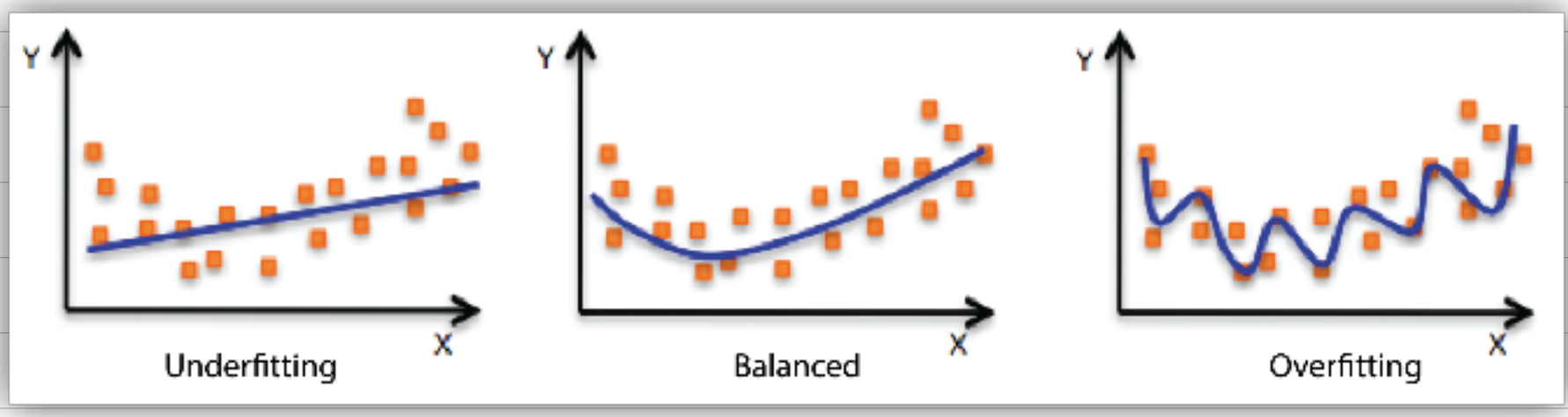
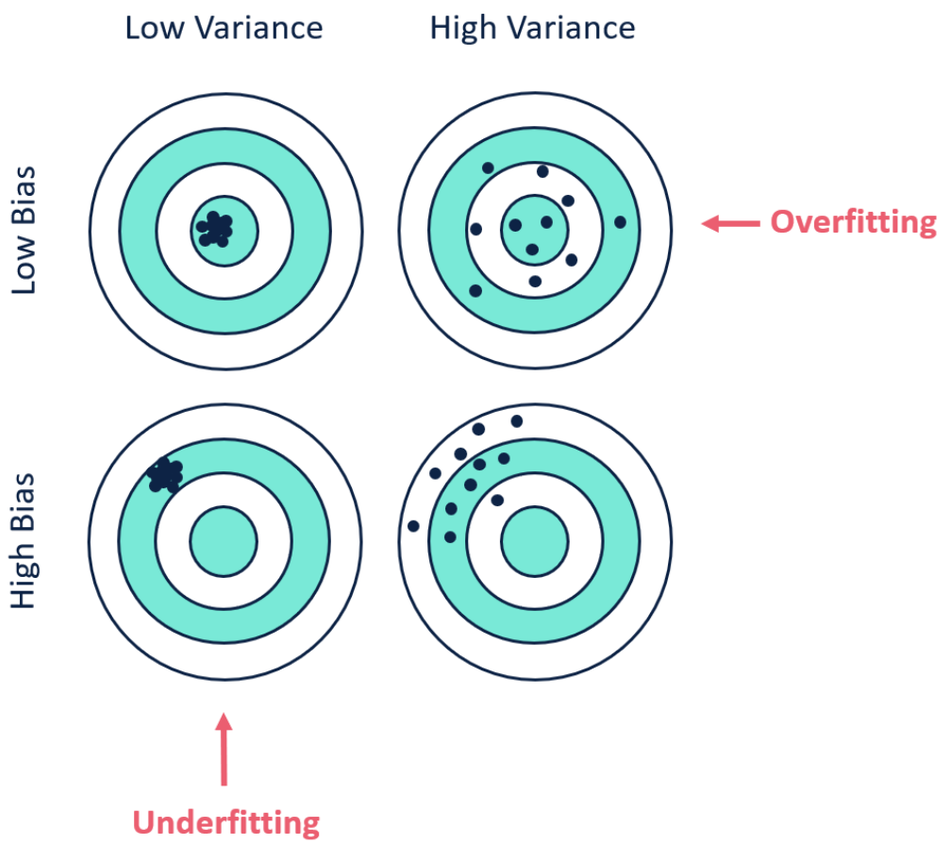
Session 6

