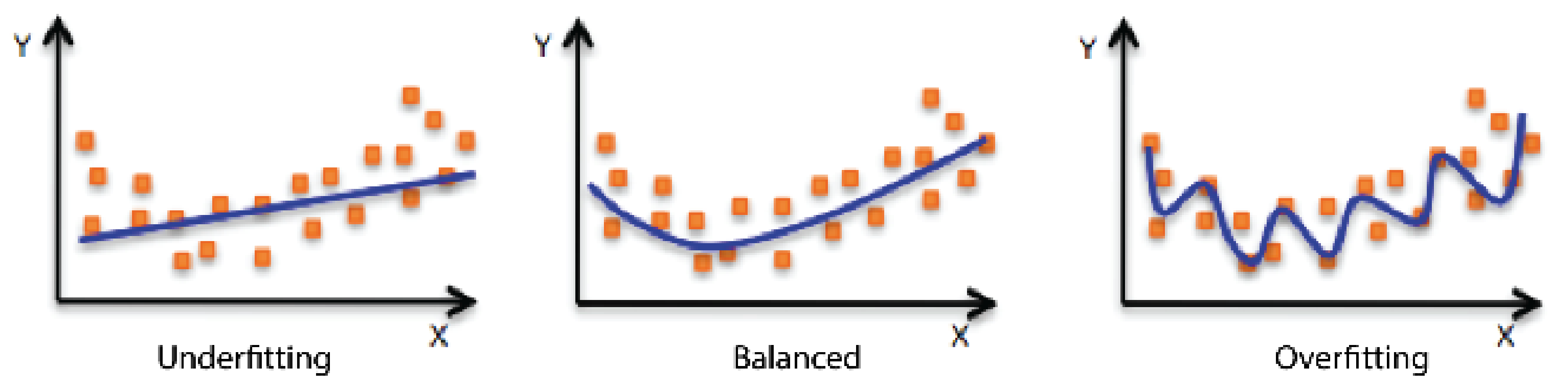
Session 13

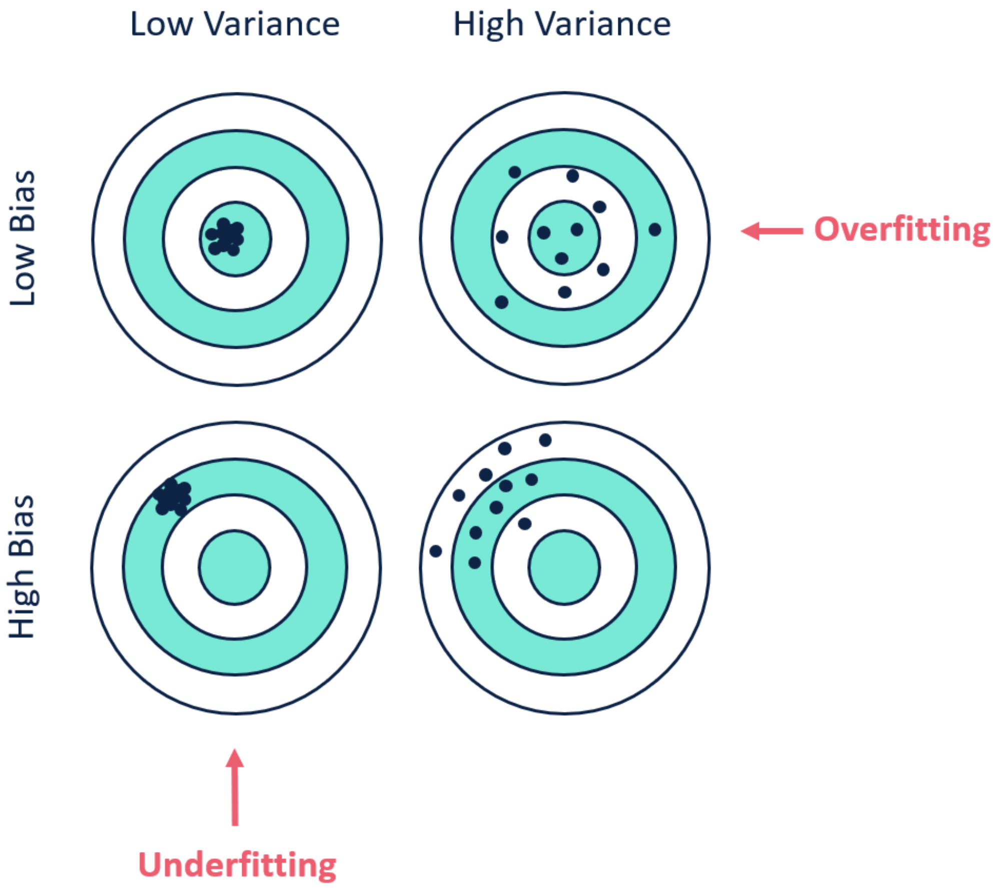
L1, L2 Regularization

Machine Learning | Hobot | Zahra Amini

Telegram: @zahraamini_ai & Instagram:@zahraamini_ai & LinkedIn: @zahraamini-ai

https://zil.ink/zahraamini





