

Session 07

Regularization

Machine Learning | Zahra Amini



Telegram: [@zahraamini_ai](https://t.me/zahraamini_ai) & Instagram: [@zahraamini_ai](https://www.instagram.com/zahraamini_ai) & LinkedIn: [@zahraamini-ai](https://www.linkedin.com/in/zahraamini-ai)

<https://zil.ink/zahraamini>

اصطلاح	تعریف
بایاس (Bias)	میزان خطای مدل به دلیل ساده‌سازی بیش از حد. مدل‌های با بایاس بالا نمی‌توانند الگوهای پیچیده داده‌ها را به خوبی یاد بگیرند.
واریانس (Variance)	میزان حساسیت مدل به تغییرات کوچک در داده‌های آموزشی. مدل‌های با واریانس بالا به جزئیات و نویزهای داده‌ها بسیار حساس هستند.
تععمیددهی (Generalization)	توانایی مدل در عملکرد خوب بر روی داده‌های جدید و نادیده. مدل‌های با تععمیددهی بالا الگوهای یادگرفته شده را به داده‌های جدید به درستی اعمال می‌کنند.

زهراء‌امینی
@zahraamini_ai

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"> • High training error • Training error close to test error • High bias 	<ul style="list-style-type: none"> • Training error slightly lower than test error 	<ul style="list-style-type: none"> • Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	<ul style="list-style-type: none"> • Complexify model • Add more features • Train longer 		<ul style="list-style-type: none"> • Perform regularization • Get more data

زنگنه

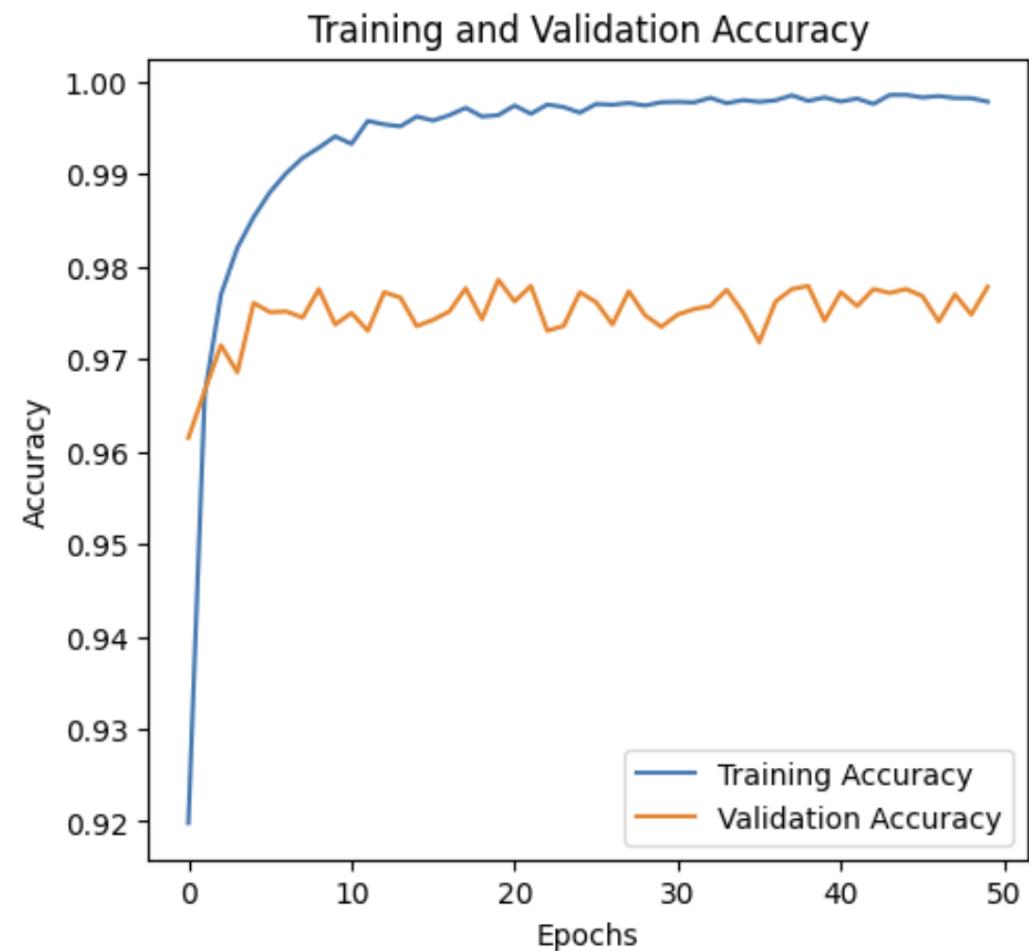
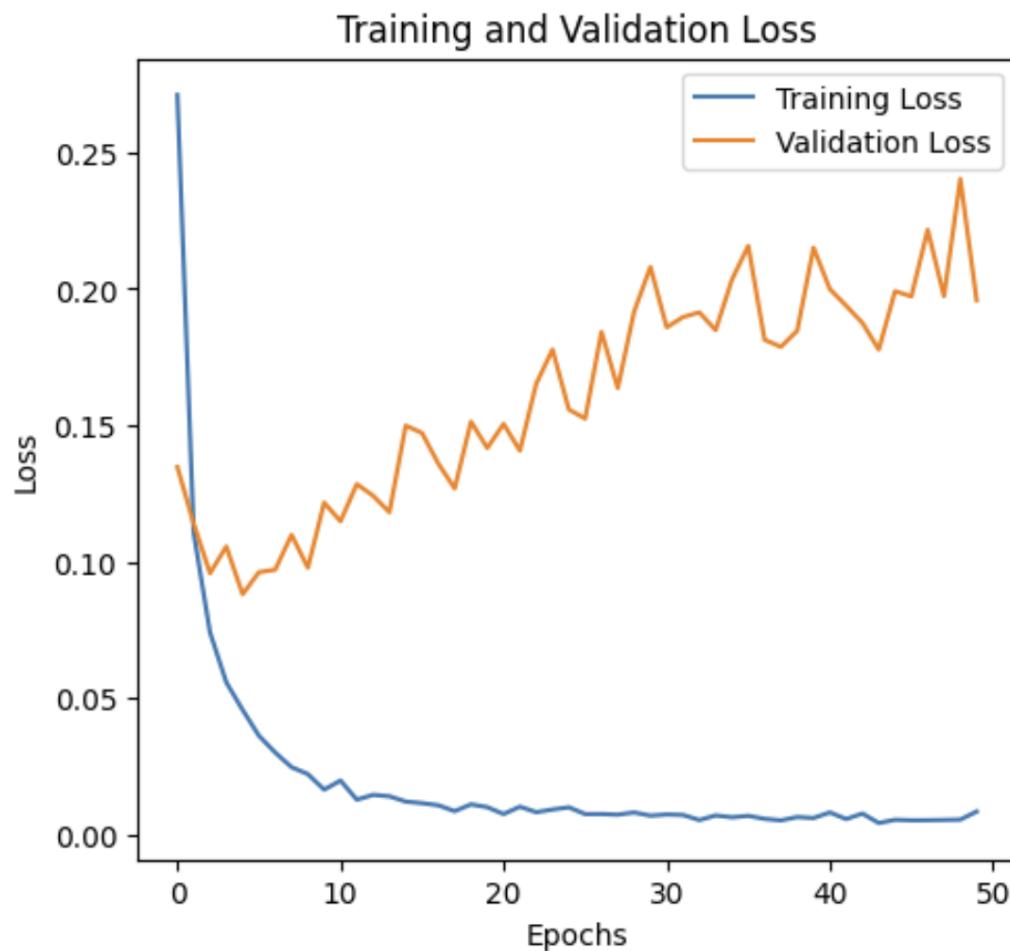
@zahraamini_ai

