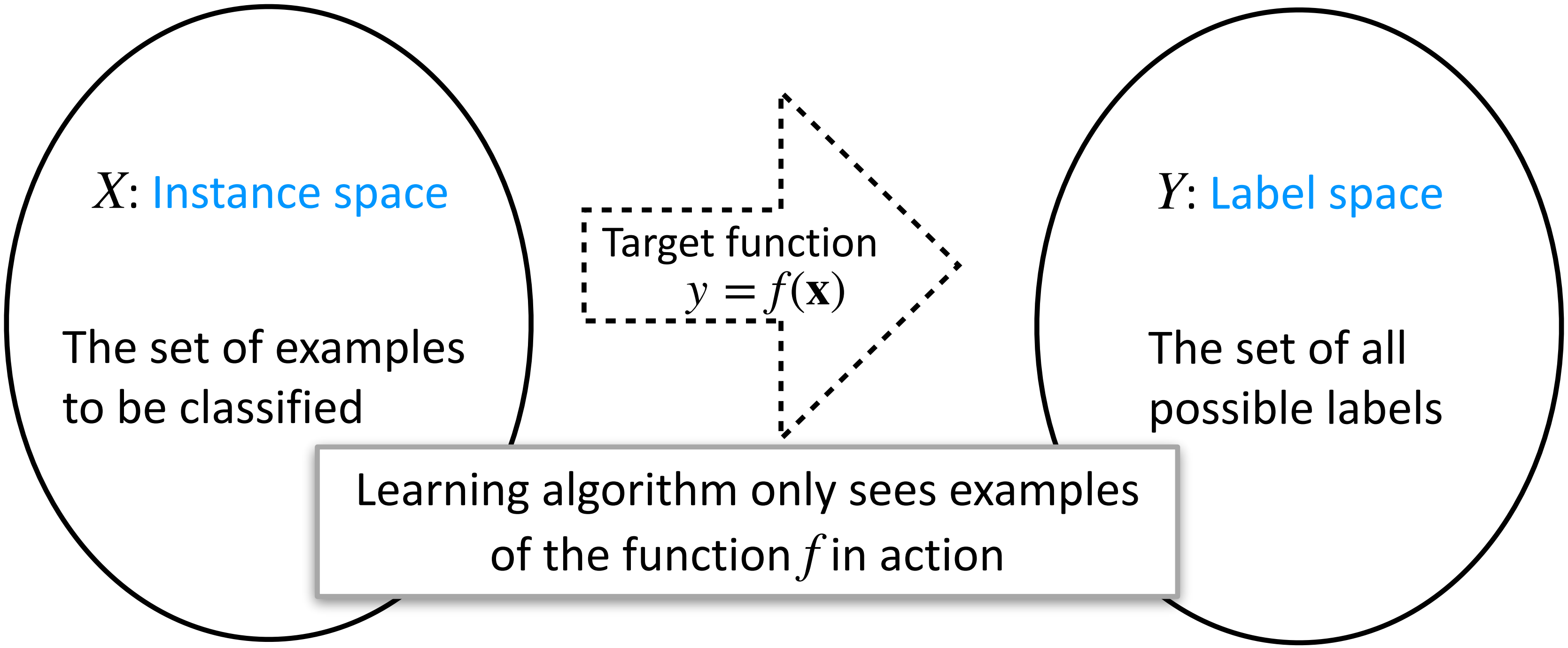
Y: Label space

The set of all possible labels

Learning algorithm only sees examples of the function f in action

We don't get to see the target function f!

Supervised learning

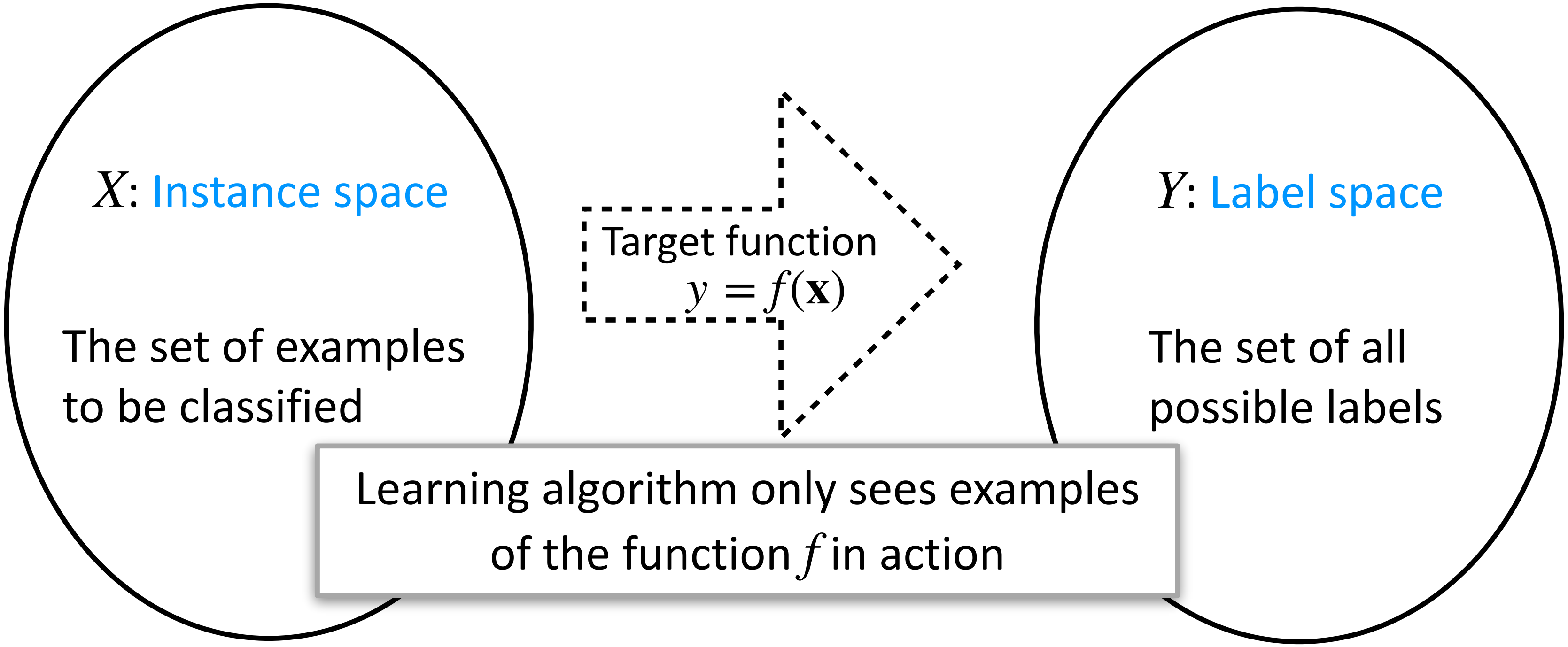


$$\begin{array}{c} \mathbf{x}_1, f(\mathbf{x}_1) \\ \mathbf{x}_2, f(\mathbf{x}_2) \\ \mathbf{x}_3, f(\mathbf{x}_3) \\ \vdots \\ \mathbf{x}_n, f(\mathbf{x}_n) \end{array}$$

Labeled training data

What might $\mathbf{x}_i$ be?