biase: عدم دقت حاصل از تخمین یک مسأله ی واقعی

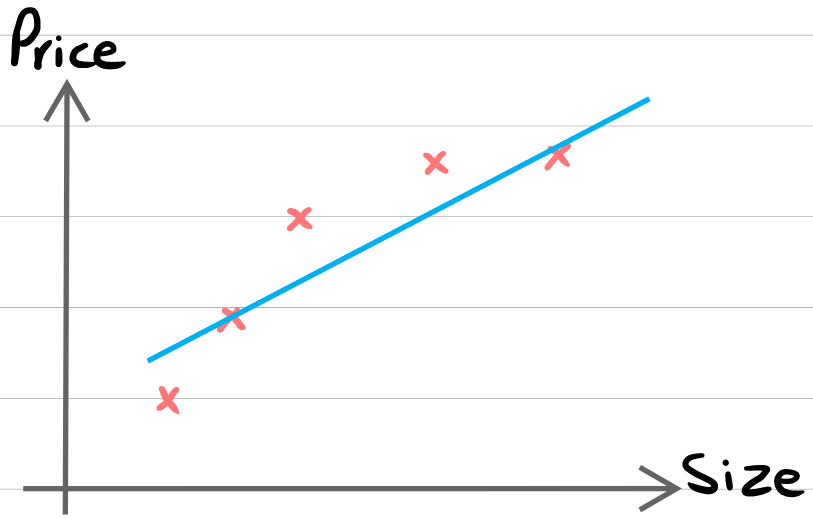
با پیچیدگی زیاد با یک مدل ساده.

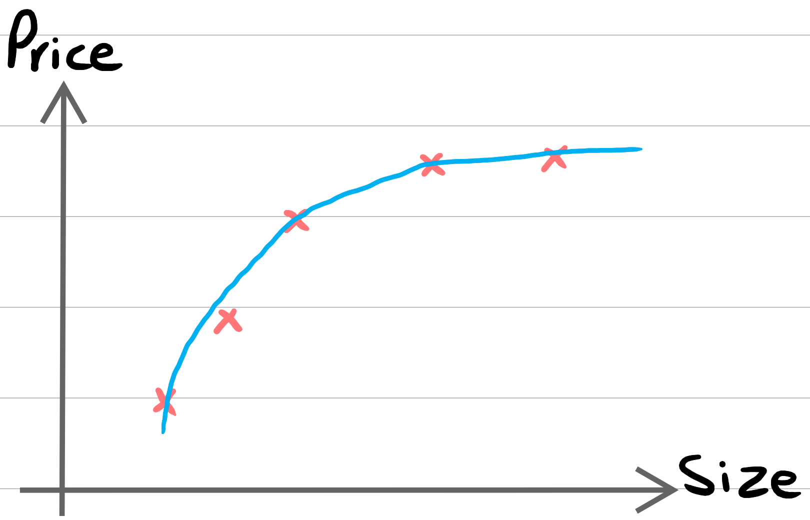
variance: اگر $f_{w,b}(x)$ را با یک مجموعه داده train

متفاوت تخمین بزنیم، چقدر تغییر می کند.

