Regression

Machine learning, 2021

Mansoor Rezghi

Department of Computer science, TMU

Ref: CB Chap3. KM: Chap7, AG

References

•C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.(CB)

$\{x_1 \ldots x_N\}$

$\{y_1 \ldots y_N\}$

Train

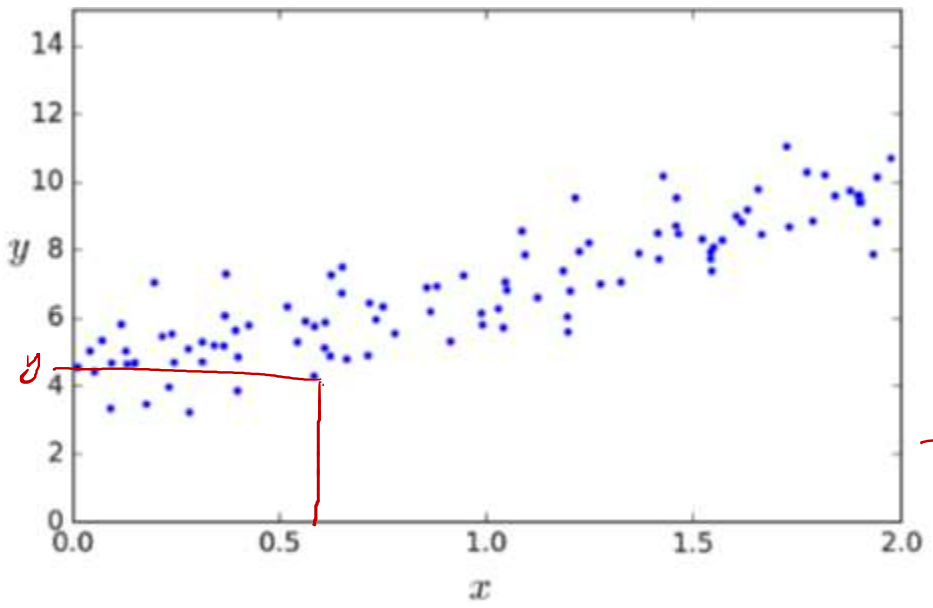
$y = f(x)$

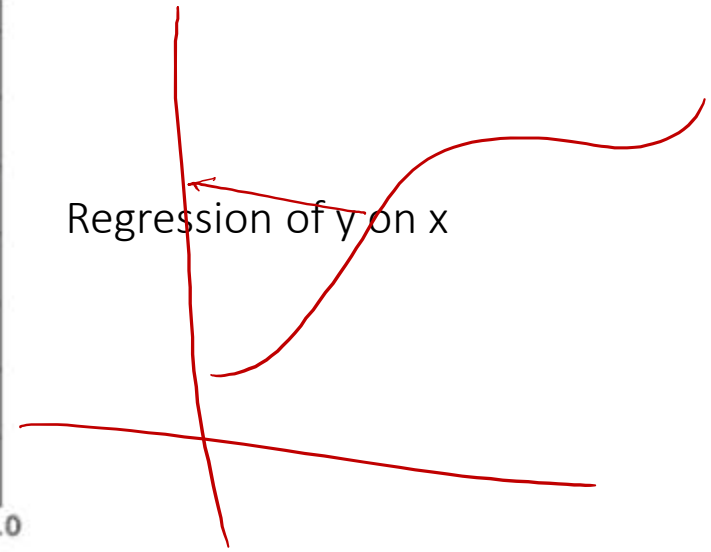
?

$y_i \simeq \bar{f}(x_i) \longrightarrow$ prediction

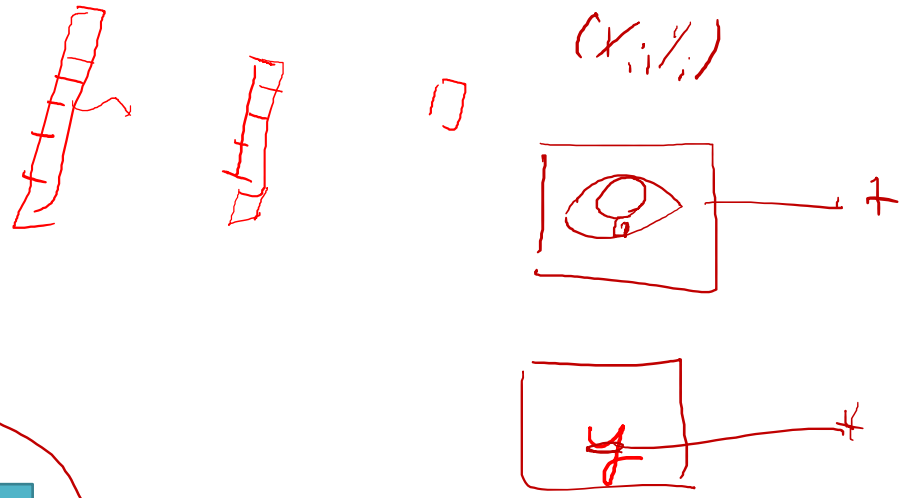
Test

The purpose of regression is to predict a variable from several other variables using training data.





Example:



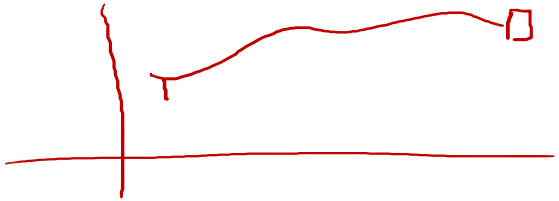
x

- Outside temperature
- Level of humidity
- Sun exposure
- #people

Temperature of a building

$x \leftarrow y$

$x_{t-1} \rightarrow y$
x_t



$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$ $\quad f(x) = w^T x + w.$

$\alpha + \beta = 1$

Regression

The purpose of regression is to predict a variable from several other variables using training data.

y: Target feature(variale)

$\{x_1, ..., x_D\}$: features

Goal: find model f, $y \approx f(x_1, ...x_D)$

x

Input data: $(\mathbf{x}^i, y_i), i = 1, ..., N$

N: ♯ Training samples

$\mathbf{x}^i = (x_1^i, ..., x_D^i)^T$

$\phi_1(x) \; - \; \phi_p(x)$

$x \in R$

Assumption about the model f

f is linear $\rightarrow$ Linear Regression

$y = f(x)$

- linear
- nonlinear

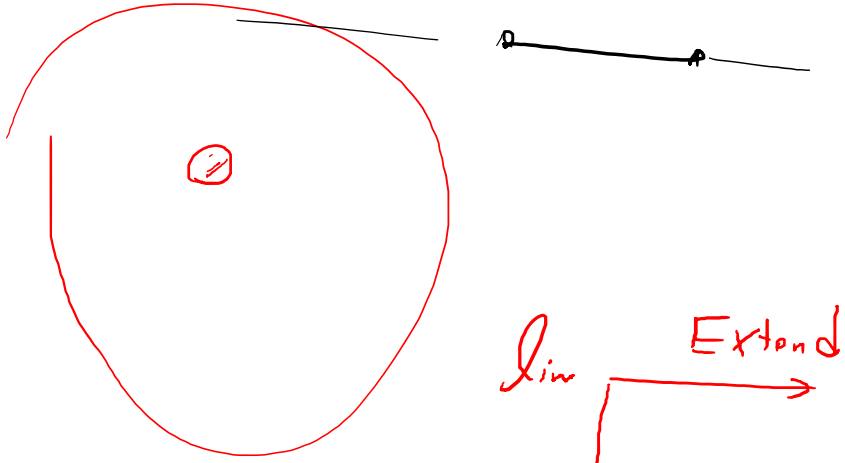
$f: \mathbb{R}^D \to \mathbb{R}$ is a lin funct.