Supervised learning



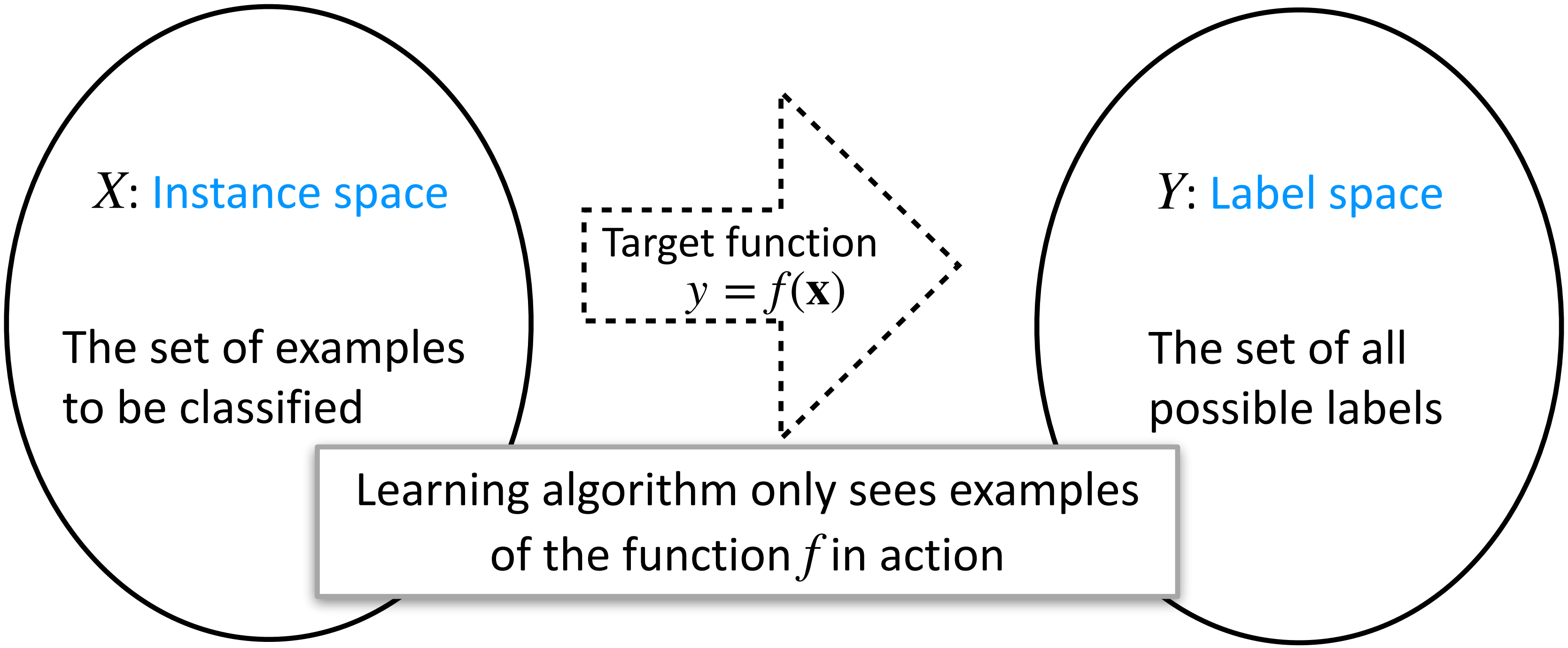
$$\begin{array}{c}\mathbf{x}_1, f(\mathbf{x}_1)\\ \mathbf{x}_2, f(\mathbf{x}_2)\\ \mathbf{x}_3, f(\mathbf{x}_3)\\ \vdots \\ \mathbf{x}_n, f(\mathbf{x}_n)\end{array}$$

Labeled training data

What might $\mathbf{x}_i$ be?

Newspaper articles, images, emails …

Supervised learning

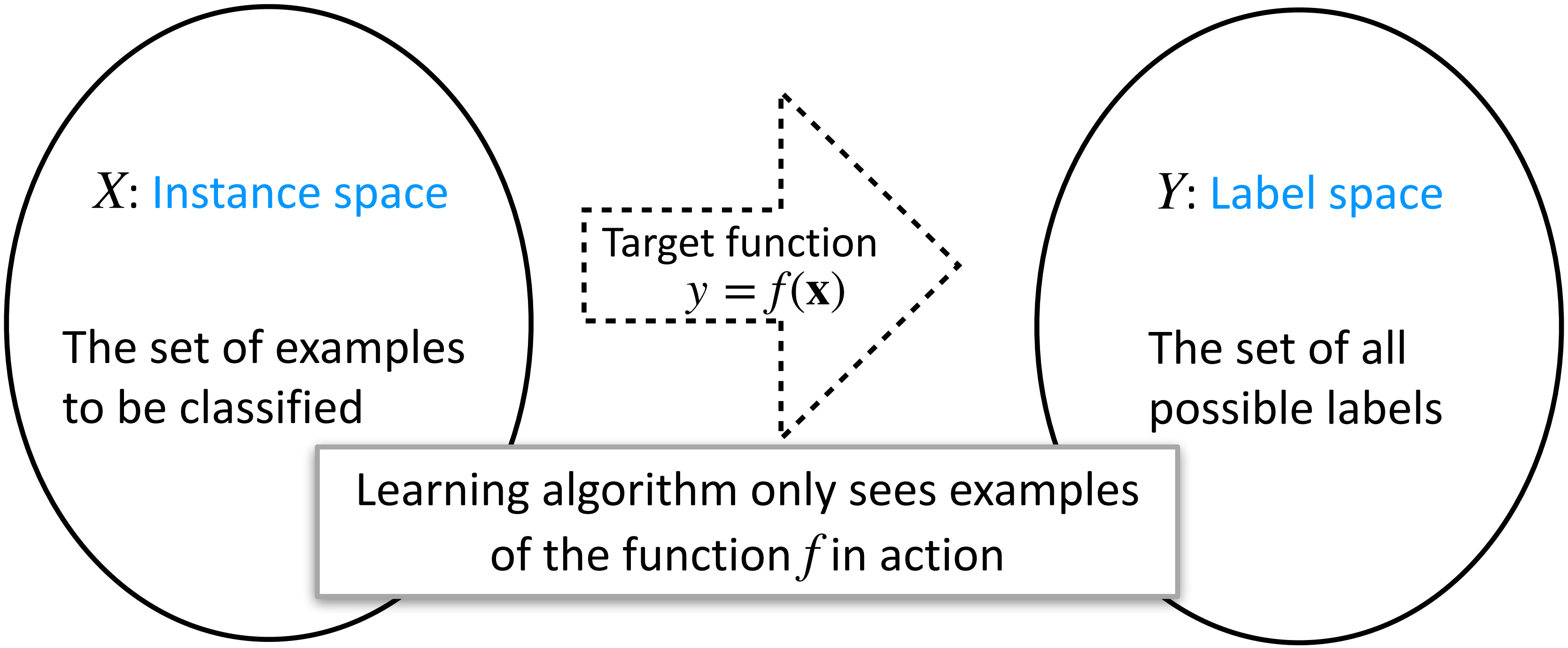


$$\begin{array}{c} \mathbf{x}_1, f(\mathbf{x}_1) \\ \mathbf{x}_2, f(\mathbf{x}_2) \\ \mathbf{x}_3, f(\mathbf{x}_3) \\ \vdots \\ \mathbf{x}_n, f(\mathbf{x}_n) \end{array}$$

Labeled training data

What are the $f(\mathbf{x}_i)$?