Regularization





biase: عدم دقت حاصل از تخمین یک مسأله ی واقعی

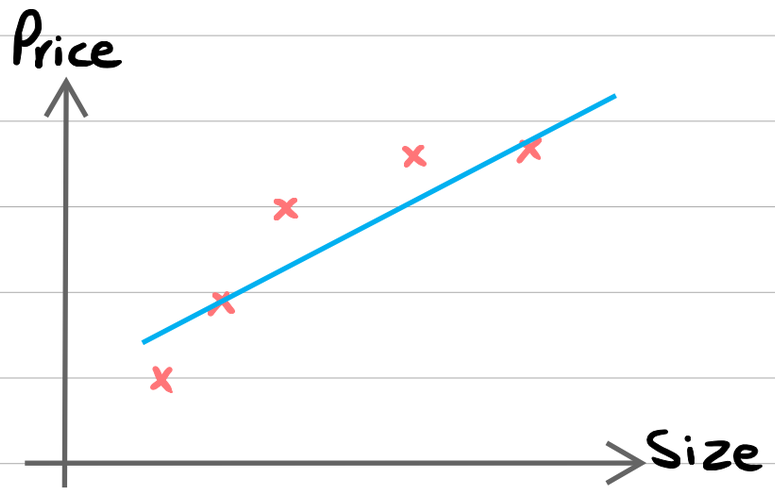
با پیچیدگی زیاد با یک مدل ساده.

variance: اگر $f_{w,b}(x)$ را با یک مجموعه داده train

متفاوت تخمین بزنیم، چقدر تغییر می‌کند.

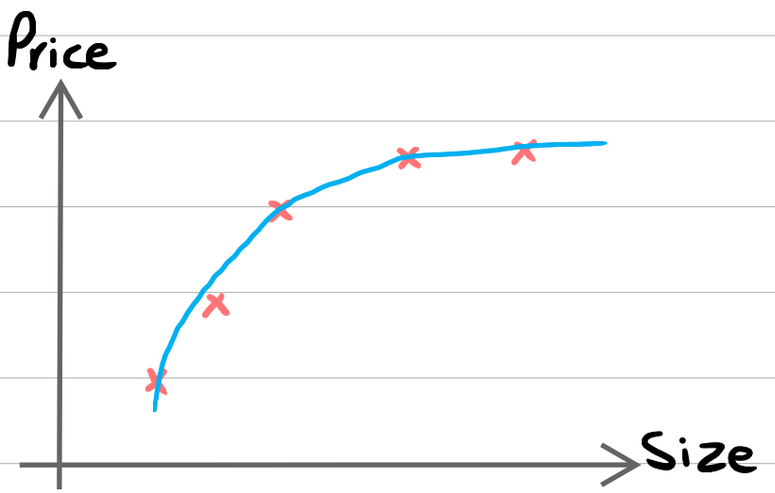
$$Error = var\ (f(x)) + Bias\ (f(x))^2 + error(\varepsilon)$$

