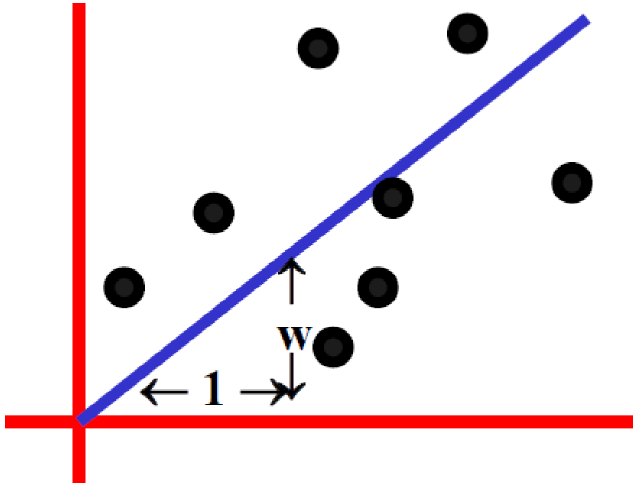
رگرسیون

مهدی شکری زاده

معرفی مسئله رگرسیون

- ❑ رگرسیون خطی تک متغیره
  - ○ ارتباط بین داده های خروجی و داده های ورودی توسط یک مدل خطی قابل توصیف است.



$$y(x) = wx + w_0$$

Single parameter linear regression

inputs	outputs
$x_1 = 1$	$y_1 = 1$
$x_2 = 3$	$y_2 = 2.2$
$x_3 = 2$	$y_3 = 2$
$x_4 = 1.5$	$y_4 = 1.9$
$x_5 = 4$	$y_5 = 3.1$

  - ○ هدف تعیین پارامترهای مدل بر اساس داده های در دسترس است.

رگرسیون خطی – تک متغیره

- ❑ **مدل رگرسیون خطی برای مسئله عنوان شده به شکل ساده یک خط به صورت زیر قابل توصیف است:**

$$\underbrace{\{x_i, y_i\}_i}_{\text{Given dataset}} \longrightarrow y_i(x) = WX_i + n_i, W = [w \quad w_0], X_i = \begin{bmatrix} x_i \\ 1 \end{bmatrix}$$

  - ○ که در آن n_i یک فرآیند سفید گوسی مستقل با میانگین 0 و واریانس نامعلوم است.

- ❑ **مسئله:**

  - ○ برای چه مقداری از بردار وزن های W احتمال رخداد زوج داده های $\{x_i, y_i\}_i$ بیشینه است؟

$$W^* = \arg\max \prod_{i=1}^{N} P(y_i | x_i, W)$$

رگرسیون خطی – تک متغیره

- توزیع شرطی داده های خروجی به شرط داده های ورودی و بردار وزن ها یک توزیع نرمال است. (چرا؟)

$$\prod_{i=1}^{N} P(y_i|x_i, W) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y - WX_i)^2}{2\sigma^2}\right)$$

  - به منظور تاثیر تمام نقاط داده شده، negative-log تابع فوق کمینه می شود. (چرا؟)
  - مسئله بهینه سازی جدید دارای پاسخ یکسان با مسئله اصلی است. (چرا؟)

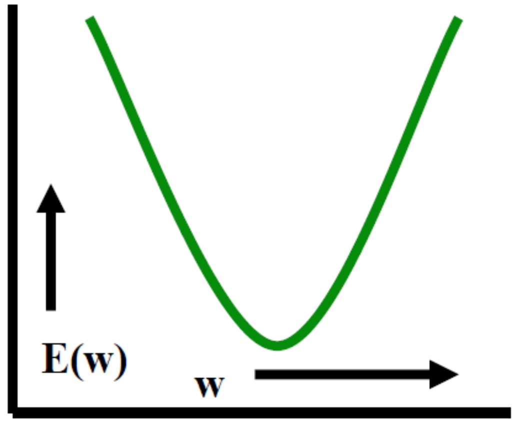
$$W^* = \arg\min -\log \prod_{i=1}^{N} P(y_i|x_i, W) \quad \Longrightarrow \quad W^* = \text{argmin} \sum_{i=1}^{N} (y - WX_i)^2$$

Least square problem ← Sum Squared Error (SSE)

رگرسیون خطی – تک متغیره

- بردار پارامترهایی که رخداد داده های خروجی به شرط داده های ورودی را بیشینه می کند پاسخ مسئله بهینه سازی حداقل مربعات است.

$$W^* = \text{argmin} \sum_{i=1}^{N} (y - WX_i)^2 \Longrightarrow W^* = \frac{\sum_{i=1}^{N} y_i X_i}{\sum_{i=1}^{N} |X_i|^2}$$