Supervised learning



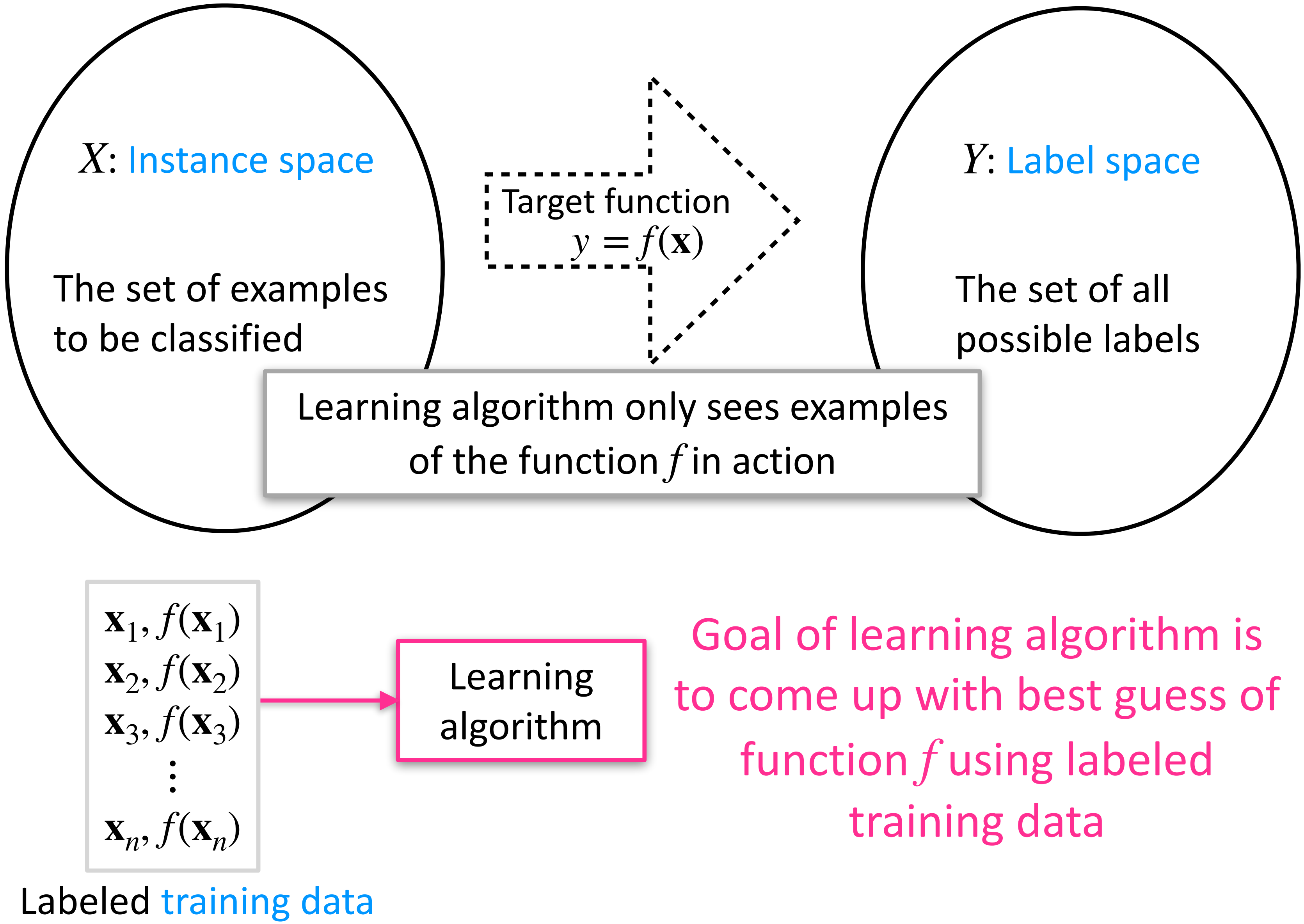
$\mathbf{x}_1, f(\mathbf{x}_1)$
$\mathbf{x}_2, f(\mathbf{x}_2)$
$\mathbf{x}_3, f(\mathbf{x}_3)$
$\vdots$
$\mathbf{x}_n, f(\mathbf{x}_n)$

Labeled training data

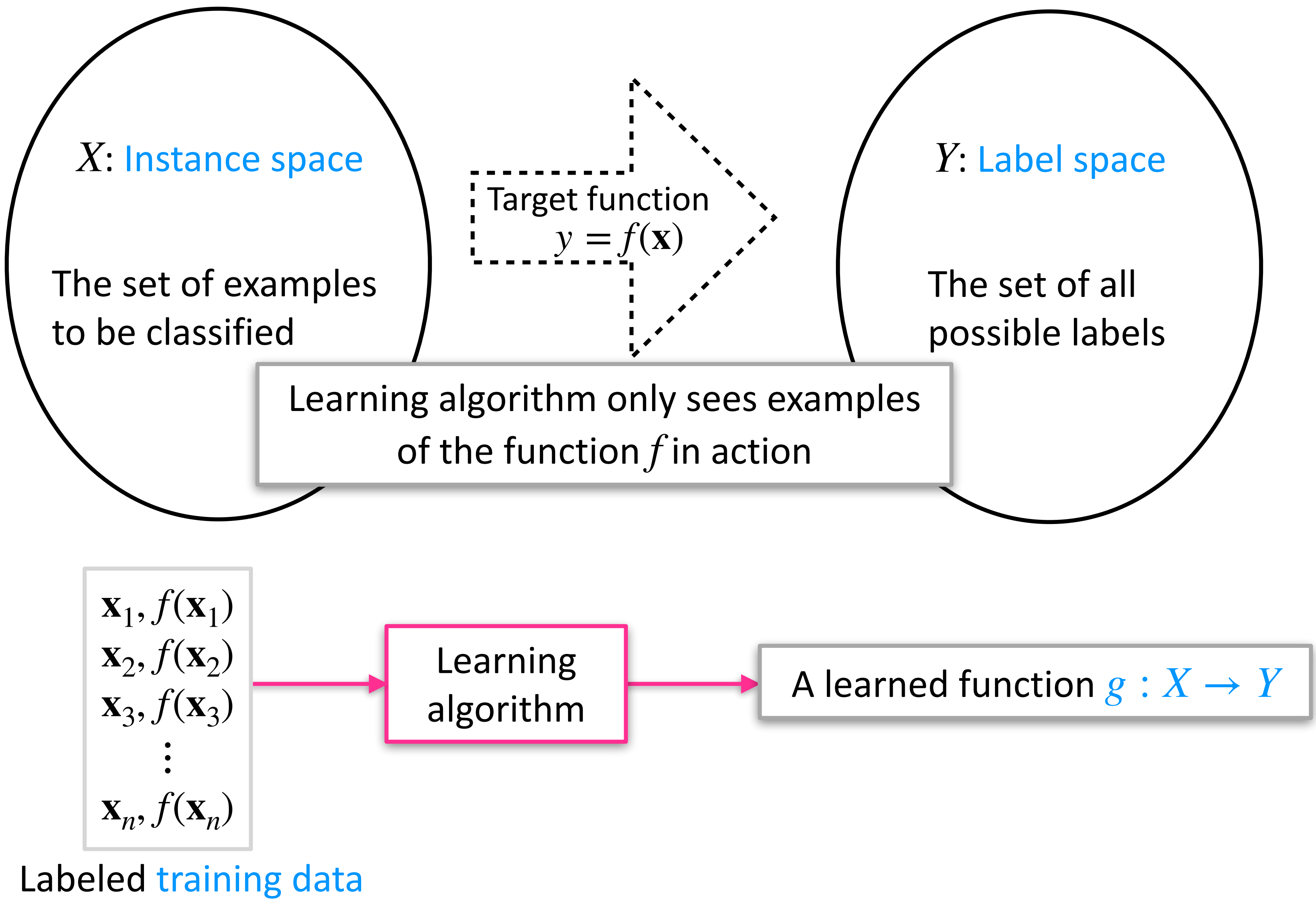
What are the $f(\mathbf{x}_i)$?

