Optimization for Data Science

Introduction into supervised learning

Robert M. Gower
&
Alexandre Gramfort

TELECOM ParisTech

Master 2 Data Science, University Paris Saclay

Core Info

- **Where** : Telecom ParisTech
- **Location** : B312
- **ECTS** : 5 ECTS
- **Volume** : 40h
- **When** : 12 weeks (including one week break for holidays + one week for exam)
- **Online:** All teaching materials on moodle: http://datascience-x-master-paris-saclay.fr/education/
- Students upload their projects / reports via moodle too.
- **All students **must** be registered on moodle.**

Who am I?

Robert M. Gower

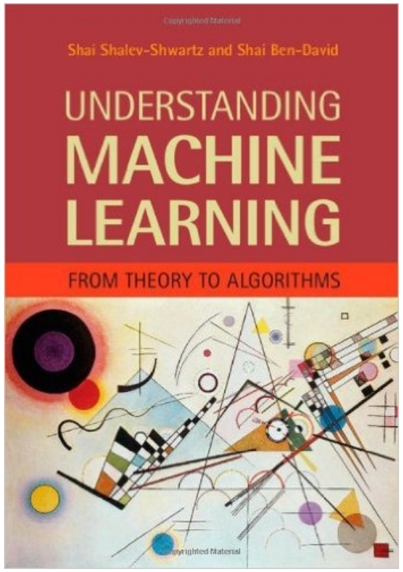
- **Assistant Prof at Telecom**
- **robert.gower@telecom-paristech.fr**
- **https://perso.telecom-paristech.fr/rgower/**
- **Research topics: Stochastic algorithms for optimization, numerical linear algebra, quasi-Newton methods and automatic differentiation (backpropagation).**

An Introduction to Supervised Learning

References classes today

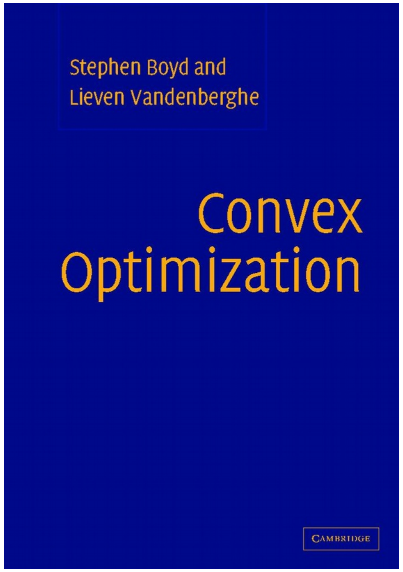
Chapter 2

Understanding Machine Learning: From Theory to Algorithms



Pages 67 to 79

Convex Optimization



Is There a Cat in the Photo?

$\longrightarrow \left\{ \begin{array}{l} \textbf{Yes} \\ \\ \textbf{No} \end{array} \right.$

Is There a Cat in the Photo?

Yes

Is There a Cat in the Photo?

Yes

Is There a Cat in the Photo?

$\longrightarrow$ **No**

Is There a Cat in the Photo?

Yes

Is There a Cat in the Photo?

$\xrightarrow{h}$ $\left\{ \begin{array}{l} \textbf{Yes} \\ \textbf{No} \end{array} \right.$

x: Input/Feature

y: Output/Target

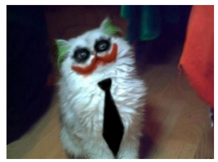
Find mapping h that assigns the "correct" target to each input

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

Labeled Data: The training set

x^1 { 

$y^1 = 1$

x^2 { 

$y^2 = 1$

x^3 { 

$y^3 = -1$

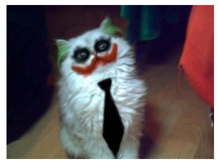
$\cdots$ x^n { 

$y^n = 1$

Labeled Data: The training set

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: The training set

x^1 { 
$y^1 = 1$

x^2 { 
$y^2 = 1$

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

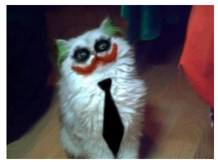
$\cdots$ x^n { 
$y^n = 1$

y= -1 means no/false

Learning Algorithm

Labeled Data: The training set

x^1 { 
$y^1 = 1$

x^2 { 
$y^2 = 1$

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

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

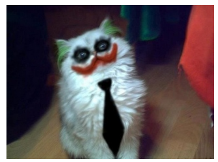
y= -1 means no/false

Learning Algorithm

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

Labeled Data: The training set

x^1 { 
$y^1 = 1$

x^2 { 
$y^2 = 1$

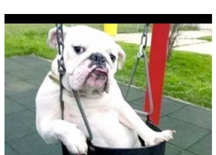
x^3 { 
$y^3 = -1$

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

y= -1 means no/false

Learning Algorithm

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

h () $\longrightarrow$ -1

Example: Linear Regression for Height

Male $= 0$
Female $= 1$

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

x_1^1	Sex	0		x_1^n	Sex	1
x_2^1	Age	30	$\cdots$	x_2^n	Age	70
y^1	Height	1,72 cm		y^n	Height	1,52 cm