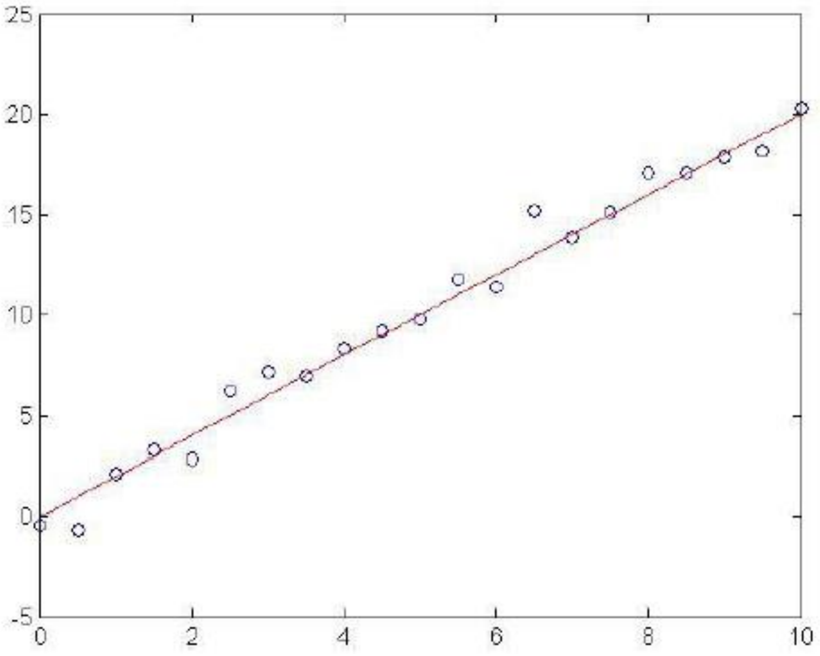
Convex function

- آیا می توان واریانس نویز را تخمین زد؟

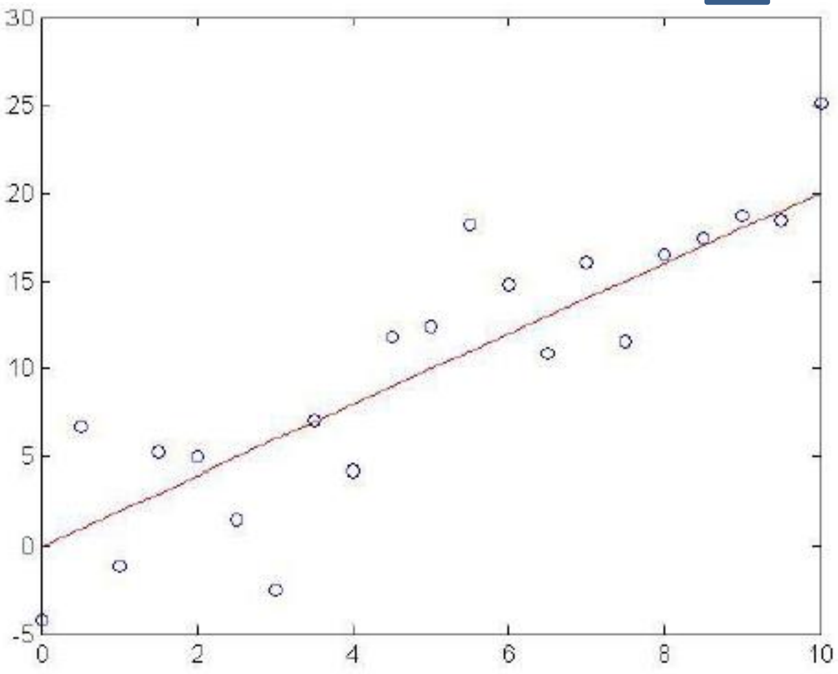
رگرسیون خطی – مثال

- داده ها توسط یک خط با شیب 2 و عرض از مبدا 0 تولید شده اند

$W^* = 2.03, \qquad \sigma = 1$



$W^* = 2.1, \qquad \sigma = 8$



رگرسیون خطی – چند متغیره

- رابطه بیان شده برای رگرسیون خطی تک متغیره، به صورت زیر قابلیت تعمیم دهی به حالت چند متغیره را داراست.

$$y(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_D x_D$$

- پارامترهای مدل در این حالت مشابه حالت اسکالر تخمین زده میشوند:

$$Y = XW + n, Y = \begin{bmatrix} y_1(x) \\ \vdots \\ y_n(x) \end{bmatrix}_{n\times 1}, X = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_D \end{bmatrix}^T_{n\times(D+1)}, W = \begin{bmatrix} w_0 \\ \vdots \\ w_D \end{bmatrix}_{(D+1)\times 1}$$

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

رگرسیون خطی با نویز متغیر

- اگر نویز خروجی اضافه شده دارای مشخصات آماری متغیر با زمان باشد، مدل رگرسیون خطی توسعه داده شده دارای بایاس است. (چرا؟)
  - برای مسئله تک متغیره

$$W^* = \operatorname{argmin} \sum_{i=1}^{N} \frac{(y - WX_i)^2}{\sigma^2} \Longrightarrow W^* = \operatorname{argmin} \sum_{i=1}^{N} \frac{(y - WX_i)^2}{\sigma_i^2}$$