تعیین‌دهی (Generalization)	واریانس (Variance)	بایاس (Bias)	تعریف	اصطلاح
ضعیف	پایین	بالا	مدل به اندازه کافی پیچیدگی ندارد تا الگوهای موجود در داده‌ها را به درستی یاد بگیرد. عملکرد ضعیف در آموزش و تست.	آندرفیت
ضعیف	بالا	پایین	مدل به قدری پیچیده است که نویزها و جزئیات غیرضروری را نیز یاد بگیرد. عملکرد عالی در آموزش و ضعیف در تست و جدید.	اورفیت
بالا	متعادل	متعادل	مدل به اندازه کافی پیچیده است تا الگوهای موجود در داده‌ها را به درستی یاد بگیرد بدون اینکه به نویزها توجه کند. عملکرد خوب در آموزش و تست و می‌تواند به خوبی تعیین دهد.	گودفیت

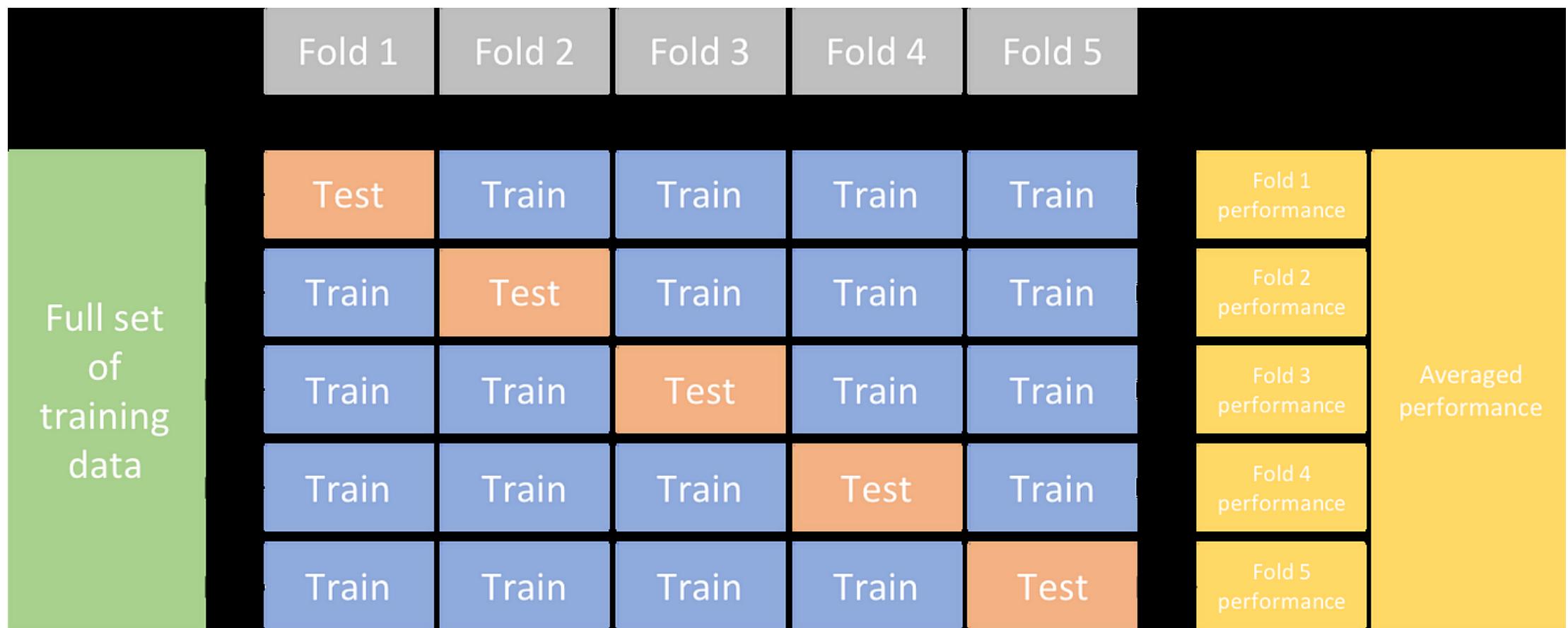
زهراء‌امینی

@zahraamini_ai

