Introduction to Machine Learning and Stochastic Optimization

Robert M. Gower

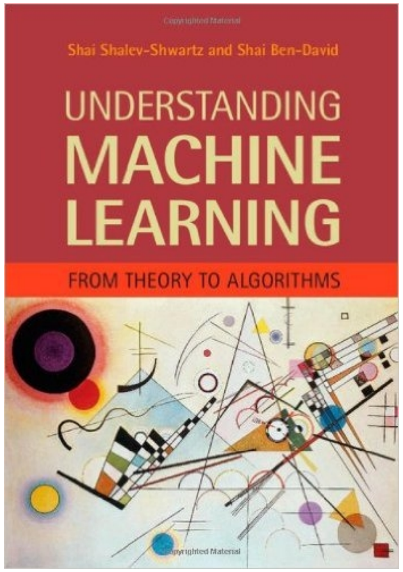
Spring School on Optimization and Data Science,
Novi Saad, March 2017

An Introduction to Supervised Learning

Some References

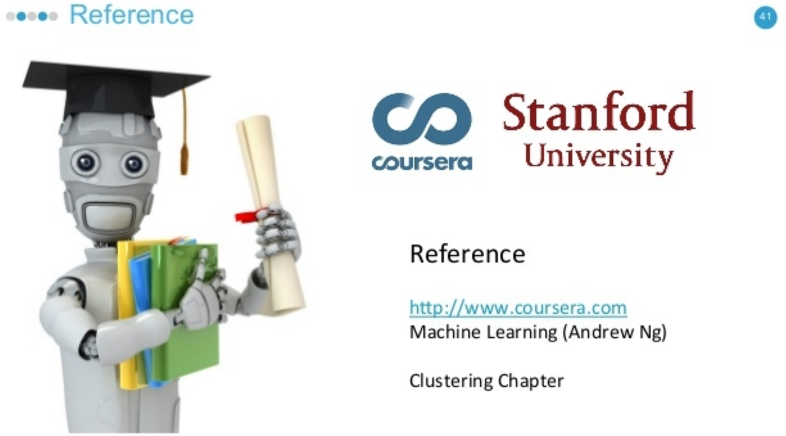
Graduate level

Understanding Machine Learning: From Theory to Algorithms



Undergraduate level

Stanford Machine Learning on Coursera by Andrew Ng



Is There a Cat in the Photo?

Yes

No

Is There a Cat in the Photo?

Yes

Is There a Cat in the Photo?

Yes

Is There a Cat in the Photo?

No

Is There a Cat in the Photo?

Yes

Is There a Cat in the Photo?

Yes

No

x: Input/Feature

y: Output/Target

Find mapping h that assigns the “correct” target to each input

$$h : x \in X \longrightarrow y \in \mathbf{R}$$

Labeled Data

x^1 

$y^1 = 1$

x^2 

$y^2 = 1$

x^3 

$y^3 = -1$

$\cdots$ x^n 

$y^n = 1$

Labeled Data

$x^1 \{$  $y^1 = 1$

$x^2 \{$  $y^2 = 1$

$x^3 \{$  $y^3 = -1$

$\cdots\ x^n \{$  $y^n = 1$

y= -1 means no/false

Labeled Data

x^1 { 
$y^1 = 1$

x^2 { 
$y^2 = 1$

x^3 { 
$y^3 = -1$

$\cdots$ x^n { 
$y^n = 1$

y= -1 means no/false

Learning Algorithm

Labeled Data

x^1 { 
$y^1 = 1$

x^2 { 
$y^2 = 1$

x^3 { 
$y^3 = -1$

$\cdots$ x^n { 
$y^n = 1$

y= *-1* means no/false

Learning Algorithm

$$h : x \in X \to y \in \mathbf{R}$$

Labeled Data

$x^1 \{$  $y^1 = 1$

$x^2 \{$  $y^2 = 1$

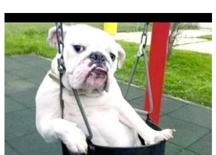
$x^3 \{$  $y^3 = -1$

$\cdots \; x^n \{$  $y^n = 1$

y= -1 means no/false

Learning Algorithm

$$h : x \in X \rightarrow y \in \mathbf{R}$$

$h\Big($  $\Big) \longrightarrow$ **-1**

Example: Linear Regression for Height

Labeled data $x \in \mathbf{R}^2, y \in \mathbf{R}_+$

x_1^1	Sex	Male
x_2^1	Age	30
y^1	Height	1,72 cm

$\cdots$

x_1^n	Sex	Female
x_2^n	Age	70
y^n	Height	1,52 cm

Example: Linear Regression for Height

Labeled data $x \in \mathbf{R}^2, y \in \mathbf{R}_+$

x_1^1 {	Sex	Male
x_2^1 {	Age	30
y^1 {	Height	1,72 cm

$\cdots$

x_1^n {	Sex	Female
x_2^n {	Age	70
y^n {	Height	1,52 cm

Example Hypothesis: Linear Model

$$h_w(x_1, x_2) = w_0 + x_1 w_1 + x_2 w_2 \overset{x_0=1}{=} \langle w, x \rangle$$

Example: Linear Regression for Height

Labeled data $x \in \mathbf{R}^2, y \in \mathbf{R}_+$

x_1^1	Sex	Male	$\dots$	x_1^n	Sex	Female
x_2^1	Age	30		x_2^n	Age	70
y^1	Height	1,72 cm		y^n	Height	1,52 cm