Example: Linear Regression for Height

Male $= 0$
Female $= 1$

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

x_1^1	Sex	0
x_2^1	Age	30
y^1	Height	1,72 cm

$\cdots$

x_1^n	Sex	1
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

Male $= 0$
Female $= 1$

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

x_1^1	Sex	0
x_2^1	Age	30
y^1	Height	1,72 cm

$\cdots$

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

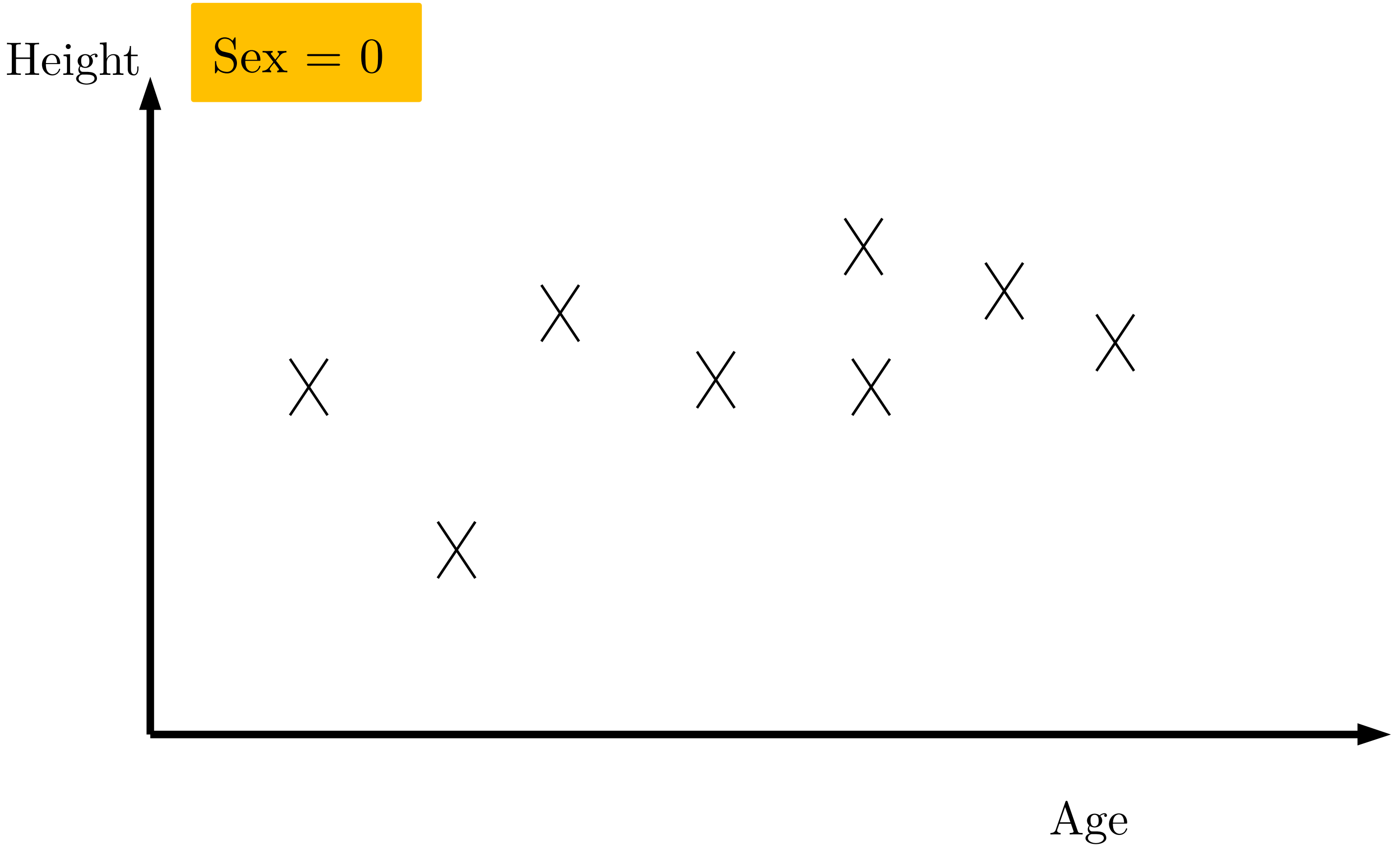
Example Hypothesis: Linear Model

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

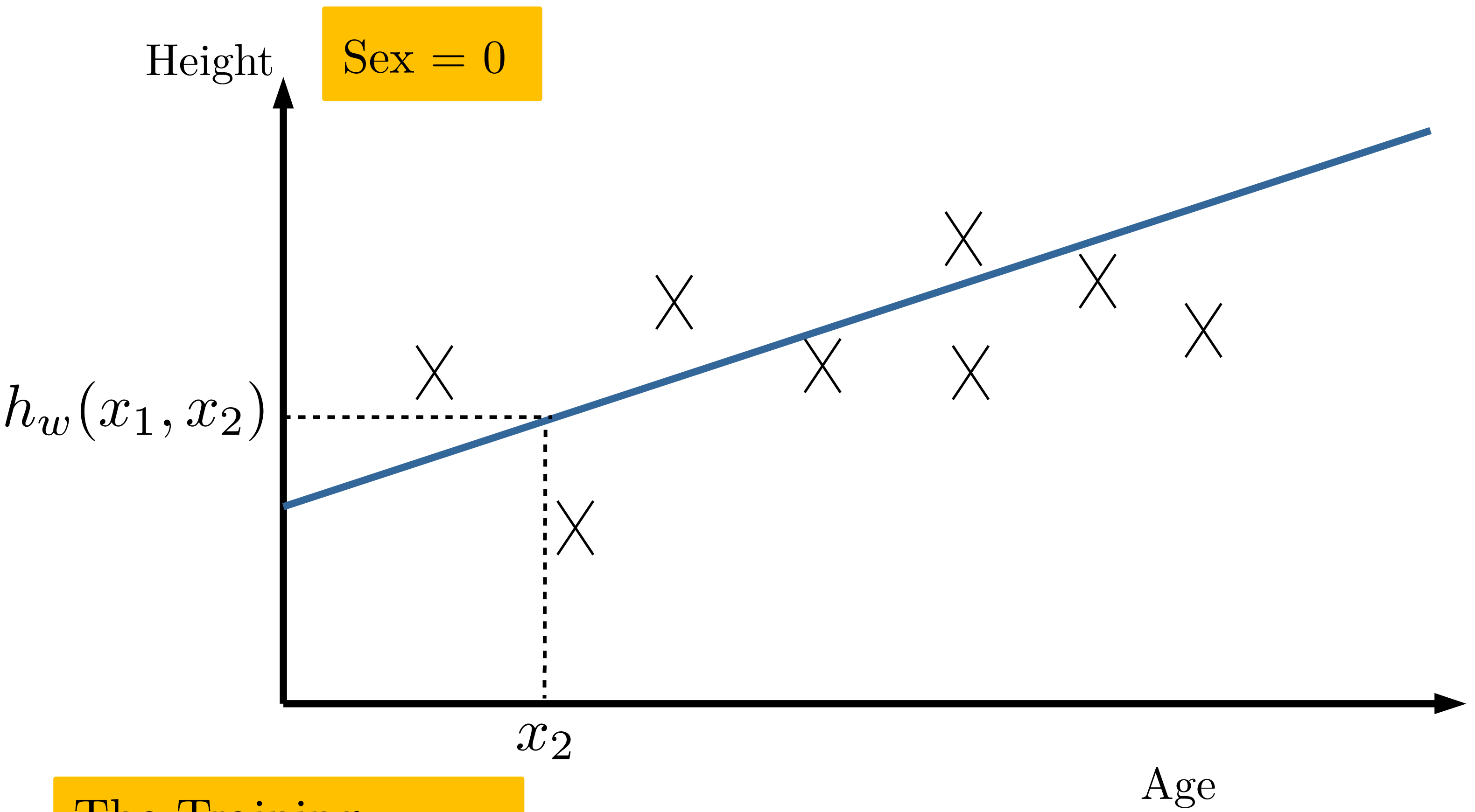
Example Training Problem:

