Machine learning

Data | **Hypothesis Space** | **Loss**

Data

- Images

feature $\underline{x} \in \mathbb{R}^n$

label y = # persons on image

$y = \begin{cases} > 0 \text{ persons} \\ 0 \text{ persons} \end{cases}$

Hypothesis Space

$w^{(1)}$, $w^{(2)}$

$h(\cdot) : \underline{x} \mapsto \hat{y} \in \mathbb{R}$

$\{0, 1\}$

{"Cat", "Dog"}

"hypothesis"

"predictor"

"classifier"

$h(\underline{x}) = \underline{w}^T \underline{x}$

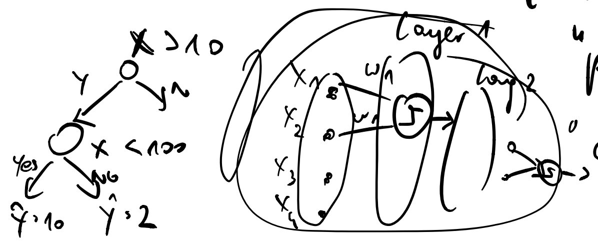
Loss

squared error $(\hat{y} - y)^2$

absolute error $|\hat{y} - y|$

$I\{y \neq \hat{y}\} = \begin{cases} 1 & y \neq \hat{y} \\ 0 & y = \hat{y} \end{cases}$

"0/1" Loss



$\underline{x}^{(m)} = \begin{pmatrix} x_1^{(m)} \\ x_2^{(m)} \\ \vdots \\ x_n^{(m)} \end{pmatrix} \in \mathbb{R}^n$

$(\underline{x}^{(1)}, y^{(1)})$ $(\underline{x}^{(2)}, y^{(2)})$ … $(\underline{x}^{(m)}, y^{(m)})$

linear hypothesis space $h^{(\underline{w})}(\underline{x}) = \underline{w}^T \underline{x} = \sum_{r=1}^{n} w_r x_r$

$\underline{w}$ … "weight vector"

squared error loss

"Linear Regression"

best linear predictor?

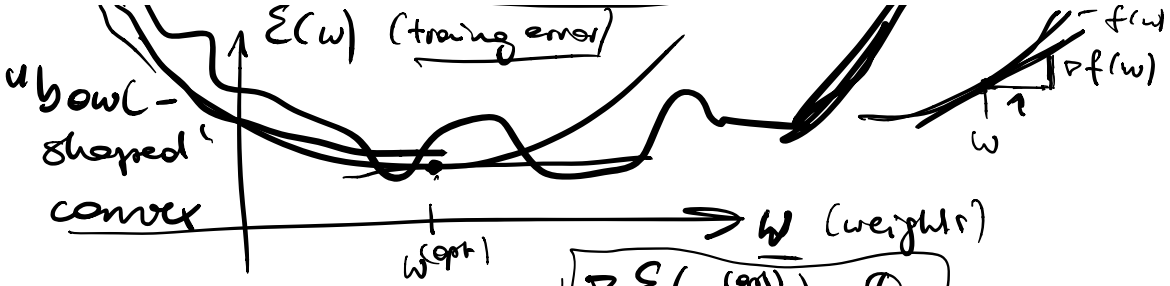
minimize $\frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)} \right)^2$

$y^{(i)}$: true label; $\hat{y}^{(i)}$: predicted label

$$\min_{\underline{w} \in \mathbb{R}^n} \mathcal{E}(w) = \frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} - \underline{w}^T \underline{x}^{(i)} \right)^2 + \lambda R(\underline{w})$$

training error of predictor $h(x) = \underline{w}^T \underline{x}$

$\lambda R(\underline{w})$: complexity $\approx \mathcal{E}^{(val)} - \mathcal{E}^{(train)}$



$$\nabla \mathcal{E}(\underline{w}^{(opt)}) = \underline{0}$$

label vector $\underline{y} = \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{pmatrix} \in \mathbb{R}^m$

feature matrix $\underline{X} = \begin{pmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(m)})^T \end{pmatrix} \in \mathbb{R}^{m \times n}$ $\quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

$$\mathcal{E}(w) = \frac{1}{m} \| \underline{y} - \underline{X}\underline{w} \|_2^2 \qquad \|\underline{x}\|_2^2 = \sum_{r=1}^{n} x_r^2$$

$$\nabla \mathcal{E}(\underline{w}^{(opt)}) = \underline{0}$$

m

$n \gg m : \mathcal{E}(w) = 0$

$$(-2/m)\, \underline{X}^T (\underline{y} - \underline{X}\underline{w}^{(opt)}) = 0$$

OVER FITTING!