$\forall\, x, y \in \mathbb{R}^D \quad f(\alpha X + \beta Y) = \alpha f(X) + \beta f(Y)$

Affine $\alpha + \beta = 1$

$\boxed{\sin, \cos, \log, e \ldots\ldots}$

$x \in \mathbb{R} \quad f(x) = ax + b \quad 0 < \alpha < 1$



$x \in \mathbb{R}^D \quad f(x) = w^T x + b$

$w \in \mathbb{R}^D \quad b \in \mathbb{R}$

lin $\xrightarrow{\text{Extend}}$ kernel $\to$ nonlinear

$= \alpha(ax + b) + \beta(ay + b)$

$\to$ ANN:

$(\alpha + \beta) b \quad \alpha f(x) + \beta f(y)$

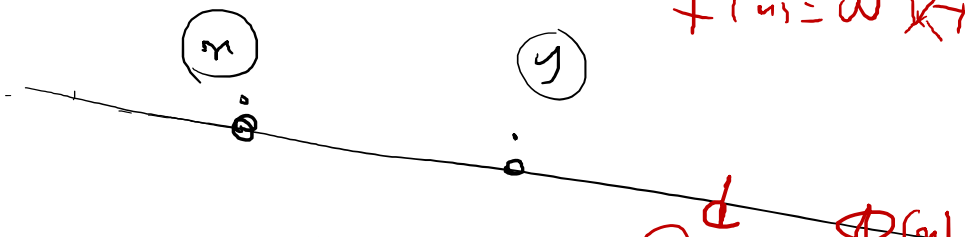
$f(\alpha x + \beta y)$

$= a(\alpha x + \beta y) + b = \cancel{\alpha a x + \alpha a y} \quad \alpha(ax) + \beta(ay) + b$

$x \quad y$

$\alpha x + \beta y$

$\alpha + \beta = 1$



$f(x) = w^T x + b$

$x \not\to \quad x \in R^d \xrightarrow{\phi(x)} \phi(x) \in R^D$

$f(x) = w^T \phi(x) + b$

$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$ Affine

$\alpha + \beta = 1$

$f(x) = ax + b$

$f(\alpha x + \beta y) = a(\alpha x + \beta y) + \overset{(\alpha+\beta)b}{b}$

$= \alpha \underbrace{(ax+b)}_{f(x)} + \beta \underbrace{(ay+b)}_{f(y)} = \alpha f(x) + \beta f(y)$

(x_i, y_i) $\quad y = f(x)$ $\quad \underline{w^T x}$

$x \in R$ $\quad f(x) = wx + b$ $\quad w^T \underline{\phi(x)}$

$$x \to \phi(x) = \begin{bmatrix} x \\ x^2 \\ x^3 \\ x^4 \end{bmatrix} \qquad w^T \phi(x) + b = [w_1 \ w_2 \ w_3 \ w_4] \begin{bmatrix} x \\ x^2 \\ x^3 \\ x^4 \end{bmatrix} + \underbrace{b}_{w_0}$$

$$= \sum_{i=1}^{4} w_i x^i + b$$

$$= \sum_{i=0}^{4} w_i x^i$$

چند جمله‌ای درجه ۴ است.

x

$x \in R \to f(x) = w_0 + wx$ $\quad$ lim

$$\to f(x) = \underbrace{w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4}_{} = \overset{w}{[w_0 \ w_1 \ w_2 \ w_3 \ w_4]} \overset{\phi(x)}{\begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \\ x^4 \end{bmatrix}} \in R^5$$

$(\phi(x_i), y_i) \to$ $\quad = w^T \phi(x)$

$x \to \phi(x)$

$x \in R$ $\quad f(x) = w_0 + w_1 x$

$x \in R$ $\quad \boxed{f(x) = \sum_{i=0}^{N} w_i x^i}$

$x \in R^n \longrightarrow \phi(x) = \begin{bmatrix} 1 \\ x \\ \vdots \\ x^N \end{bmatrix} \in R^{N+1}$

$f(\cdot) = w^T \phi(x)$

$\underline{w^T \phi(x)}$

$f(x) = \frac{a_0}{2} + \sum_{i=1}^{N} a_n \sin nx + b_n \cos nx$

$= w^T \phi(x)$

$\phi(x) = \begin{bmatrix} \frac{1}{2} \\ \sin x \\ \cancel{\sin 2x} \\ \sin \\ \cos \\ \vdots \end{bmatrix}$

$w = \begin{bmatrix} a_0 \\ a_1 \\ b_1 \\ a_2 \\ b_2 \\ \vdots \end{bmatrix}$

$x \in \mathbb{R}^2 \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$w^T x$

$w^T \phi(x)$

$$f(x) = \boxed{w_0 + w_1 x_1} + w_2 x_2 + w_{12} x_1 x_2 + w_{11} x_1^2 + w_{22} x_2^2$$

$$x \in \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \xrightarrow{\phi} \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix} = \phi(x) \in \mathbb{R}^6$$

$$w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_{12} \\ w_{11} \\ w_{22} \end{bmatrix}$$

$$f = w^T \phi(x)$$

Linear Regression

$$y \approx w_0 + w_1 x_1 + \dots . w_D x_D$$

$x \in R^D$

$$y \approx w^T \mathbf{x} + w_0$$

$y \simeq \sum w_i x_i$

Finding the parameters $\{w_i\}, i = 0, .., D$

$y_j \simeq \sum w_i x_i^{(j)}$

These parameters should approximate the values of y_i for $X^i, i = 1, ..., N$ (Training data),

$$y_i \approx a_0 + \sum_{j=1}^{D} w_j x_j^{(i)} = w_0 + w^T \mathbf{x}^i, \quad i = 1, ..., N$$

$\bar{y}_i$ تخمین

$\min \sum_{i=1}^{N} \text{loss}(y_i, \bar{y}_i)$

challenges

Modeling the closeness of $\sum_{j=1}^{D} w_j x_j^{(i)}$ and y_i for all training data.

How to obtain the parameters (optimization)

Generalization:$Train, Test$

Modelling

$f = w^T x + w_0$

$\frac{1}{N}\sum_{i=1}^{N} loss(\hat{y}_i, y_i)$

$\hat{y}_i = w^T x_i + w_0$

$= \underbrace{[w_0 \; w^T]}_{\bar{w}^T} \underbrace{\begin{bmatrix} 1 \\ x_i \end{bmatrix}}_{\bar{x}_i}$

$=$

$= \bar{w}^T x_i$

$f(x) = w^T x + w_0$

- Generalization: $Train, Test$

$$min_{w,w_0} \sum_{i=1}^{N} Loss(\underbrace{w_0 + w^T \mathbf{x}^i}_{\hat{y}_i}, y_i)$$

$w^T x_i$

w

- Examples: Losss($\bar{y}$,y)=$(y - \hat{y})^2$ $\quad y \in \mathbb{R}$
- Losss($\hat{y}$,y)=$|y - \hat{y}|$

$$f(\ ) = w^T x = w^T x + w_0$$

$$\begin{bmatrix} w^T & 1 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & w^T \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix}$$

$$f(\ ) = w^T \phi(x)$$

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_N \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$= \begin{bmatrix} w^T x_1 \\ \\ w^T x_N \end{bmatrix} = \begin{bmatrix} x_1^T w \\ \\ x_N^T w \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix}}_{X^T} w$$

$x_i \in \mathbb{R}^D, \quad y_i \in \mathbb{R}$

$f(x, w) = w^T x \qquad \{x_i, y_i\}$

$$\frac{1}{N} \sum_{i=1}^{N} \text{loss}(\hat{y}_i, \hat{y}_i)$$

$$\text{loss}(\underbrace{w^T x_i}_{\hat{y}_i}, y_i)$$

$$\text{loss}(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2$$

$$\sum_{i=1}^{N} (\hat{y}_i - y_i)^2 = \| \hat{y} - y \|_2^2$$

$$= \| X^T w - y \|_2^2$$

Least squares problem

LS

$$y = w^T x \xrightarrow{\text{لوس}} \boxed{\min \| X^T w - y \|_2^2}$$

$$\sum \| y_i - \hat{y}_i \|_2^2$$

$$\min \| Ax - b \| \rightarrow x = (A^T A)^{-1} A^T b$$

$$x_i \in \mathbb{R}^D \qquad y_i \in \mathbb{R}^d \qquad W \in \mathbb{R}^{D \times d} \qquad y_i \simeq \underbrace{W^T x_i}_{\hat{y}_i}$$

$$\underset{\mathbb{R}^d}{\mathcal{Y}} = \underline{W}^T \underset{\mathbb{R}^D}{\underline{X}}$$

$$\underset{d}{Y} = d \overset{D}{\boxed{W^T}} \boxed{X} n$$

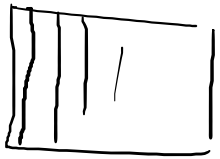
$$W = (X X^T)^{-1} X Y^T$$

$$\sum \text{loss}(\hat{y}_i, y_i) = \sum \| W^T x_i - y_i \|_2^2 = \left\| \overbrace{[W^T x_1, \dots, W^T x_N]}^{W^T X} - \overbrace{[y_1, \dots, y_N]}^{Y} \right\|_F^2$$

$$= \| W^T \underset{\text{known}}{X} - \underset{\text{known}}{Y} \|_F^2 = \| X^T \underline{W} - \underline{Y}^T \|_F^2$$