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

Linear Regression for Height



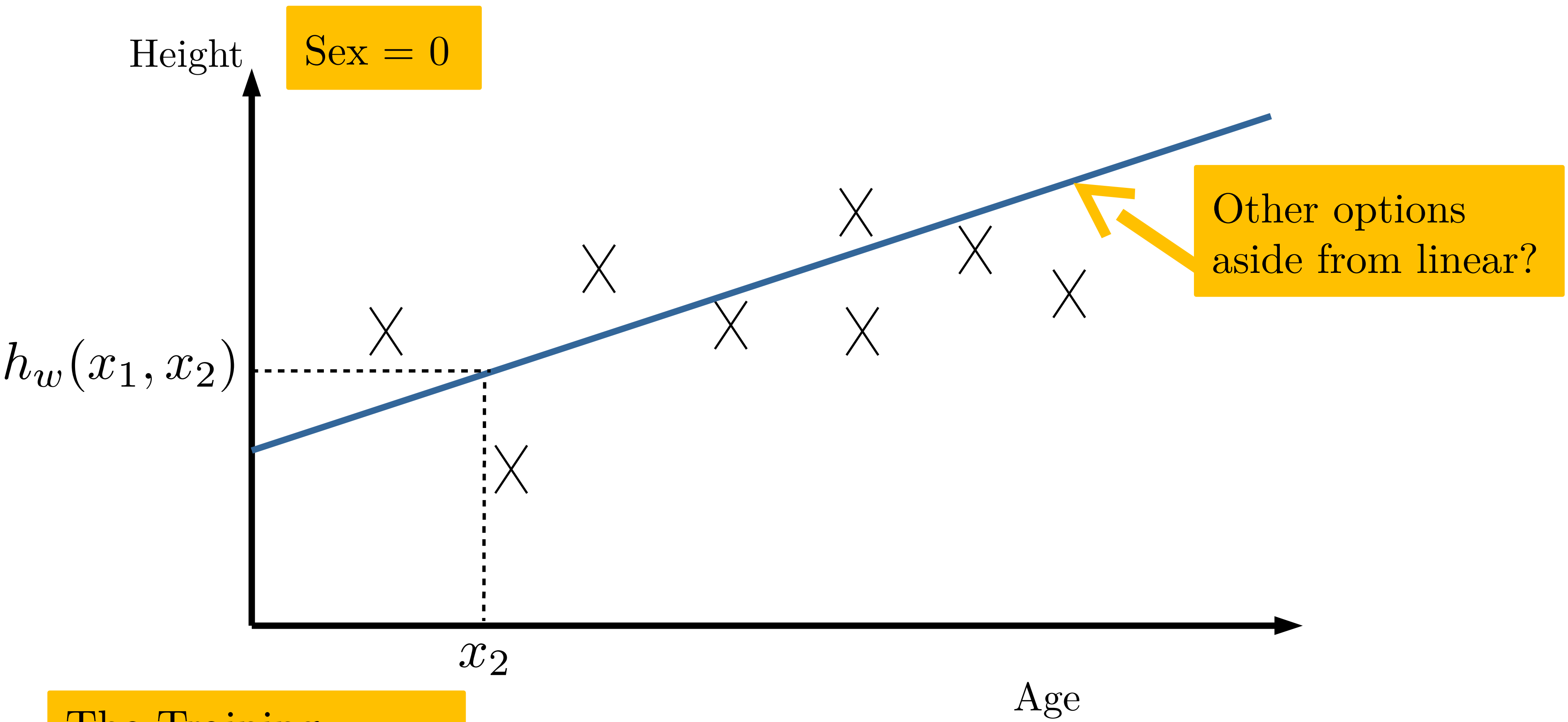
Linear Regression for Height



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



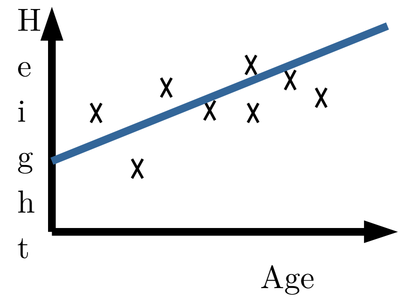
The Training Algorithm

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

Parametrizing the Hypothesis

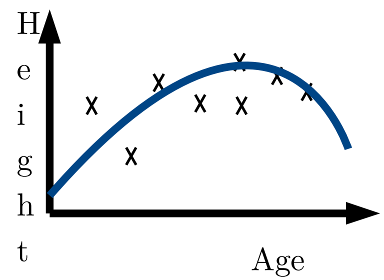
Linear:

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

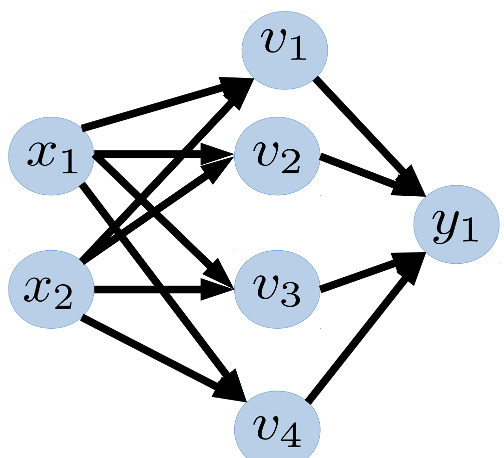


Polinomial:

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



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?

$$\text{Let } y_h := h_w(x)$$

Loss Functions

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

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

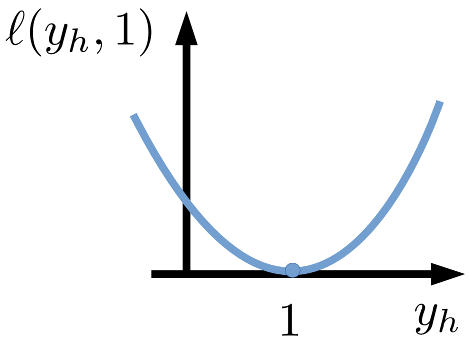
The Training Problem

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

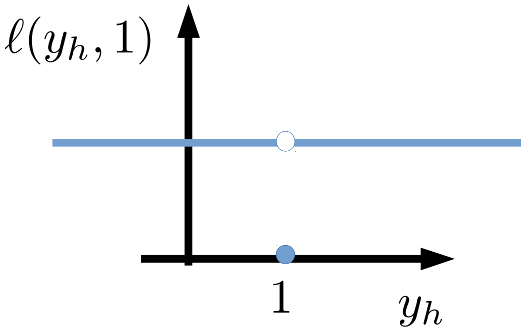
Choosing the Loss Function

Let $y_h := h_w(x)$

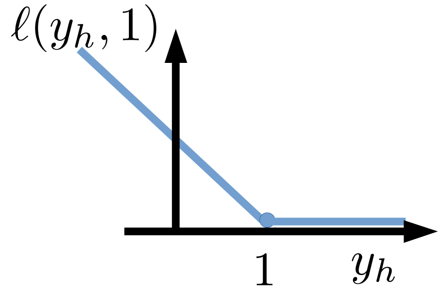
Quadratic Loss $\ell(y_h, y) = (y_h - y)^2$