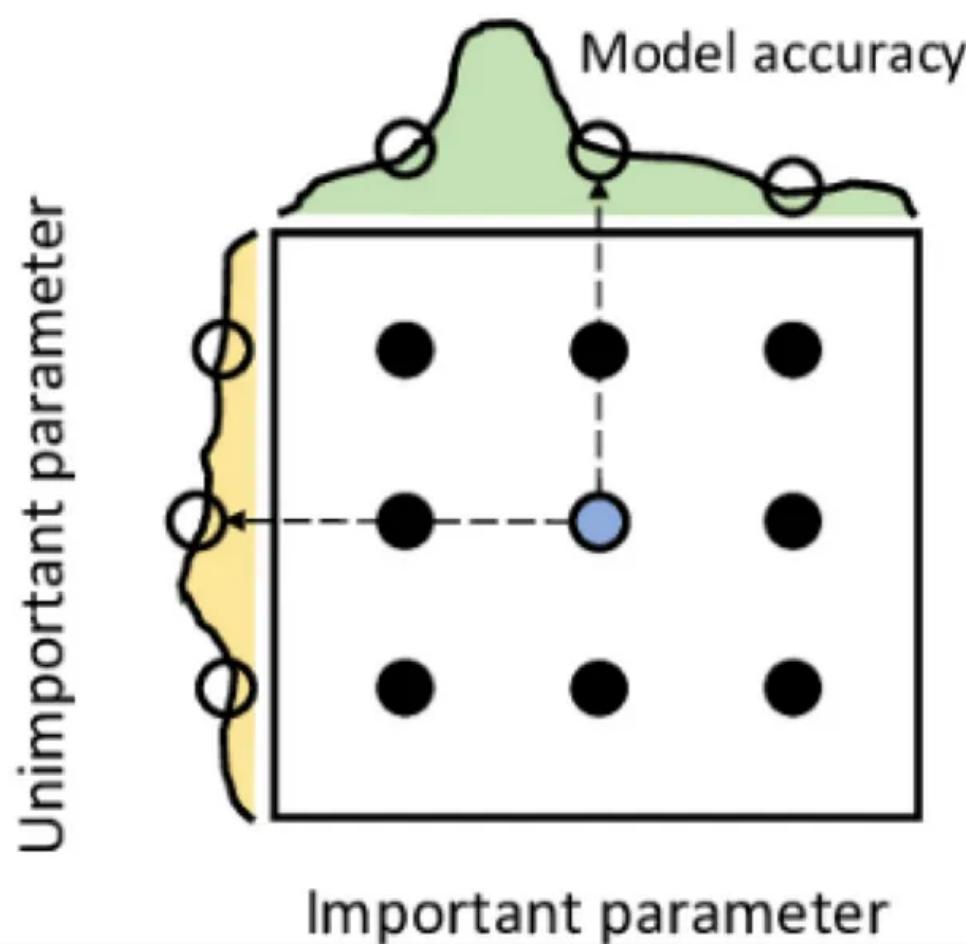
تحلیل نتایج کد



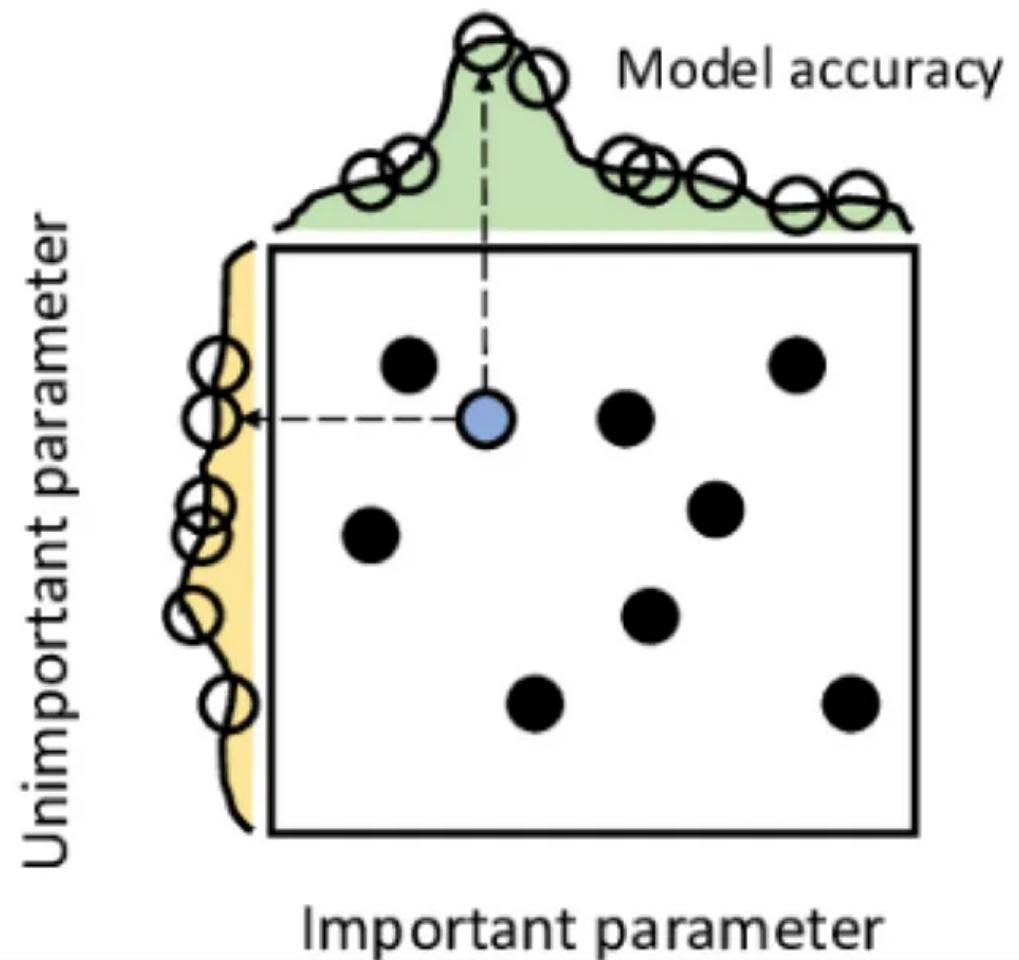
k-fold cross-validation



Exhaustive Grid search



Random Grid search

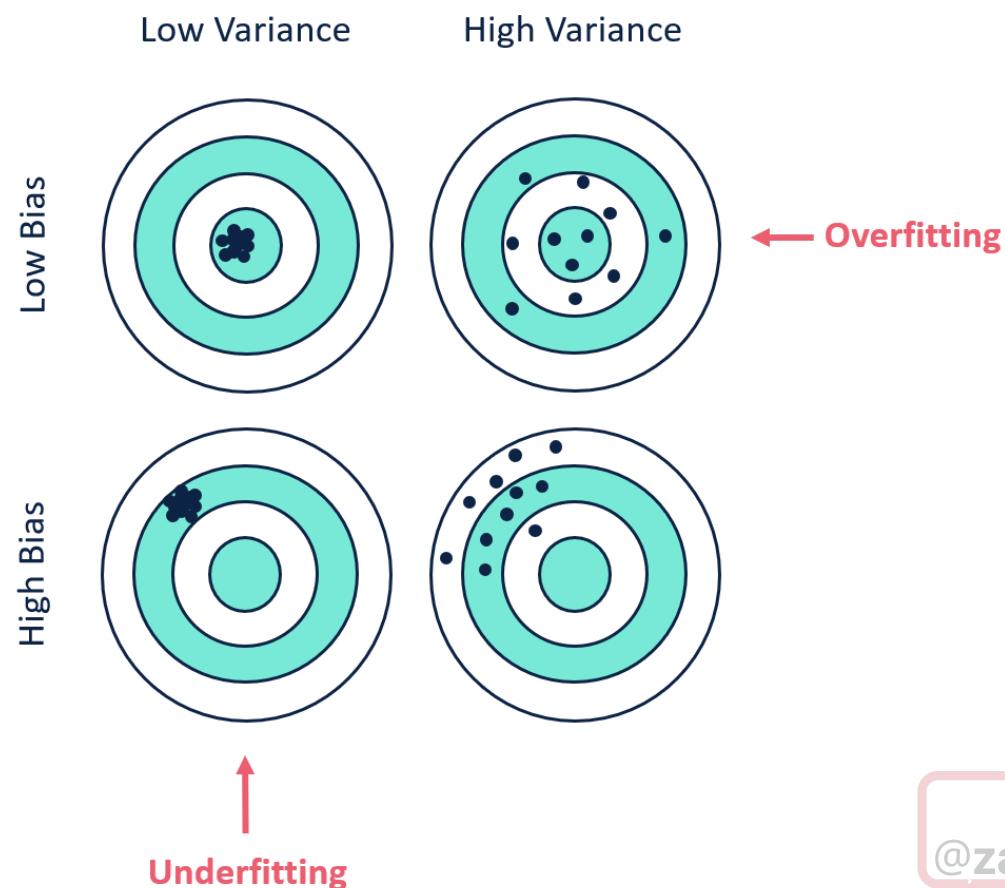
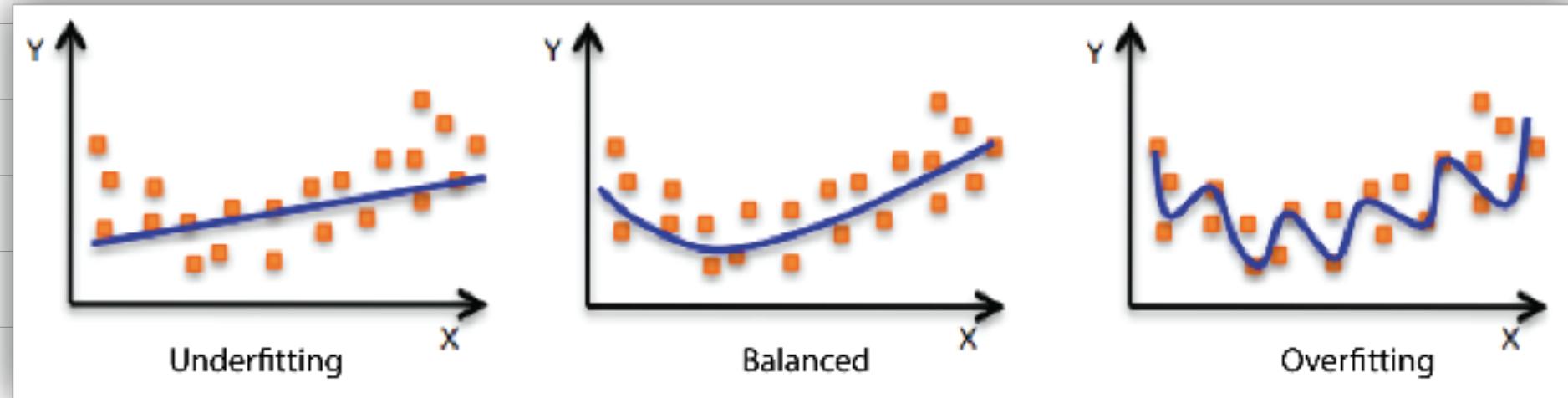


این روش تمامی ترکیب‌های ممکن از پارامترها را آزمایش می‌کند هر نقطه سیاه در شبکه نشان‌دهنده یک ترکیب از پارامترهای است که آزمایش شده است

ناحیه سبز در بالای نمودار نشان‌دهنده دقت مدل در ترکیب‌های مختلف است
این روش می‌تواند زمان بر و محاسباتی سنگین باشد زیرا تمامی ترکیب‌ها باید آزمایش شوند