Regression



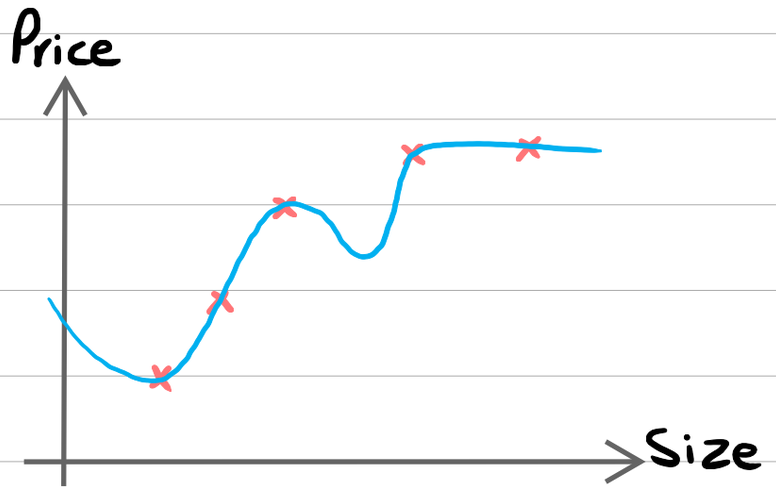
$w_1 x + b$

Underfit ⇝ high bias



$w_1 x + w_2 x^2 + b$

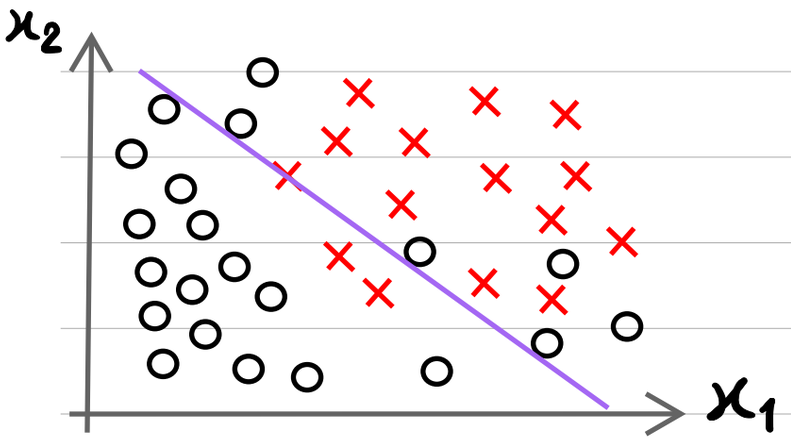
Just right ⇝ generalization



$w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + b$

Overfit ⇝ high variance

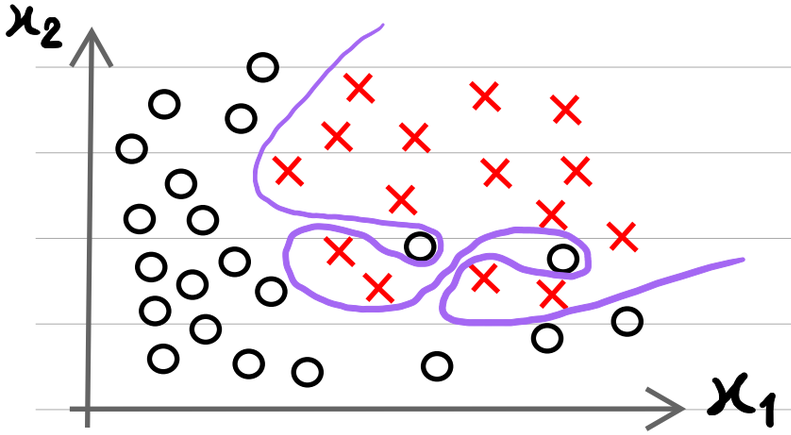
Classification

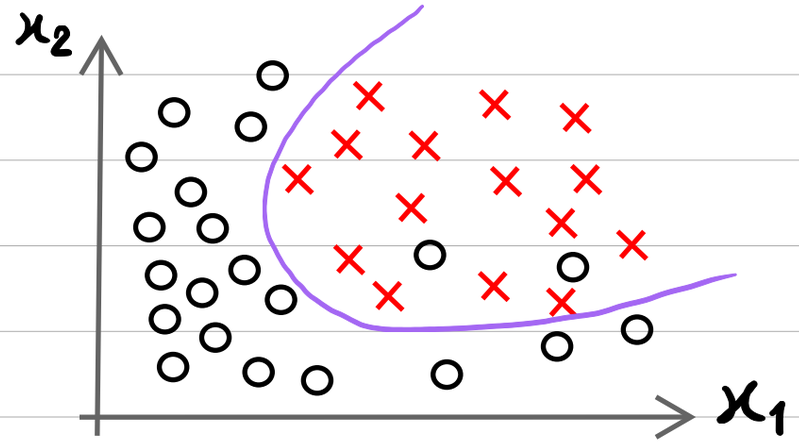


$$z = w_1 x_1 + w_2 x_2 + b$$

$$f_{w,b}(x) = g(z) \rightsquigarrow \text{g:sigmoid}$$

Underfit $\rightsquigarrow$ high bias





$$z = w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2$$

$$+ w_5 x_1 x_2 + b$$

Just right $\rightsquigarrow$ generalization

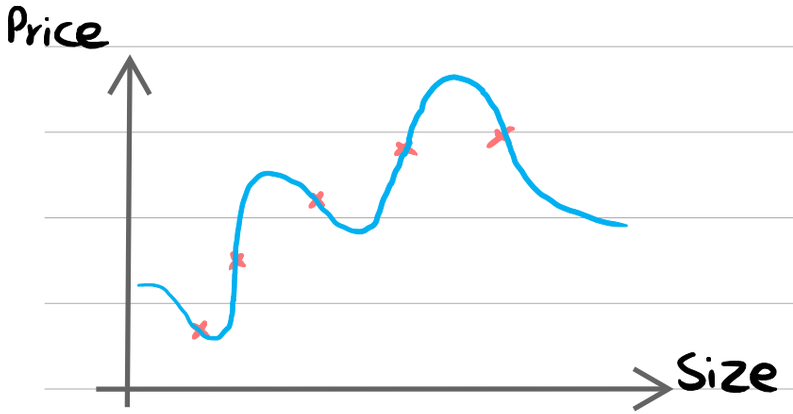
$$z = w_1 x_1 + w_2 x_2 + w_3 x_1^2 x_2 + w_4 x_1^2 x_2^2$$

$$+ w_5 x_1^2 x_2^3 + w_6 x_1^3 x_2 + \cdots + b$$

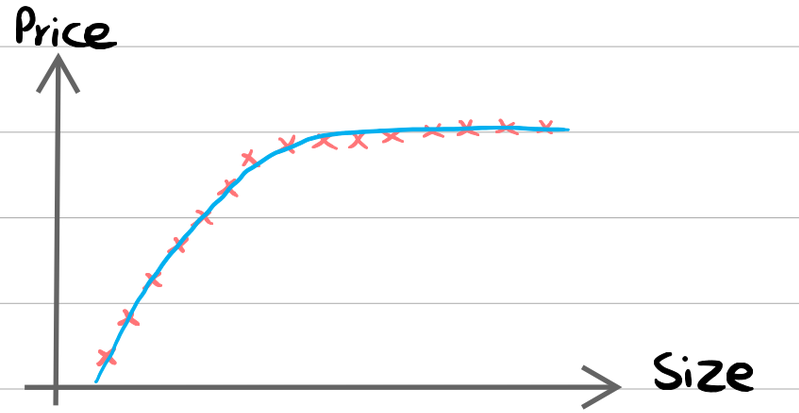
Overfit $\rightsquigarrow$ high variance