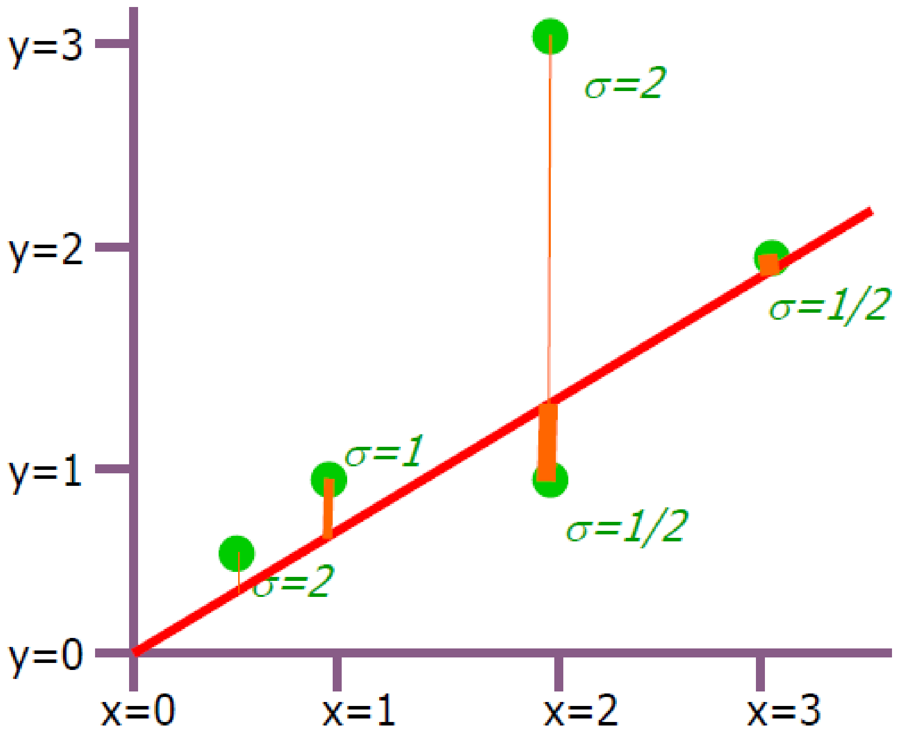
Weighted Least Square (WLS) ⟵ Weighted Sum square error

$$W^* = \frac{\sum_{i=1}^{N} \frac{y_i X_i}{\sigma_i}}{\sum_{i=1}^{N} \frac{|X_i|^2}{\sigma_i}}$$

For Gaussian colored noise

رگرسیون خطی با نویز متغیر

- هر نمونه متناسب با نویزی که می گیرد وزن دهی می شود.



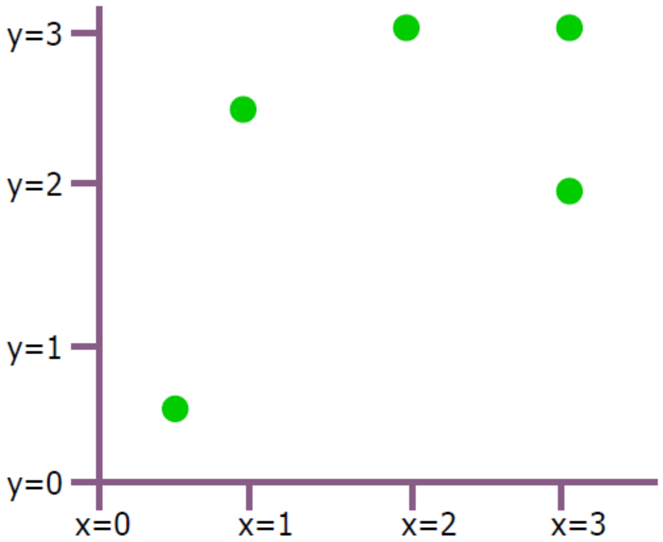
$$W^* = (X^T Q^{-1} X)^{-1} X^T Q^{-1} Y, \qquad Q^{-1} = \begin{pmatrix} \frac{1}{\sigma_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_n^2 \end{pmatrix}$$

For multivariate case

رگرسیون غیرخطی

❑ داده های زیر را در نظر بگیرید:

x_i	y_i
½	½
1	2.5
2	3
3	2
3	3



$$y_i(x_i) = \sqrt{w + x_i} + n_i$$

$$w^* = \arg\min \sum_{i=1}^{N} \left(y_i - \sqrt{w + x_i}\right)^2$$

For white Gaussian noise

❑ حل بسته ریاضی برای مسئله فوق وجود ندارد. (چرا؟) راه حل چیست؟

Numerical Optimization

رگرسیون غیرخطی

- **در بسیاری از موارد تابع هزینه بدست آمده برای مسئله رگرسیون غیرخطی محدب نیست.**