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

Regression



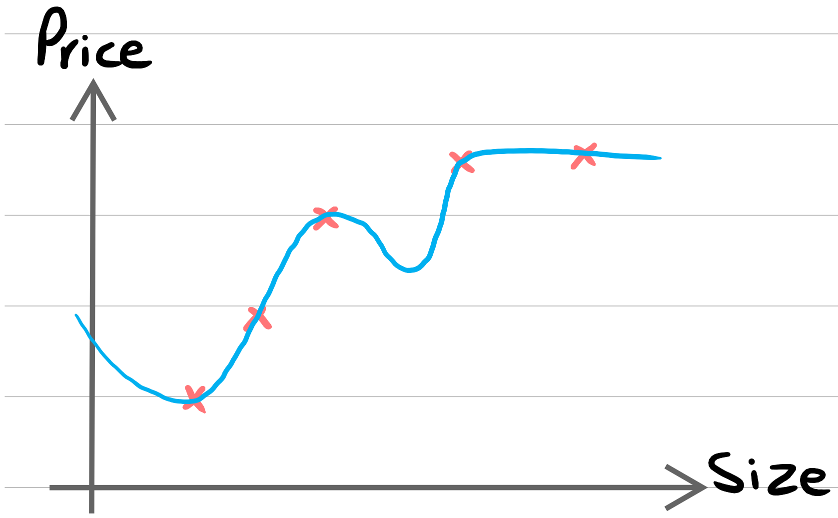


$w_1 x + b$

$w_1 x + w_2 x^2 + b$

Underfit ⇝ high bias ↑

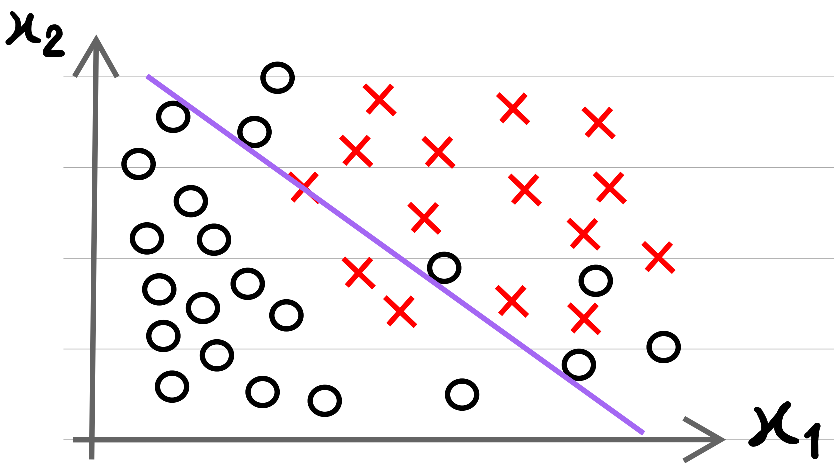
Just right ⇝ generalization



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

Overfit ⇝ high variance ↑

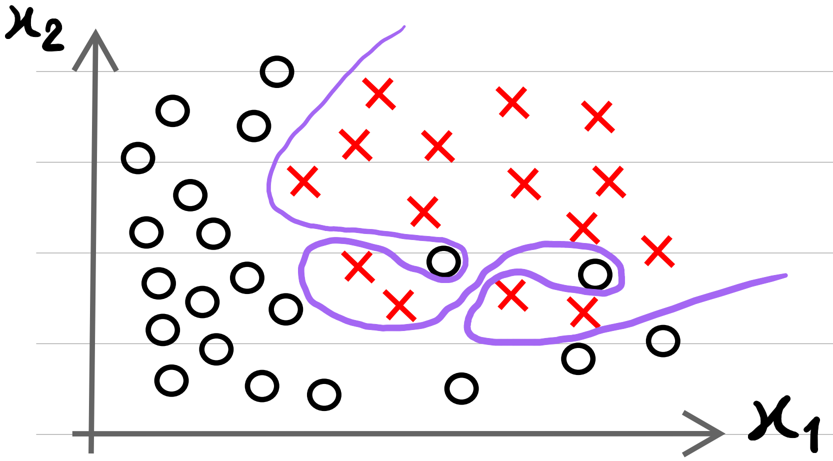
Classification

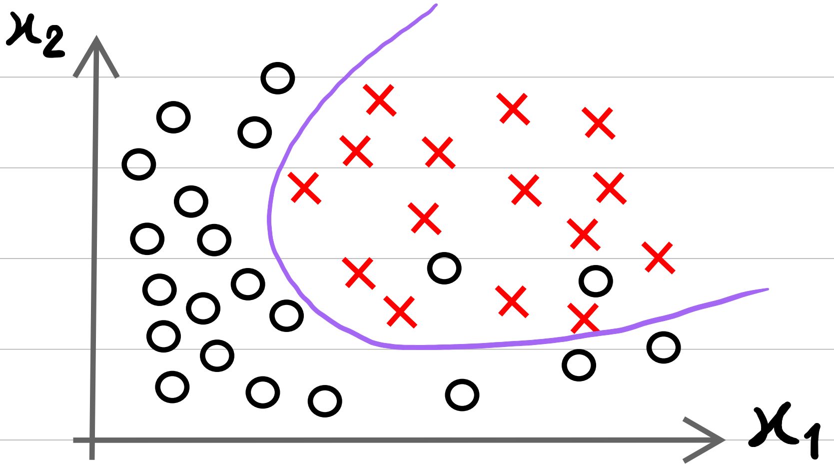


$$z = w_1 x_1 + w_2 x_2 + b$$

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

Underfit $\rightsquigarrow$ high bias ↑





