Machine Learning

Lecture 2: Empirical Risk Minimization

Sibei Yang

SIST, ShanghaiTech

Email: yangsb@shanghaitech.edu.cn

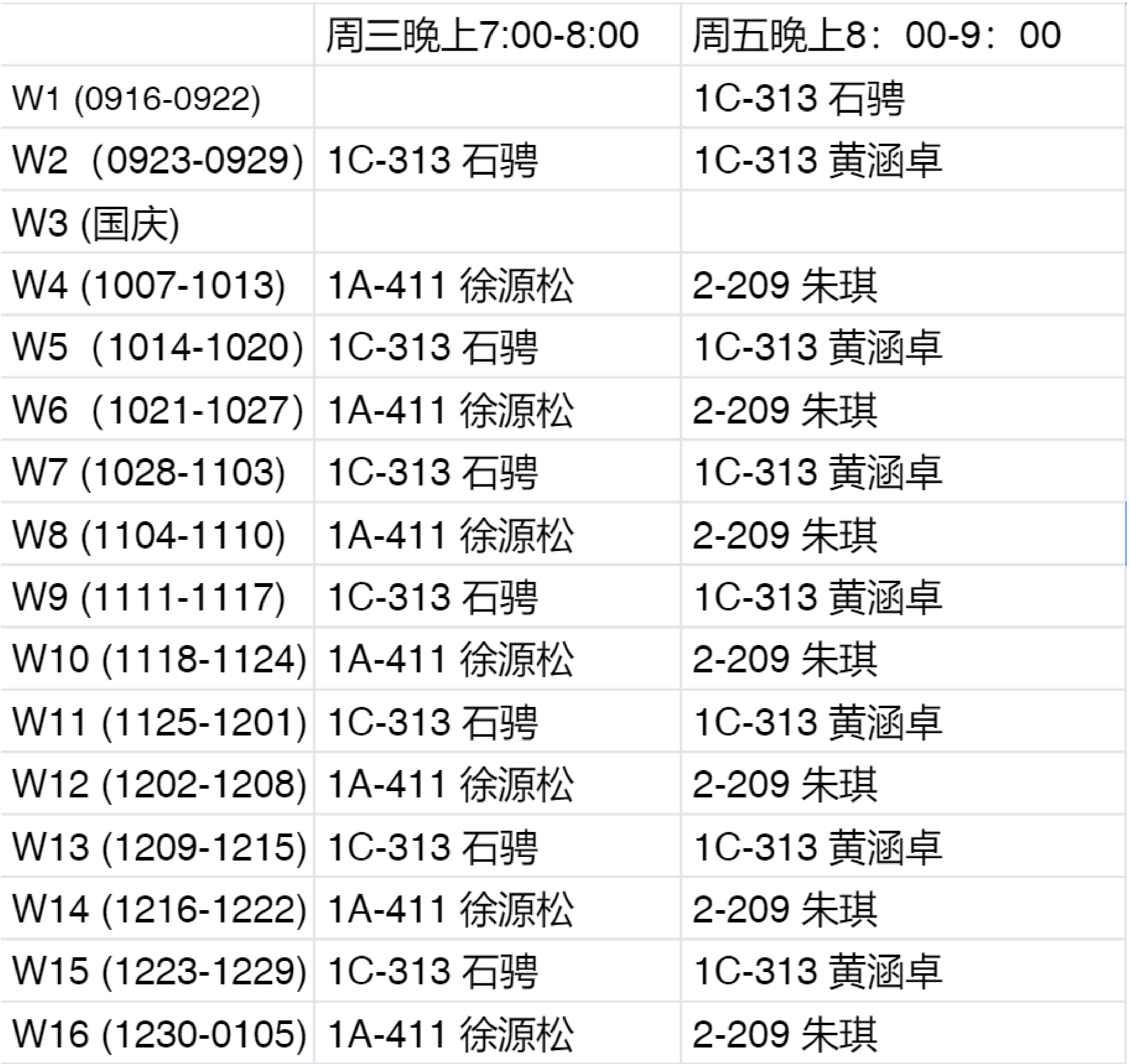
Outline

- Hypothesis class
- Joint Distribution of Data
- Expected Risk Minimization
- Empirical Risk Minimization