  (با فرض شناخته بودن ارتباط ورودی و خروجی)

- **نگاشت فضای ورودی**
  - ارتباط بین ورودی و خروجی در فضای داده های در دسترس خطی نیست.
  - فضای ورودی داده ها را به فضایی نگاشت می دهیم تا ارتباط بین ورودی و خروجی خطی شود (یا با خط قابل تقریب باشد)

$$y(x) = w_0 + w_1\phi_1(x_1) + w_2\phi(x_2) + \cdots + w_D\phi(x_D)$$

رگرسیون غیرخطی

- **نگاشت فضای ورودی**

$$y(x) = w_0 + w_1\phi_1(x_1) + w_2\phi(x_2) + \cdots + w_D\phi(x_D)$$

  - کلیه مطالب ذکر شده بر اساس فضای ورودی قابل تعمیم دهی به فضای جدید هستند.

- **نگاشت های رایج**
  - چند جمله ای
  - توابع شعاعی پایه
  - تبدیلات مختلف بر روی فضای ورودی (تبدیل فوریه، تبدیل موجک، و ...)

رگرسیون چندجمله ای

- **رابطه بین ورودی و خروجی توسط یک چند جمله ای مشخص می شود**
  - مدل رگرسیون چندجمله ای درجه 2 برای یک متغیر

$$y(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots + w_n x^n$$

- **برای حالت چند متغیر مدل چند جمله ای از بسط نیوتن محاسبه می شود**
  - مدل رگرسیون چندجمله ای درجه 2 برای دو متغیر

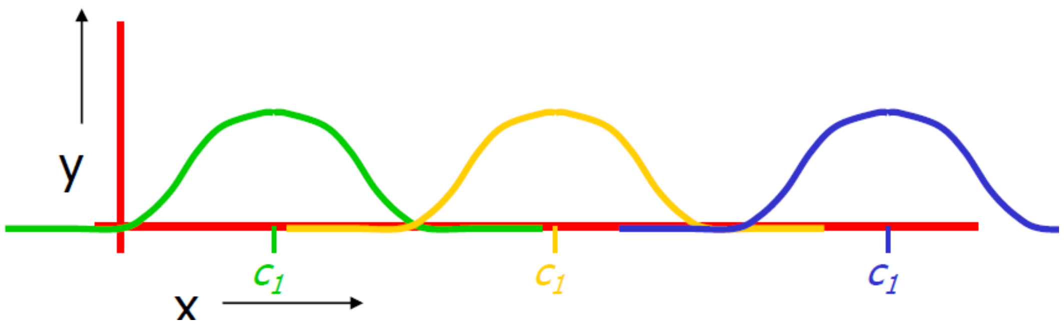
$$y(x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + w_5 x_1 x_2$$

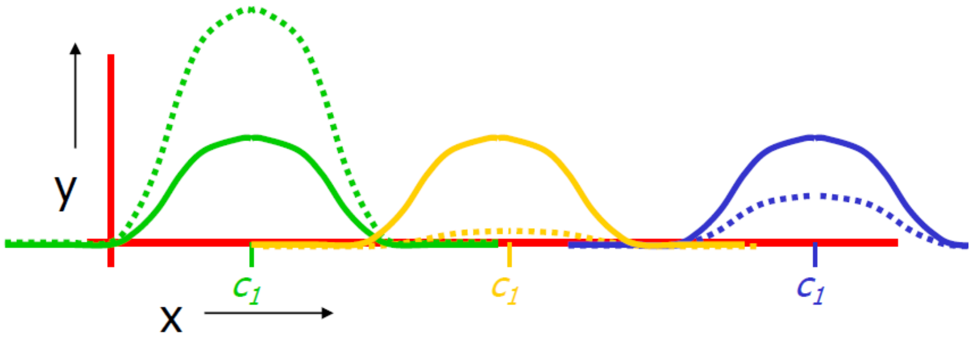
  - آیا می توان از مدلهای فوق برای رگرسیون خطی استفاده کرد؟

توابع پایه شعاعی

- این توابع عمدتا برای سنجش فاصله مورد استفاده قرار می گیرند. بنابراین مدل بدست آمده بر اساس فاصله نقاط فضای ورودی از مرکز این توابع است.