این روش تعدادی ترکیب از پارامترها را به صورت تصادفی انتخاب می‌کند نقاط سیاه نشان‌دهنده ترکیب‌های انتخاب شده به صورت تصادفی هستند

ناحیه سبز در بالای نمودار همچنان نشان‌دهنده دقت مدل است
این روش معمولاً سریع‌تر و کارآمدتر است زیرا ترکیب‌های کمتری آزمایش می‌شوند ولی می‌توانند همچنان به دقت بالایی برسد



: عدم دقت حاصل از تخمین یک مدل های واقعی **biase**

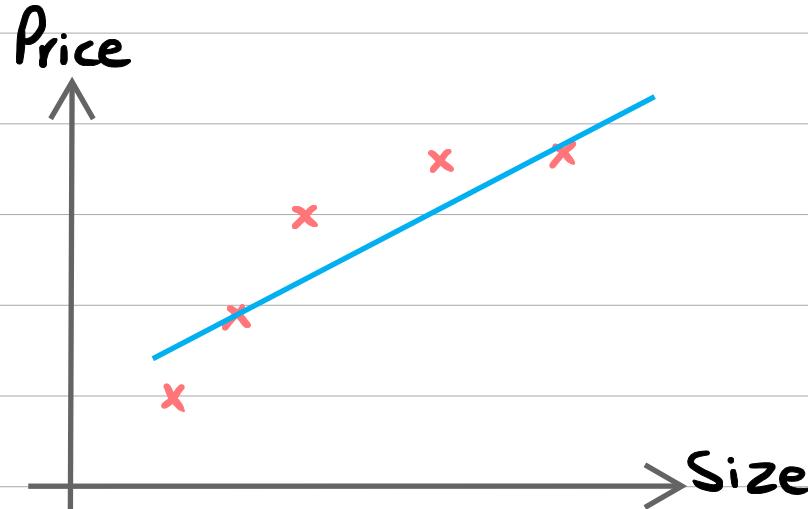
با چیزی زیاد با یک مدل ساده.

train او را با یک گروه داده $f_{w,b}(x)$ ار: **Variance**

متغیرات تخمین بزرگ، چند تغییر می کند.

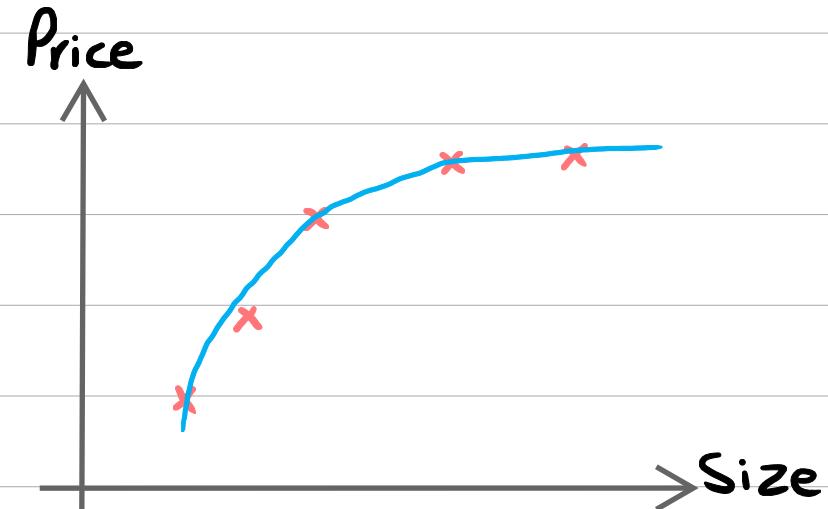
$$\text{Error} = \text{var}(f(x)) + \text{Bias}(f(x))^2 + \text{error}(\epsilon)$$