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

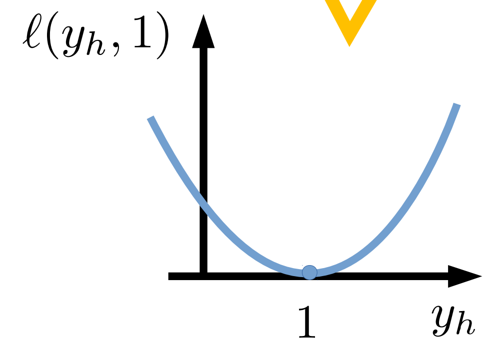


Choosing the Loss Function

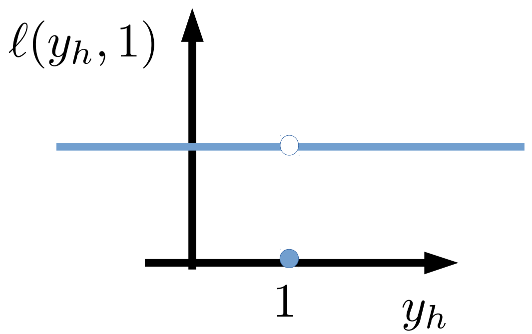
y=1 in all figures

Let $y_h := h_w(x)$

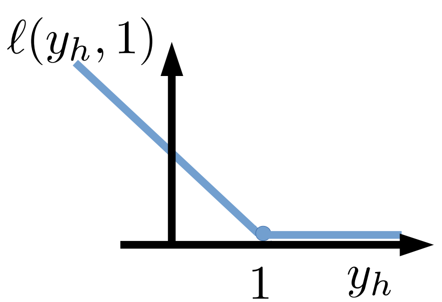
Quadratic Loss $\quad \ell(y_h, y) = (y_h - y)^2$



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

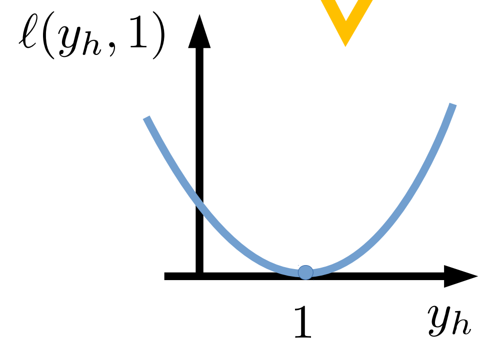


Choosing the Loss Function

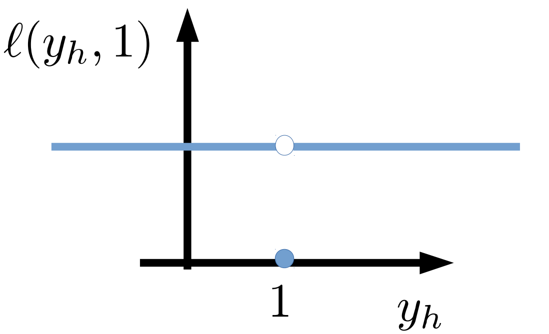
y=*1* in all figures

Let $y_h := h_w(x)$

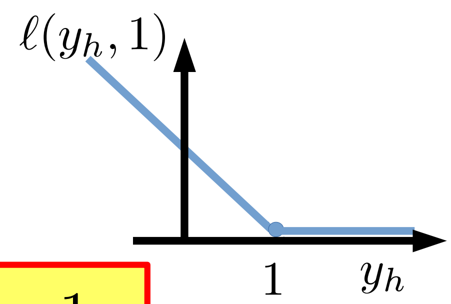
Quadratic Loss $\quad \ell(y_h, y) = (y_h - y)^2$