Least square problem with multiple Right h

$$\|A\|_F^2 = \sum \|a_i\|_2^2 = \sum \|b_i^T\|_2^2 = \sum_{i,j=1}^{N} a_{ij}^2 = tr(A^T A)$$



$$A = [a_1, \dots, a_n] = \begin{bmatrix} b_1^T \\ \vdots \\ b_m^T \end{bmatrix}$$

$$\underbrace{A^T}_{A^T A} \underbrace{A}_{A^T A} = \begin{bmatrix} a_1^T \\ \vdots \\ a_N^T \end{bmatrix} [a_1 \dots a_N] = \begin{bmatrix} \overset{\|a_1\|_2^2}{a_1^T a_1} & & \\ & a_2^T a_2 & \\ & & \\ & & a_n^T a_n \end{bmatrix}$$

$$\min \|W^T \Phi(X) - Y\|_F^2 = \|\Phi(X)^T W - Y^T\|$$

Least squares

- Least squares Case: $\min_{w} \|X\bar{\mathbf{w}} - \mathbf{y}\|_2^2$ $\boxed{\text{opt. value?}}$
- $\bar{\mathbf{w}} = (w_0, \ldots, w_D)$
- $X = \begin{pmatrix} 1 & \mathbf{x}_1^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^T \end{pmatrix}$

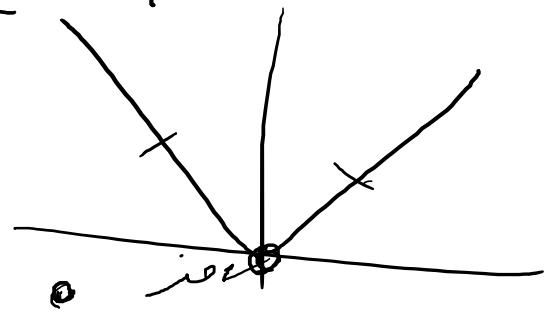
Regression: 1D case

If the data points were on the line, the parameters would satisfy the equations

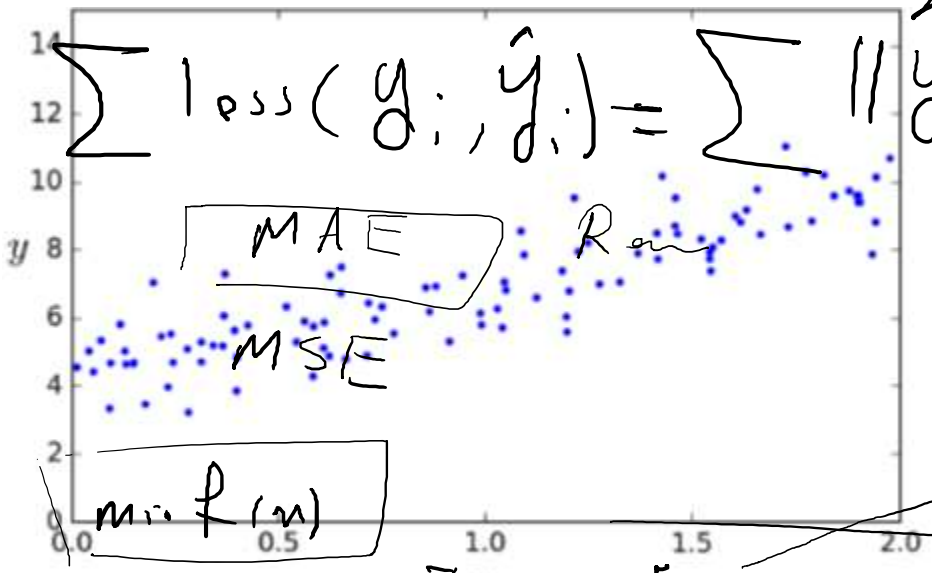
Predicted y-value		Observed y-value
$\beta_0 + \beta_1 x_1$	$=$	y_1
$\beta_0 + \beta_1 x_2$	$=$	y_2
$\vdots$		$\vdots$
$\beta_0 + \beta_1 x_n$	$=$	y_n

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$\text{loss}(y_i, \hat{y}_i) = |y_i - \hat{y}_i|$ $\quad y_i \in \mathbb{R}$ $\quad \|x\|_1 = \sum |x_i|$



$\text{loss}(y_i, \hat{y}_i) = \|y_i - \hat{y}_i\|_1$ $\quad y_i \in \mathbb{R}^d$

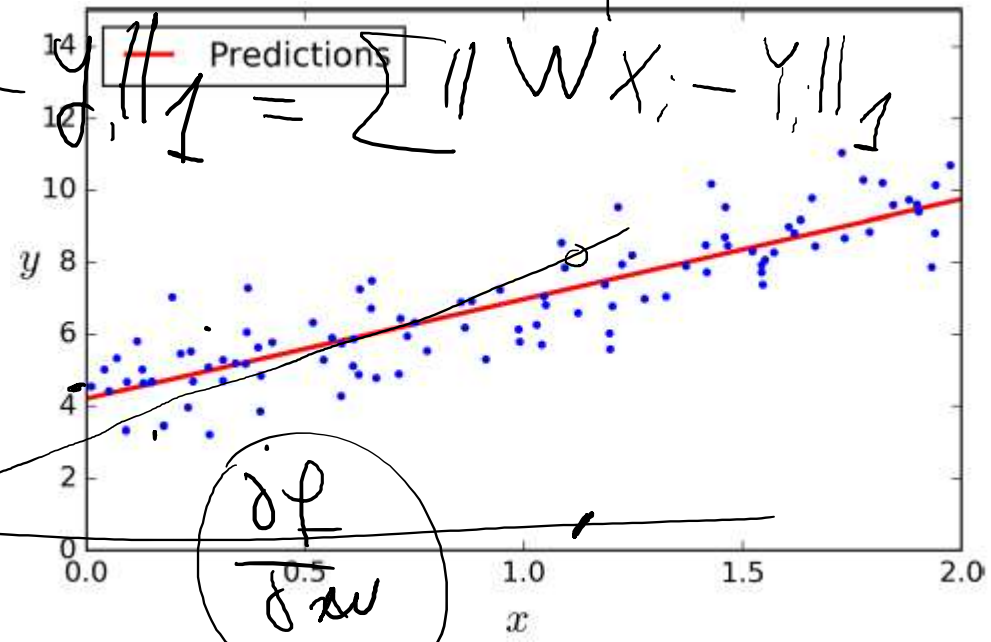


$\sum \text{loss}(y_i, \hat{y}_i) = \sum \|\hat{y}_i - y_i\|_1 = \sum \|W^T X_i - Y_i\|_1$