$$z = w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + w_5 x_1 x_2 + b$$

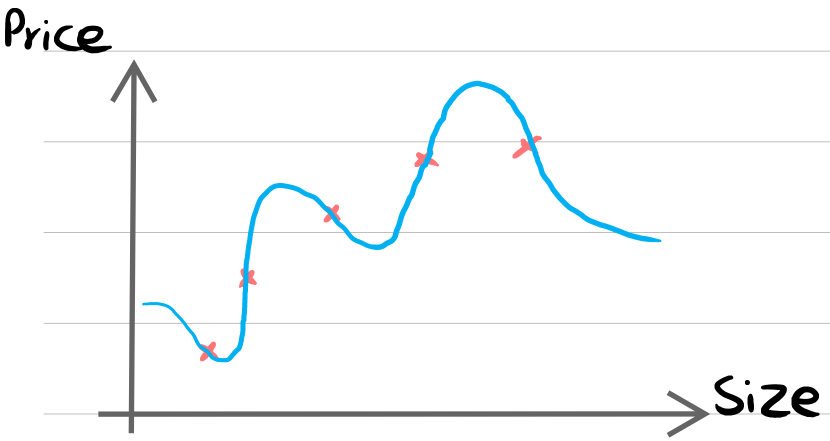
Just right $\rightsquigarrow$ generalization

$$z = w_1 x_1 + w_2 x_2 + w_3 x_1^2 x_2 + w_4 x_1^2 x_2^2 + w_5 x_1^2 x_2^3 + w_6 x_1^3 x_2 + \dots + b$$

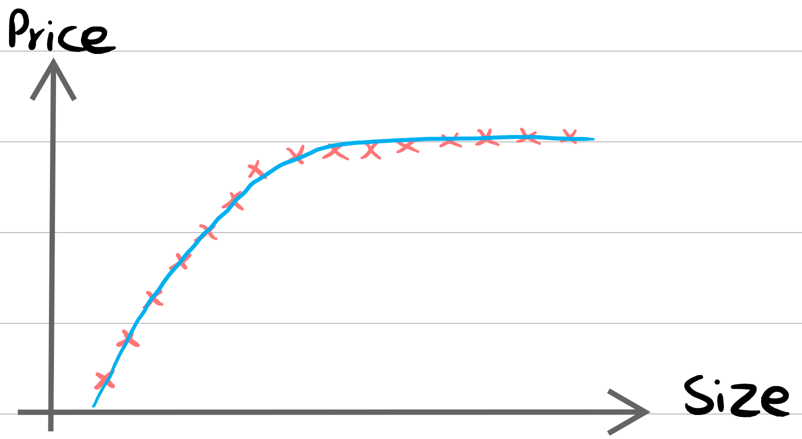
Overfit $\rightsquigarrow$ high variance ↑

ج: برای رفع مشکل overfit چه کنیم؟

1. Collect more training examples



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

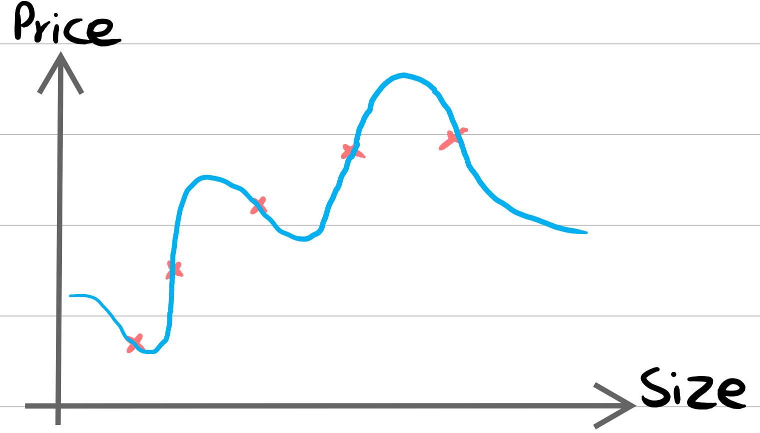
all features + Insufficient sample ⇝ overfit