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

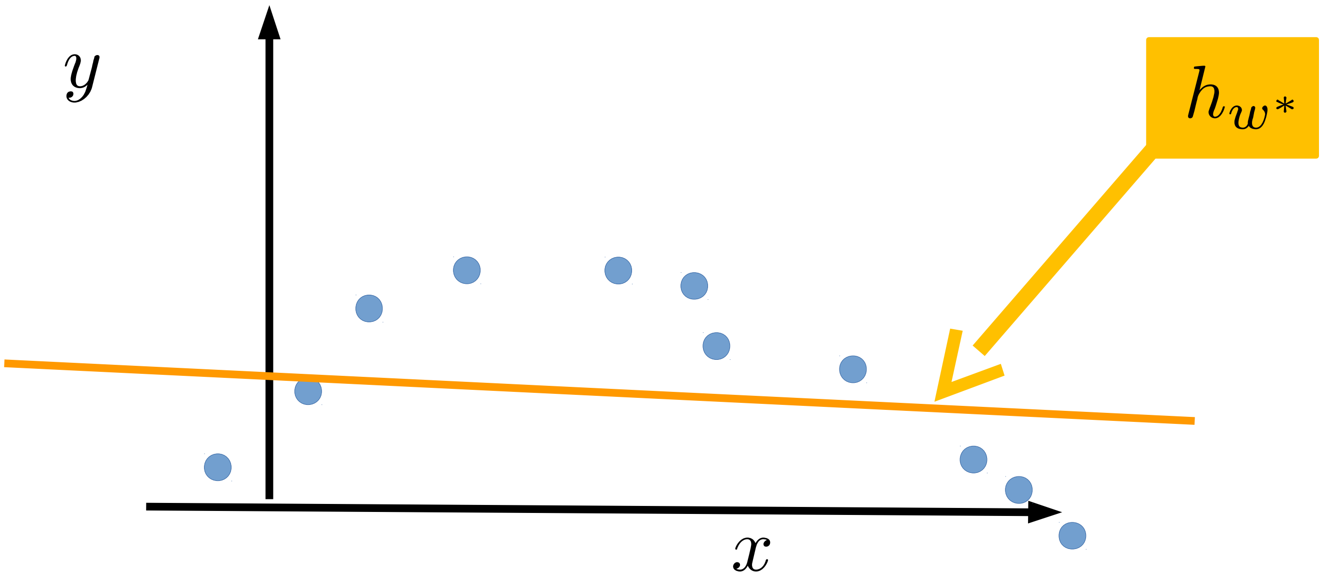
The Training Problem

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

Is a notion of Loss enough?

What happens when we do not have enough data?