MAE

MSE

$\min f(x)$



$\frac{\partial f}{\partial w}$

$f(x) = \|X^T W - y\|_2^2$

$= \sum \|w^T x_i - y_i\|_1$

$f($ $\{x_i\}, \{y_i\}$

$y = f(x_i)$

$\downarrow$

f: lineal

$f(x) = W^T x + b$

$x \in \mathbb{R}^D$, $y \in \mathbb{R}^d$

$W \in \mathbb{R}^{D \times d}$, $b \in \mathbb{R}^d$

$$f(x) = \boxed{W^T x}$$

$$= \underline{[W^T \; b]} \begin{bmatrix} x \\ 1 \end{bmatrix}$$

$x \longrightarrow \Phi(x)$ | Poly fun

$y_i \approx f(x_i) = \hat{y}_i$ $\quad = \underline{W}^T \Phi(x)$

proxima

$\| \min \sum \text{loss}(\hat{y}_i, y_i)$

$\text{loss}(y, \hat{y}) = \| y - \hat{y} \|_2^2$

$\text{loss}(y, \hat{y}) = \| y - \hat{y} \|_1$

$\vdots$

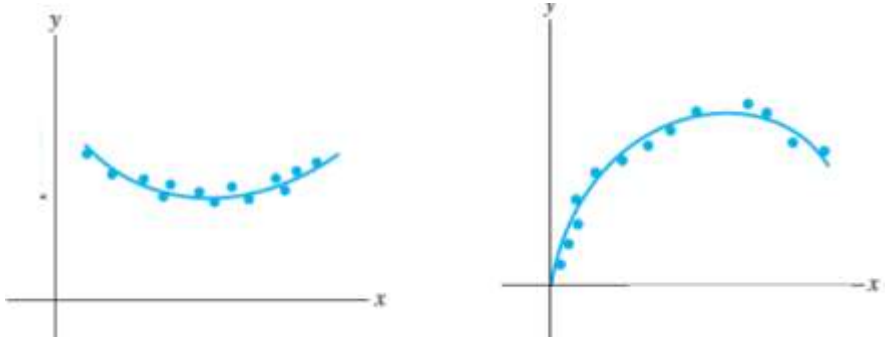
RSS $\hat{y}_i, y_i$ $\sum \|\hat{y}_i - w^T x_i\|_2^2$

1D Curve fitting: with base function

$$\phi(x) = \begin{bmatrix} \phi_1(x) \\ \vdots \\ \phi_{M-1}(x) \end{bmatrix}$$

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = w^T \phi(x)$$

$\phi_j(\mathbf{x})$ are known as *basis functions*.



Polynomial $\phi_j(x) = x^j$

Gaussian $\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$

Sigmoid $\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$

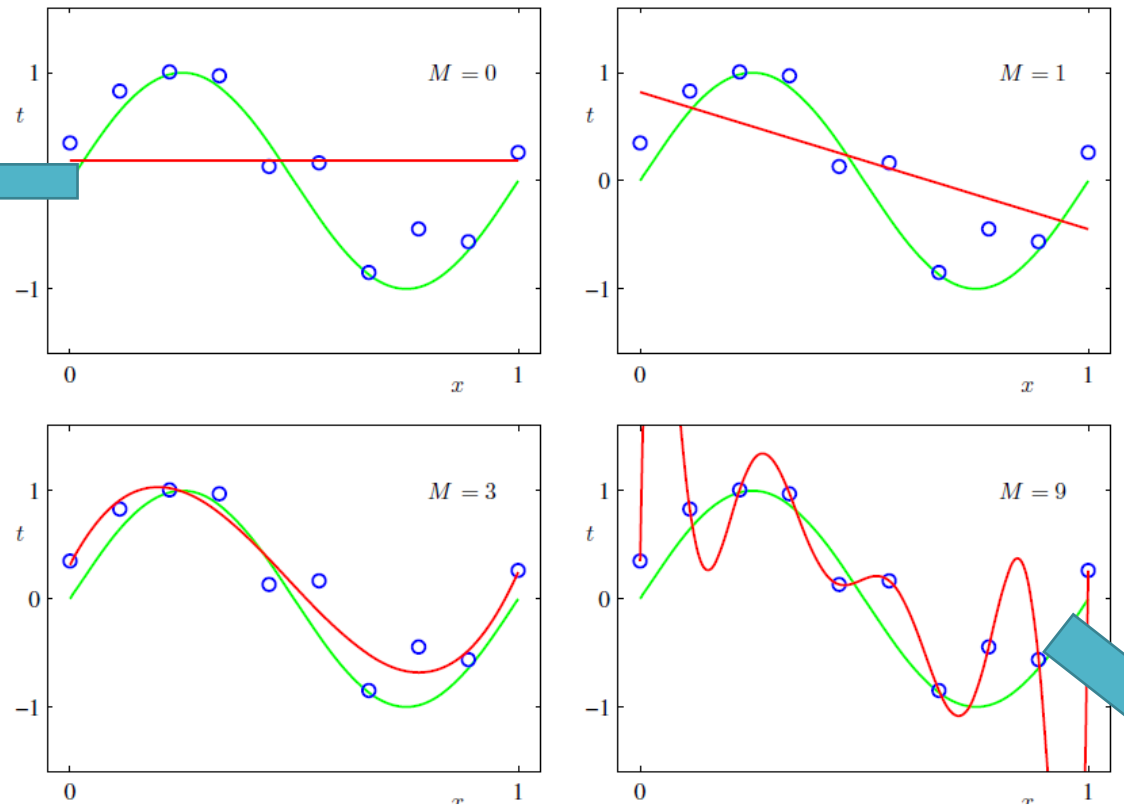
wavelet

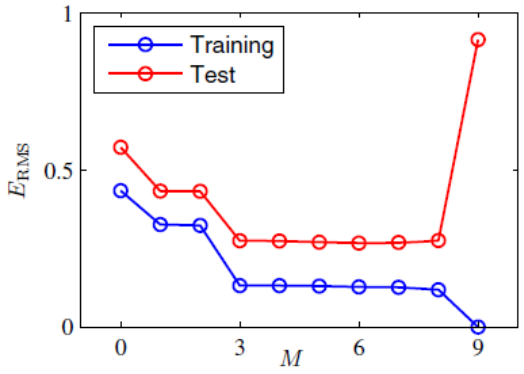
B-spline

Underfitting-Overfitting

Model performs well on the training data, but it does not generalize well(weak generalization)

Under-fitting





Over-fitting

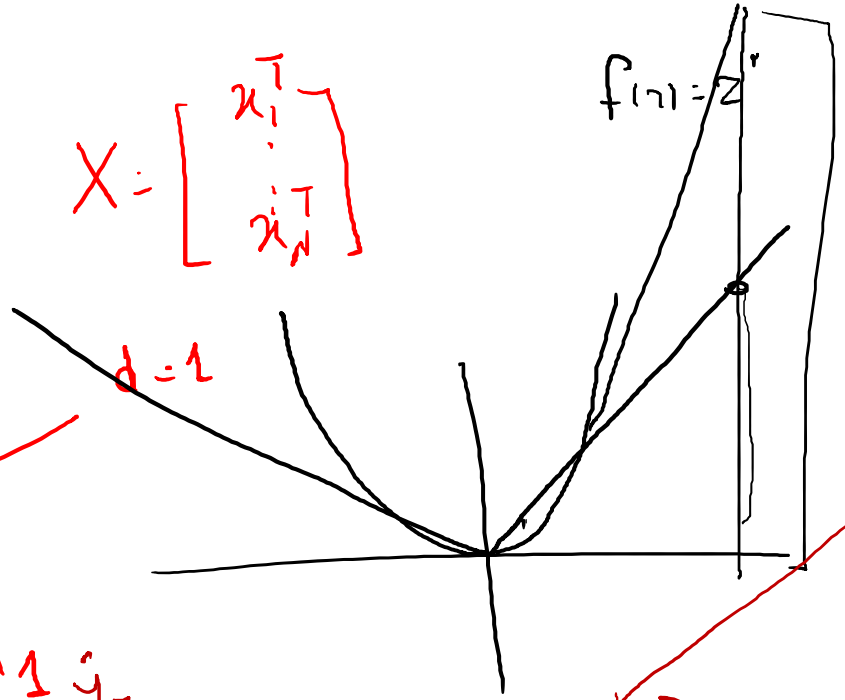
$\{x_i, y_i\}$ $\quad f(x) = w^T x + b := w^T x$