$\cdot \underline{C}^{-1}$

$$(2/m)(\underline{X}^T \underline{X})\underline{w}^{(opt)} = (+2/m)\underline{X}^T \underline{y}$$

$$\underline{w}^{(opt)} = \underline{C}^{-1} \underline{X}^T \underline{y}$$

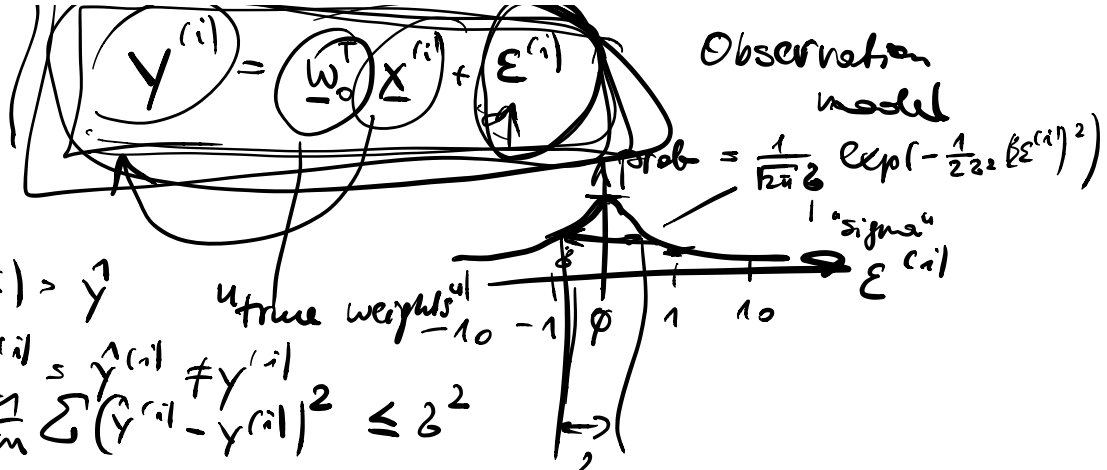
$\underline{C} = \underline{X}^T \underline{X}$ ... non-invertible / singular

$n > m$

training $\rightarrow \underline{w}^{(opt)}$ $\rightarrow$ validate!

$$\mathcal{E}^{(val)}(\underline{w}^{(opt)}) = \sum_{i=m+1}^{m+m_{val}} (\underline{w}^T \underline{x}^{(i)} - y^{(i)})^2$$

$\mathcal{E} \approx \mathcal{E}^{(val)}$ $\sim$ $\mathcal{E}^{reference}$



$(x^{(1)}, y^{(1)}) \; \cdots \; (x^{(m)}, y^{(m)})$

$y^{(1)} = w_0^T x^{(1)} + \varepsilon^{(1)} \quad \cdots \quad y^{(m)} = w_0^T x^{(m)} + \varepsilon^{(m)}$

$y^{(1)} \in \mathbb{R}$, $\varepsilon^{(1)} \in \mathbb{R}$

$x^{(1)} = x^{(2)} = \cdots = x^{(m)}$

empirical mean $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} y^{(i)}$

$\hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m} (y^{(i)} - \hat{\mu})^2 \quad (\approx \sigma^2)$

Optimizer **SGD** — Stochastic Gradient Descent

$\min_{w} \frac{1}{m}\sum (y^{(i)} - w^T x^{(i)})^2 = \min_{w} \frac{1}{m}\|y - Xw\|_2^2$