Sections 1-3 from Introduction to Statistical Learning Theory

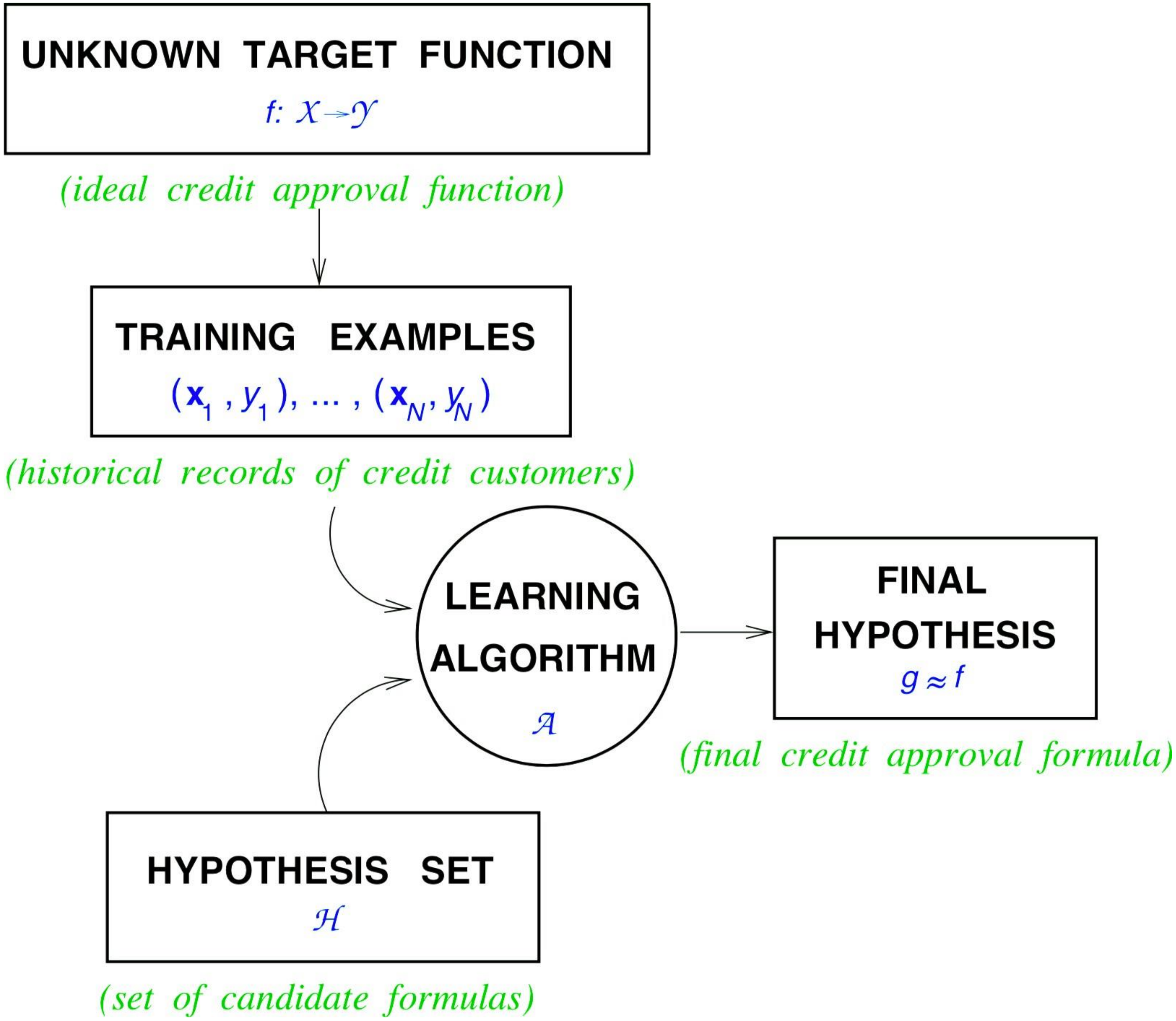
Announcement - Office Hour

- Tuesday: 10:00-12:00
  杨思蓓（ 1C-403D ）



	周三晚上7:00-8:00	周五晚上8：00-9：00
W1 (0916-0922)		1C-313 石骋
W2（0923-0929）	1C-313 石骋	1C-313 黄涵卓
W3 (国庆)		
W4 (1007-1013)	1A-411 徐源松	2-209 朱琪
W5（1014-1020）	1C-313 石骋	1C-313 黄涵卓
W6（1021-1027）	1A-411 徐源松	2-209 朱琪
W7 (1028-1103)	1C-313 石骋	1C-313 黄涵卓
W8 (1104-1110)	1A-411 徐源松	2-209 朱琪
W9 (1111-1117)	1C-313 石骋	1C-313 黄涵卓
W10 (1118-1124)	1A-411 徐源松	2-209 朱琪
W11 (1125-1201)	1C-313 石骋	1C-313 黄涵卓
W12 (1202-1208)	1A-411 徐源松	2-209 朱琪
W13 (1209-1215)	1C-313 石骋	1C-313 黄涵卓
W14 (1216-1222)	1A-411 徐源松	2-209 朱琪
W15 (1223-1229)	1C-313 石骋	1C-313 黄涵卓
W16 (1230-0105)	1A-411 徐源松	2-209 朱琪