Output of the target function f on $\mathbf{x}_i$

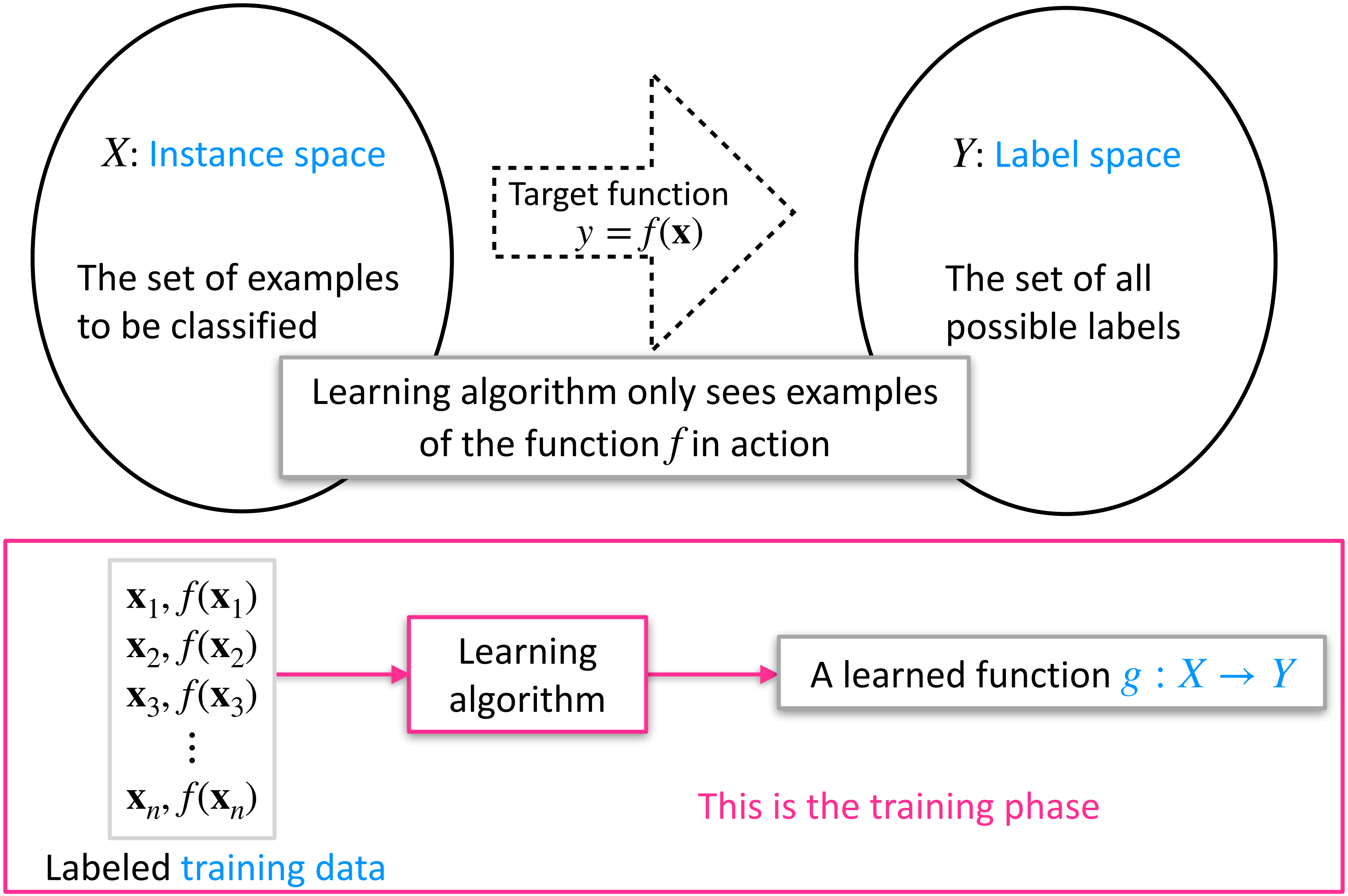
Supervised learning



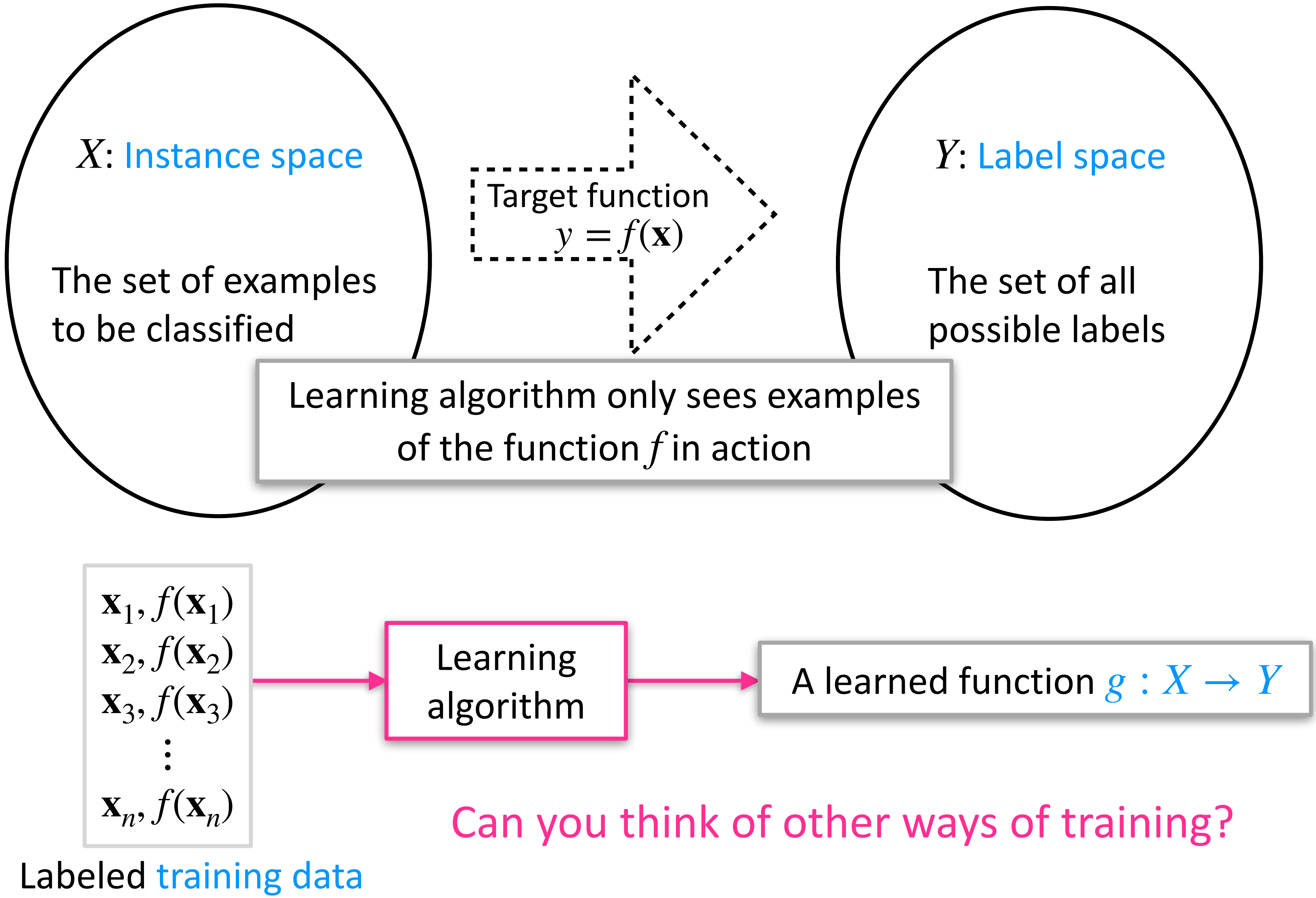
Supervised learning



Supervised learning



Supervised learning



Supervised learning

X: Instance space

The set of examples to be classified

Target function
$y = f(\mathbf{x})$

Y: Label space

The set of all possible labels

Learning algorithm only sees examples of the function f in action

$\mathbf{x}_1, f(\mathbf{x}_1)$
$\mathbf{x}_2, f(\mathbf{x}_2)$
$\mathbf{x}_3, f(\mathbf{x}_3)$
$\vdots$
$\mathbf{x}_n, f(\mathbf{x}_n)$