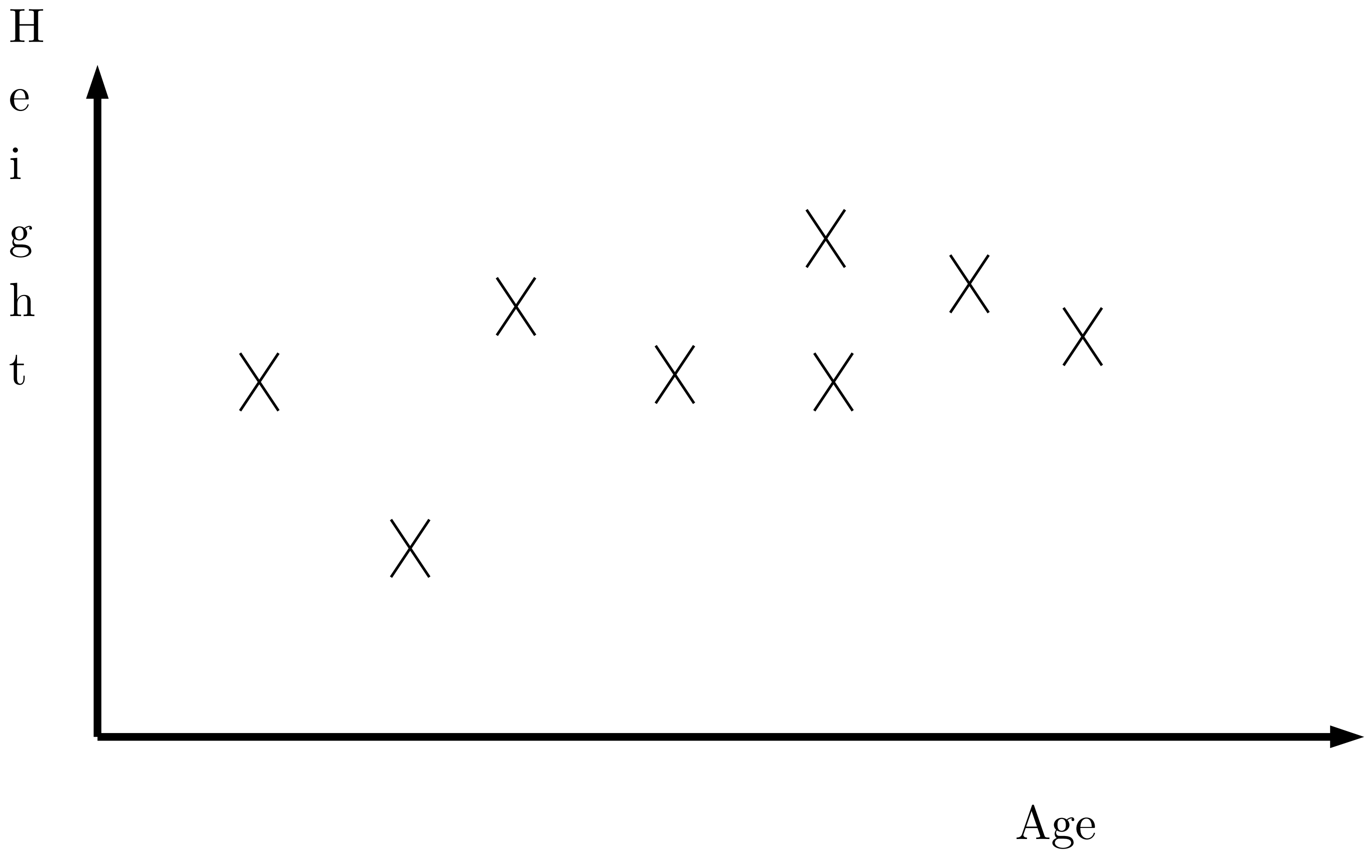
Example Hypothesis: Linear Model

$$h_w(x_1, x_2) = w_0 + x_1 w_1 + x_2 w_2 \overset{x_0=1}{=} \langle w, x \rangle$$

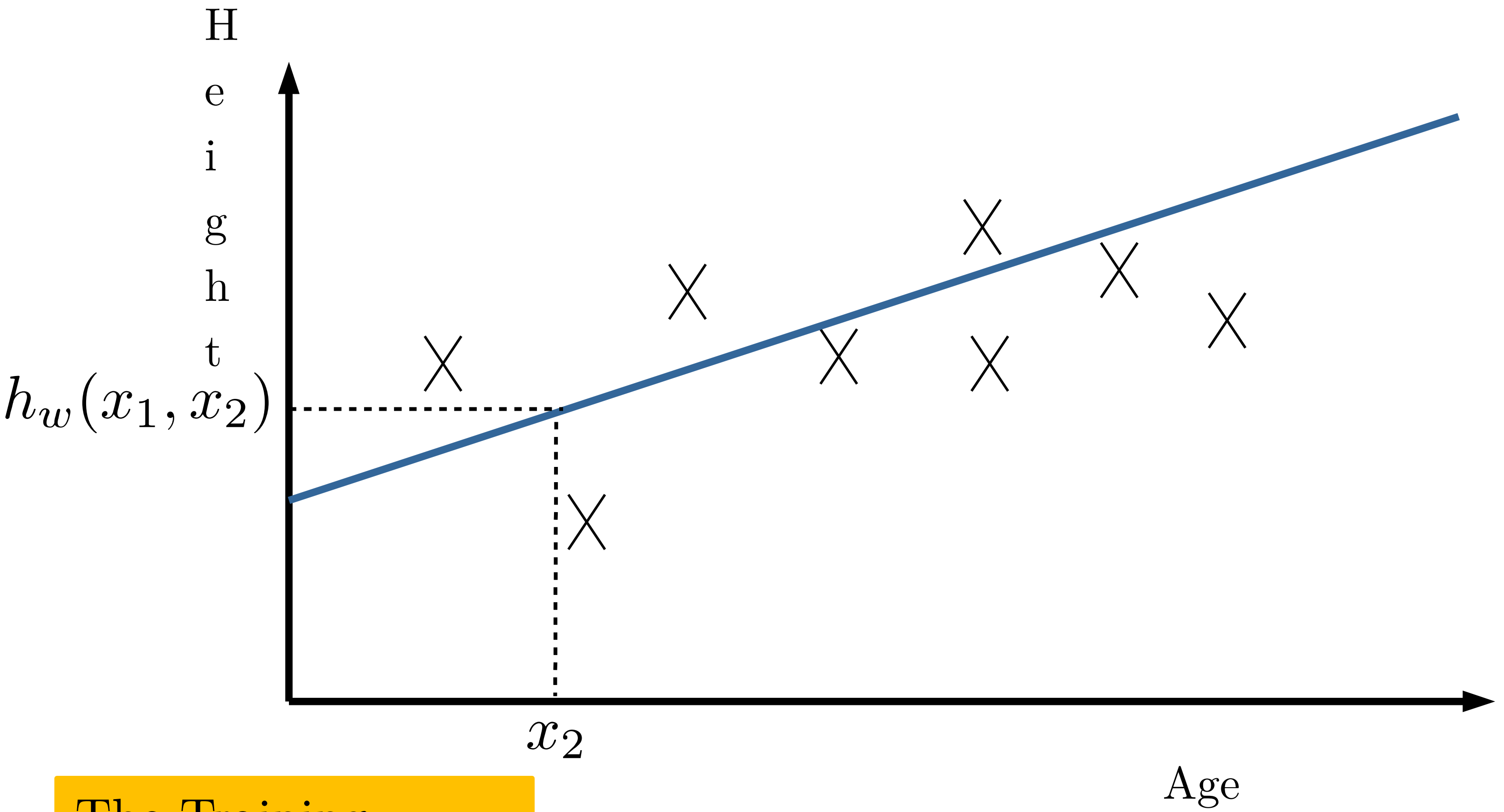
Example Training Problem:

$$\min_{w \in \mathbf{R}^3} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x_1^i, x_2^i) - y^i \right)^2$$

Linear Regression for Height



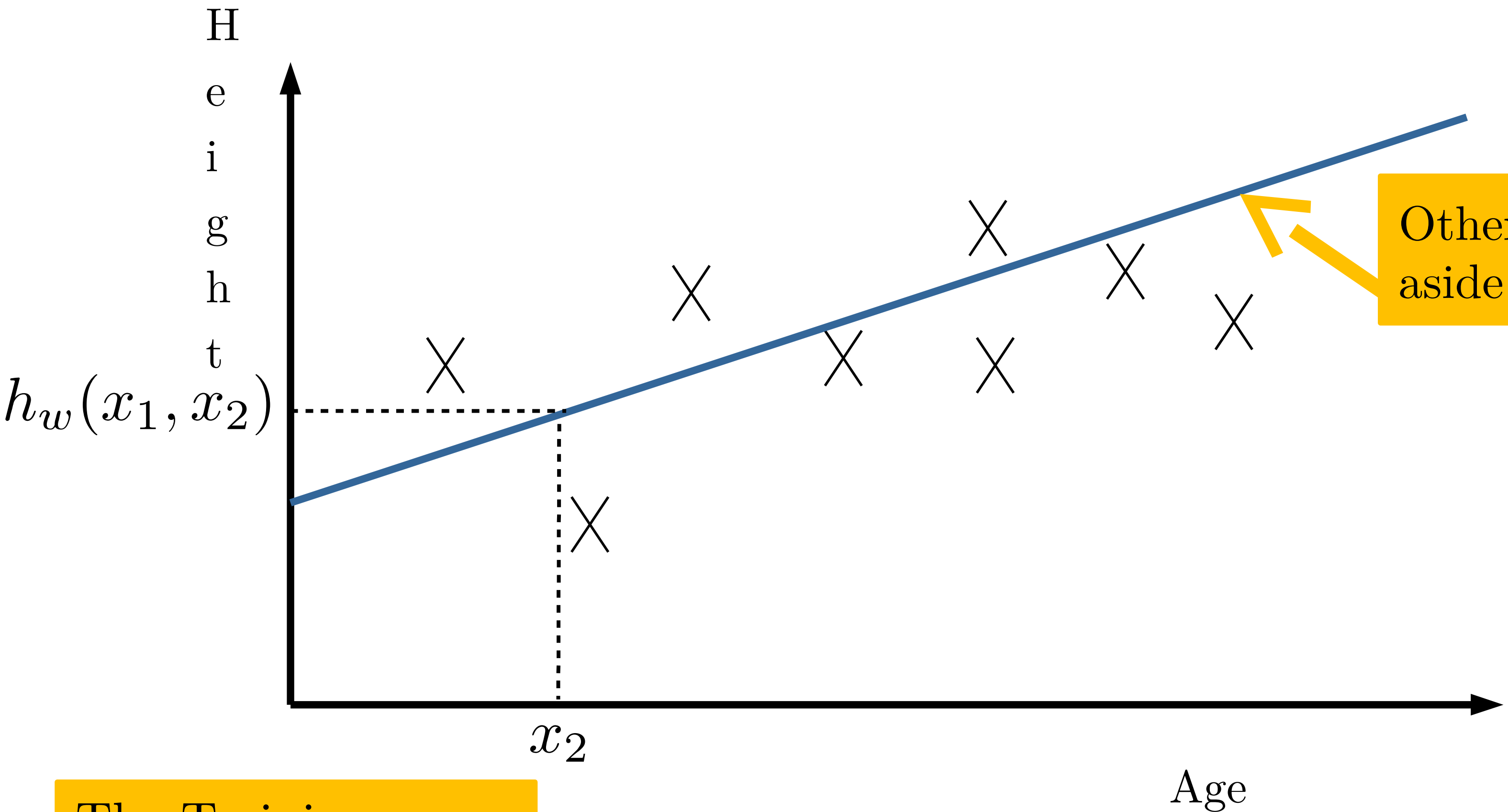
Linear Regression for Height



The Training Algorithm

$$\min_{w \in \mathbf{R}^3} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x_1^i, x_2^i) - y^i\right)^2$$

Linear Regression for Height



Other options aside from linear?