Components of learning



UNKNOWN TARGET FUNCTION

$f: \mathcal{X} \to \mathcal{Y}$

(ideal credit approval function)

TRAINING EXAMPLES

$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$

(historical records of credit customers)

LEARNING ALGORITHM

$\mathcal{A}$

FINAL HYPOTHESIS

$g \approx f$

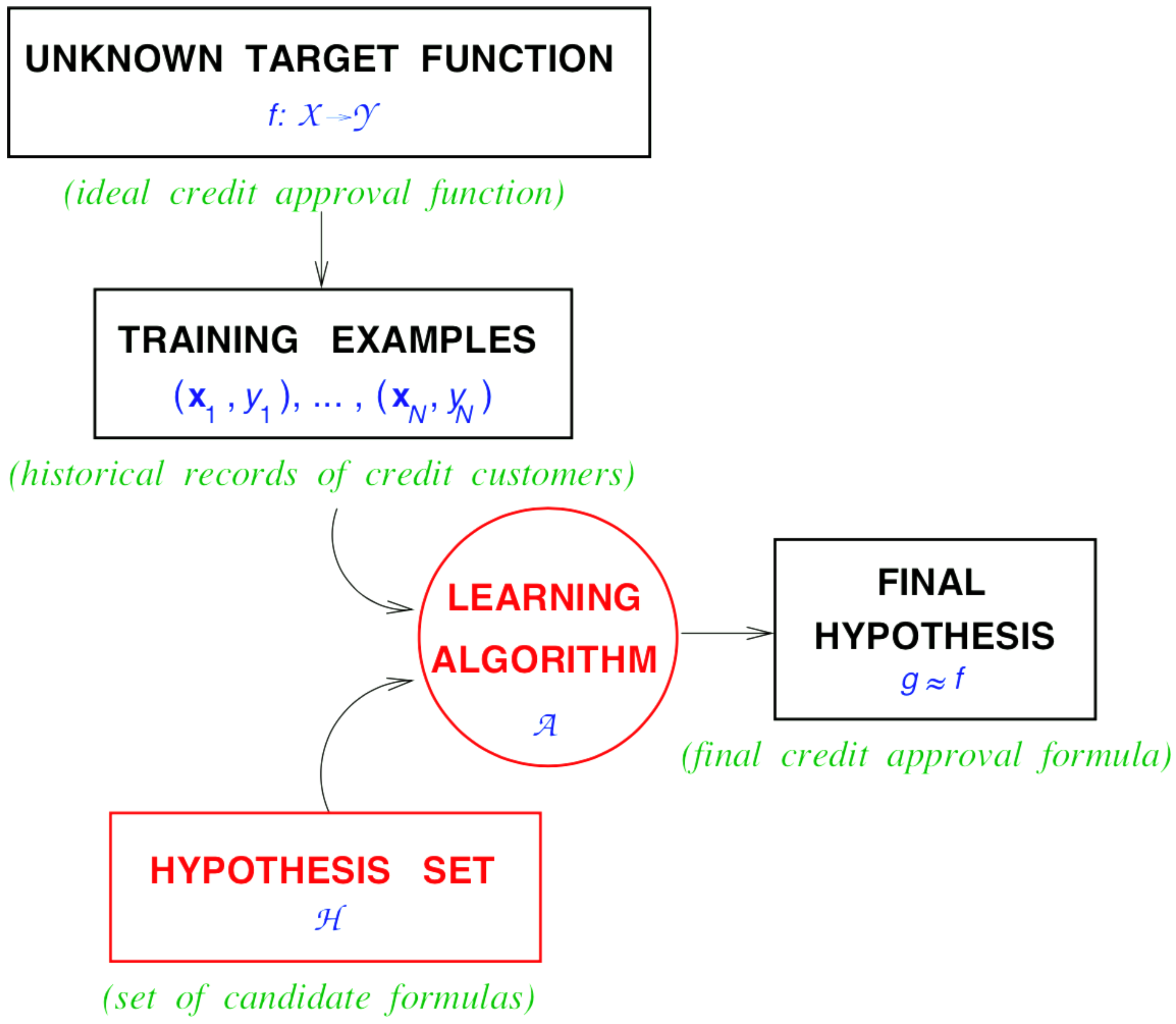
(final credit approval formula)

HYPOTHESIS SET

$\mathcal{H}$

(set of candidate formulas)

Components of learning


UNKNOWN TARGET FUNCTION
$f: \mathcal{X} \to \mathcal{Y}$
(ideal credit approval function)
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
(historical records of credit customers)
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g \approx f$
(final credit approval formula)
HYPOTHESIS SET
$\mathcal{H}$
(set of candidate formulas)

Two solution components of the learning problem:

- The Hypothesis Set:
  $\mathcal{H} = \{h\}, \quad g \in \mathcal{H}$
- The learning algorithm

Together, they are referred to as the learning model.