$$y(x) = 2\phi_1(x) + 0.05\phi_2(x) + 0.5\phi_3(x)$$

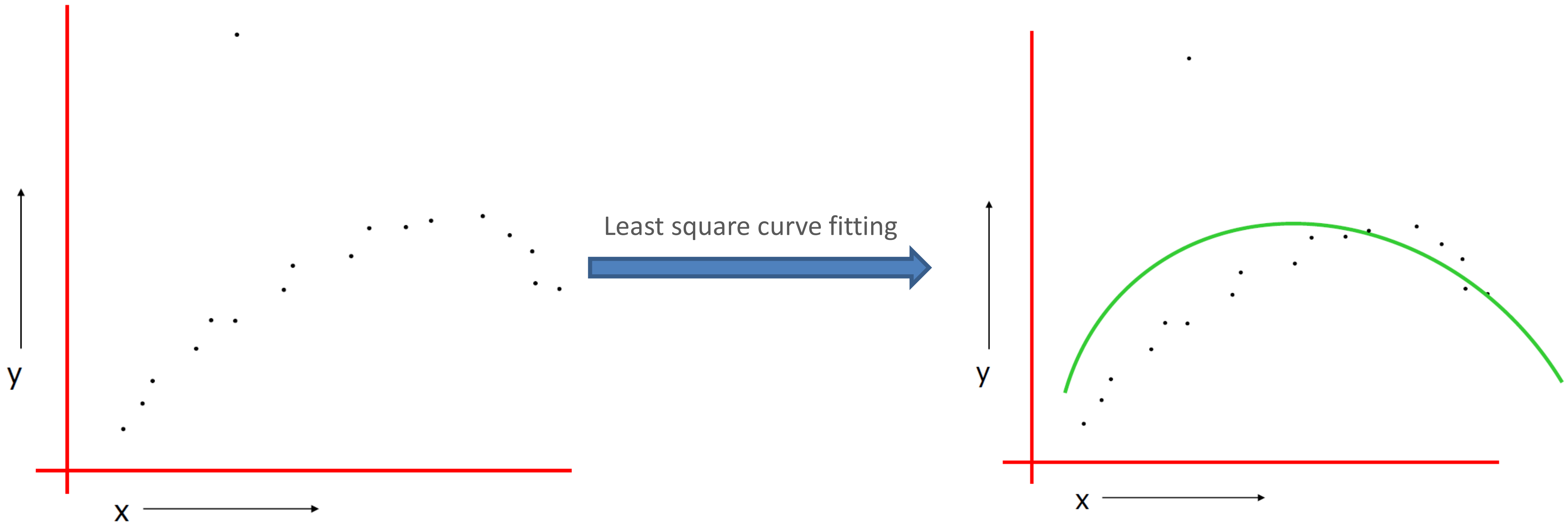




- پارامترهای کرنل چگونه تعیین می شوند؟
- در فضای n بعدی توصیف این توابع به چه شکل است؟