Labeled training data

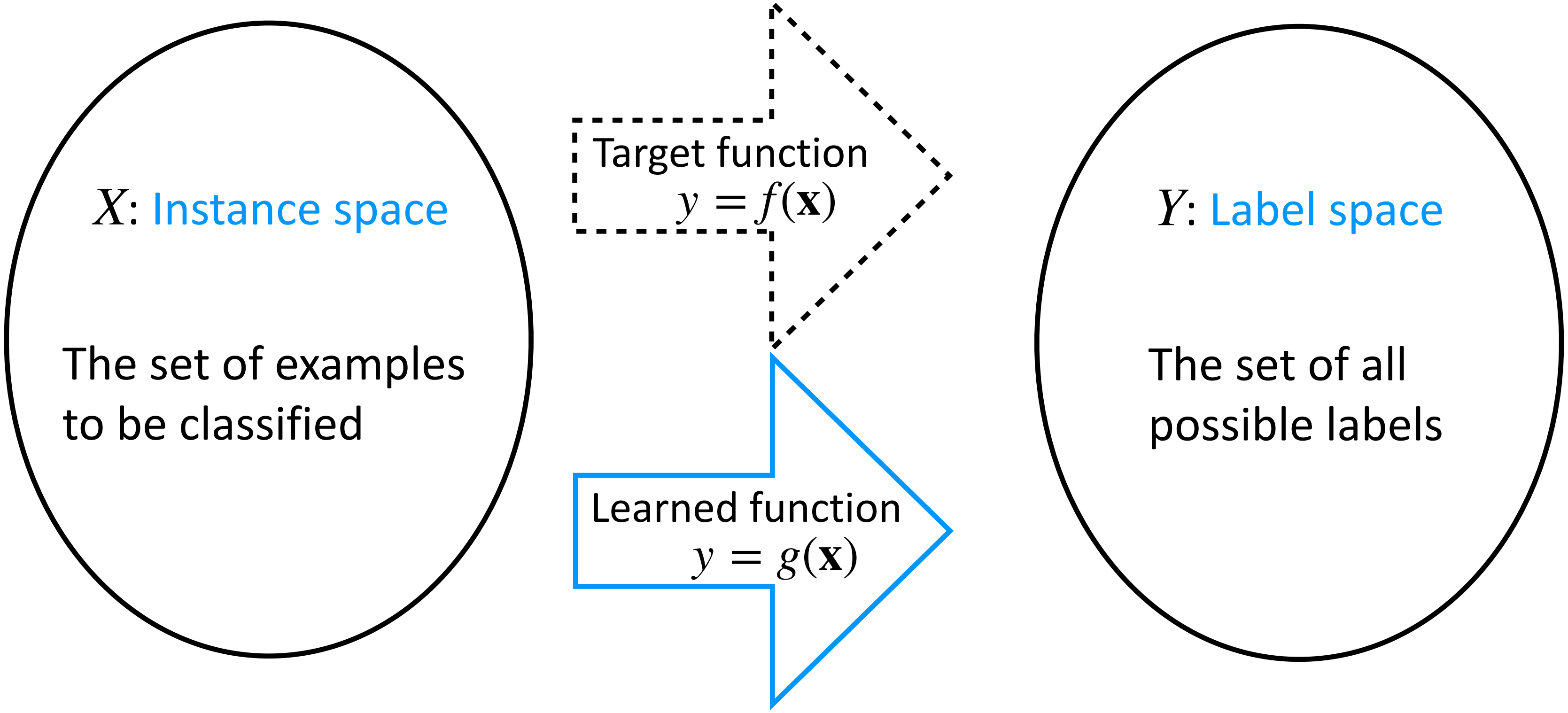
$\rightarrow$ Learning algorithm $\rightarrow$ A learned function $g : X \rightarrow Y$

Can you think of other ways of training?

What if we saw only one example at a time (online learning)?

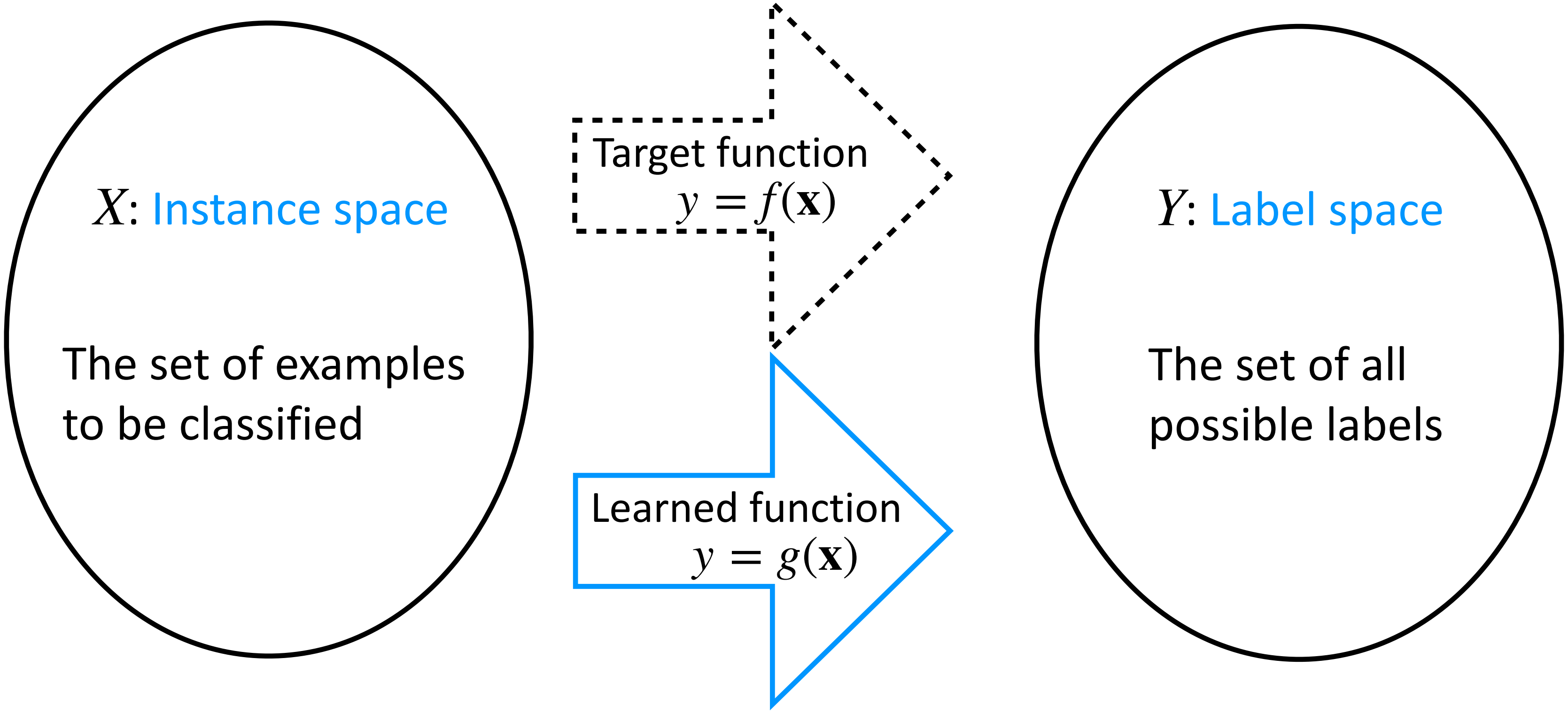
What if we didn't have labels (unsupervised learning)?

Supervised learning: evaluation



How do you know whether g is a good function?