Regression



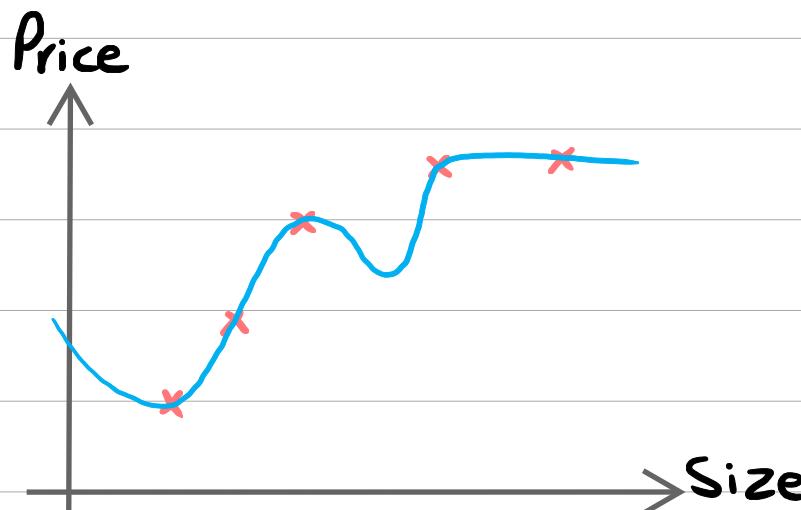
$$w_1x + b$$

Underfit \rightsquigarrow high bias \uparrow



$$w_1x + w_2x^2 + b$$

Just right \rightsquigarrow generalization



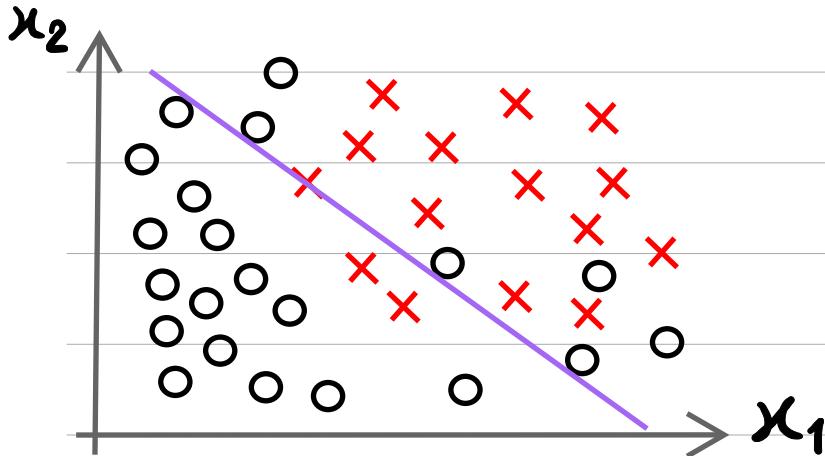
$$w_1x + w_2x^2 + w_3x^3 + w_4x^4 + b$$

Overfit \rightsquigarrow high variance \uparrow

Classification

زهراء امینی

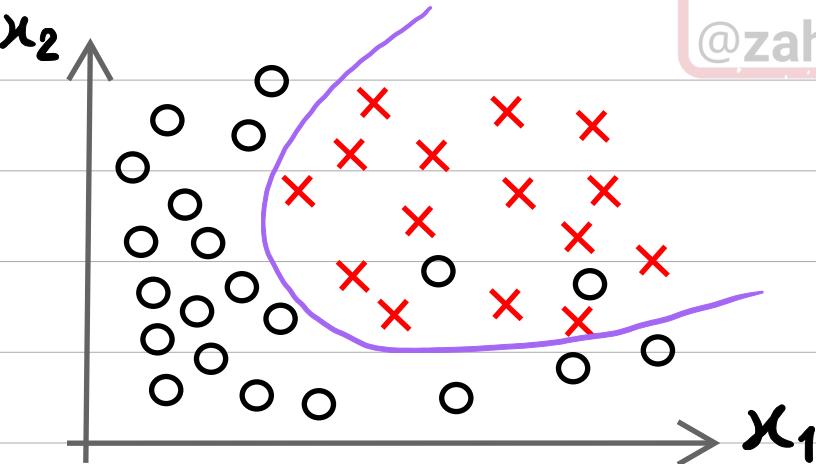
@zahraamini_ai



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

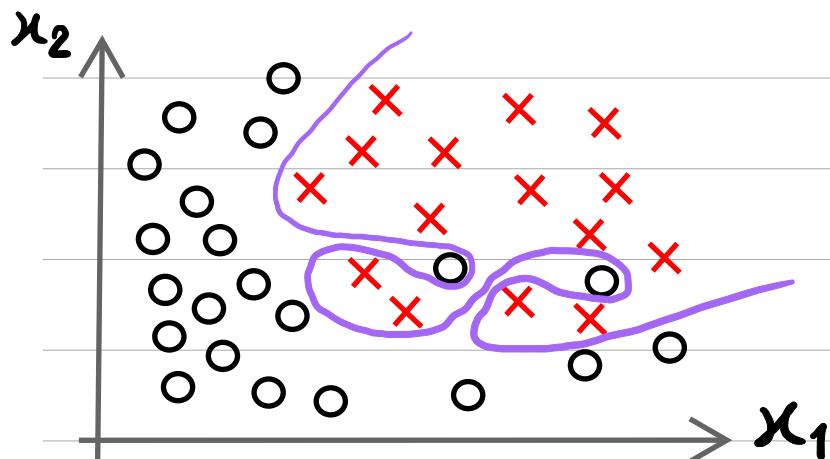
$$f_{w,b}(x) = g(Z) \rightarrow g: \text{Sigmoid}$$

Underfit \rightsquigarrow high bias \uparrow



$$\begin{aligned} Z = & w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 \\ & + w_5 x_1 x_2 + b \end{aligned}$$