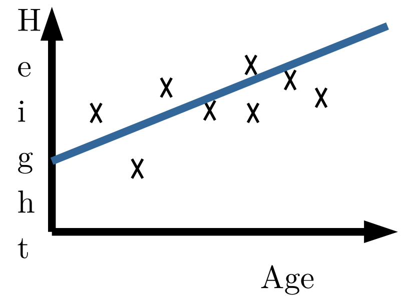
The Training Algorithm

$$\min_{w \in \mathbf{R}^3} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x_1^i, x_2^i) - y^i\right)^2$$

Parametrizing the Hypothesis

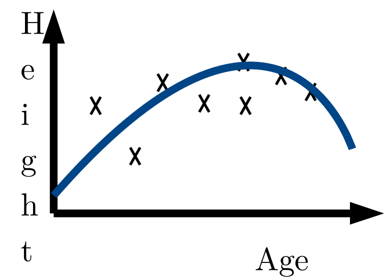
Linear:

$$h_w(x) = \sum_{i=0}^{d} w_i x_i$$

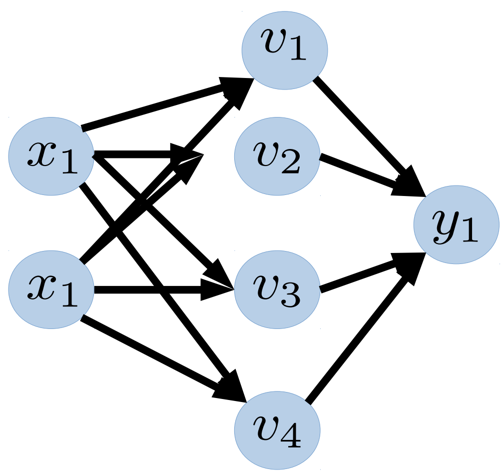


Polinomial:

$$h_w(x) = \sum_{i,j=0}^{d} w_{ij} x_i x_j$$



Neural Net:



exe :

$$v_1 = \mathrm{sign}(w_{11}x_1 + w_{12}x_2)$$

$$v_4 = 1/\left(1 + \exp(w_{41}x_1 + w_{42}x_2)\right)$$

Loss Functions

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i\right)^2$$

Why a Squared Loss?

Loss Functions

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i\right)^2$$

Why a Squared Loss?

Let $y_h := h_w(x)$

Loss Functions

$$\begin{array}{rccc} \ell : & \mathbf{R} \times \mathbf{R} & \to & \mathbf{R}_+ \\ & (y_h, y) & \to & \ell(y_h, y) \end{array}$$

The Training Problem

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)$$

Loss Functions

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i \right)^2$$

Why a Squared Loss?

Let $y_h := h_w(x)$

Loss Functions

$$\begin{array}{rccc} \ell : & \mathbf{R} \times \mathbf{R} & \rightarrow & \mathbf{R}_+ \\ & (y_h, y) & \rightarrow & \ell(y_h, y) \end{array}$$

Typically a convex function

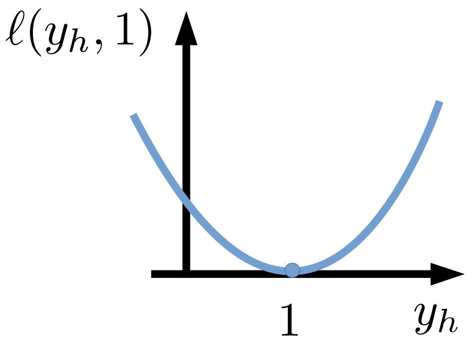
The Training Problem

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell \left(h_w(x^i), y^i \right)$$

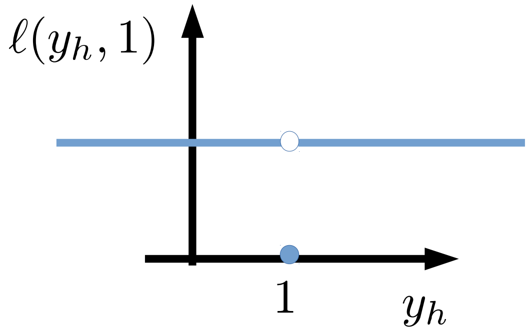
Choosing the Loss Function

Let $y_h := h_w(x)$

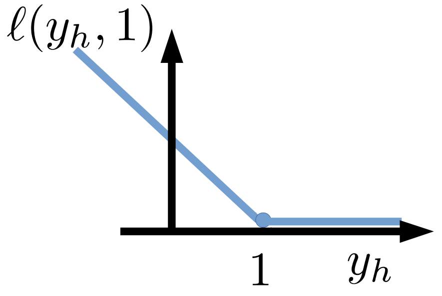
Quadratic Loss $\ell(y_h, y) = (y_h - y)^2$



Binary Loss $\ell(y_h, y) = \begin{cases} 0 & \text{if } y_h = y \\ 1 & \text{if } y_h \neq y \end{cases}$



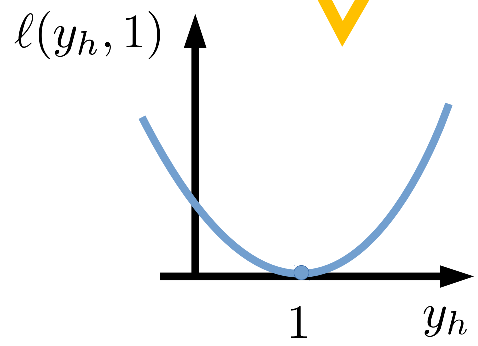
Hinge Loss $\ell(y_h, y) = \max\{0, 1 - y_h y\}$



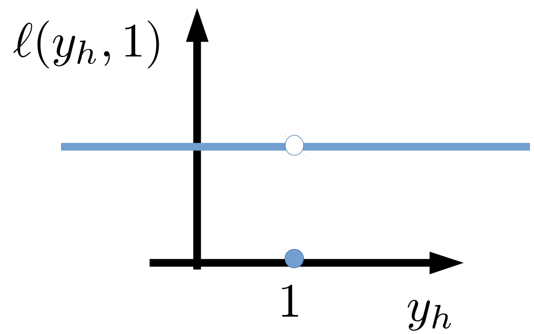
Choosing the Loss Function

y=*1* in all figures

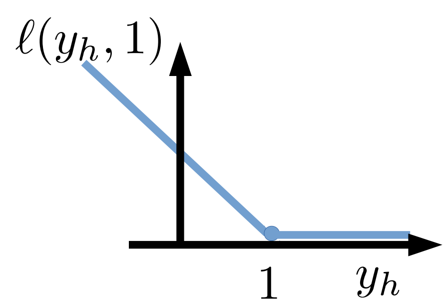
Let $y_h := h_w(x)$



Quadratic Loss $\ell(y_h, y) = (y_h - y)^2$



Binary Loss $\ell(y_h, y) = \begin{cases} 0 & \text{if } y_h = y \\ 1 & \text{if } y_h \neq y \end{cases}$