1. Collect more training examples

ح: برای رفع مشکل overfit چه کنیم؟



overfit



Good fit

2. Select features to include/exclude

x_1	x_2	x_3	x_4	x_5		x_{100}	y
Size	bedroom	floors	age	avg income	...	distance coffee shop	Price

all features + Insufficient sample ⟿ overfit

کافی نبودن نمونه‌ها

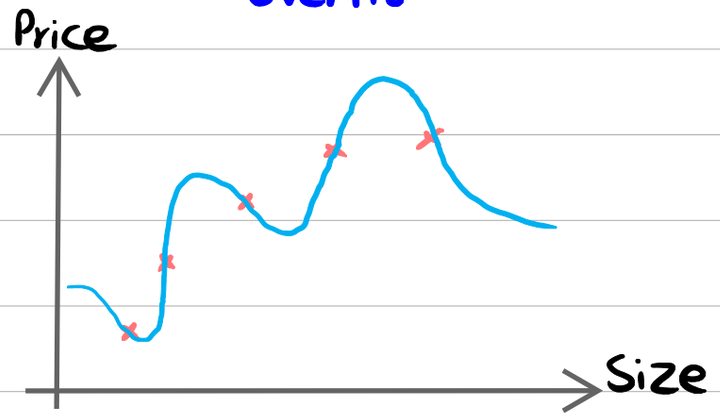
selected features

③ → Size, #bedrooms, age

ممکن است ویژگی‌های مفید حذف شوند

3. Reduce the size of parameters W_j

Overfit

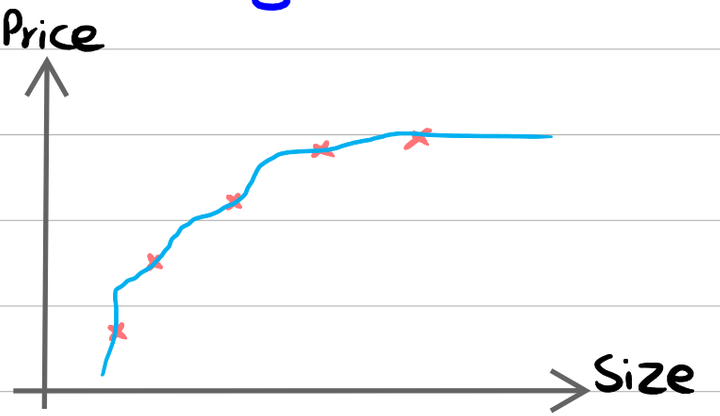


$f(x) = 28x - 385x^2 + 39x^3$

$-174x^4 + 100$

Large values for W_j

Regularization



$f(x) = 13x - 0.23x^2 + 0.0000014x^3$

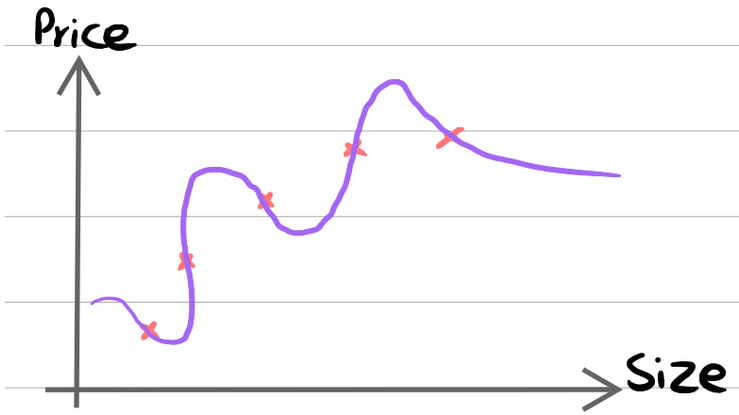
$-0.0001x^4 + 10$

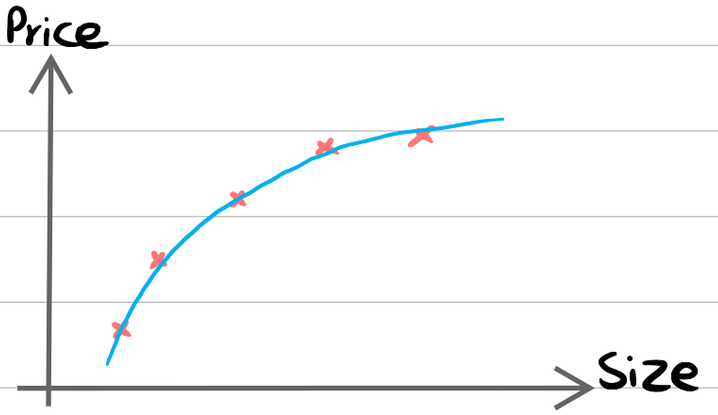
Small values for W_j

Regularization:

Reg. باعث کاهش خطای شود و همچنین می‌تواند از underfit و overfit جلوگیری کند. ← چه طور؟

کمک می‌کند ضرایب را کوچک یا حذف کنیم. ← همچنین می‌تواند روش خوبی برای انتخاب ویژگی باشد، چه طور؟





$$w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + b$$

$$w_1 x + w_2 x^2 + \underbrace{w_3 x^3}_{\approx 0} + \underbrace{w_4 x^4}_{\approx 0} + b$$