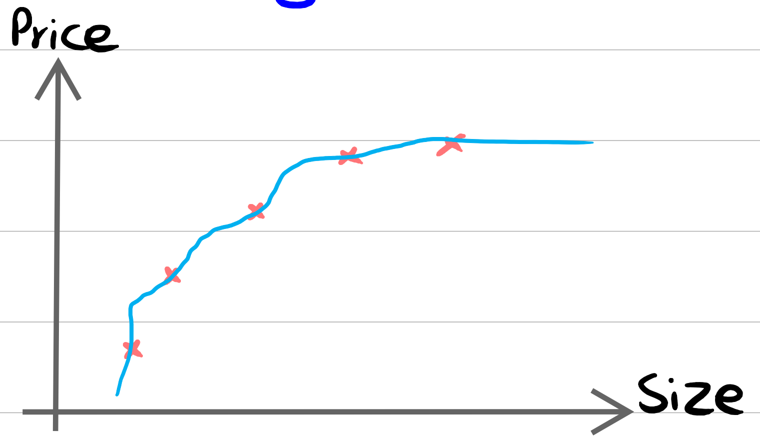
کافی نبودن نمونه‌ها

selected features ⇝ ③ → Size, #bedrooms, age

ممکن است ویژگی های مفید حذف شوند

3. Reduce the size of parameters w_j


Overfit


Regularization

$$f(x) = 28x - 385x^2 + 39x^3 - 174x^4 + 100$$

$$f(x) = 13x - 0.23x^2 + 0.000014x^3 - 0.0001x^4 + 10$$

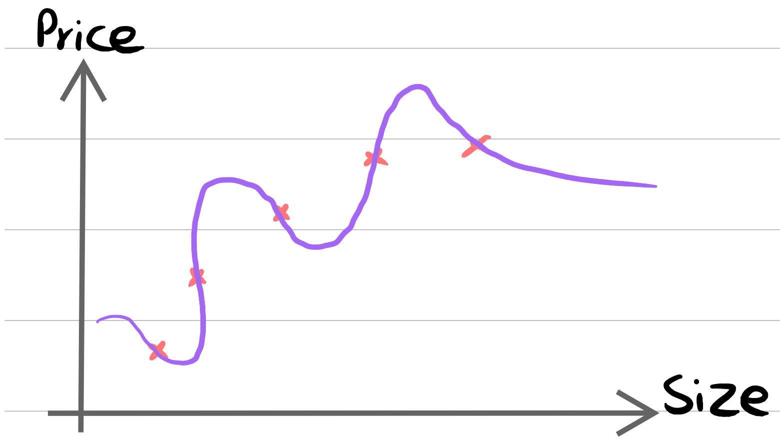
Large values for w_j

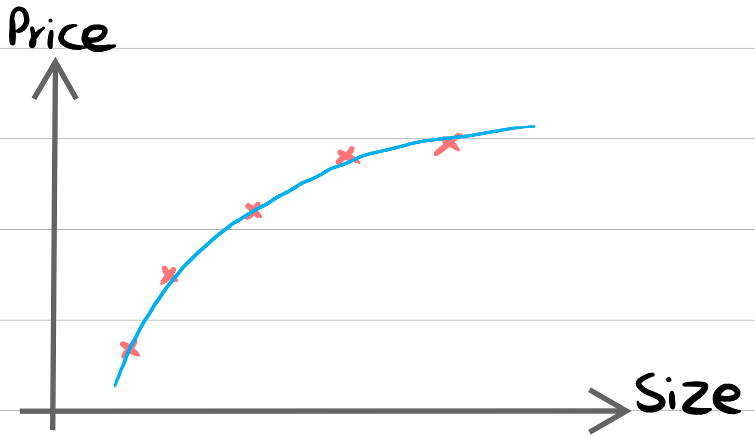
Small values for w_j

Regularization:

Reg. باعث کاهش خطا می شود و همچنین می تواند از underfit , overfit جلوگیری کند. ← چه طور؟

کمک می کند ضرایب را کوچک یا حذف کنید. ← همچنین می تواند روش خوبی برای انتخاب ویژگی باشد، چه طور؟





$$w_1x + w_2x^2 + w_3x^3 + w_4x^4 + b$$

$$w_1x + w_2x^2 + \underbrace{w_3x^3}_{\approx 0} + \underbrace{w_4x^4}_{\approx 0} + b$$

Make w_3, w_4 really small (≈ 0)

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

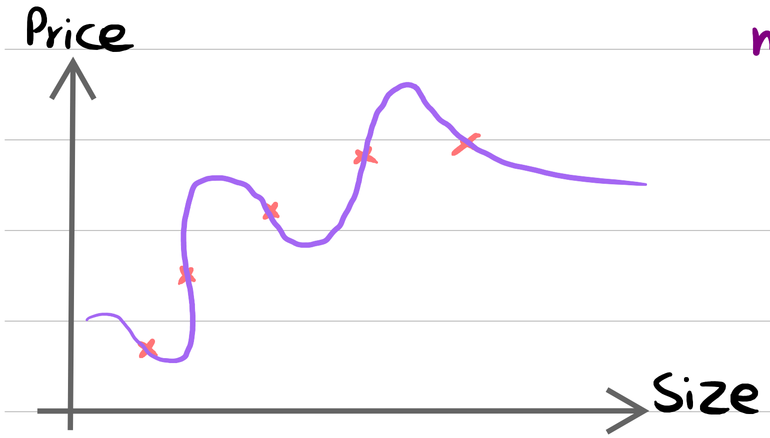
$$\underbrace{1000\,\overset{0.001}{w_3^2}}_{1} + \underbrace{1000\,\overset{0.002}{w_4^2}}_{2}$$

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:

↗ Regularizer term

repeat

{

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

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

}

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

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

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

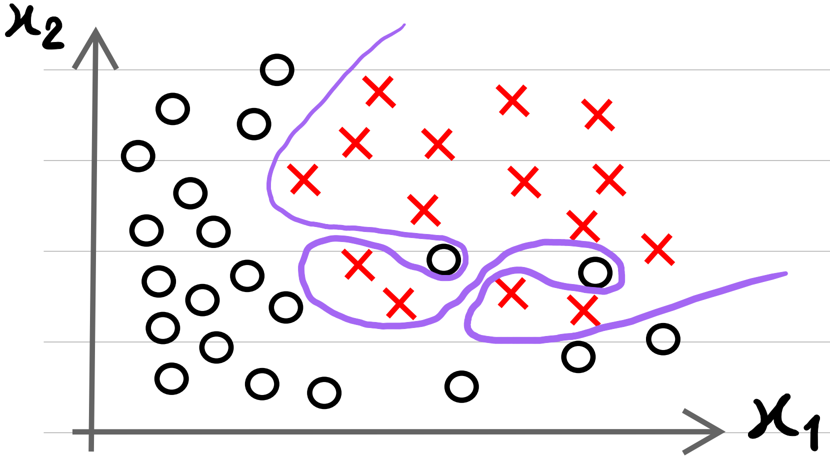
$$w_j = \underbrace{w_j - \alpha \frac{\lambda}{m} w_j}_{w_j\left(1-\alpha\frac{\lambda}{m}\right)} - \underbrace{\alpha \frac{1}{m} \sum_{i=1}^{m} \left(f_{w,b}(x^{(i)}) - y^{(i)}\right) x_j^{(i)}}_{\text{usual update}}$$

$\hookrightarrow \alpha = 0.01,\ \lambda = 1 \rightsquigarrow \alpha \frac{\lambda}{m} = 0.01 \times \frac{1}{50} = 0.0002$ $\qquad w_j\left(\overset{0.9998}{\cancel{1-0.0002}}\right)$

$m = 50$

Classification $\longrightarrow$ Logistic Regression

Overfit $\rightsquigarrow$ high variance $\uparrow$



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

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

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

$\hookrightarrow$ $\min J(w,b) \rightsquigarrow w_j \downarrow$

GD:

repeat

{

$$w_j = w_j - \alpha \boxed{\frac{\partial}{\partial w_j} J(w,b)} \longrightarrow = \frac{1}{m} \sum_{i=1}^{m} \Big(\underbrace{f_{w,b}(x^{(i)})}_{\text{Logistic Regression} \rightarrow f_{w,b}(x) = \frac{1}{1+e^{-wx+b}}} - y^{(i)} \Big) x_j^{(i)} + \frac{\lambda}{m} w_j$$

$$b = b - \alpha \boxed{\frac{\partial}{\partial b} J(w,b)} \longrightarrow = \frac{1}{m} \sum_{i=1}^{m} \Big(f_{w,b}(x^{(i)}) - y^{(i)} \Big)$$

}

Regularization → Ridge $L2 = \sum_{j=0}^{n} w_j^2$

Regularization → Lasso $L1 = \sum_{j=0}^{n} |w_j|$

Regularization → Elastic Net L1, L2

alpha ⟶ α

Ridge → همه ضرایب کوچک می شوند.

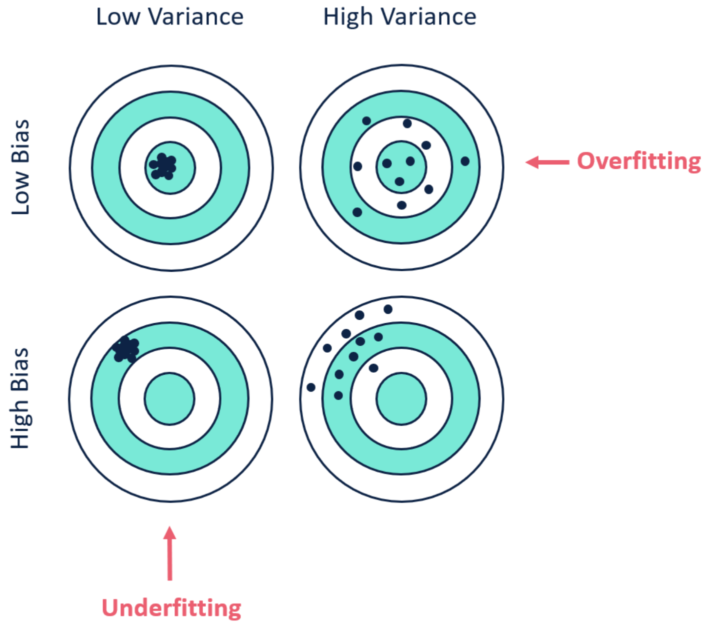
$$Loss = \sum_{i=1}^{m} \left(f_{w,b}(x) - y^{(i)}\right)^2 + \lambda \sum_{j=0}^{n} w_j^2$$