Why

- How good are different algorithms on unknown test sets?
- How many training samples do we need to achieve small error?
- What is the smallest possible error we can achieve?
- …

Learning Theory

Data and Labels

- In learning we seek a mapping from the initial data $\mathcal{X}$ (the domain of abstract input objects) to some label set $\mathcal{Y}$ (anything we want to predict)

- Hypothesis: $h : \mathcal{X} \rightarrow \mathcal{Y}$

- Example: in character recognition, $\mathcal{X}$ consists of possible images of letters and $\mathcal{Y}$ consists of the twenty-six letters of the Latin alphabet.

- Note: For simplicity we will use binary labels $\{+1, -\ 1\}$. Whether something is the letter "G" (+1) or not the letter "G" (−1), or whether given image contains a face (+1) or does not contain a face (−1).

Joint Distribution

- Joint distribution $p_{X,Y}(x, y)$
- Future data is coming from some unknown source joint distribution $p_{X,Y}$ over input objects and their corresponding labels, which we write as the joint distribution $p_{X,Y}$, where $X \in \mathcal{X}, Y \in \mathcal{Y}$
- Example: Character recognition source distribution would assign much more probability to (“image containing a circular shape”, “O”) than to (“image containing a circular shape”, “T”).

Conditional Probability

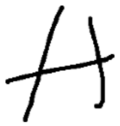
- Conditional distribution $p_{Y|X}(y \mid x)$

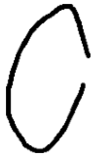
- We can define course joint distribution as really having two components

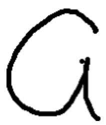
$$p_{XY}(x, y) = p_{Y|X}(y \mid x) \cdot p_X(x)$$

- where $p_{Y|X}(y \mid x)$ is the conditional probability of the label random variable y given the appearance random variable and $p_X(x)$ is marginal probability of the input image.

- Example: In character recognition we may have

$p_{Y|X}(Y = \text{"A"} \mid X =$  $) = 0.9$

$p_{Y|X}(Y = \text{"O"} \mid X =$  $) = 0.6$

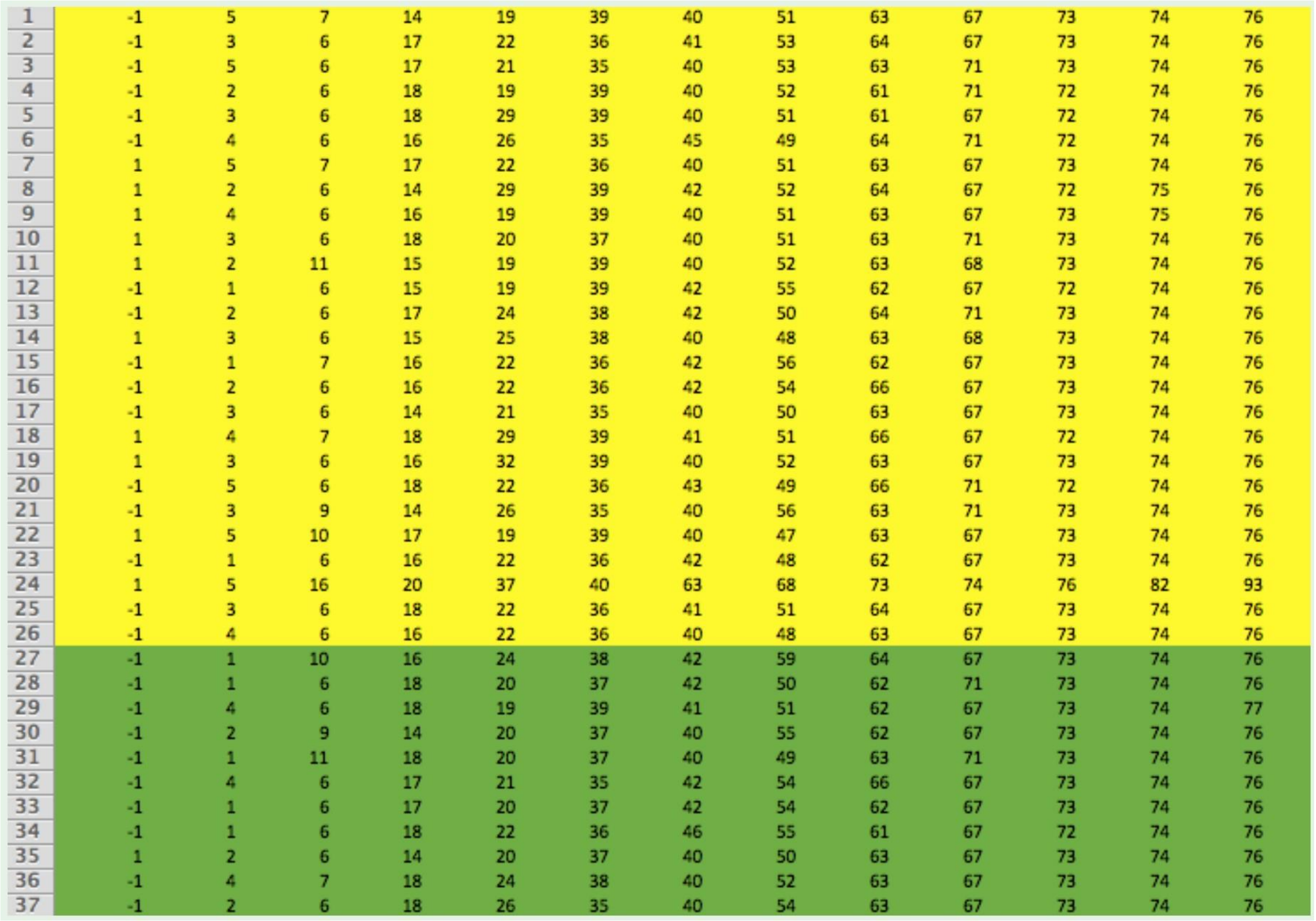
$p_{Y|X}(Y = \text{"a"} \mid X =$  $) = 0.4$