Make w_3, w_4 really small (≈ 0)

$$\min \left[\frac{1}{2m} \sum_{i=1}^{m} \left(f_{w,b}(x^{(i)}) - y^{(i)} \right)^2 \right]$$

$$\underbrace{1000\, w_3^2}_{1} + \underbrace{1000\, w_4^2}_{2}$$

0.001 ← w_3^2

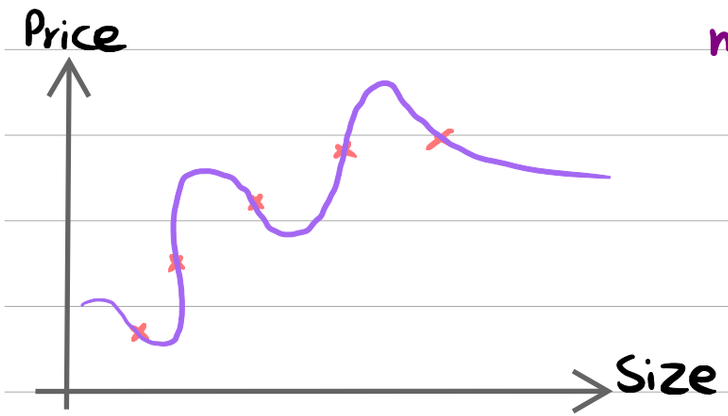
0.002 ← w_4^2

x_1	x_2	x_3	x_4	x_5		x_{100}	y
Size	bedroom	floors	age	avg income	...	distance coffee shop	Price

$w_1, w_2, w_3, \dots, w_{100}, b$ $\quad n = 100$

$$J(w,b) = \frac{1}{2m} \sum_{i=1}^{m} \left(f_{w,b}(x^{(i)}) - y^{(i)} \right)^2 + \frac{\lambda}{2m} \sum_{j=1}^{n} w_j^2 + \frac{\lambda}{2m} b^2$$

λ: lambda Regularization Parameter



$$\min J(w,b) = \min \left[\frac{1}{2m} \sum_{i=1}^{m} \left(f_{w,b}(x^{(i)}) - y^{(i)} \right)^2 + \frac{\lambda}{2m} \sum_{j=1}^{n} w_j^2 \right]$$

$w, b = ?$

Gradient Descent:

repeat

{

$$w_j = w_j - \alpha \left[\frac{\partial}{\partial w_j} J(w,b) \right] = \frac{1}{m} \sum_{i=1}^{m} \left(f_{w,b}(x^{(i)}) - y^{(i)} \right) x_j^{(i)} + \frac{\lambda}{m} w_j$$

Regularizer term

$$b = b - \alpha \left[\frac{\partial}{\partial b} J(w,b) \right] = \frac{1}{m} \sum_{i=1}^{m} \left(f_{w,b}(x^{(i)}) - y^{(i)} \right)$$

}

$$\frac{\partial}{\partial w_j} \left[\frac{1}{2m} \sum_{i=1}^{m} \left(\underbrace{f_{w,b}(x^{(i)})}_{w \cdot x^{(i)} + b} - y^{(i)} \right)^2 + \frac{\lambda}{2m} \sum_{j=1}^{n} w_j^2 \right] = \frac{1}{2m} \sum_{i=1}^{m} \left[(wx^{(i)} + b - y^{(i)}) \times 2 \times x_j^{(i)} \right] + \sum_{j=1}^{n} \frac{\lambda}{2m} \times 2 \, w_j$$

$$= \frac{1}{m} \sum_{i=1}^{m} \left[\underbrace{(w \cdot x^{(i)} + b}_{f_{w,b}(x)} - y^{(i)}) x_j^{(i)} \right] + \sum_{i=1}^{m} \frac{\lambda}{m} w_j = \frac{1}{m} \sum_{i=1}^{m} \left[(f_{w,b}(x^{(i)}) - y^{(i)}) x_j^{(i)} \right] + \frac{\lambda}{m} w_j$$

$$w_j = w_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(f_{w,b}(x^{(i)}) - y^{(i)}\right) x_j^{(i)} + \frac{\lambda}{m} w_j$$

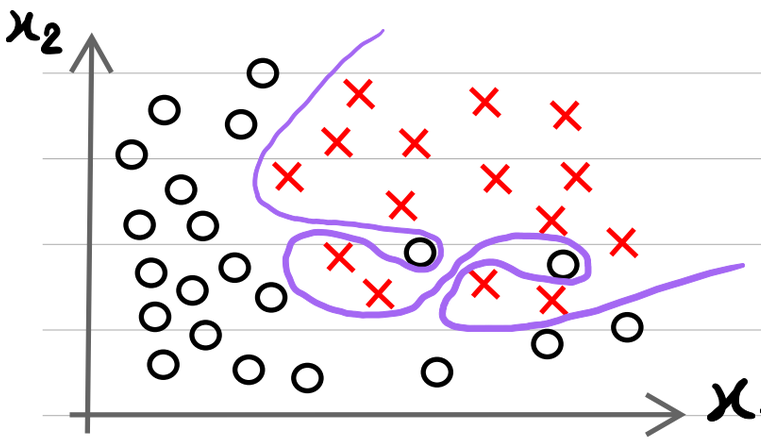
$$w_j = \underbrace{w_j - \alpha \frac{\lambda}{m} w_j}_{w_j(1-\alpha\frac{\lambda}{m})} - \underbrace{\alpha \frac{1}{m} \sum_{i=1}^{m} \left(f_{w,b}(x^{(i)}) - y^{(i)}\right) x_j^{(i)}}_{\text{usual update}}$$