$$X := \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix}$$

Norm: (RMSE): $\min \| X w - y \|_2^2$

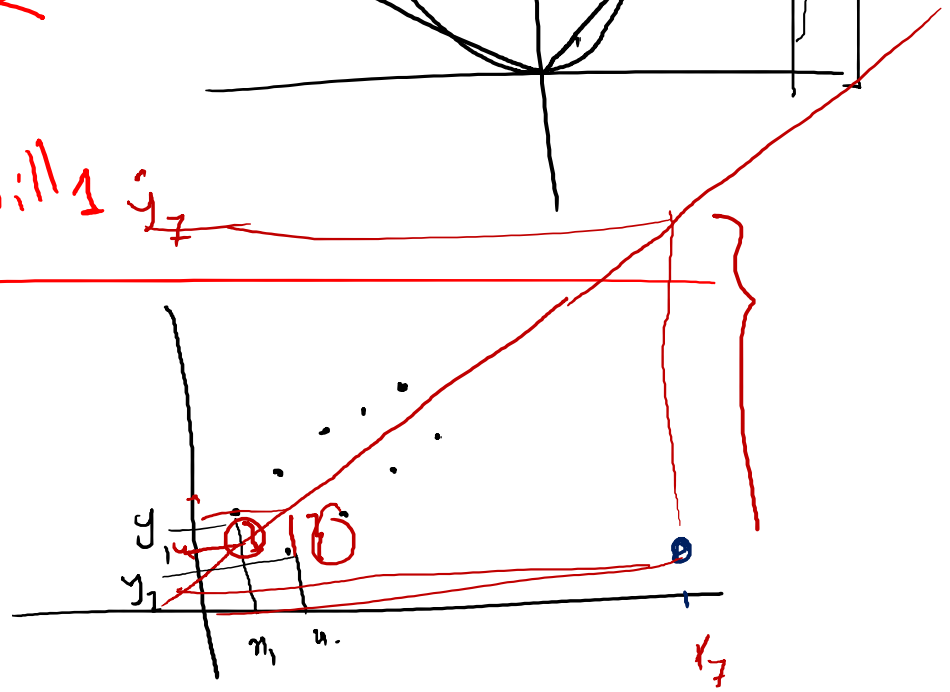
$\min \| X w - y \|_2^2$

$d = 1$



Norm1: $\Rightarrow$ $\min \| X w - y \|_1$

$\min \sum \| X w_i - y_i \|_1$



$\frac{|y - \hat{y}|^2}{2} = z^2 \qquad G(z) = z^2$

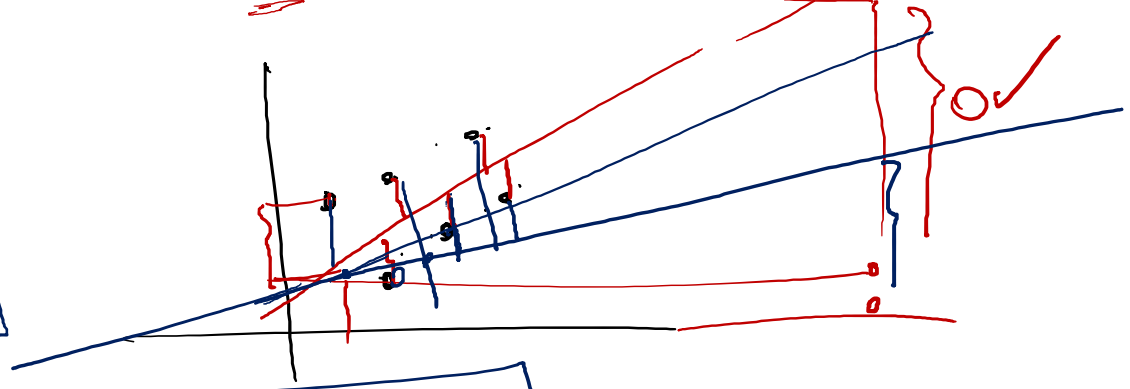
$\frac{|y - \hat{y}|}{2} = |z| \qquad G(z) = |z|$

$y = 1, \ \hat{y} = 2 \Rightarrow \begin{cases} \Rightarrow 9 \\ \Rightarrow 3 \end{cases}$

$y = 1, \ \hat{y} = 11 \mapsto \begin{cases} 100 \\ 10 \end{cases}$

$$\sum_{i=1}^{6} (y_i - \hat{y}_i)^2 + \boxed{(y_7 - \hat{y}_7)^2}$$

$\| \ \|_2^2 \rightarrow$ حساسیت به داده پرت



$\|Xw - y\|_2$

$\boxed{\text{Robust}}$

$\boxed{\text{Early Stopping}}$

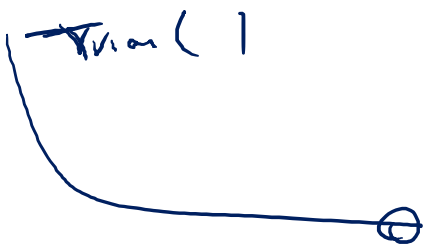
$\|Xw - y\|_1$

AME

$y = f(x)$

$y_i \sim f(x_i)$

$\min \dfrac{\sum (y_i - \hat{y}_i(w))^2}{G(w)}$



$\boxed{\min_w G(w)}$

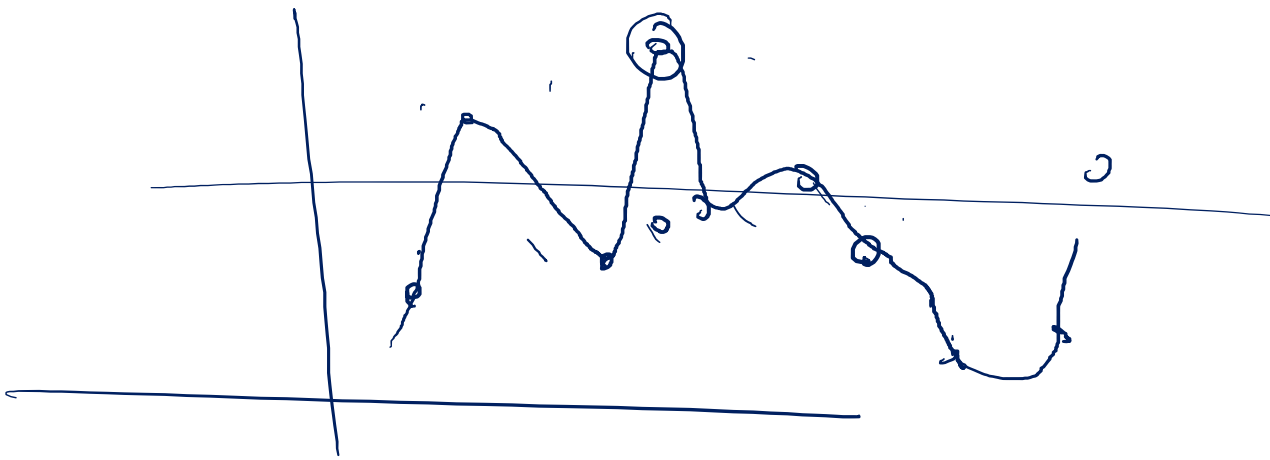
$w_1 \qquad w_2$

$G(w_1) < G(w_2)$

Overfitting

Overfitting happens when the model is too complex relative to the amount and noisiness of the training data