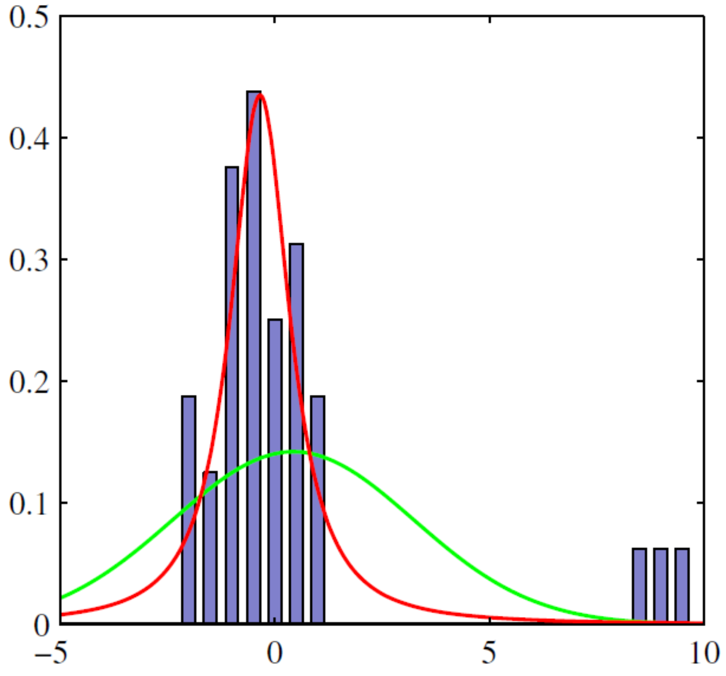
رگرسیون مقاوم

- وجود داده های پرت چه تاثیری بر مدل خواهند داشت؟

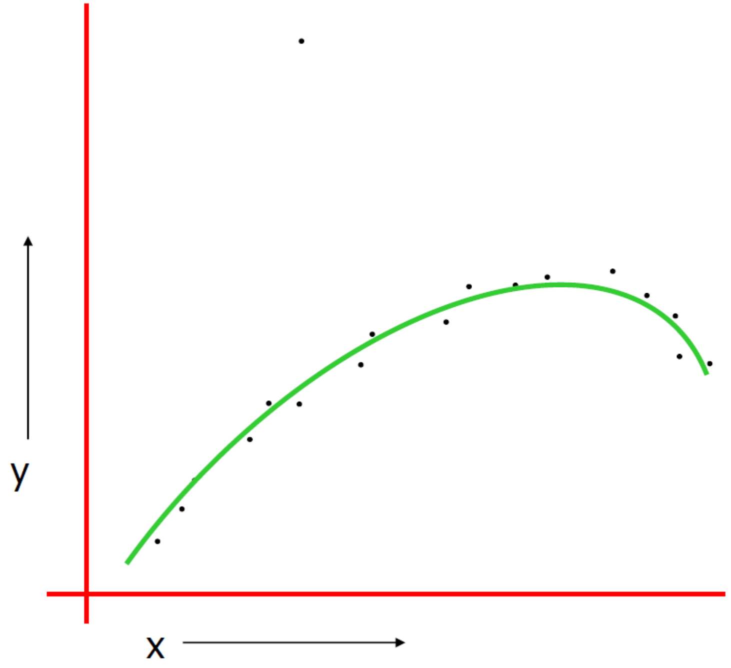


رگرسیون مقاوم

- ❑ مدل سازی رگرسیون بر اساس فرض نرمال بودن نویز انجام شده است که نسبت به داده های پرت حساسیت بالایی دارد.
  - ○ یک راه حل استفاده از توزیع های long tail است.



LS for Laplacian noise



رگرسیون مقاوم

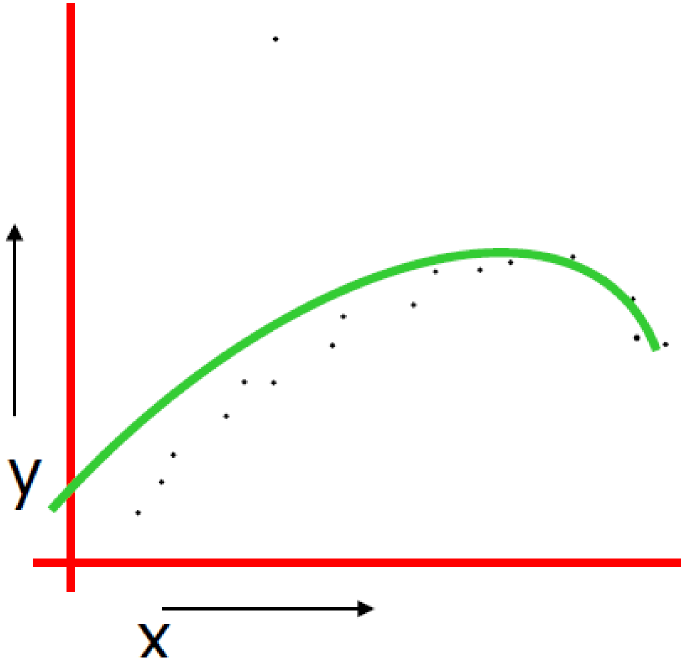
- **مشکلی که فرض نرمال بودن نویز ایجاد می کند، هم ارز کردن تابع بیشینه شباهت با تابع هزینه حداقل مربعات است.**
  - در صورتی که داده پرت وجود داشته باشد، مربع فاصله این داده تا منحنی برازش شده باعث افزایش تابع هزینه خواهد شد.
  - یک راه حل استفاده از تابع هزینه Huber است.
  - حل این مسئله سریعتر از فرض توزیع لاپلاس برای نویز است

$$L_H(r,\delta) = \begin{cases} \frac{r^2}{2}, & |r| \leq \delta \\ \delta|r| - \frac{\delta^2}{2}, & |r| > \delta \end{cases}$$

رگرسیون مقاوم

- رویکرد ساده دیگری نیز وجود دارد که نیازی به تغییر تابع هزینه و یا فرضیات مسئله ندارد.
  - استفاده از کرنل برای وزن دهی بردار خطا

$$\alpha_k = \exp\left((y_k - y_{est})^2\right)$$



- الگوریتم:
  - تخمین حداقل مربعات بر اساس داده های موجود
  - تعیین بردار تخمین خروجی y_{est} بر اساس مدل آموزش دیده شده
  - تشکیل بردار وزن بر اساس بردار تخمین زده شده α_k
  - تکرار مرحله 1 با بردار خروجی $\alpha_k y_k$

بیش برازشی در مدل رگرسیون

- ❑ **همانطور که گفته شد، برای یک مدل رگرسیون بردار پارامترها به نحوی تعیین می شود تا احتمال تولید دنباله ورودی و خروجی بر اساس داده های داده شده بیشینه شود.**