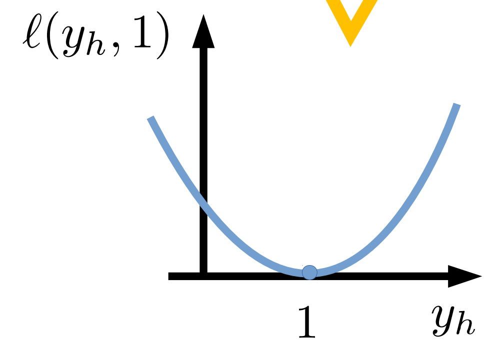
Hinge Loss $\ell(y_h, y) = \max\{0, 1 - y_h y\}$

Choosing the Loss Function

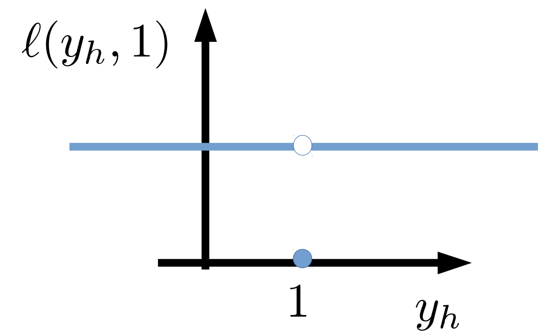
y=1 in all figures

Let $y_h := h_w(x)$

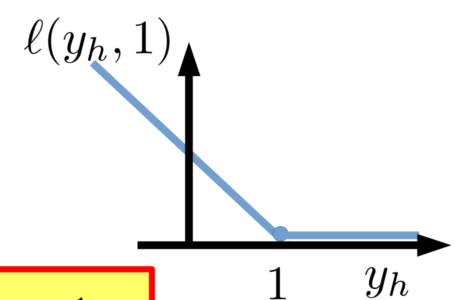
Quadratic Loss $\quad \ell(y_h, y) = (y_h - y)^2$



Binary Loss $\quad \ell(y_h, y) = \begin{cases} 0 & \text{if } y_h = y \\ 1 & \text{if } y_h \neq y \end{cases}$



Hinge Loss $\quad \ell(y_h, y) = \max\{0, 1 - y_h y\}$



EXE: Plot the binary and hinge loss function in when $y = -1$

Loss Functions

Is a notion of Loss enough?

What happens when we do not have enough data?

Loss Functions

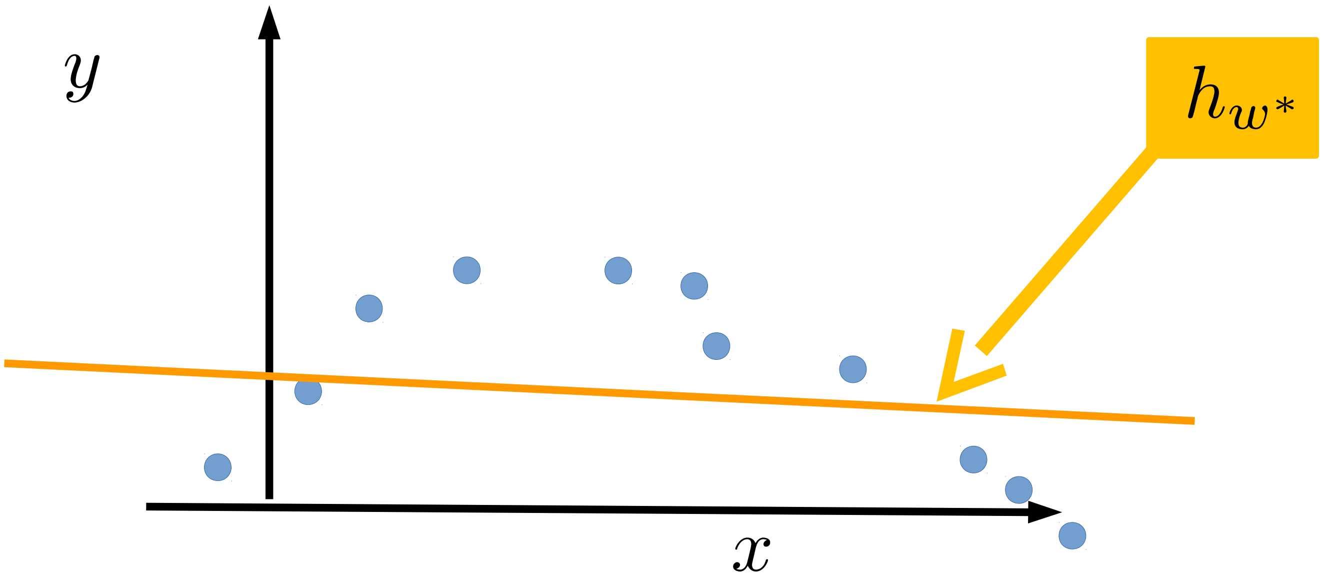
The Training Problem

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)$$

Is a notion of Loss enough?

What happens when we do not have enough data?