↳ $\alpha = 0.01$, $\lambda = 1$ $\rightsquigarrow$ $\alpha \frac{\lambda}{m} = 0.01 \times \frac{1}{50} = 0.0002$

$m = 50$

$w_j (1 - 0.0002) \rightarrow 0.9998$

Classification ⟶ Logistic Regression



Overfit ⟿ high variance ↕

$$z = w_1 x_1 + w_2 x_2 + w_3 x_1^2 x_2 + w_4 x_1^2 x_2^2$$

$$+ w_5 x_1^2 x_2^3 + w_6 x_1^3 x_2 + \cdots + b$$

$$J(w,b) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log\left(f_{w,b}(x^{(i)})\right) + (1-y^{(i)}) \log\left(1 - f_{w,b}(x^{(i)})\right) \right] + \frac{\lambda}{2m} \sum_{j=1}^{n} w_j^2$$

↳ $\min J(w,b)$ ⟿ w_j ↓

GD:

repeat

{

$$w_j = w_j - \alpha \left[\frac{\partial}{\partial w_j} J(w,b) \right] \rightarrow = \frac{1}{m} \sum_{i=1}^{m} \left(f_{w,b}(x^{(i)}) - y^{(i)} \right) x_j^{(i)} + \frac{\lambda}{m} w_j$$

Logistic Regression $\rightarrow f_{w,b}(x) = \frac{1}{1+e^{-wx+b}}$

$$b = b - \alpha \left[\frac{\partial}{\partial b} J(w,b) \right] \rightarrow = \frac{1}{m} \sum_{i=1}^{m} \left(f_{w,b}(x^{(i)}) - y^{(i)} \right)$$

}

Regularization

- Ridge — $L2 = \sum_{j=0}^{n} w_j^2$
- Lasso — $L1 = \sum_{j=0}^{n} |w_j|$
- Elastic Net — L1, L2

alpha ⟶ α

Ridge ⟶ همه ضرایب کوچک می شوند.

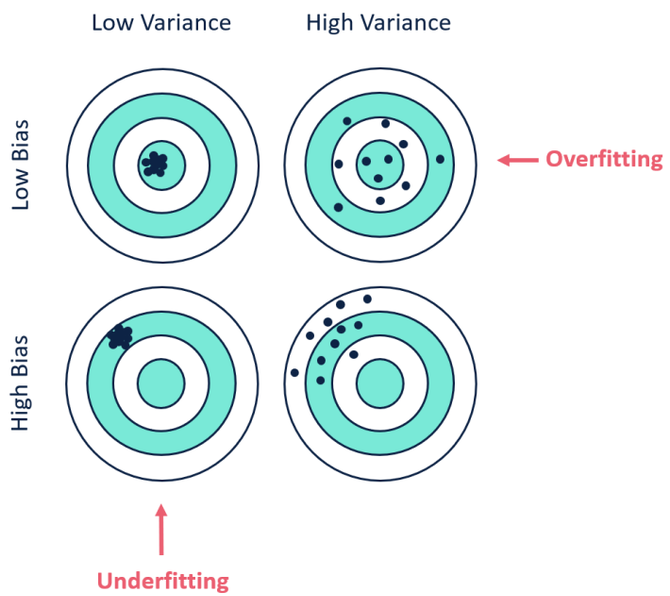
$$Loss = \sum_{i=1}^{m} \left(f_{w,b}(x) - y^{(i)}\right)^2 + \lambda \sum_{j=0}^{n} w_j^2$$

$\alpha = 0$ ⟶ Ridge ✗

$\alpha = \infty$ ⟶ همه ضرایب کوچک و نزدیک به صفر می شوند.

$\alpha \uparrow$ bias $\uparrow$ variance $\downarrow$

$\alpha \downarrow$ bias $\downarrow$ variance $\uparrow$



Lasso:

ضرایب را کوچک می‌کند و برخی از ضرایب را به ۰ می‌رساند.

$$Loss = \sum_{i=1}^{m} \left(f_{w,b}(x) - y^{(i)}\right)^2 + \lambda \sum_{j=0}^{n} |w_j|$$

$\alpha = 0 \longrightarrow$ هیچ وزنی حذف نمی‌شود.

$\alpha = \infty \longrightarrow$ تمام وزن‌ها را حذف می‌کند.