Just right \rightsquigarrow generalization

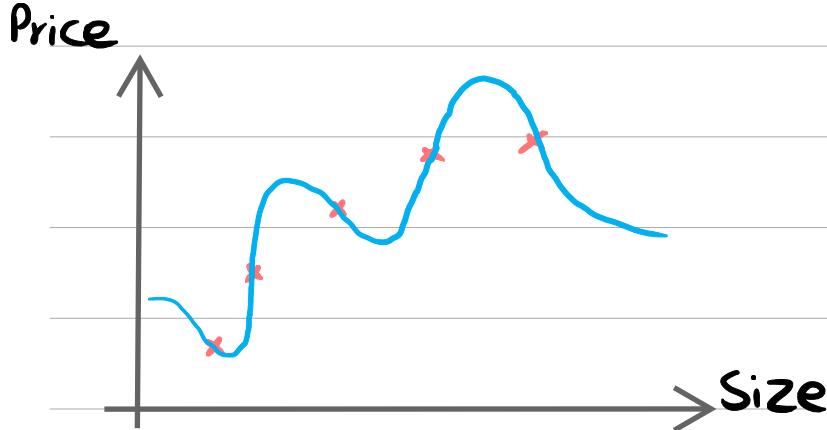


$$\begin{aligned} Z = & w_1 x_1 + w_2 x_2 + w_3 x_1^2 x_2 + w_4 x_1^2 x_2^2 \\ & + w_5 x_1^2 x_2^3 + w_6 x_1^3 x_2 + \dots + b \end{aligned}$$