$\mathcal{E}(w)$ training error

optimal weight vector

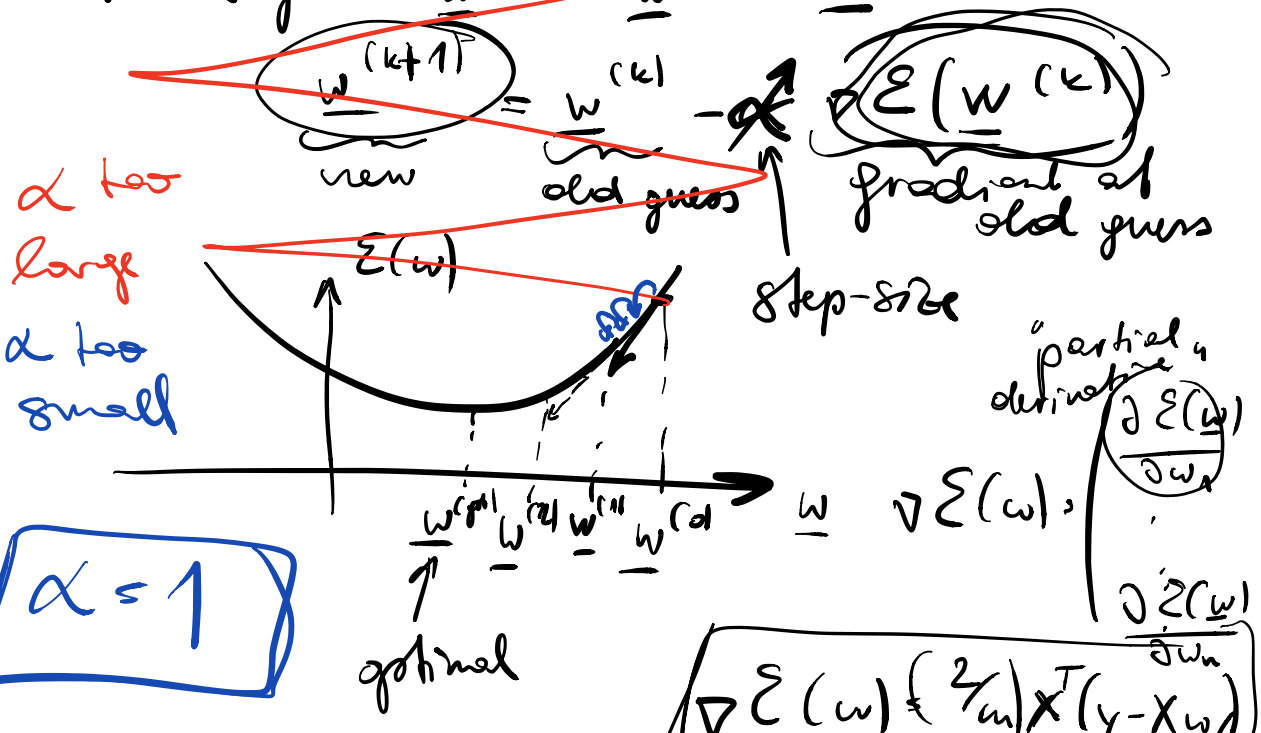
$\nabla \mathcal{E}(w^{(opt)}) = 0$

Gradient Descent

initial guess $w^{(0)} \rightarrow w^{(1)} \rightarrow w^{(2)}$

$$w^{(k+1)} = w^{(k)} - \alpha \nabla \mathcal{E}(w^{(k)})$$

new = old guess − step-size · gradient of old guess



$\alpha = 1$

"partial derivative"

$$\nabla \mathcal{E}(w) = \left(\frac{\partial \mathcal{E}(w)}{\partial w_0}, \frac{\partial \mathcal{E}(w)}{\partial w_1}, \ldots \right)$$

$$\nabla \mathcal{E}(w) = \tfrac{2}{m} X^T (y - Xw)$$

linear regression

$$\nabla \mathcal{E}(w^{(k)}) = \text{error of } k\text{-th iterate}$$

$$\tfrac{2}{m} X^T \underbrace{(y - Xw^{(k)})}_{\begin{pmatrix} y^{(1)} - \hat{y}^{(1)} \\ \vdots \\ y^{(m)} - \hat{y}^{(m)} \end{pmatrix}}$$

squared error loss + linear predictor