Elastic Net:

ترکیبی از lasso و Ridge

↳ L1, L2

$$Loss = \frac{1}{2m} \sum_{i=1}^{m} \left(f_{w,b}(x) - y_i\right)^2 + (1-\lambda) \times \frac{\alpha}{2} \times \sum_{i=1}^{n} \underbrace{w_j^2}_{L2} + \lambda\alpha \underbrace{|w_j|}_{L1}$$

L1_rate = 1 $\longrightarrow$ L1 → lasso

L1_rate = 0 $\longrightarrow$ L2 → Ridge

x	y
1	2
2	3
3	5

$$\hat{y} = wx + b$$

$$Loss = MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

$w=1.5$, $b=0.5$

$$Loss = Reg.\ MSE = MSE + \lambda \sum_{j=1}^{m} w_j^2$$

$\lambda = 0.1$

$$\hat{y}_1 = wx + b = 1.5 \times 1 + 0.5 = 2 \qquad \hat{y}_2 = 1.5 \times 2 + 0.5 = 3.5 \qquad \hat{y}_3 = 1.5 \times 3 + 0.5 = 5$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y - \hat{y})^2 = \frac{1}{3}\left[(2-2)^2 + (3-3.5)^2 + (5-5)^2\right] = \frac{1}{3}\left[0 + 0.25 + 0\right] = 0.083$$

$$Reg.\ MSE = MSE + \lambda w^2 = 0.083 + (0.1 \times (1.5)^2) = 0.308$$

Reg. MSE > MSE

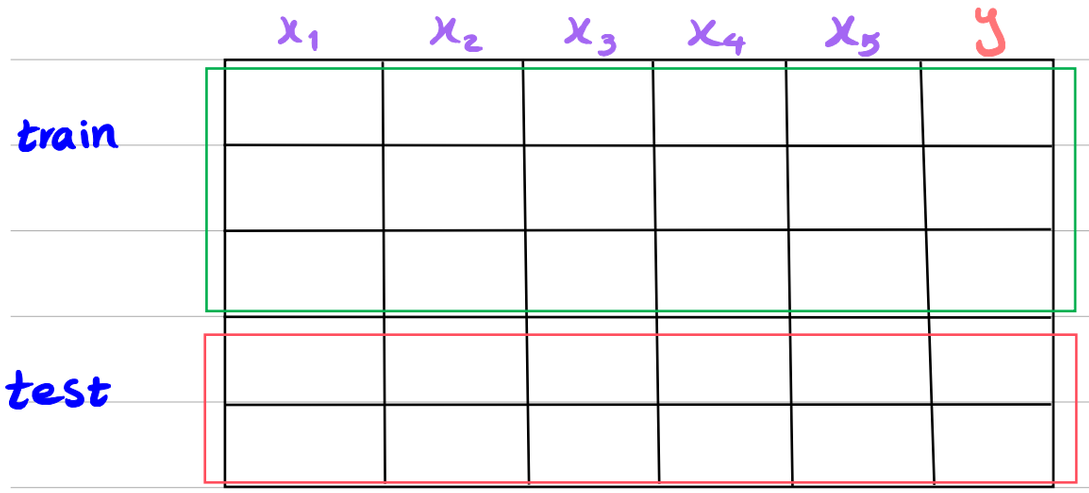
برای کاهش MSE چه باید کرد؟

اگر w را کاهش دهیم MSE کاهش می‌یابد و از طرفی با کاهش w مدل از بیش برازش نجات می‌یابد.

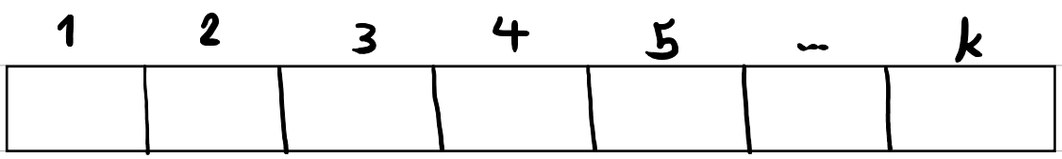
* Reg. یک تعادل ایجاد می‌کند بین fit-well و وزنهای کوچک که می‌تواند باعث Generalization بهتر داده های دیده نشده شود.

Cross Validation

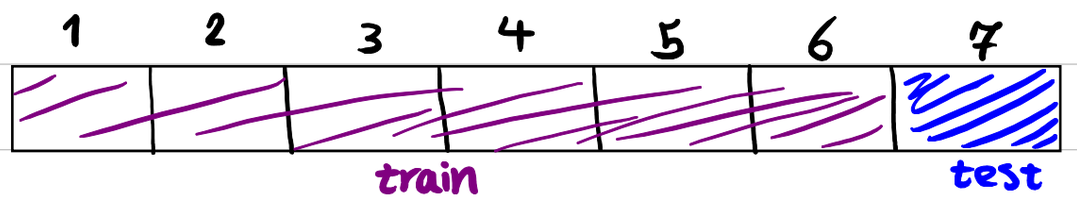
؟: آیا روشی برای پیدا کردن هایپر پارامتر (α) مناسبی هست؟

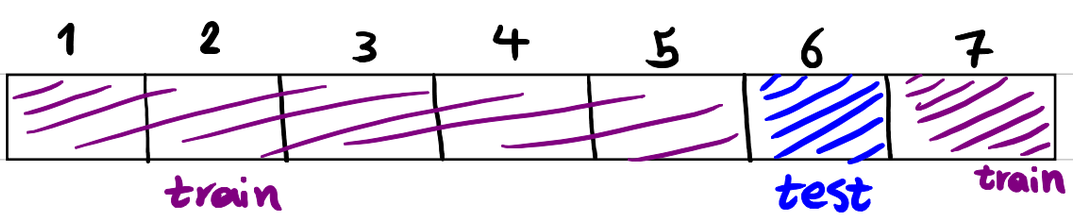


k-fold:

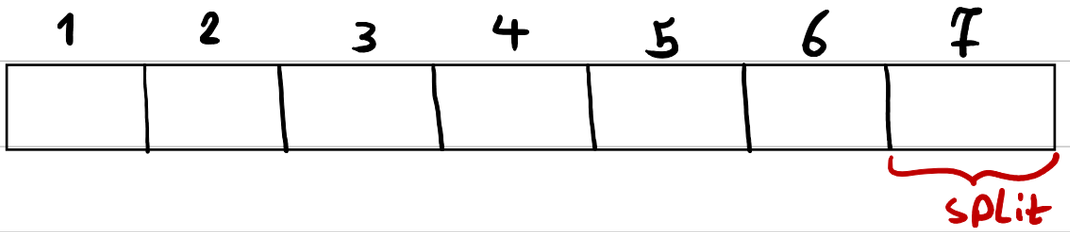


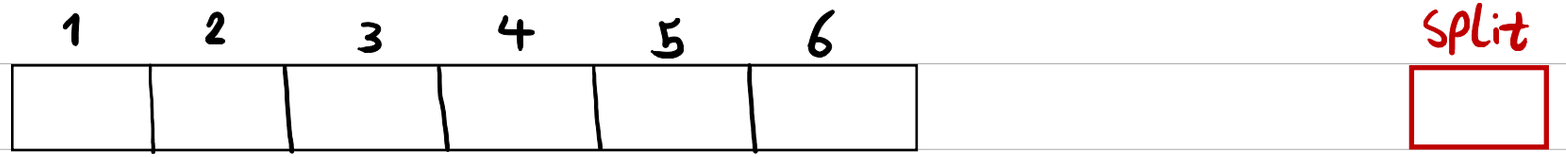
k=7 , CV=2

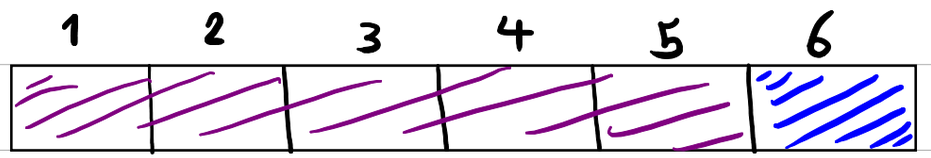




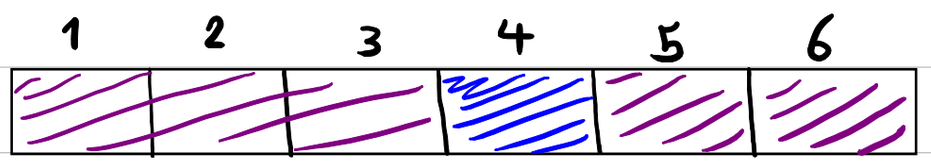
Mean error

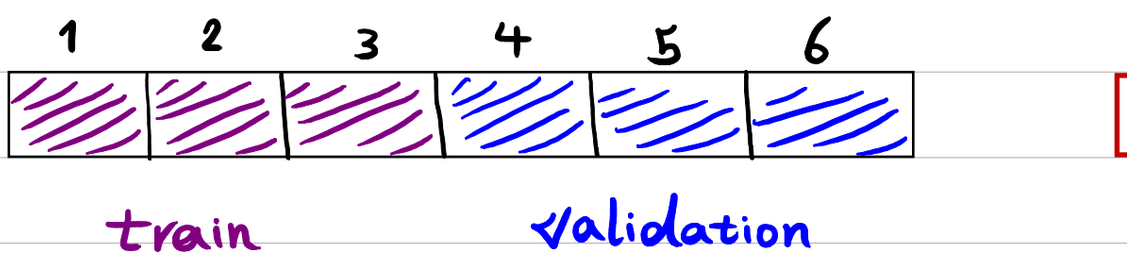

1
2
3
4
5
6
7
split


1
2
3
4
5
6
Split


1
2
3
4
5
6


1
2
3
4
5
6


1
2
3
4
5
6


1
2
3
4
5
6
train
validation
test

x	y
1	2
2	4
3	6
4	8
5	10

K-fold cross-validation (k = 5)

① | 1 | 2 | 3 | 4 | 5 | → MSE = 0

test → x = [1], y = [2]

train → x = [2, 3, 4, 5], y = [4, 6, 8, 10]

② | 1 | 2 | 3 | 4 | 5 | → MSE ≃ 79 × 10^-31

③ | 1 | 2 | 3 | 4 | 5 | → MSE = 0

④ | 1 | 2 | 3 | 4 | 5 | → MSE = 0

⑤ | 1 | 2 | 3 | 4 | 5 | → MSE = 0

$$avg\_mse = \frac{1}{m} \sum^{k} MSE = 1.5 \times 10^{-31}$$