Overfitting and Model Complexity

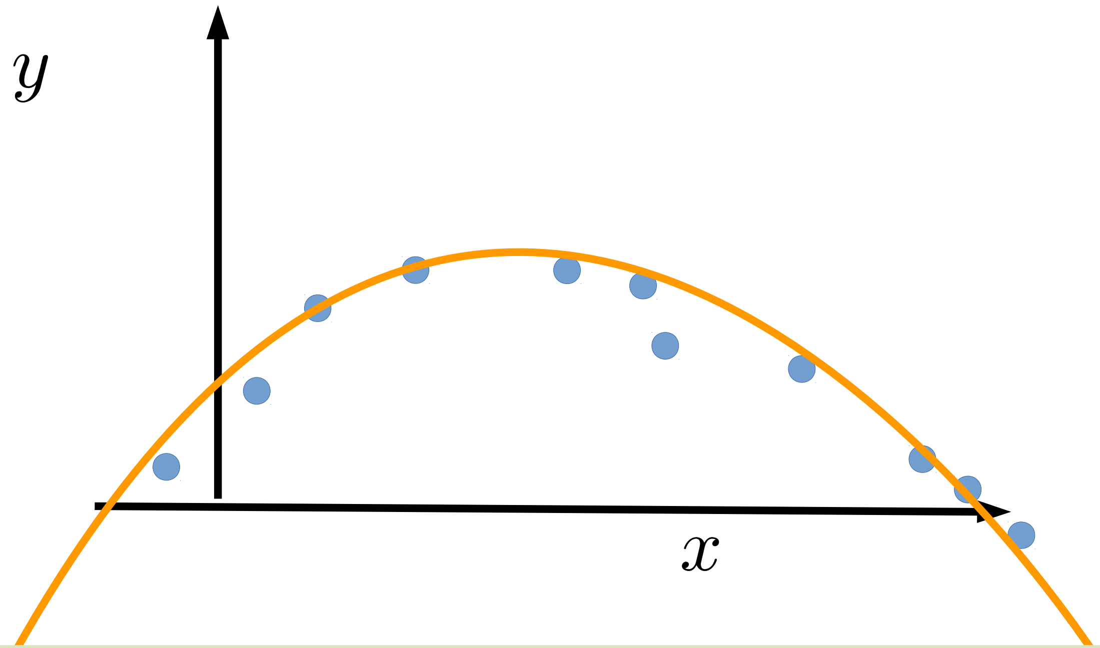


Fitting 1st order polynomial

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

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

Overfitting and Model Complexity

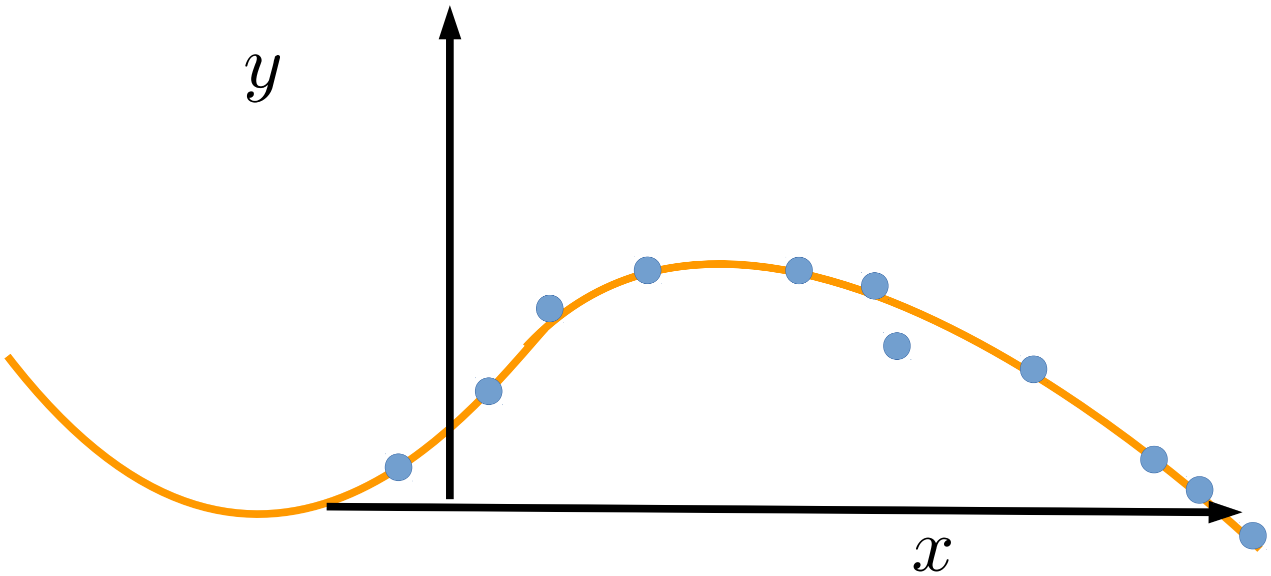


Fitting 1st order polynomial

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

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

Overfitting and Model Complexity

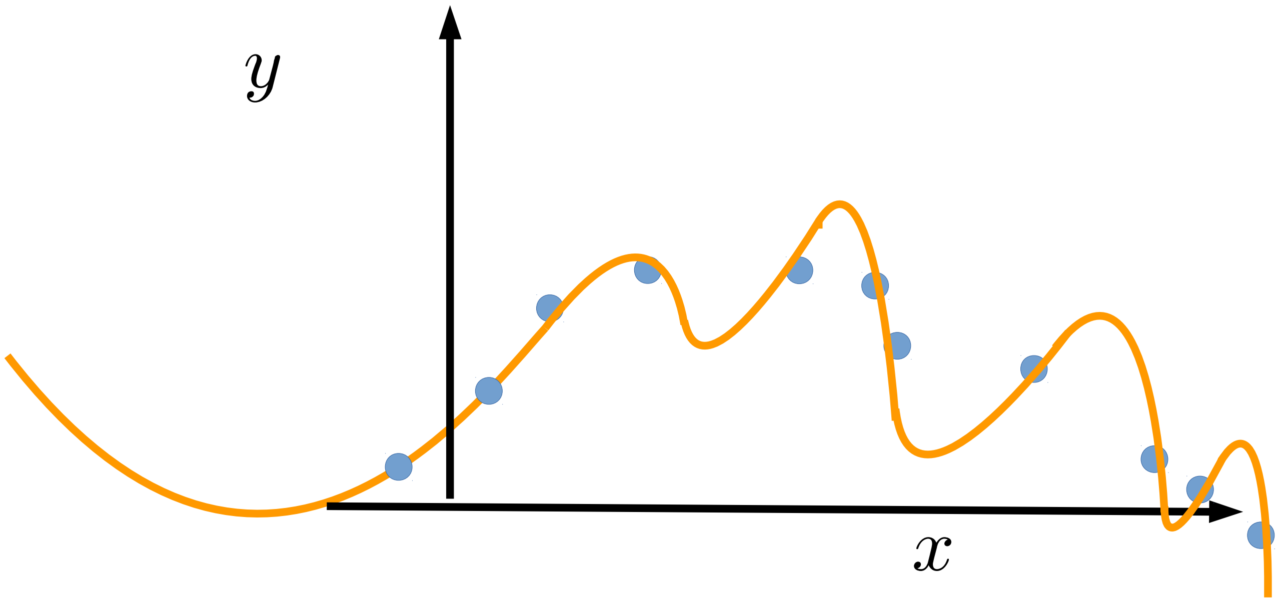


Fitting 3[rd] order polynomial

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

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

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

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

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

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

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

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

General Training Problem

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

Regularization