Training Set and Test Set

- A training set is a set data used to discover potentially predictive relationship.

- A test set is a set of data used to assess the strength and utility of a predictive relationship

- Random split



1	-1	5	7	14	19	39	40	51	63	67	73	74	76
2	-1	3	6	17	22	36	41	53	64	67	73	74	76
3	-1	5	6	17	21	35	40	53	63	71	73	74	76
4	-1	2	6	18	19	39	40	52	61	71	72	74	76
5	-1	3	6	18	29	39	40	51	61	67	72	74	76
6	-1	4	6	16	26	35	45	49	64	71	72	74	76
7	1	5	7	17	22	36	40	51	63	67	73	74	76
8	1	2	6	14	29	39	42	52	64	67	72	75	76
9	1	4	6	16	19	39	40	51	63	67	73	75	76
10	1	3	6	18	20	37	40	51	63	71	73	74	76
11	1	2	11	15	19	39	40	52	63	68	73	74	76
12	-1	1	6	15	19	39	42	55	62	67	72	74	76
13	-1	2	6	17	24	38	42	50	64	71	73	74	76
14	1	3	6	15	25	38	40	48	63	68	73	74	76
15	-1	1	7	16	22	36	42	56	62	67	73	74	76
16	-1	2	6	16	22	36	42	54	66	67	73	74	76
17	-1	3	6	14	21	35	40	50	63	67	73	74	76
18	1	4	7	18	29	39	41	51	66	67	72	74	76
19	1	3	6	16	32	39	40	52	63	67	73	74	76
20	-1	5	6	18	22	36	43	49	66	71	72	74	76
21	-1	3	9	14	26	35	40	56	63	71	73	74	76
22	1	5	10	17	19	39	40	47	63	67	73	74	76
23	-1	1	6	16	22	36	42	48	62	67	73	74	76
24	1	5	16	20	37	40	63	68	73	74	76	82	93
25	-1	3	6	18	22	36	41	51	64	67	73	74	76
26	-1	4	6	16	22	36	40	48	63	67	73	74	76
27	-1	1	10	16	24	38	42	59	64	67	73	74	76
28	-1	1	6	18	20	37	42	50	62	71	73	74	76
29	-1	4	6	18	19	39	41	51	62	67	73	74	77
30	-1	2	9	14	20	37	40	55	62	67	73	74	76
31	-1	1	11	18	20	37	40	49	63	71	73	74	76
32	-1	4	6	17	21	35	42	54	66	67	73	74	76
33	-1	1	6	17	20	37	42	54	62	67	73	74	76
34	-1	1	6	18	22	36	46	55	61	67	72	74	76
35	1	2	6	14	20	37	40	50	63	67	73	74	76
36	-1	4	7	18	24	38	40	52	63	67	73	74	76
37	-1	2	6	18	26	35	40	54	63	67	73	74	76

Hypothesis and loss (cost/risk) function

- Hypothesis : A hypothesis (a predictor) h is a function from $\mathcal{X}$ to $\mathcal{Y}$, $h : \mathcal{X} \to \mathcal{Y}$
- Loss function $\text{loss}(h(x), y)$: How we can evaluate the performance of h on a given (input, label) pair (x, y)
- Example: If the label $h(x)$ does not match the provided label y, we incur a loss of 1 and if the prediction $h(x)$ does match the provided label y, we incur 0 loss.
- The loss function that represents this measure of performance is called the 0-1 loss and defined as

$$\text{loss}(h(x), y) = \begin{cases} 1 & \text{if } h(x) \neq y \\ 0 & \text{if } h(x) = y \end{cases}$$