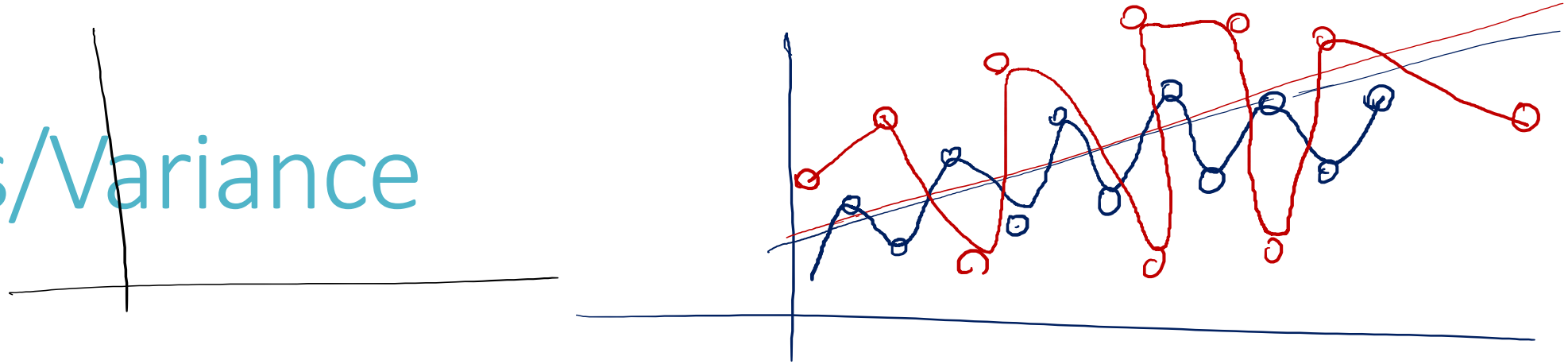
- simplifying the model
- More training data
- Reducing the noise in the training data
- Insert extra information(Regularization)



Underfitting

- underfitting is the opposite of overfitting
- it occurs when your model is too simple to learn the underlying structure of the data
- more powerful model

Bias/Variance



Bias

- Due to wrong assumptions
- Assuming that the data is linear when it is actually quadratic.
- A high-bias model is most likely to underfit the training data.

$\boxed{4}$

$$f(x) = \sum_{i=0}^{9} w_i x^i$$

$\boxed{10}$

Variance

- Model's excessive sensitivity to small variations in the training data.
- A model with many degrees of freedom is likely to have high variance
- Overfit the training data.

Overfitting: More training data

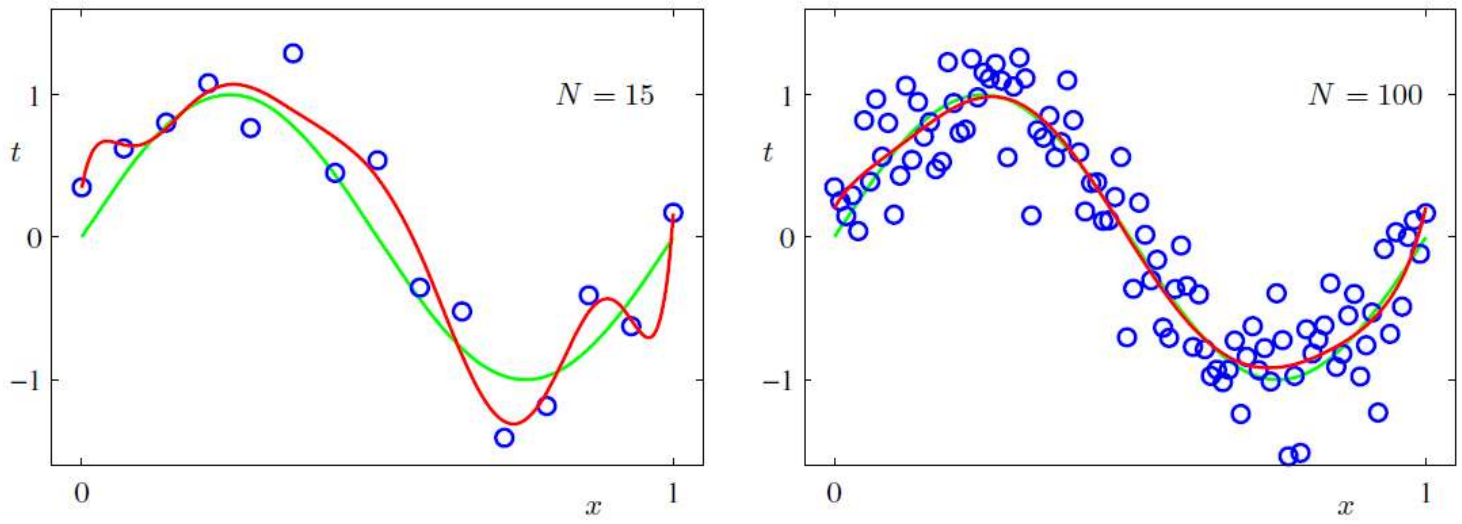


Figure 1.6 Plots of the solutions obtained by minimizing the sum-of-squares error function using the $M = 9$ polynomial for $N = 15$ data points (left plot) and $N = 100$ data points (right plot). We see that increasing the size of the data set reduces the over-fitting problem.

Large number of data

$$\begin{cases} \min \|Xw - y\|_2^2 \\ \min \Omega(w) \end{cases} \Longrightarrow \min \|Xw - y\|_2^2 + \lambda \Omega(w)$$

Overfitting: Regularization

Table 1.1 Table of the coefficients $\mathbf{w}^\star$ for polynomials of various order. Observe how the typical magnitude of the coefficients increases dramatically as the order of the polynomial increases.

	$M = 0$	$M = 1$	$M = 6$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

M(parameters) increases $\Longrightarrow$ Coefficients with higher magnitute

$w^T \phi(x) = \begin{bmatrix} \phi_1(x) \\ \phi_M(x) \end{bmatrix}$ $\quad w^T \phi(x) = \sum w_i \phi_i(x)$

$\Phi X w \quad \min \|Ax - y\|_2^2 \rightarrow x = \boxed{(A^T A)^{-1} A^T} y$

$A^{-1} (A^{-T} \alpha A^{-1})$

Regularization

$\hat{y} = \sum_{j=0}^{M} w_j \phi_j(x), \boxed{\phi_j(x) = x^j} \Rightarrow \min_{\mathbf{w}} \|\Phi \mathbf{w} - \mathbf{y}\|_2^2 \Rightarrow \mathbf{w} = \boxed{(\Phi^T \Phi)^{-1} \Phi^T} \mathbf{y}$

$\phi^{\dagger}$

Noisy target value $\quad y = y_{\text{exact}} + n, \; n : \text{noise} \Rightarrow \boxed{\mathbf{w}} = \mathbf{w}_{exact} + \boxed{\mathbf{w}_{noise}}$

$\|\Phi w - y\|$

$w = (\Phi^T \Phi)^{-1} \Phi^T (y_e + n)$

$= (\Phi^T \Phi)^{-1} \Phi^T y_e + (\Phi^T \Phi)^{-1} \Phi^T n$

$$\mathbf{w}_{\text{exact}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}_{\text{exact}}$$

$$\boxed{\mathbf{w}_{\text{noise}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{n}}$$

ill-cond

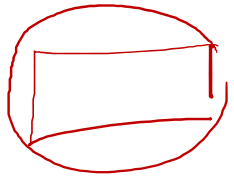
Condition number of $\Phi^T \Phi$: The sensitivity of the answer to the presence of noise in the data

$\min \|\Phi w - y\|_2^2 + \lambda \Omega(w)$

Regularization:

$\Omega(w)$

Adding more information into the model



Penalty Terms	Projections
○ Tikhonov-Ridge ○ L1(Lasso)	TSVD Early stopping

Truncated SVD

Early stopping (Semi - Con

Tikhonov-Ridge

$\|x\|_\infty = \max_{1\le i\le n} |x_i|$

$\|p\|_p^p = \sum_{i=1}^{m} |p_i|^p$

$\|x\|_p^p = 1$

$\|x\|_p^p \le 1$

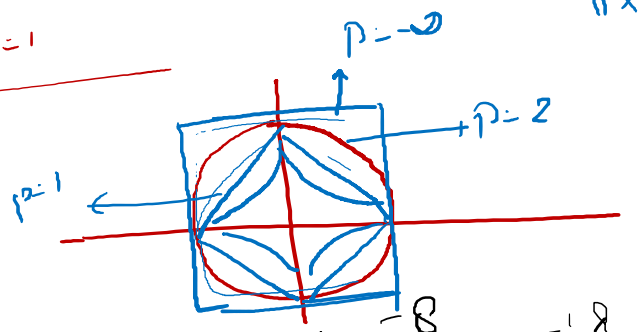
$x \in \mathbb{R}^2 \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \|x\|_r^r = 1 \Rightarrow x_1^r + x_2^r = 1$