Regularizor Functions

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

Controls tradeoff between fit and complexity

General Training Problem

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

Regularization

Regularizor Functions

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

Controls tradeoff between fit and complexity

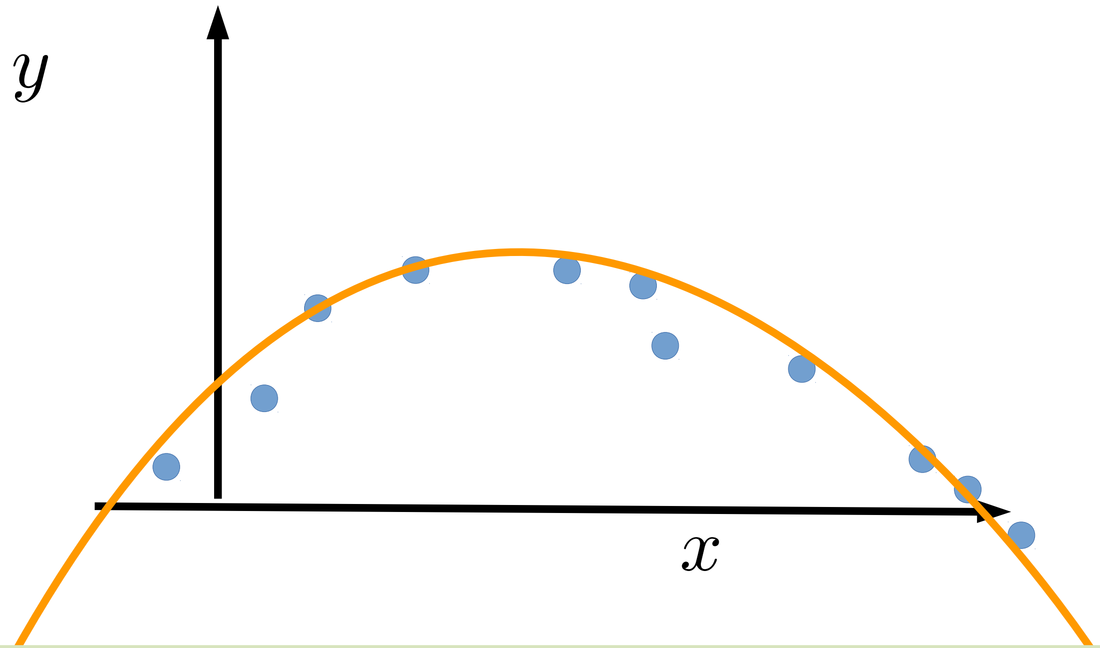
General Training Problem

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

Exe:

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

Overfitting and Model Complexity

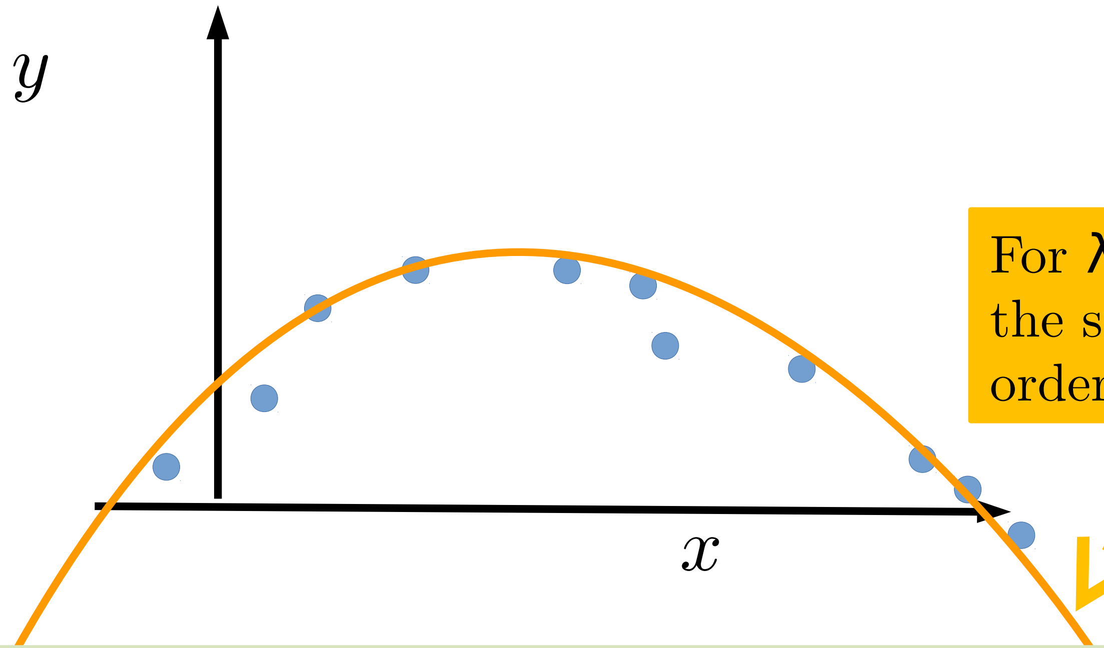


Fitting k$^{\text{th}}$ order polynomial

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

$$w^* = \arg\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i\right)^2 + \lambda ||w||_1$$

Overfitting and Model Complexity



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

Fitting k^{th} order polynomial

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

$$w^* = \arg\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i \right)^2 + \lambda ||w||_1$$

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

$$\ell(y_h, y) = \ln(1 + e^{-y y_h})$$

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*:

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

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

The Statistical Learning Problem: The hard truth

Do we really care if the loss $\ell\left(h_w(x^i), y^i\right)$ is small on the ***known*** labelled data paris (x^i, y^i) ? **Nope**

We really want to have a small loss on new unlabelled Observations!

Assume data sampled $(x, y) \sim \mathcal{D}$ where $\mathcal{D}$ is an unknown distribution

The Statistical Learning Problem: The hard truth

The statistical learning problem:

Minimize the expected loss over an *unknown* expectation

$$\min_{w \in \mathbf{R}^d} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\left(h_w(x), y\right)\right]$$

Variance of sample mean:

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\left(h_w(x), y\right)\right] - \frac{1}{n}\sum_{i=1}^{n} \ell\left(h_w(x_i), y_i\right)\right| = O\left(\frac{1}{n}\right)$$