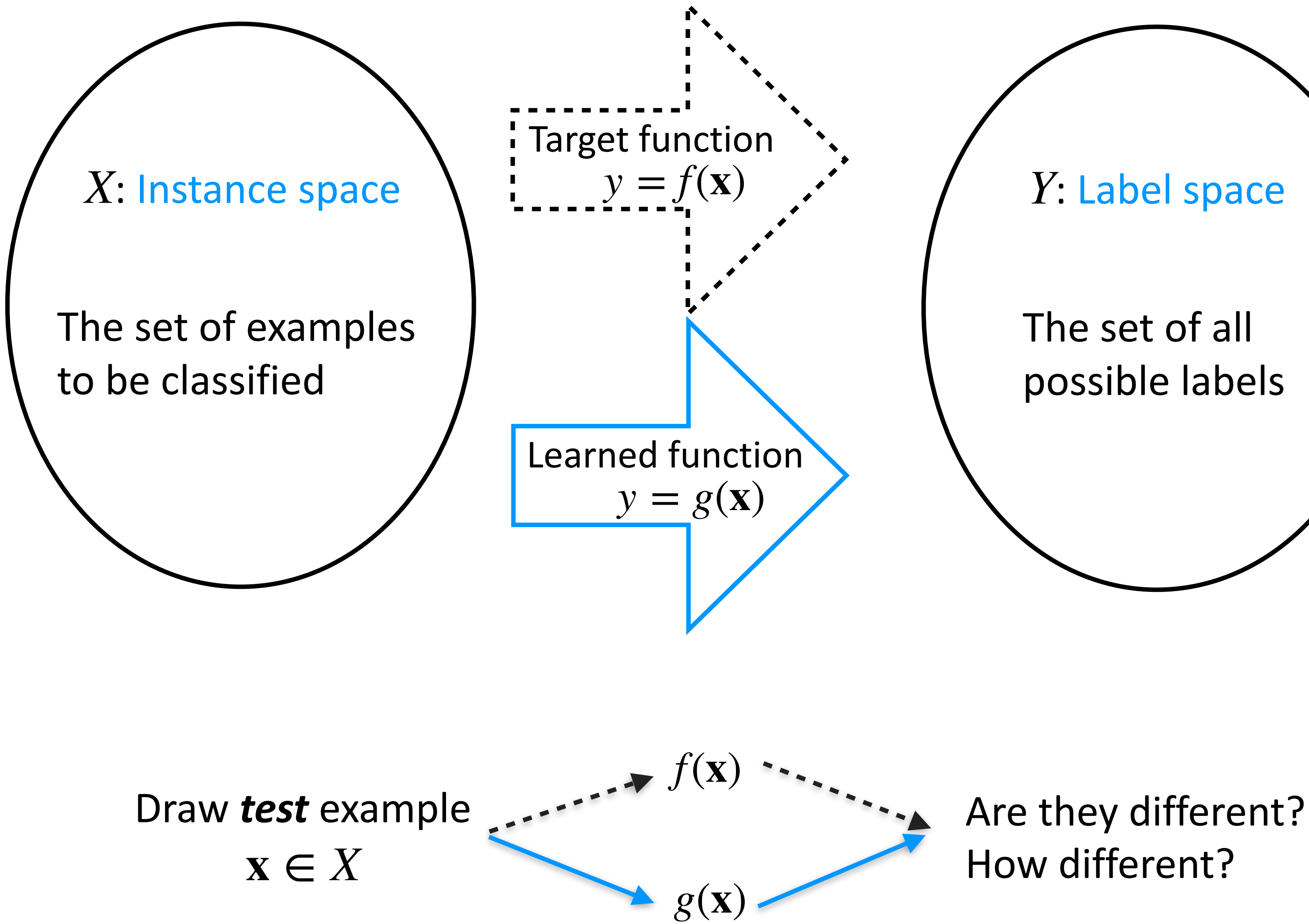
Supervised learning: evaluation



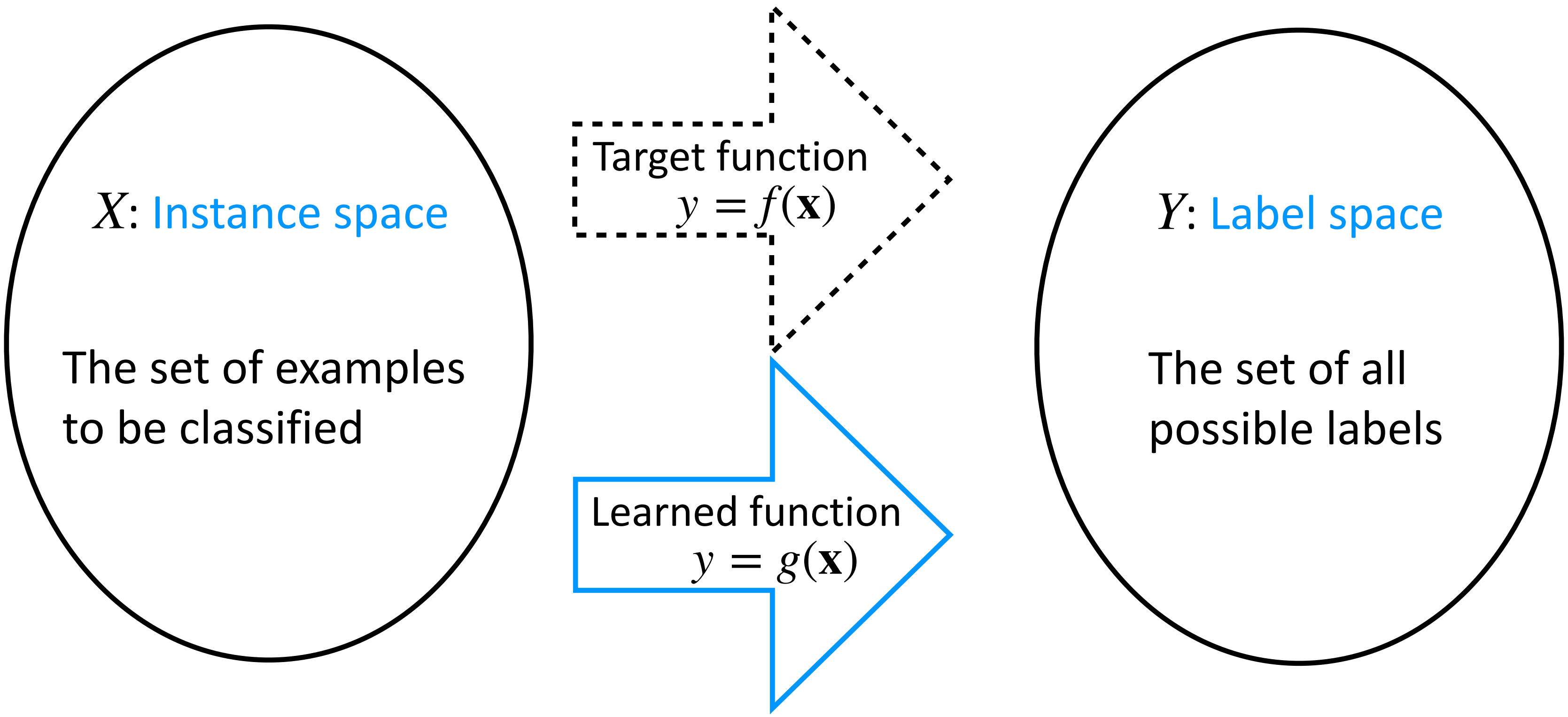
How do you know whether g is a good function?

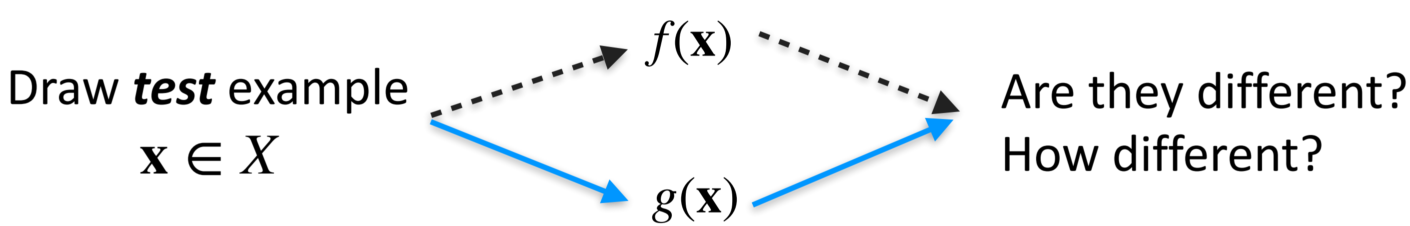
Use examples and labels g has not seen to test.