Overfitting and Model Complexity

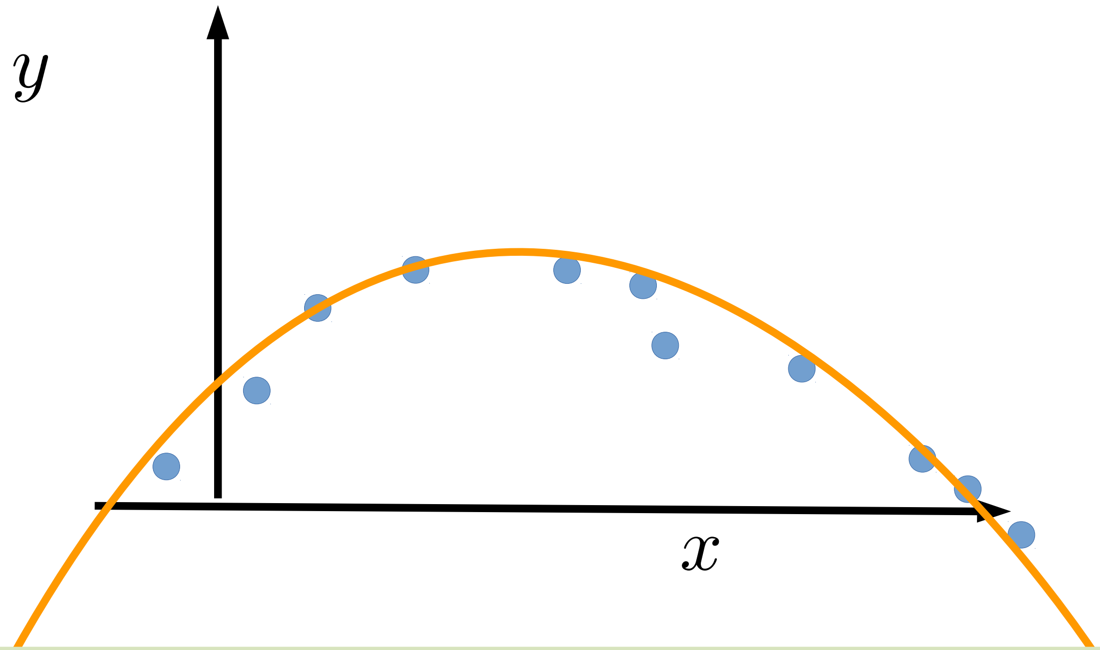


Fitting 1st order polynomial

$$h_w = \langle w, x \rangle$$

$$w^* = \arg\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i\right)^2$$

Overfitting and Model Complexity

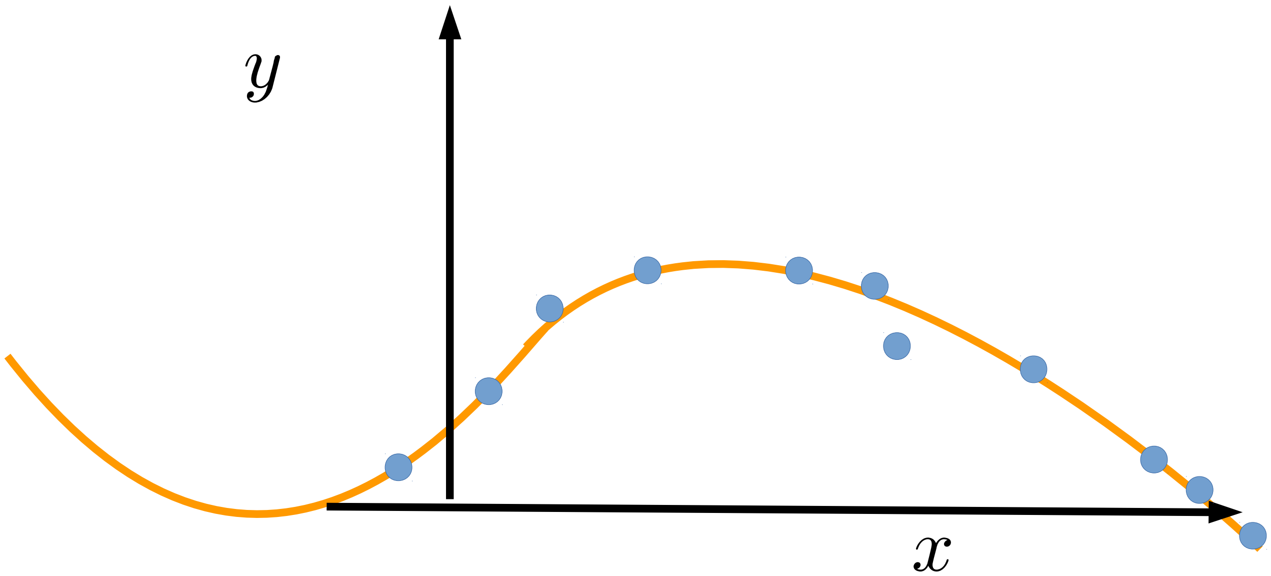


Fitting 1st order polynomial

$$h_w = w_0 + w_1 x + w_2 x^2$$

$$w^* = \arg \min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i\right)^2$$

