Optimization for Machine Learning

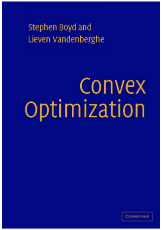
Part I : An Introduction to Supervised Learning

Lionel Tondji

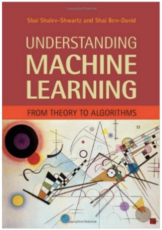
African Master's in Machine Intelligence

July 15, 2024

References for the lectures

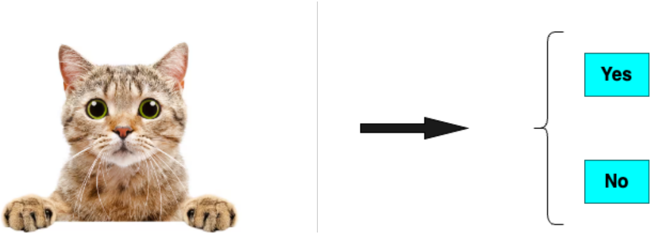


Chapter 2. **"Understanding Machine Learning: From Theory to Algorithms"**.



Pages 67 to 79. **"Convex Optimization, Stephen Boyd"**.

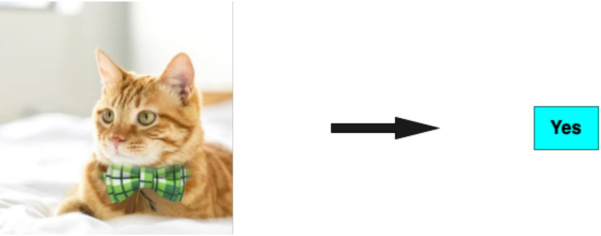
Is There a Cat in the Photo?



Is There a Cat in the Photo?

Yes

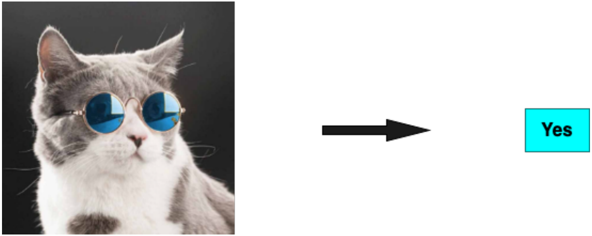
Is There a Cat in the Photo?



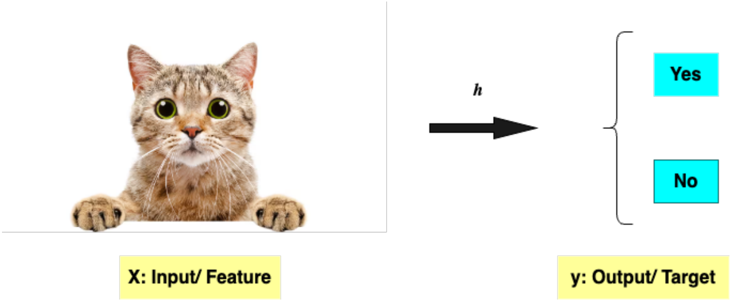
Is There a Cat in the Photo?

No

Is There a Cat in the Photo?



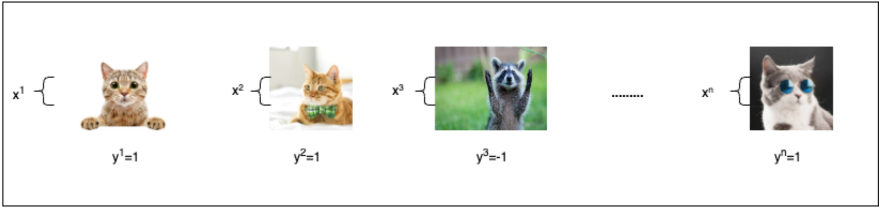
Is There a Cat in the Photo?