$\|x\|_1 = 1 \quad |x_1| + |x_2| = 1$

$\max\{|x_1|, |x_2|\} = 1$ Sup

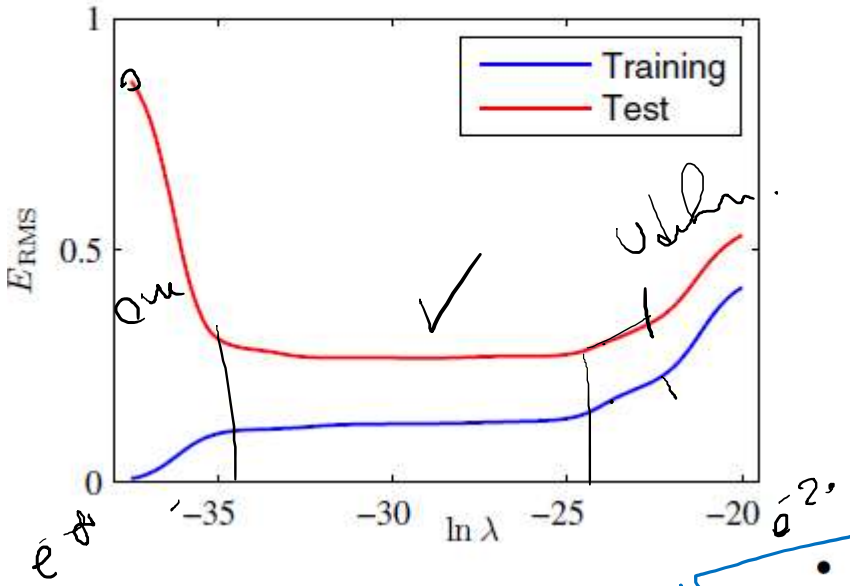
$\|x\|_0 = \#\{x_i \neq 0\}$

$\|x\|_0 \le 1$



$\lambda = e^{-8}$ $\quad e^{-18}$ $\quad e^{0}$

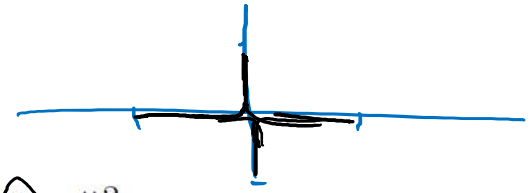
Graph of the root-mean-square error (1.3) versus $\ln \lambda$ for the $M = 9$ polynomial.



$$\min_{\mathbf{w}} \|\Phi\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

- Smooth solution
- Easy to solve

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01



$= \max_{i} |p_i|$

- General Form

$$\min_{\mathbf{w}} \|\Phi\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|L\mathbf{w}\|_2^2$$

$\|x\|_\infty = \lim_{p\to\infty} \|x\|_p = \lim_{p\to 0} \sqrt[p]{\sum |p_i|^p}$

$$\min \|\Phi w - y\|_2^2 \longrightarrow w = (\Phi^T \Phi)^{-1} \Phi^T y$$

$$\min \|\Phi w - y\|_2^2 + \underbrace{\lambda \|w\|_2^2} = \left\| \begin{pmatrix} \Phi w - y \\ \sqrt{\lambda} w \end{pmatrix} \right\|_2^2$$

$$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \qquad \|z\|_2^2 = \|z_1\|_2^2 + \|z_2\|_2^2$$

$$= \left\| \underbrace{\begin{pmatrix} \Phi \\ \sqrt{\lambda} I \end{pmatrix}}_{B} w - \underbrace{\begin{pmatrix} y \\ 0 \end{pmatrix}}_{\bar{y}} \right\|_2^2$$

$$= \underline{\|Bw - \bar{y}\|_2^2} \Rightarrow$$

$$w = (B^T B)^{-1} B^T \bar{y}$$

$$w = \left[\begin{pmatrix} \Phi^T & \sqrt{\lambda} I \end{pmatrix} \begin{pmatrix} \Phi \\ \sqrt{\lambda} I \end{pmatrix} \right]^{-1} \underbrace{\begin{pmatrix} \Phi^T & \sqrt{\lambda} I \end{pmatrix} \begin{pmatrix} y \\ 0 \end{pmatrix}}_{\Phi^T y}$$

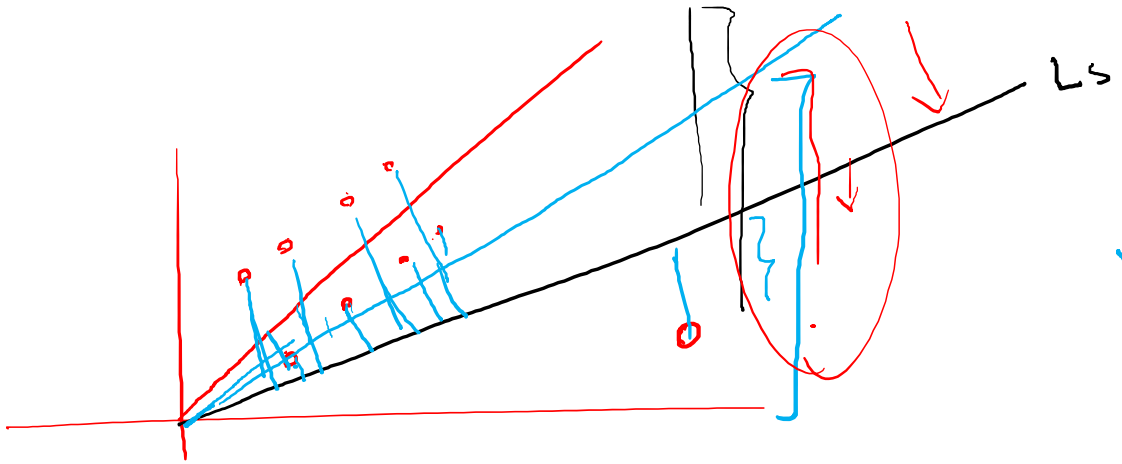
$$= \left[\Phi^T \Phi + \lambda I \right]^{-1} \Phi^T y$$

$$\|\Phi w - y\|_2^2 + \lambda \|w\|_2^2 \longrightarrow \boxed{\|\sqrt{\lambda}\, w\|_2^2 = \|\sqrt{\lambda}\, I\, w\|_2^2}$$

$$= \left\| \begin{matrix} \Phi w - y \\ \sqrt{\lambda}\, w \end{matrix} \right\|_2 = \left\| \begin{pmatrix} \Phi w \\ \sqrt{\lambda}\, w \end{pmatrix} - \begin{pmatrix} y \\ 0 \end{pmatrix} \right\|$$

$$\begin{pmatrix} \Phi \\ \sqrt{\lambda}\, I \end{pmatrix} w = \begin{pmatrix} \Phi w \\ \sqrt{\lambda}\, I w \end{pmatrix} = \begin{pmatrix} \Phi w \\ \sqrt{\lambda}\, w \end{pmatrix}$$

$$\left\| \begin{pmatrix} \Phi \\ \sqrt{\lambda}\, I \end{pmatrix} w - \begin{pmatrix} y \\ 0 \end{pmatrix} \right\|$$



$+\|\underline{w}\|$

$$\min \|\overset{r}{\Phi w - y}\|_2^2 = \sum r_i^2$$

$$\min \|\Phi w - y\|_1 = \sum \boxed{|r_i|}$$

$r = \Phi w - y$? $r_i = ?$

span.

r

$$r_i = \Phi_i^T w - y_i = \underline{w^T \Phi_i} - \underline{y_i}$$

$f_1 \to f_2$

L1(sparse)regularization-Lasso

$m \; \|\Phi w - y\|_1$

$$\min_{\mathbf{w}} \|\Phi\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1$$