Overfitting and Model Complexity

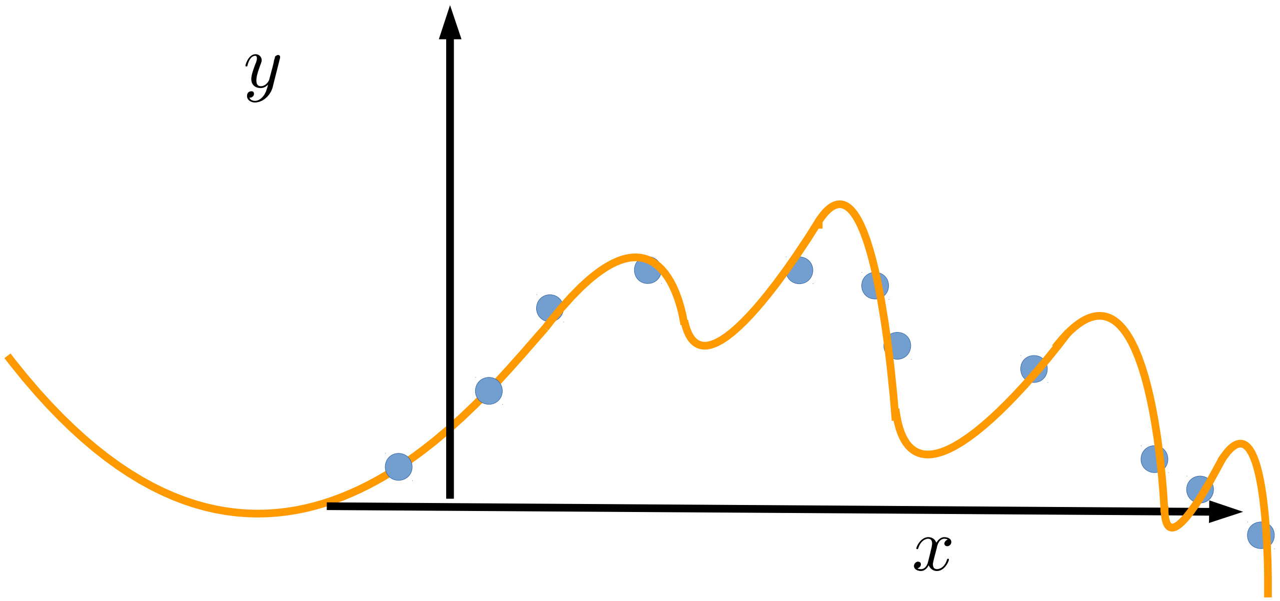


Fitting 3rd order polynomial

$$h_w = \sum_{i=0}^{3} w_i x^i$$

$$w^* = \arg \min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i \right)^2$$

Overfitting and Model Complexity



Fitting 9th order polynomial

$$h_w = \sum_{i=0}^{9} w_i x^i$$

$$w^* = \arg \min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i\right)^2$$

Regularization

Regularizor Functions

$$\begin{aligned} R : \quad \mathbf{R}^d \quad &\rightarrow \quad \mathbf{R}_+ \\ w \quad &\rightarrow \quad R(w) \end{aligned}$$

General Training Problem

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell \left(h_w(x^i), y^i \right) + \lambda R(w)$$

Regularization

Regularizor Functions

$$\begin{aligned} R: \quad \mathbf{R}^d &\to \mathbf{R}_+ \\ w &\to R(w) \end{aligned}$$

General Training Problem

$$\min_{w \in \mathbf{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)}_{\text{Goodness of fit, fidelity term ...etc}} + \lambda R(w)$$

Regularization

Regularizor Functions

$$\begin{aligned} R: \quad \mathbf{R}^d &\to \mathbf{R}_+ \\ w &\to R(w) \end{aligned}$$

General Training Problem

$$\min_{w \in \mathbf{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)}_{\text{Goodness of fit, fidelity term ...etc}} + \lambda \underbrace{R(w)}_{\text{Penlizes complexity}}$$

Regularization

Regularizor Functions

$$R : \mathbf{R}^d \rightarrow \mathbf{R}_+$$
$$w \rightarrow R(w)$$

Controls tradeoff between fit and complexity

General Training Problem

$$\min_{w \in \mathbf{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell \left(h_w(x^i), y^i \right)}_{\text{Goodness of fit, fidelity term ...etc}} + \lambda \underbrace{R(w)}_{\text{Penlizes complexity}}$$

Goodness of fit, fidelity term ...etc

Penlizes complexity

Regularization

Regularizor Functions

$$R: \quad \mathbf{R}^d \rightarrow \mathbf{R}_+$$
$$\qquad w \rightarrow R(w)$$

Controls tradeoff between fit and complexity

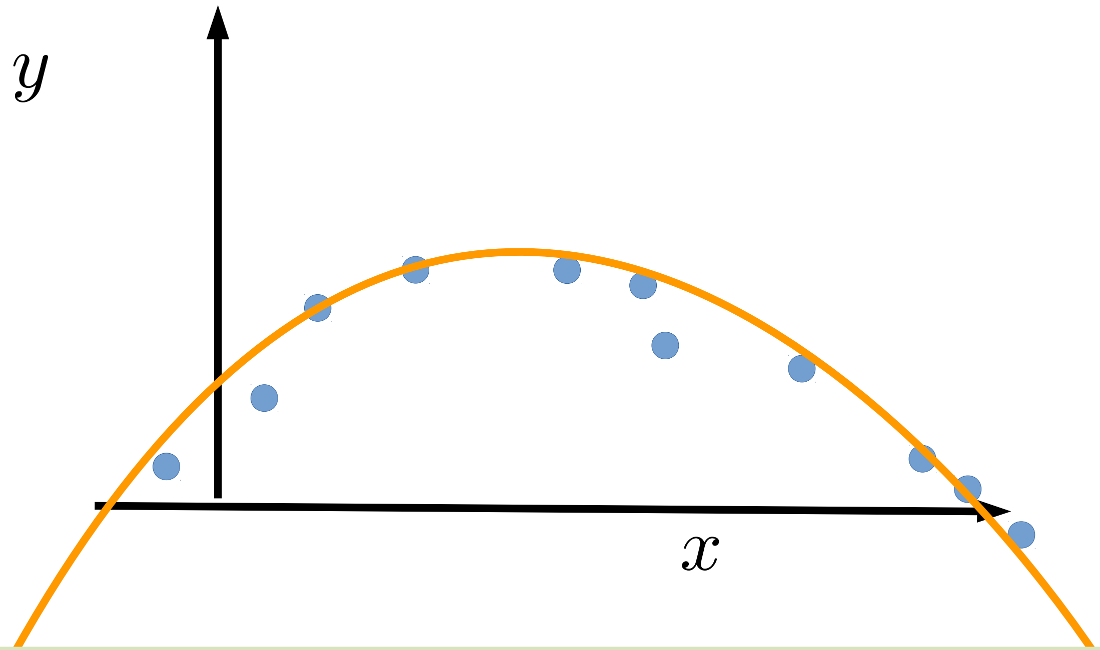
General Training Problem

$$\min_{w \in \mathbf{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)}_{\text{Goodness of fit, fidelity term ...etc}} + \underbrace{\lambda R(w)}_{\text{Penlizes complexity}}$$