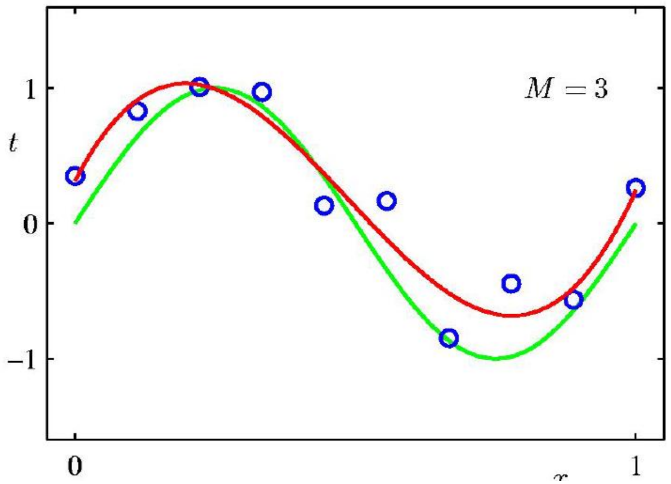
تکرار آزمایش	پیشامد
1	رو
2	رو
. . .	. . .
100	رو

- ○ مسئله پرتاب سکه را در نظر بگیرید:
  - ○ اگر سکه را برای بار 101 ام پرتاب کنیم، بر اساس مشاهدات بدست آمده چه پیشامدی رخ خواهد داد؟
  - ○ آیا نتیجه آزمایش با دانش پیشین ما از پرتاب سکه همخوانی دارد؟
- ○ برای جلوگیری از مسئله بیش برازش، نیاز داریم تا عدم قطعیت ناشی از آزمایش را نیز در مدل سازی دخیل کنیم.

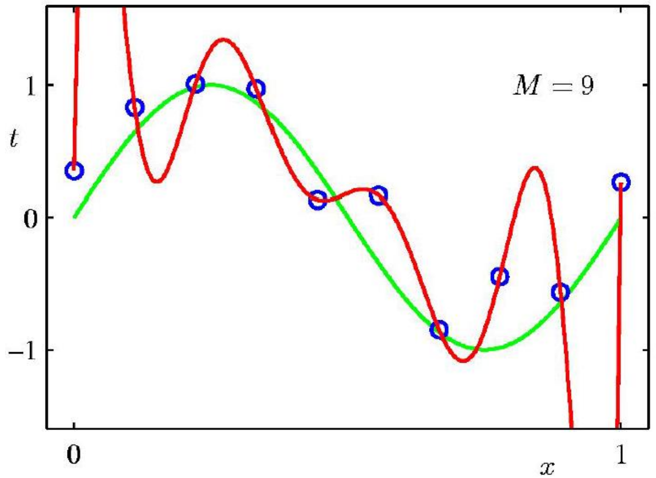
بیش برازشی در مدل رگرسیون

- در مدل های یادگیری ماشین بیش برازشی به دلایل زیر اتفاق می افتد:
  - تعداد ناکافی داده ها در فاز آموزش مدل
  - پیچیدگی مدل

رگرسیون چند جمله ای با درجه 3



رگرسیون چند جمله ای با درجه 9



از دیدگاه آماری، بیش برازشی یک مدل، هم ارز افزایش واریانس تخمین است

تنظیم کنندگی مدل Regularization

- **افزودن نویز خروجی**
  - این اقدام هم ارز لحاظ کردن مقدار عدم قطعیت در برچسب داده هاست.

- **جمع آوری داده**
  - داده های حجیم نقش تنظیم کنندگی در مدل را دارند.
  - در شرایطی که داده های زیادی در دست باشد، رویکرد بیشینه شباهت قابلیت تعمیم دهی مدل را افزایش می دهد

تنظیم کنندگی مدل Regularization

- رویکرد بیزین

$$y(x) = W^T\phi(x) + n \rightarrow \underbrace{P(y(x)|W)}_{\text{Evidence}}, \underbrace{P(W)}_{\text{Prior}}$$

  - توزیع پارامترها هر توزیع دلخواهی می تواند باشد
(بر اساس دانش ما از مسئله تعیین می شود)

$$P(Y,W) = P(Y|W)P(W) \rightarrow W^* = \arg\max \log P(Y,W)$$ Maximum a Posterior

- بهترین توزیع پیشین
  - خانواده توزیع های نمایی

تنظیم کنندگی مدل Regularization

- ❑ تنظیم کنندگی بر اساس نرم - 2 پارامترها

$$y(x) = W^T\phi(x) + n \rightarrow P(y(x)|W) \sim N(W^T\phi(x), \sigma^2), P(W) \sim N(\mu, \sigma_w^2)$$

- ○ اگر میانگین توزیع پیشین را صفر در نظر بگیریم:

$$W^* = \operatorname{argmin} \sum_{i=1}^{N} (y - WX_i)^2 + \underbrace{\frac{\sigma^2}{\sigma_w^2}}_{\lambda} |W|^2$$

L2 Regularization (Ridge Regression)

- ○ مسئله فوق را می توان به صورت یک مسئله بهینه سازی مقید بیان کرد (چگونه؟)
- ○ رابطه بردار وزن ها چگونه محاسبه می شود؟