$\alpha = 0 \longrightarrow$ Ridge X

$\alpha = \infty \longrightarrow$ 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 0 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 Ridge و lasso

$\hookrightarrow L1, L2$

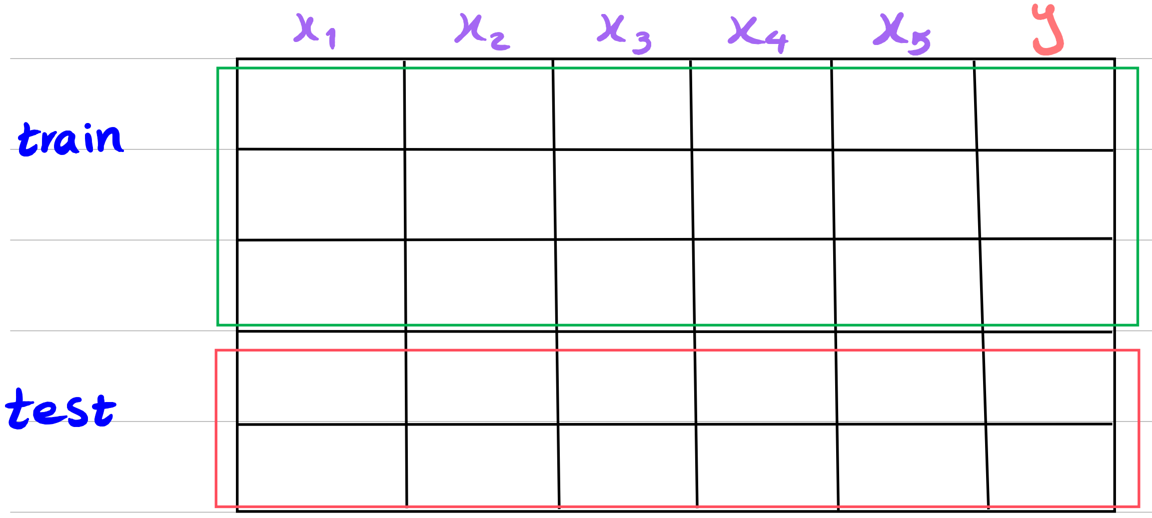
$$Loss = \frac{1}{2m} \sum_{i=1}^{m} \left(f_{w,b}(x) - y_i \right)^2 + (1-\lambda) \times \frac{\alpha}{2} \times \sum_{i=1}^{n} \underbrace{w_j^2}_{L2} + \boxed{\lambda}\,\alpha \underbrace{|w_j|}_{L1}$$

L1_rate = 1 $\longrightarrow$ L1 → lasso

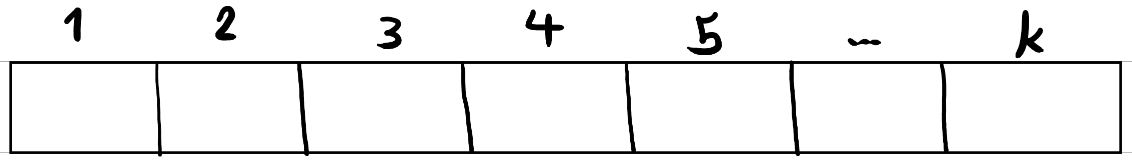
L1_rate = 0 $\longrightarrow$ L2 → Ridge

Cross Validation

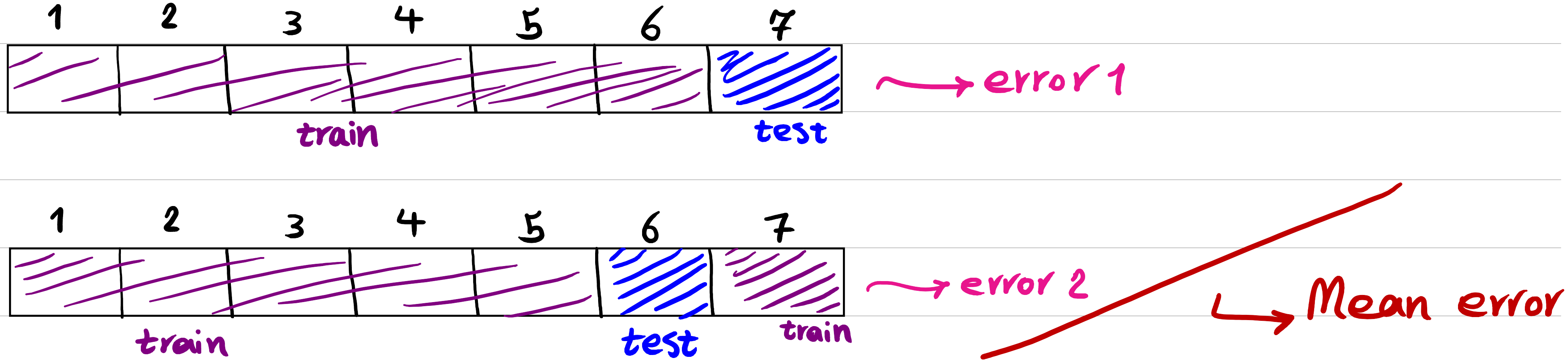
آیا روشی برای پیدا کردن هایپرپارامتر (α) مناسبی هست؟