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

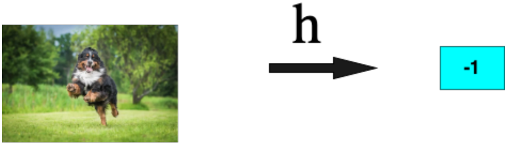
$$h : x \in \mathbb{R}^n \rightarrow y \in \mathbb{R}$$

Labeled Data: The training set



$y = -1$ means no/false $\Downarrow$

Training Algorithm $\Longrightarrow$ $h : x \in \mathbb{R}^n \rightarrow y \in \mathbb{R}$



Example: Linear Regression for Height

Labelled data: $x \in \mathbb{R}^2, y \in \mathbb{R}_+$, Male $= 0$, Female $=1$.

x_1^1	Sex	0
x_2^1	Age	30
y^1	Height	1.82 cm

......

x_1^n	Sex	1
x_2^n	Age	70
y^n	Height	1.52 cm

Example Hypothesis: Linear Model

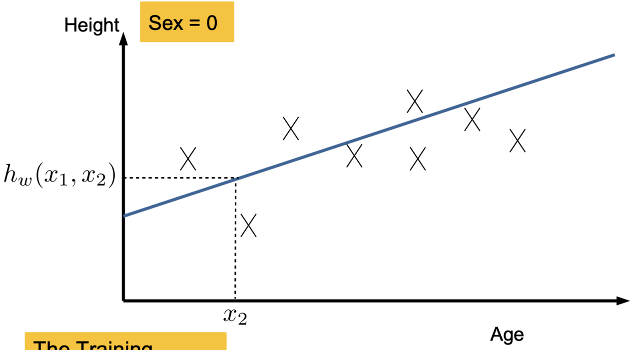
$$h_w(x_1, x_2) = w_0 + x_1 w_1 + x_2 w_2 = \langle w, x \rangle$$

with $x_0 = 1$.

Example Training Problem:

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

Linear Regression for Height



The Training Algorithm

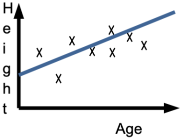
$$\min_{w \in \mathbf{R}^3} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x_1^i, x_2^i) - y^i \right)^2$$

Other options aside from linear?

Parametrizing the Hypothesis

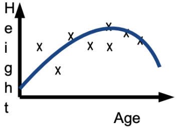
Linear:

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

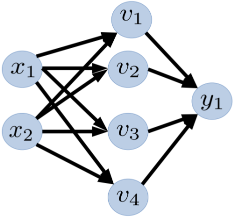


Polinomial:

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



Neural Net:



exe :

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

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

Loss Functions

$$\min_{w\in\mathbb{R}^3} \frac{1}{n}\sum_{i=1}^{n}(h_w(x^i) - y^i)^2$$

Let $y_h := h_w(x)$.
Loss Functions

$$\begin{aligned} \ell : \quad & \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+ \\ & (x, y) \mapsto \ell(y_h, y) \end{aligned}$$

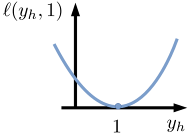
The Training Problem

$$\min_{w\in\mathbb{R}^3} \frac{1}{n}\sum_{i=1}^{n}\ell(h_w(x^i), y^i)$$

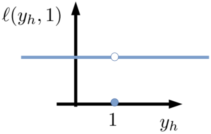
Different the Loss Functions

Let $y_h := h_w(x)$.

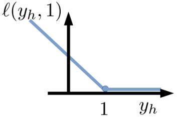
* Square Loss :
  $\ell(y_h, y) = (y_h - y)^2$



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



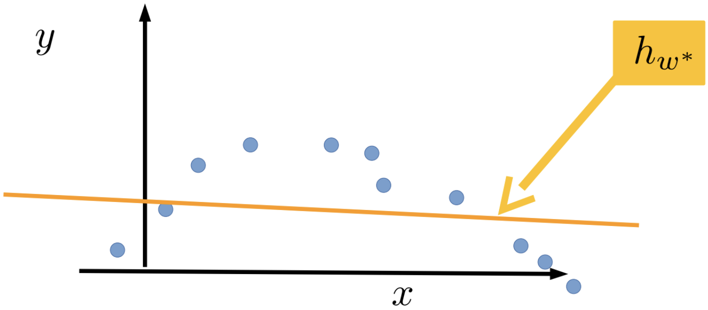
* Hinge Loss :
  $\ell(y_h, y) = \max\{0, 1 - y_h y\}$



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

Is a notion of Loss enough?
What happens when we do not have enough data?