Exe:

$$R(w) = ||w||_2^2, \quad ||w||_1, \quad ||w||_p, \quad \text{other norms} \ \ldots$$

Overfitting and Model Complexity

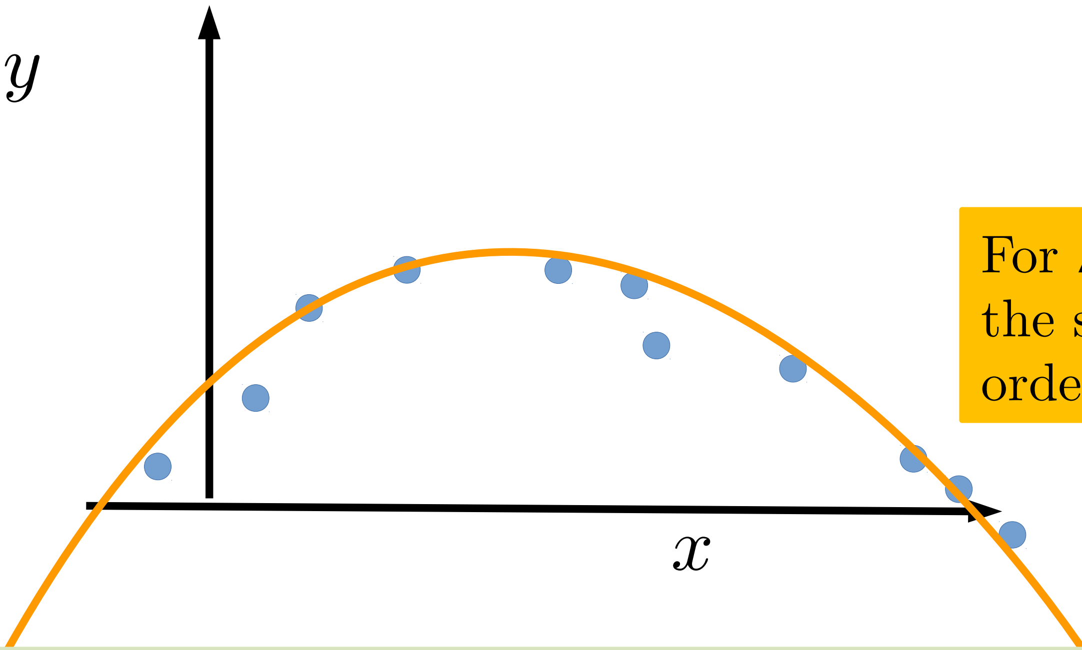


Fitting k^th order polynomial

$$h_w = \sum_{i=0}^{k} w_i x^i$$

$$w^* = \arg\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i\right)^2 + \lambda ||w||_2^2$$

Overfitting and Model Complexity



For λ big enough, the solution is a 2nd order polynomial

Fitting kth order polynomial

$$h_w = \sum_{i=0}^{k} w_i x^i$$

$$w^* = \arg\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i\right)^2 + \lambda ||w||_2^2$$

Exe: Ridge Regression

Linear hypothesis

$$h_w(x) = \langle w, x \rangle$$

L2 regularizor

$$R(w) = ||w||_2^2$$

L2 loss

$$\ell(y_h, y) = (y_h - y)^2$$

Ridge Regression

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y^i - \langle w, x^i \rangle)^2 + \lambda ||w||_2^2$$

Exe: Support Vector Machines

Linear hypothesis

$$h_w(x) = \langle w, x \rangle$$

L2 regularizor

$$R(w) = ||w||_2^2$$

Hinge loss

$$\ell(y_h, y) = \max\{0, 1 - y_h y\}$$

SVM with soft margin

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y^i \langle w, x^i \rangle\} + \lambda ||w||_2^2$$

Exe: Logistic Regression

Linear hypothesis

$$h_w(x) = \langle w, x \rangle$$

+

L2 regularizor

$$R(w) = ||w||_2^2$$

Logistic loss

$$\ell(y_h, y) = \max\{0, 1 - y_h y\}$$

Logistic Regression

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ln(1 + e^{-y^i \langle w, x^i \rangle}) + \lambda ||w||_2^2$$

The Machine Learners Job

(1) Get the labeled data: $(x^1, y^1), \ldots, (x^n, y^n)$

The Machine Learners Job

(1) Get the labeled data: $(x^1, y^1), \ldots, (x^n, y^n)$

(2) Choose a parametrization for hypothesis: $h_w(x)$

The Machine Learners Job

(1) Get the labeled data: $(x^1, y^1), \ldots, (x^n, y^n)$

(2) Choose a parametrization for hypothesis: $h_w(x)$

(3) Choose a loss function: $\ell(h_w(x), y) \geq 0$

The Machine Learners Job

(1) Get the labeled data: $(x^1, y^1), \ldots, (x^n, y^n)$

(2) Choose a parametrization for hypothesis: $h_w(x)$

(3) Choose a loss function: $\ell(h_w(x), y) \geq 0$

(4) Solve the *training problem*:

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right) + \lambda R(w)$$

The Machine Learners Job

(1) Get the labeled data: $(x^1, y^1), \ldots, (x^n, y^n)$

(2) Choose a parametrization for hypothesis: $h_w(x)$

(3) Choose a loss function: $\ell(h_w(x), y) \geq 0$

(4) Solve the *training problem*:

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right) + \lambda R(w)$$

(5) Test and cross-validate. If fail, go back a few steps

The Machine Learners Job

(1) Get the labeled data: $(x^1, y^1), \ldots, (x^n, y^n)$

(2) Choose a parametrization for hypothesis: $h_w(x)$

(3) Choose a loss function: $\ell(h_w(x), y) \geq 0$

(4) Solve the *training problem*:

$$\boxed{\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right) + \lambda R(w)}$$

(5) Test and cross-validate. If fail, go back a few steps