Overfit \rightsquigarrow high variance \uparrow

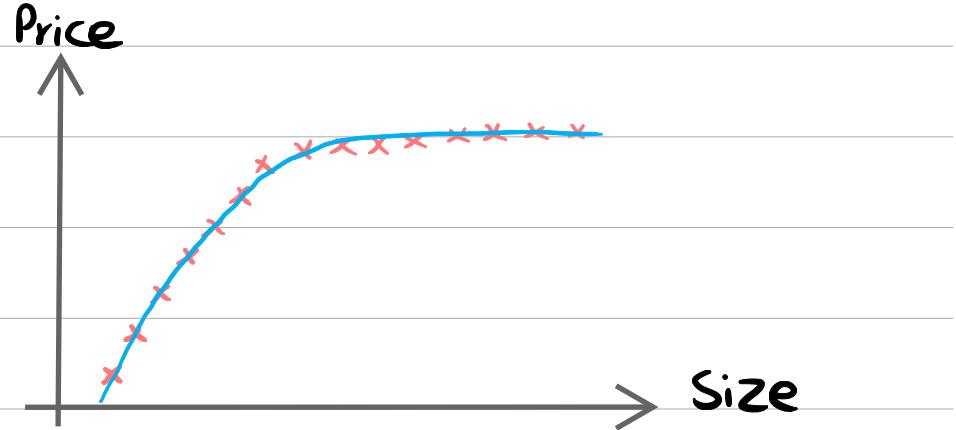
1. Collect more training examples

؟ برای رفع مشکل overfit چه کنم؟



Overfit

زهراء الميني
@zahraamini_ai



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

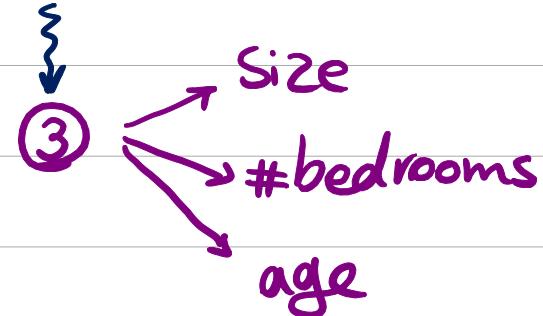
all features

Overfit

+
Insufficient sample

کافی نبودن ممونها

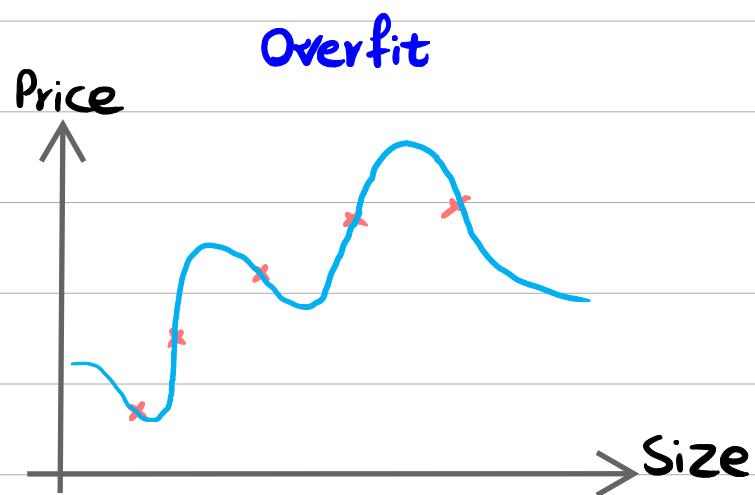
Selected features



مسکن است دیگر کمی های

غایی حذف شوند

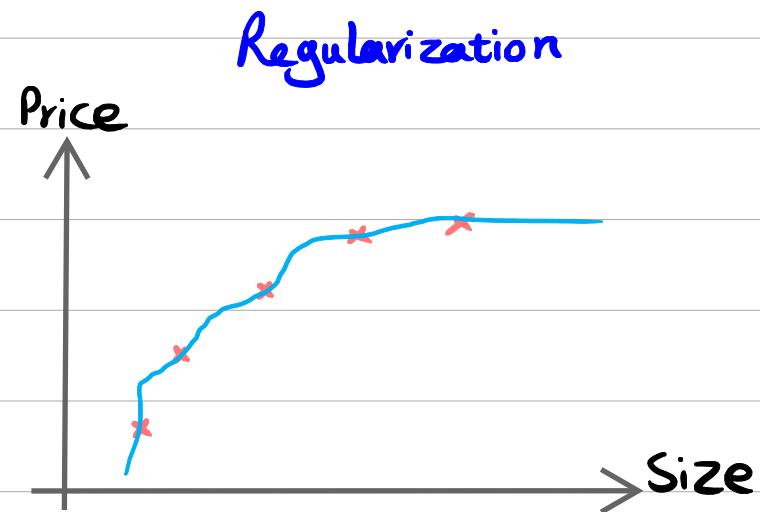
3. Reduce the size of parameters W_j



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

$$-174x^4 + 100$$

Large values for W_j



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

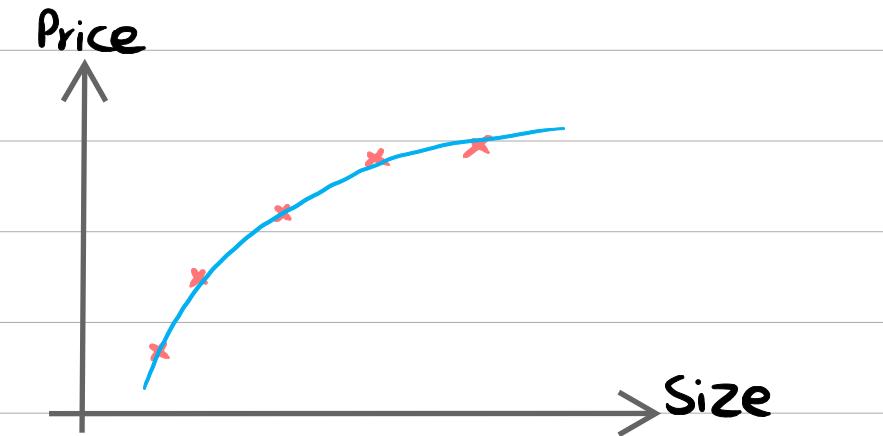
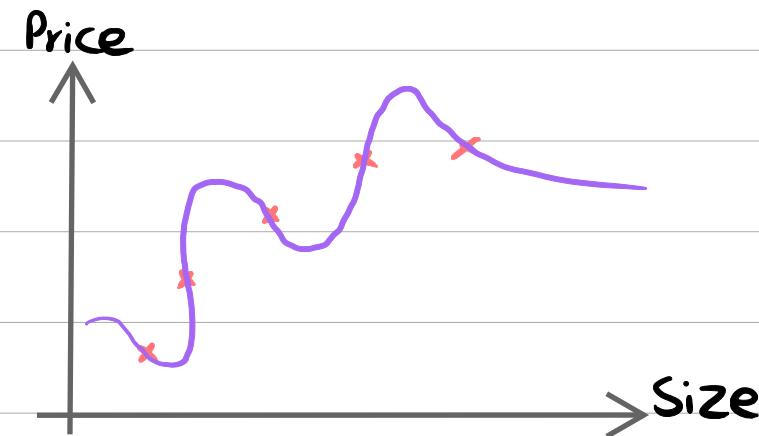
$$-0.0001x^4 + 10$$

Small values for W_j

Regularization:

بسیار کاهش خطای شود و همین تواند overfit, underfit جلوگیری کند. چه طور؟

همین تواند روش خوبی برای انتخاب دیتلری 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)

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

$$\underbrace{1000 W_3^2}_1 + \underbrace{1000 W_4^2}_2$$

0.001 0.002

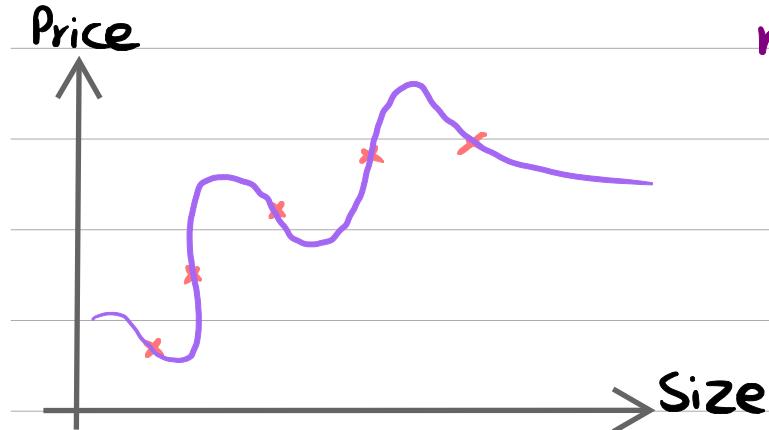
x_1	x_2	x_3	x_4	x_5	..	x_{100}	y
Size	bedroom	floors	age	avg income	..	distance coffee shop	Price

$W_1, W_2, W_3, \dots, W_{100}, b$

$n=100$

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

λ : Lambda Regularization Parameter



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

$W, b = ?$

زهراء امینی

@zahraamini_ai

Gradient Descent:

repeat

{

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

Regularizer term

}

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

زنگنه
@zahraamini_ai

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

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

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

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

$$\hookrightarrow \alpha = 0.01, \lambda = 1 \implies \alpha \frac{\lambda}{m} = 0.01 \times \frac{1}{50} = 0.0002$$

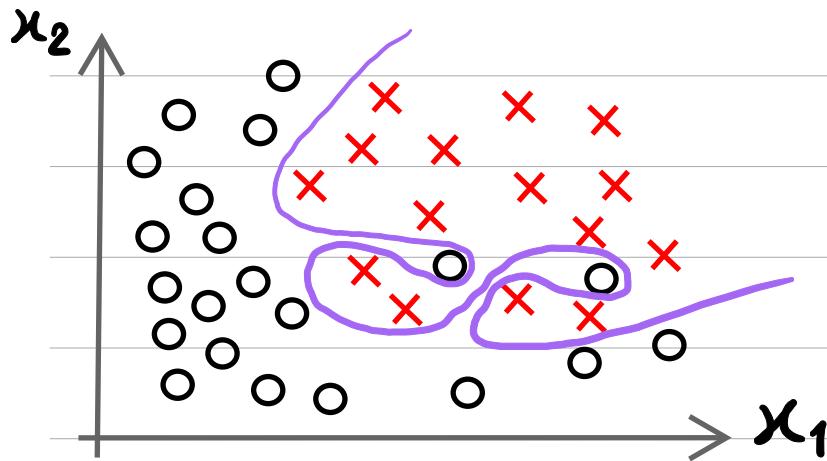
$m = 50$

$$w_j \cancel{(1-0.0002)}^{\text{0.9998}}$$



Classification \rightarrow Logistic Regression

Overfit \leadsto high variance \uparrow



$$Z = W_1 x_1 + W_2 x_2 + W_3 x_1^2 x_2 + W_4 x_1^2 x_2^2 \\ + W_5 x_1^2 x_2^3 + W_6 x_1^3 x_2 + \dots + b$$

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

$\min J(w, b) \leadsto w_j \downarrow$

GD:

Repeat

{

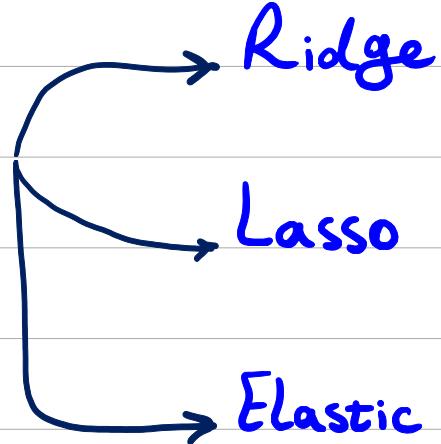
$$w_j = w_j - \alpha \frac{\partial}{\partial w_j} J(w, b) \rightarrow = \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)}) x_j^{(i)} + \frac{\lambda}{m} w_j$$