Hypothesis class/set/space

- Define $\mathcal{H}$ as a set of predictors, written as $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$.
- Example: Consider binary labels, so the label set is $\mathcal{Y} = \{+1, -1\}$. In addition, we may consider $\mathcal{X} = \mathbb{R}^2$
- Some specific examples of hypothesis classes:
  $\mathcal{H} = \{\text{sign}(w^T x + b) \mid w \in \mathbb{R}^2, b \in \mathbb{R}\}$
  $\mathcal{H} = \{\sum_{i=1}^{2} x_i \leq \theta \mid \theta \in \mathbb{R}_+\}$
- In a learning problem, we restrict hypotheses in a certain class.

Expected risk/loss/cost

- Expected Risk $R[h]$

- How well we expect to do (on average) over the entire (admittedly known) source joint distribution $p_{X,Y}(x,y)$?

- The expected risk $R[h]$ of a hypothesis h on that distribution, measures the performance of this hypothesis by evaluating its expected loss over pairs (x,y) drawn from the distribution

$$\boxed{R[h] = \mathbb{E}_{(X,Y)\sim p_{X,Y}}[\text{loss}(h(x),y)] = \sum_{X,Y} p(x,y)\text{loss}(h(x),y)}$$

- Note that the randomness comes from the data...

- Other terms with the same meaning are *expected loss, generalization error, or source-distribution risk*

Additional property of expected risk

- A predictor h is "good" on a particular source joint distribution if it has low risk $R[h]$ on that distribution

- The expected risk $R[h]$ is the probability that the predictor h will incorrectly predict the label for any pair (x, y) drawn at random from the source joint distribution:

$$R[h] = \mathbb{E}_{(X,Y)\sim p_{X,Y}}[\mathrm{loss}(h(x), y)] = \mathbb{P}_{(X,Y)\sim p_{X,Y}}\{h(X) \neq Y\}$$

- This equivalence between the risk and the probability of incorrect label prediction holds only for this 0,1 loss

- We will be assuming that the source joint distribution is fixed

The Learning Process (Ideal Case)

Learning from a function from examples! Given

- Domain $\mathcal{X}, \mathcal{Y}$
- The target function f. (unknown)
- $\mathcal{H}$: Hypothesis set; the set of all possible hypotheses
- Extrapolated observed ys over all x
- Final hypothesis (your predictor/model): $g \approx f$
- Ideal case: g is obtained by minimizing $R[h]$

$$g = \arg\min_{h} R[h], \quad h \in \mathcal{H}$$

Expected risk minimizer

$$g = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{(X,Y)\sim p_{X,Y}}[\mathrm{loss}(h(x), y)]$$

- We want to find a predictor that minimizes the expected loss on the true joint distribution, but⋯
- We do not have complete knowledge of the true source joint distribution
- We should choose our predictor to minimize the expected loss on what we do have access to
- We hope that this predictor will do well on the true source joint distribution (However⋯)

Empirical risk minimizer

- The expected loss of a predictor h on a particular observed sample data set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ could also be referred to as the empirical risk

$$\hat{R}_{\mathcal{D}}[h] = \frac{1}{m}\sum_{i=1}^{m}[\text{loss}(h(x_i), y_i)] = \frac{1}{m}\sum_{i=1}^{m}\{h(x_i) \neq y_i\}$$

- Usually, the target predictor is found by

$$\hat{h} = \text{arg}\min_{h} \ \hat{R}_{\mathcal{D}}[h], \qquad h \in \mathcal{H}$$

- Parameterize $h(\,\cdot\,; \theta) \iff \theta$

$$\hat{\theta} = \text{arg}\min_{\theta} \ R_{\mathcal{D}}[\theta], \qquad \theta \in \ \Theta$$

Goal of Machine Learning

- The core of machine learning deals with representation and generalization:
- Representation (Explanation) of data instances and functions evaluated on these instances are part of all machine learning systems
- Generalization (Prediction) is the property that the system will perform well on unseen data instances

Notes:

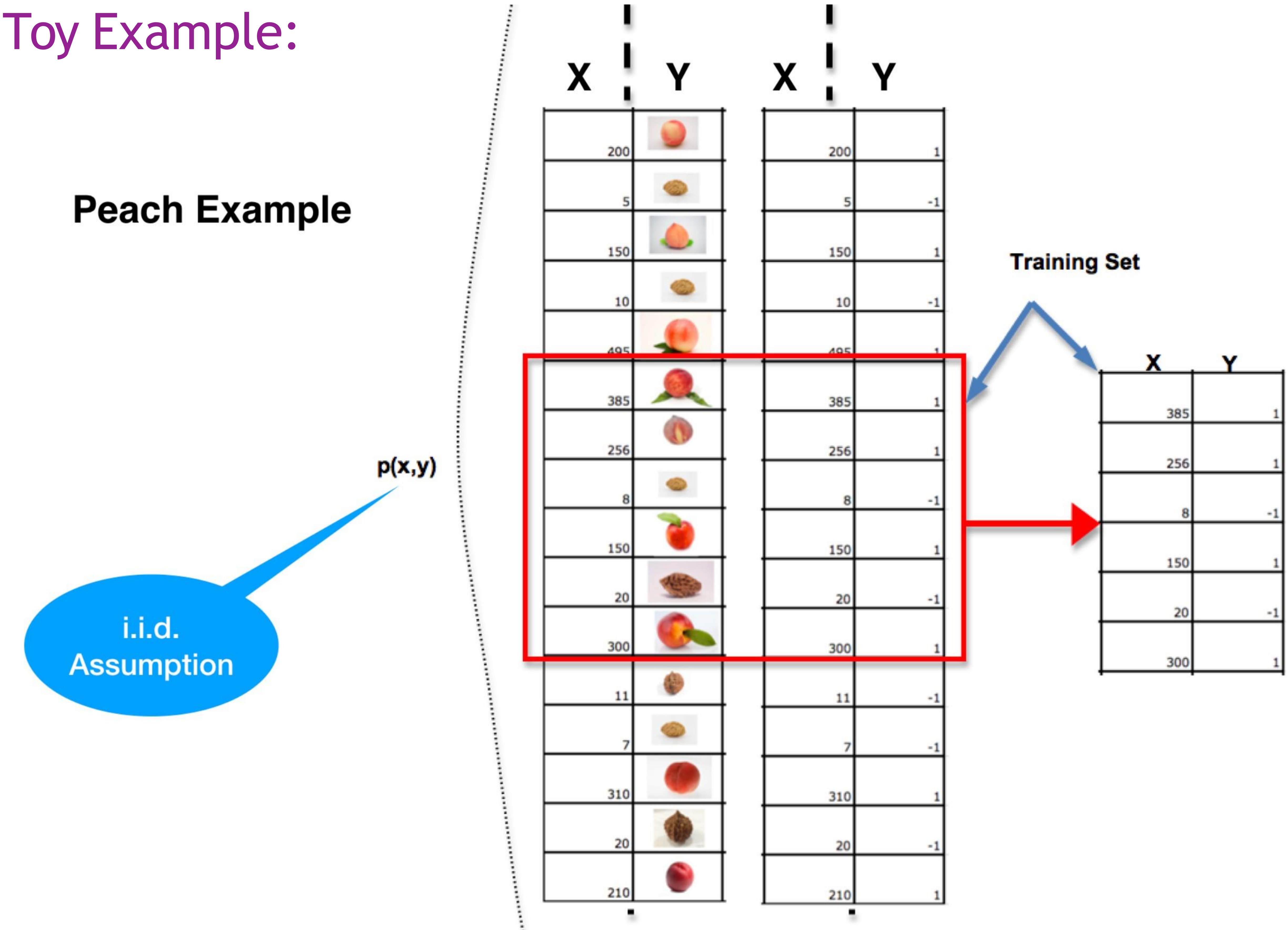
- Are we done with Empirical Risk Minimization?
- Remember the ultimate goal is ***Expected Risk Minimization***
- In **Empirical Risk Minimization**, we use

$$\boxed{\sum_{i=1}^{m} \text{loss}(h(x_i), y_i; \theta) \text{ to approximate } \mathbb{E}[\text{loss}(h(x), y)]}$$

- When is this a good approximation (good generalization)?
- What if it’s not a good approximation (bad generalization)?

Toy Example:

Peach Example



X	Y
200	1
5	-1
150	1
10	-1
495	1
385	1
256	1
8	-1
150	1
20	-1
300	1
11	-1
7	-1
310	1
20	-1
210	1

Training Set

X	Y
385	1
256	1
8	-1
150	1
20	-1
300	1

Question: Why we need i.i.d. throughout this course?