تنظیم کنندگی مدل Regularization

- تنظیم کنندگی بر اساس نرم - 1 پارامترها

$$y(x) = W^T\phi(x) + n \rightarrow P(y(x)|W) \sim N(W^T\phi(x), \sigma^2)$$

- اگر توزیع پیشین پارامترها را یک توزیع لاپلاس با مقدار پارامتر location صفر فرض کنیم:

$$W^* = \operatorname{argmin} \sum_{i=1}^{N} (y - WX_i)^2 + \lambda|W|$$

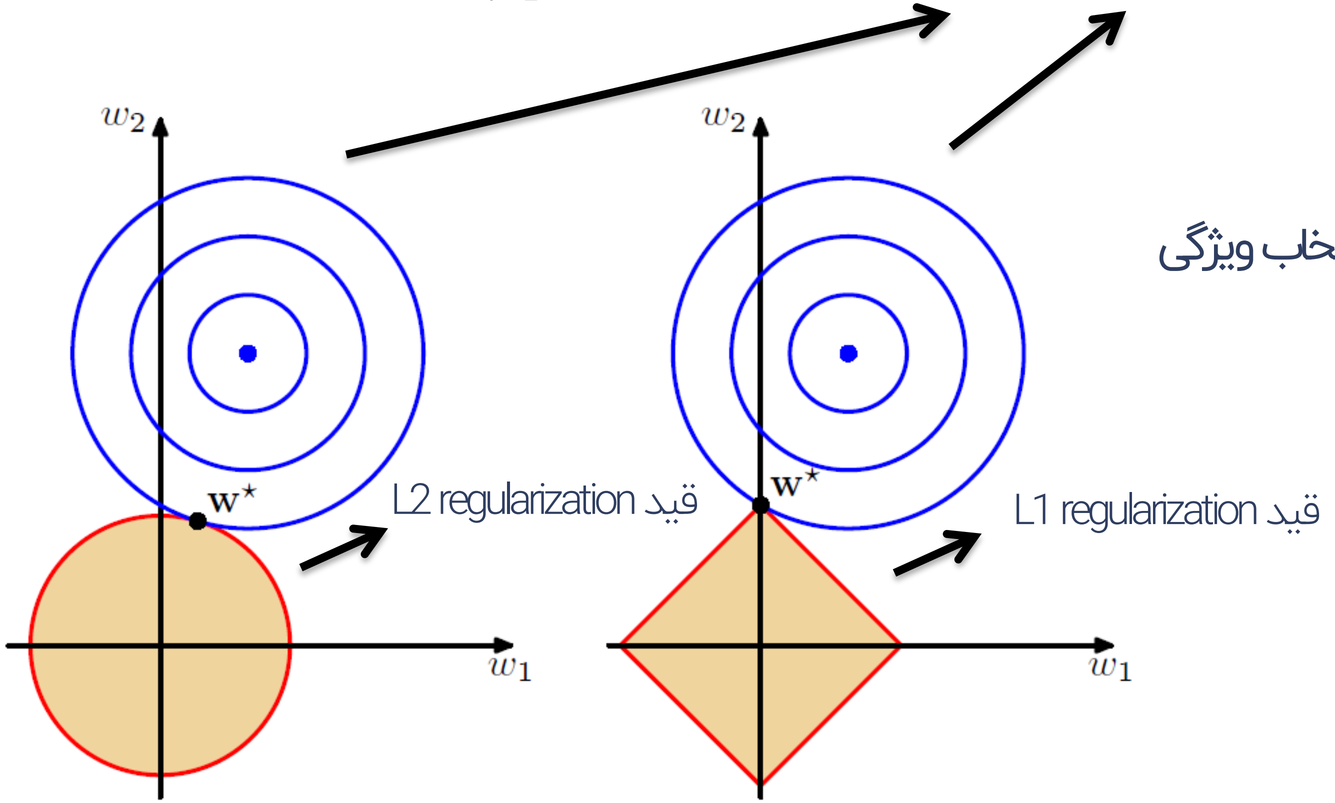
L1 Regularization

- مسئله فوق را می توان به صورت یک مسئله بهینه سازی مقید بیان کرد (چگونه؟)

تنظیم کنندگی – دید هندسی

$$W^* = \operatorname{argmin} \sum_{i=1}^{N} (y - WX_i)^2$$

کانتورهای سطوح ثابت تابع هزینه (؟)



- ❑ کدام مدل بهتر است؟
- ❑ از کدام مدل می توان به عنوان مدل انتخاب ویژگی استفاده کرد؟

حداقل مربعات گام به گام

- پاسخ مسئله حداقل مربعات چند متغیره به تابع هزینه حداقل مربعات به صورت زیر محاسبه شد:

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

- اگر ابعاد ماتریس X بزرگ باشد، اپراتور معکوس گیری با مشکلات زمانی جدی موجه خواهد شد. راه حل چیست؟
  - الگوریتم های آنلاین آموزش – گرادیان نزولی
  - پاسخی که الگوریتم گرادیان نزولی برای مسئله رگرسیون با فرضیات داده شده بدست می دهد یک بهینه سراسری است (چرا؟)