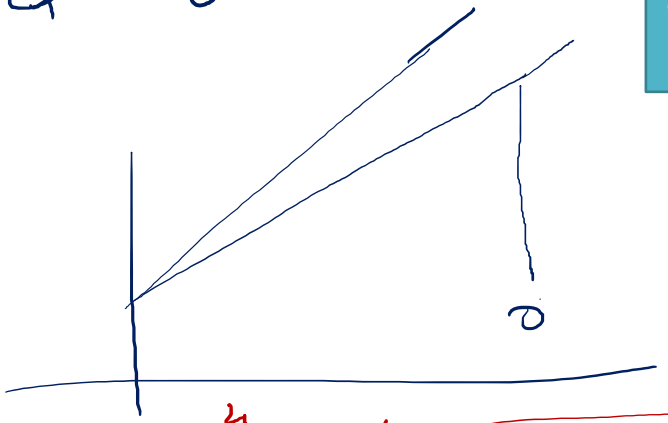
$r = \Phi w - y$

$r_i = \Phi_i^T w - y_i$

$= w^T \Phi_i - y_i$

$r = \Phi w - y$

- Sparse solution
- Able to construct edges



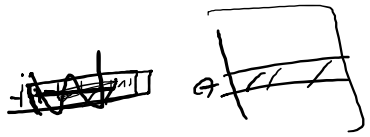
$m \; \|\Phi w - y\|_2^2 + \lambda \|w\|_2^2$

$m \; \|\Phi w - y\|_2^2 + \lambda \|w\|_1$

$y = \sum_{i=1}^{4} w_i x^i = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4$

$XW - X$

$X - X$



Multi-task Lasso

TFIDF

Word Embedding

$\|w\|_1 = ?$

$$\min_W \; ||XW - Y||_{\text{Fro}}^2 + \alpha ||W||_{21}$$

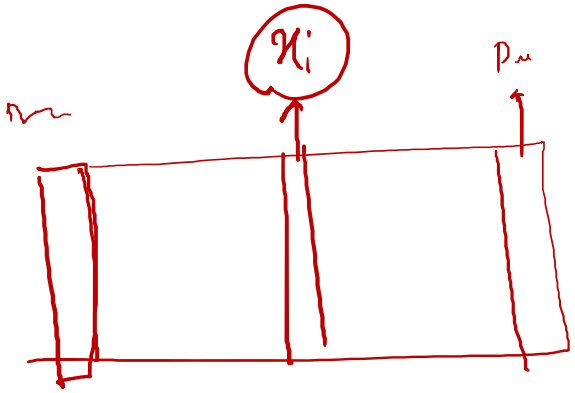
$$||A||_{\text{Fro}} = \sqrt{\sum_{ij} a_{ij}^2}$$

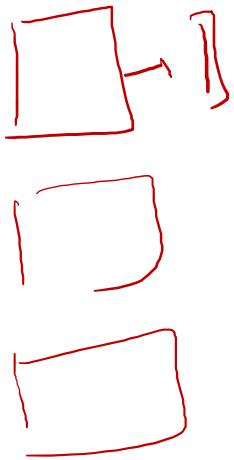
$$||A||_{21} = \sum_i \boxed{\sqrt{\sum_j a_{ij}^2}}$$

$$\min \; ||XW - X||_F^2$$

Text → a i

Sparse Subspace Clustering (SCC)





Text

Summerization

$$x_i \sim \sum_{j \neq i}^{3} \alpha_j x_j$$

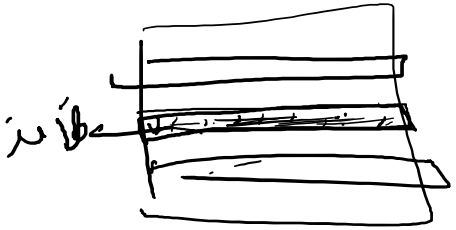
$$X_i \simeq X w^i$$

$$X_1 \simeq X w_1$$

$$X_2 \simeq X w_2$$

$$\vdots$$

$$X_n \simeq X w_N$$

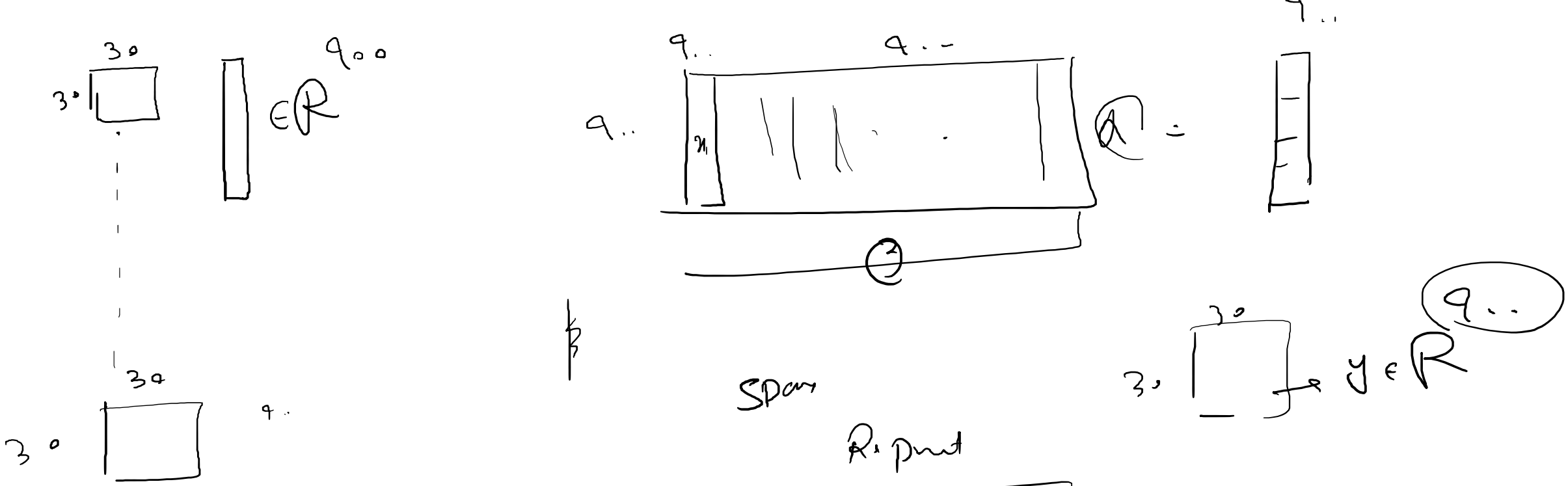


$$A = \begin{bmatrix} a_1^T \\ a_2^T \\ a_3^T \\ \vdots \\ a_N^T \end{bmatrix} \qquad \begin{pmatrix} \|a_1\|_2 \\ \|a_2\|_2 \\ \vdots \\ \|a_N\| \end{pmatrix}$$

$$S$$

$$\sum_{i=1}^{N} \|a_i^T\|_2$$

$$\sum_i \|a_i^T\|_2 = \|S\|_1$$

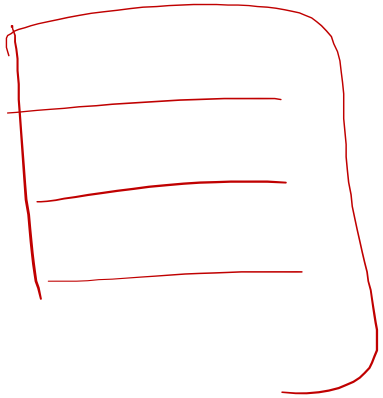


Span

R. product

Compress Sens

$f = Xw$ $\quad f = w^T x$



Hyperparameter

Elastic-Net

$$\min_{w} \; ||Xw - y||_2^2 + \rho||w||_1 + \alpha \cdot ||w||_2^2$$

Quality of Regression

$y_i - y$ w

$f_1 = w_1^T x$ $f_1 = w_2^T x$

- RMSE(Root Mean Squares error)
- MAE(Mean Absolut Error)
- MAPE(Mean absolute percentage error)

Test

$$\frac{1}{N} \sum \frac{|y_i - \hat{y}_i|}{|y_i|}$$

$$\frac{1}{N} \sum (y_i - \hat{y}_i)^2$$

$$\frac{1}{N} \sum |y_i - \hat{y}_i|$$

Quiz (L_2 Vs L1 error term)

What's the main advantage of using the absolute error term(norm)over the mean squares error (l_2) in the objective function of regression?