Machine Learning Procedures

Hypothesis (Prediction function): $h_\theta(x) = \text{sign}(x - \theta)$, parameterize by θ

Training set

385	1
256	1
8	-1
150	1
20	-1
300	1

$h_{30}(x) = \text{sign}(x - 30)$

1	0
1	0
-1	0
1	0
-1	0
1	0

$h_{200}(x) = \text{sign}(x - 200)$

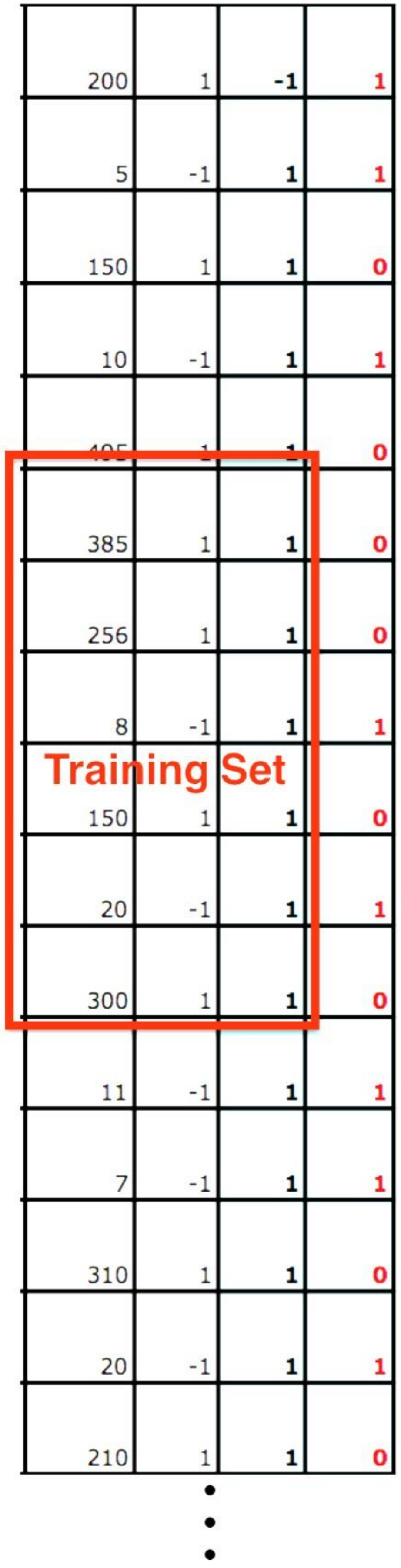
1	0
1	0
-1	0
-1	1
-1	0
1	0

Loss Function: $\text{loss}(h(x), y) = \quad h(x) \neq y$

On average over training set: $\text{loss}(h(x), y) = \frac{1}{m} \sum_{i=1}^{m} (h(x_i) \neq y_i)$
(training error, empirical risk, in-sample error)

$$\text{loss}(h_{30}(x), y) = 0 \quad \text{loss}(h_{200}(x), y) = 1/6$$

Sample Space $p(x, y)$


Training Set

200	1	-1	1
5	-1	1	1
150	1	1	0
10	-1	1	1
405	1	1	0
385	1	1	0
256	1	1	0
8	-1	1	1
150	1	1	0
20	-1	1	1
300	1	1	0
11	-1	1	1
7	-1	1	1
310	1	1	0
20	-1	1	1
210	1	1	0

⋮

$$\theta = 200, \quad h_{200}(x) = \text{sign}(x - 200)$$

Expected loss on the entire sample space:

$$\mathbb{E}[\text{loss}(h(X), Y)] = \int \text{loss}(h(x), y) \cdot p(x, y) \; d(x, y)$$

(generalization error, expected risk, out-of-sample error)

What is our Hypothesis Set?

$$\mathcal{H} = \{h(x; \theta) \mid 0 \leq \theta \leq 500\}$$

What is the best (optimal) θ value?

- The “key (ultimate) goal” in machine learning is to answer this question.
- Obviously, the “best θ” should be the value minimizing $\mathbb{E}[\text{loss}(h(X; \theta), Y)]$.
- This is called *expected risk minimization*

Expected Risk Min v.s. Empirical Risk Min

- Expected Risk Minimization

$$\min_h \ \mathbb{E}[\text{loss}(h(X), Y)] \quad \text{s.t. } h \in \mathcal{H} \quad \rightarrow \quad \text{parameterize...}$$

$$\min_\theta \ \mathbb{E}[\text{loss}(h(X;\theta), Y)] = \min_\theta \int \text{loss}(h(x;\theta), y) \cdot p(x, y) \ d(x, y)$$

However, generally we can’t do this….why?

- Empirical Risk Minimization

$$\min_h \ \frac{1}{m} \sum_{i=1}^{m} [\text{loss}(h(x_i), y_i)] \quad \text{s.t. } h \in \mathcal{H} \quad \rightarrow \quad \text{parameterize...}$$

$$\min_\theta \ \frac{1}{m} \sum_{i=1}^{m} [\text{loss}(h(X;\theta), Y)] \quad \text{s.t. } 0 \leq \theta \leq 500$$