Supervised learning: evaluation



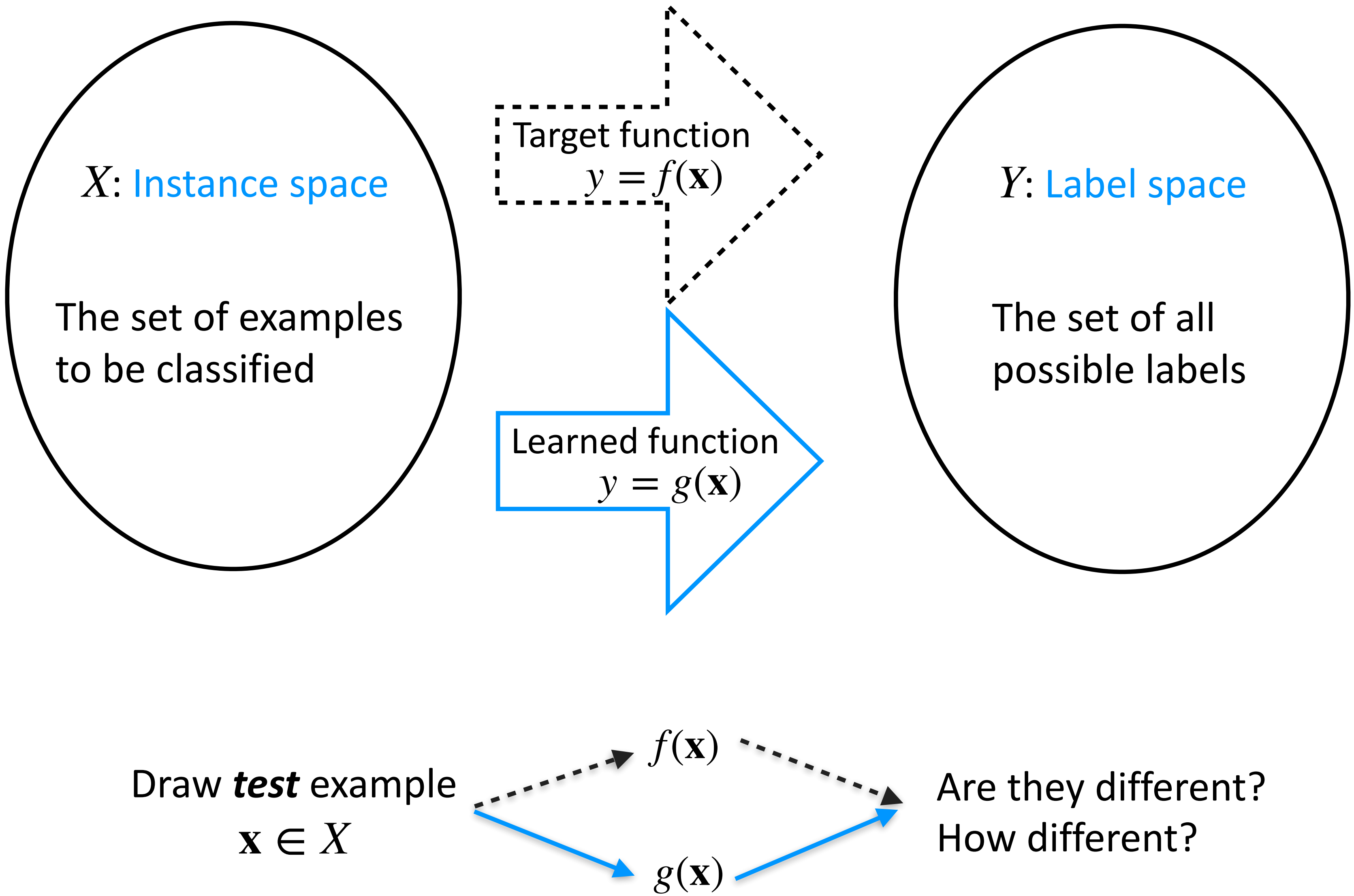
Supervised learning: evaluation





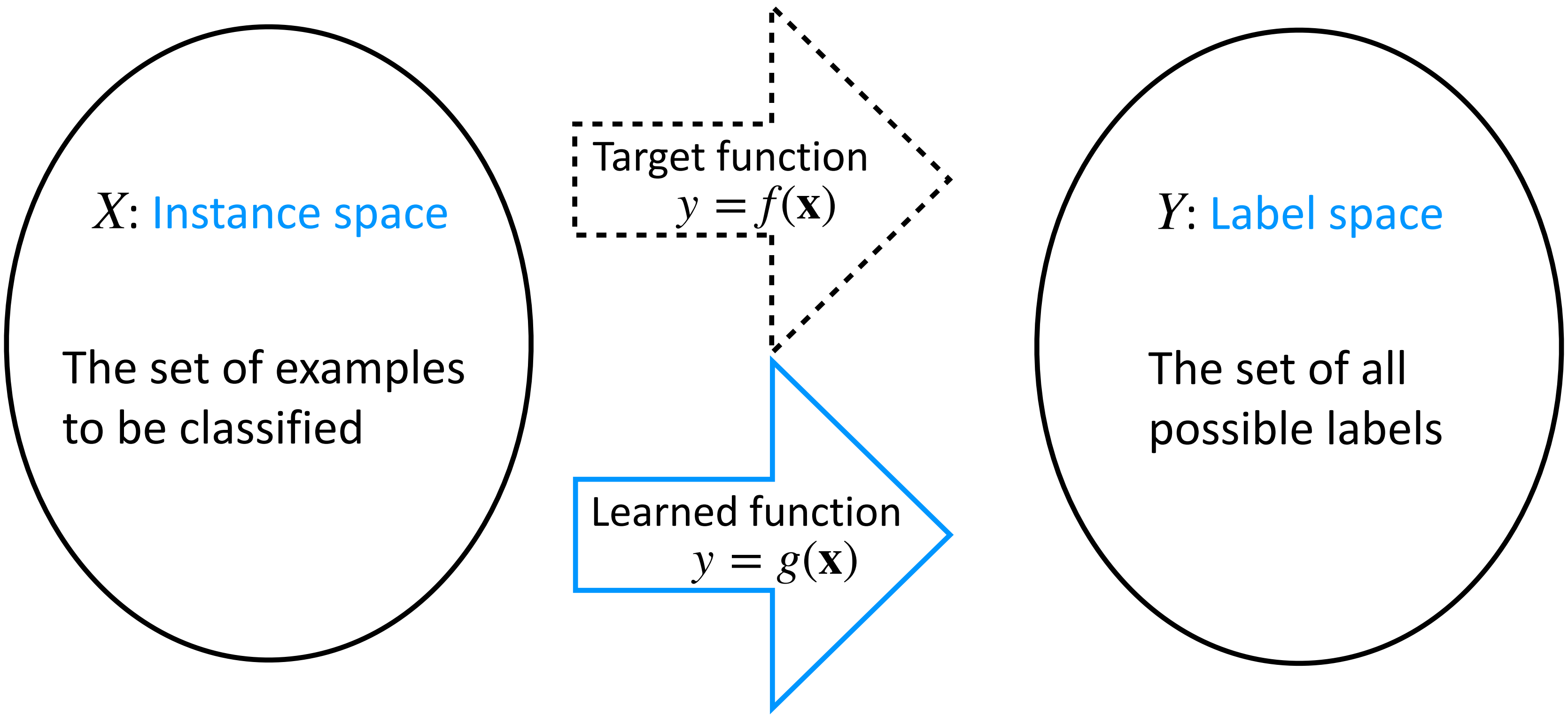
Apply model to many test examples and compare to the target's prediction
Aggregate these results to get a quality measure

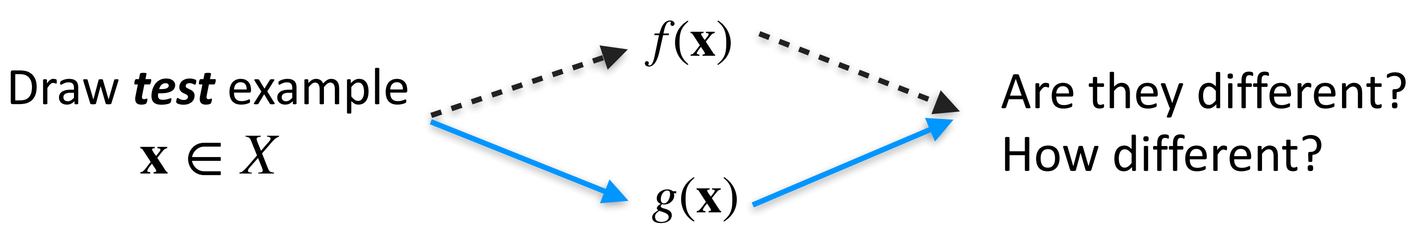
Supervised learning: evaluation



Can we use these test examples during the training phase?