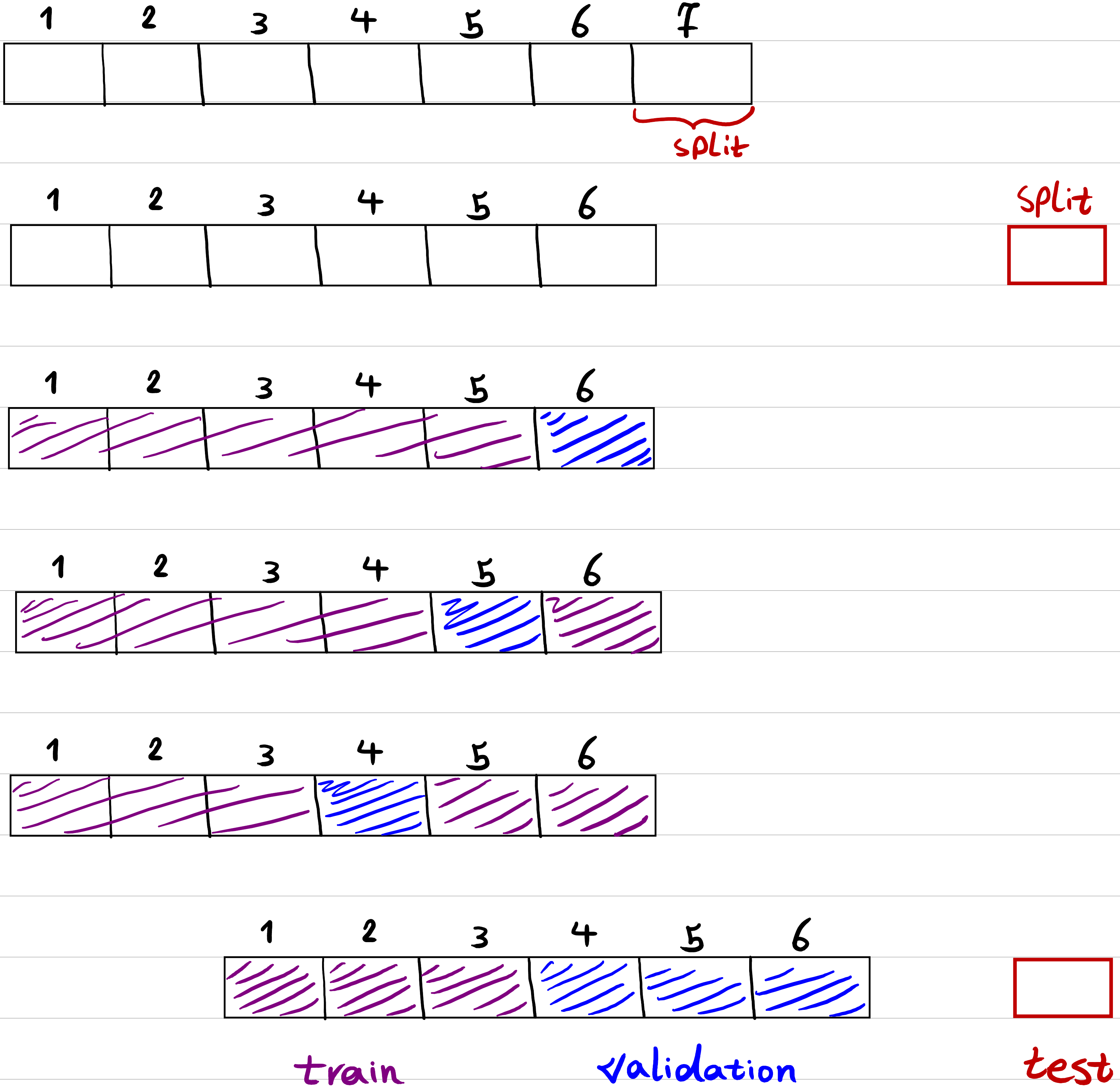


k-fold:



k=7 , cv=2




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