Overfitting and Model Complexity

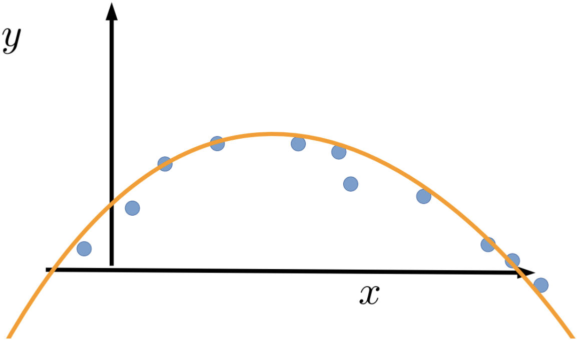


Fitting 1st order polynomial

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

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

Overfitting and Model Complexity

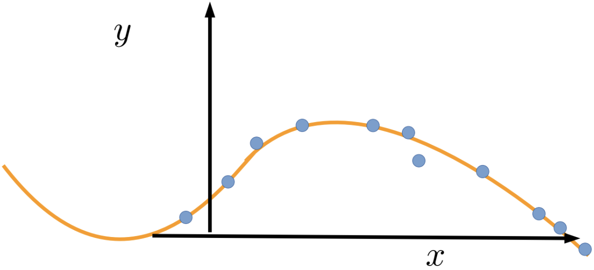


Fitting 2nd order polynomial

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

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

Overfitting and Model Complexity

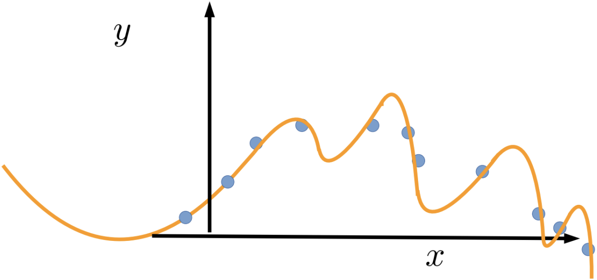


Fitting 3rd order polynomial

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

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

Overfitting and Model Complexity



Fitting 9th order polynomial

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

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

Regularization/Prior

Regularizor Functions

$$R: \quad \mathbb{R}^d \to \mathbb{R}_+$$
$$w \mapsto R(w)$$

General Training Problem

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

- First term : Goodness of fit, fidelity term ...etc
- Second term: Penalizes complexity
- The constant λ : Controls tradeoff between fit and complexity

Example : $R(w) = \|w\|_2^2, \|w\|_1, \|w\|_p$, other norms

Example: Ridge Regression

1. Linear hypothesis : $h_w(x) = \langle w, x \rangle$
2. L2 Regularizer : $R(w) = \|w\|_2^2$
3. L2 Loss : $\ell(y_h, y) = (y_h - y)^2$

$\Rightarrow$ Ridge Regression :

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

Example: Support Vector Machines

1. Linear hypothesis : $h_w(x) = \langle w, x \rangle$
2. L2 Regularizer : $R(w) = \|w\|_2^2$
3. Hinge Loss : $\ell(y_h, y) = \max\{0, 1 - y_h y\}$

$\Rightarrow$ SVM with soft margin :

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

Example: Logistic Regression

1. Linear hypothesis : $h_w(x) = \langle w, x \rangle$
2. L2 Regularizer : $R(w) = \|w\|_2^2$
3. Logistic Loss : $\ell(y_h, y) = ln(1 + e^{-yy_h})$

$\Rightarrow$ Logistic Regression :

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

ML as seen by Optimizer

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 :

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

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 $\ell(h_w(x^i), y^i)$
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

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

Variance of sample mean:

$$\left| E_{(x,y)\sim\mathcal{D}}\left[\ell(h_w(x), y)\right] - \frac{1}{n}\sum_{i=1}^{n} \ell(h_w(x^i), y^i) \right|^2 = \mathcal{O}(\frac{1}{n})$$

Acknowledgements

Thanks for your attention!