Supervised learning: evaluation





If we train on test examples, essentially memorizing labels for examples. No generalization as a result.

The general setting for supervised learning

Given: training examples that are pairs of the form $(\mathbf{x}, f(\mathbf{x}))$

The general setting for supervised learning

Given: training examples that are pairs of the form $(\mathbf{x}, f(\mathbf{x}))$

The function f is unknown

This is what we want to discover!

The general setting for supervised learning

Given: training examples that are pairs of the form $(\mathbf{x}, f(\mathbf{x}))$

Typically the input x is represented as feature vectors

The function f is unknown

The general setting for supervised learning

Given: training examples that are pairs of the form $(\mathbf{x}, f(\mathbf{x}))$

Typically the input x is represented as feature vectors

- E.g.,: $\mathbf{x} \in \{0,1\}^d$ or $\mathbf{x} \in \Re^d$ (d-dimensional vectors)

Basically an array with d elements!

The function f is unknown

The general setting for supervised learning

Given: training examples that are pairs of the form $(\mathbf{x}, f(\mathbf{x}))$

Typically the input x is represented as feature vectors

- E.g.,: $\mathbf{x} \in \{0,1\}^d$ or $\mathbf{x} \in \Re^d$ (d-dimensional vectors)
- A deterministic mapping from instances in your problem (e.g., news articles) to features

The function f is unknown

The general setting for supervised learning

Given: training examples that are pairs of the form $(\mathbf{x}, f(\mathbf{x}))$

Typically the input x is represented as feature vectors

- E.g.,: $\mathbf{x} \in \{0,1\}^d$ or $\mathbf{x} \in \Re^d$ (d-dimensional vectors)
- A deterministic mapping from instances in your problem (e.g., news articles) to features

The function f is unknown

Instances: real things to categorize, like emails, articles, pictures

Features: “interesting” attributes of the instances, like 1 if email contains word free, 0 otherwise, pixel patterns, …

Features form a vector space: d-dimensional vectors form a d-dimensional vector space. Each number in vector is a single dimension that captures one feature

The general setting for supervised learning

Given: training examples that are pairs of the form $(\mathbf{x}, f(\mathbf{x}))$

Typically the input x is represented as feature vectors

- E.g.,: $\mathbf{x} \in \{0,1\}^d$ or $\mathbf{x} \in \Re^d$ (d-dimensional vectors)
- A deterministic mapping from instances in your problem (e.g., news articles) to features

The function f is unknown

For a training example $(\mathbf{x}, f(\mathbf{x}))$, the value of $f(\mathbf{x})$ is called its label

The general setting for supervised learning

Given: training examples that are pairs of the form $(\mathbf{x}, f(\mathbf{x}))$

The goal of learning: use the training examples to find a good approximation for f

The label determines the kind of problem we have

- **Binary classification**: label space = $\{-1,1\}$
- **Multiclass classification**: label space = $\{1,2,3,\ldots,K\}$
- **Regression**: label space $=\Re$

The general setting for supervised learning

Given: training examples that are pairs of the form $(\mathbf{x}, f(\mathbf{x}))$

The goal of learning: use the training examples to find a good approximation for f

The label determines the kind of problem we have

- **Binary classification**: label space = $\{-1, 1\}$

The general setting for supervised learning

Given: training examples that are pairs of the form $(\mathbf{x}, f(\mathbf{x}))$

The goal of learning: use the training examples to find a good approximation for f

The label determines the kind of problem we have

- **Binary classification**: label space = $\{-1,1\}$
- **Multiclass classification**: label space = $\{1,2,3,\ldots,K\}$

The general setting for supervised learning

Given: training examples that are pairs of the form $(\mathbf{x}, f(\mathbf{x}))$

The goal of learning: use the training examples to find a good approximation for f

The label determines the kind of problem we have

- **Binary classification**: label space = $\{-1, 1\}$
- **Multiclass classification**: label space = $\{1, 2, 3, \ldots, K\}$
- **Regression**: label space $= \Re$

What are good applications for supervised learning?

Or, when should we use (supervised learning) and when should we not?

Talk to your neighbor and discuss

What are good applications for supervised learning?

Or, when should we use (supervised learning) and when should we not?

There is no human expert

- e.g., identify DNA binding sites

Humans can perform a task, but can't describe how they do it

- e.g., object recognition, is there a cat in the image?

The desired function is hard to obtain in closed form

- e.g., will stock market go up or down tomorrow?

Examples of binary classification

The label space consists of 2 elements

Spam filtering

- Is an email spam or not?

Recommendation systems

- Given user’s movie preferences, will she like a new movie

Anomaly or malware detection

- Is a smartphone app malicious?
- Is a Twitter user a bot?

Authorship identification

- Were the two documents written by the same person?

Times series prediction

- Will the future value of a stock increase or decrease with respect to its current value?

Examples of binary classification

The label space consists of 2 elements

Spam filtering

- Is an email spam or not?

Recommendation systems

- Given user's movie preferences, will she like a new movie

Anomaly or malware detection

- Is a smartphone app malicious?
- Is a Twitter user a bot?

Authorship identification

- Were the two documents written by the same person?

Times series prediction

- Will the future value of a stock increase or decrease with respect to its current value?

Many tasks can be reduced to binary classification

Using Supervised Learning

EXAMPLES

Using supervised learning

We should be able to specify

1. What is our instance space?
   - What are the inputs to the problem? What are the features?

2. What is our label space?
   - What kind of learning task are we dealing with?

3. What is our hypothesis space?
   - What functions should the learning algorithm search over?

4. What is our learning algorithm?
   - How do we learn the model from the labeled data?

5. What is our loss function or evaluation metric?
   - How do we measure success? What drives learning?

Using supervised learning

We should be able to specify

1. What is our instance space?
   - What are the inputs to the problem? What are the features?

2. What is our label space?
   - What kind of learning task are we dealing with?

3. What is our hypothesis space?
   - What functions should the learning algorithm search over?

4. What is our learning algorithm?
   - How do we learn the model from the labeled data?

5. What is our loss function or evaluation metric?
   - How do we measure success? What drives learning?

Much of the rest of the semester

1. The instance space X

X: Instance space

The set of examples to be classified

Target function
$y = f(x)$

Y: Label space

The set of all possible labels

The goal of learning: find this target function

Learning is *search* over functions

E.g., the set of all possible names, documents, sentences, images, emails, …

E.g., {Spam, Not-Spam}, {+, -}, …

1. The instance space X

X: Instance space

The set of examples to be classified

E.g., the set of all possible names, documents, sentences, images, emails, …

Designing an appropriate *feature representation* of the instance space is crucial

Instances $x \in X$ are defined by features/ attributes

What might features be?

vumanfredi@wesleyan.edu

1. The instance space X

X: Instance space

The set of examples to be classified

E.g., the set of all possible names, documents, sentences, images, emails, …

Designing an appropriate *feature representation* of the instance space is crucial

Instances $x \in X$ are defined by features/ attributes

Features could be Boolean

- Example: does the email contain the word "free"

Features could be real-valued

- Example: what is the height of the person?
- Example: what was the stock price yesterday?

Features could be hand-crafted or themselves learned

Instances as feature vectors

An input to the problem
(e.g., emails, names, images)

Feature
function

A feature vector

Instances as feature vectors

An input to the problem (e.g., emails, names, images) → Feature function → A feature vector

Feature functions, also known as feature extractors

- Often deterministic, but could also be learned
- Convert the examples to a collection of attributes (typically thought of as high-dimensional vectors)

Important part of the design of a learning based solution