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

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

}

Regularization



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

alpha $\rightarrow \alpha$

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

L1, L2

زهراء امینی
@zahraamini-ai

Ridge \rightarrow مکان ضرایب کوچک ہی شوند

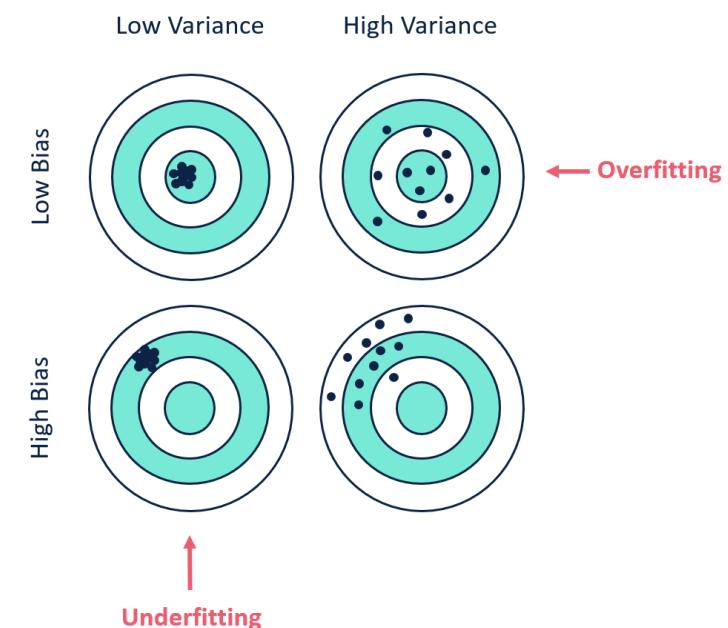
$\alpha = 0 \rightarrow$ Ridge X

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

$\alpha = \infty \rightarrow$ مکان ضرایب کوچک و نزدیک به صفر ہی شوند.

$\alpha \uparrow$ bias ↑
 $\alpha \downarrow$ bias ↓

$\alpha \uparrow$ variance ↑
 $\alpha \downarrow$ variance ↓



هنرایب را که جگ های لند و برقی از هنرایب را به ۵ هی رساند.

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

$\alpha = 0 \rightarrow$ همچنانچه وزنی حذف نمی شود.

$\alpha = \infty \rightarrow$ تمام وزنها حذف می کند.

زهراءامینی
@zahraamini_ai

Elastic Net:

↳ L1, L2

ترکیبی از Lasso, Ridge

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

L1_rate = 1 \rightarrow L1 \rightarrow Lasso

L1_rate = 0 \rightarrow L2 \rightarrow Ridge

Cross Validation

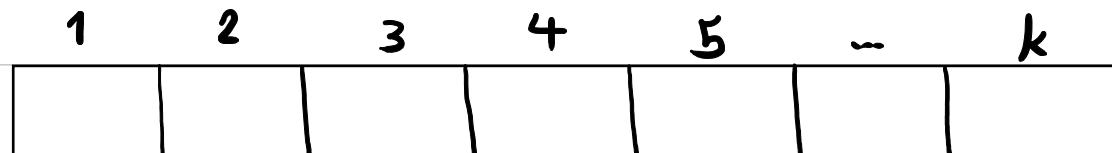
۸. آیا روشی برای پیدا کردن هایپر پارامتر (α)

	x_1	x_2	x_3	x_4	x_5	y
train						
test						

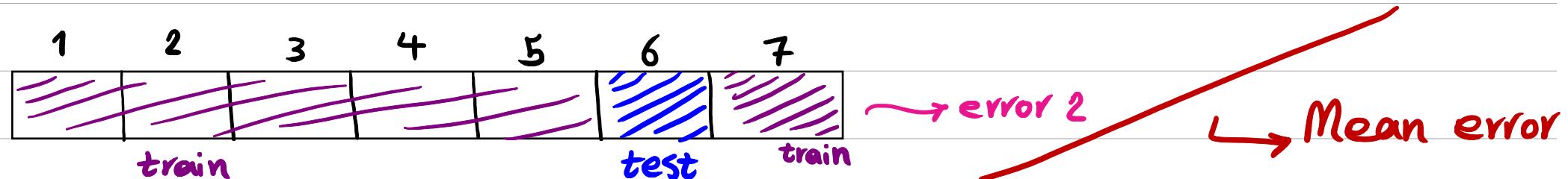
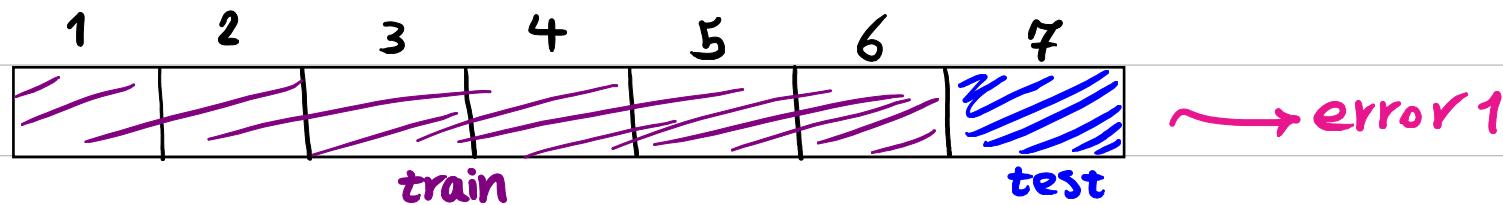
نمایی ملت ۶

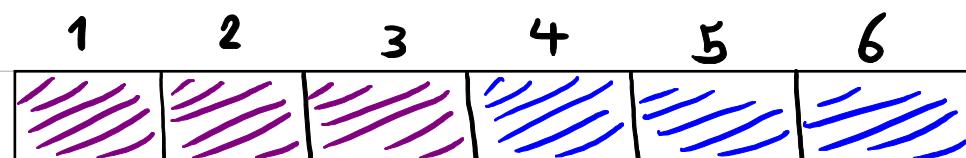
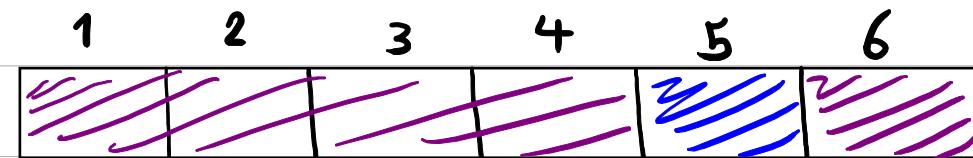
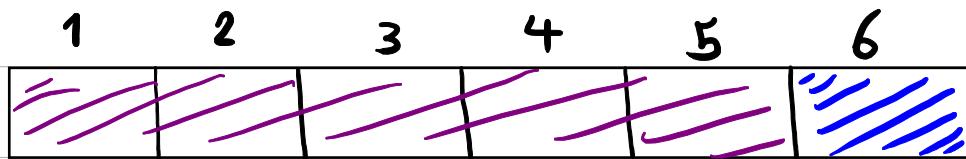
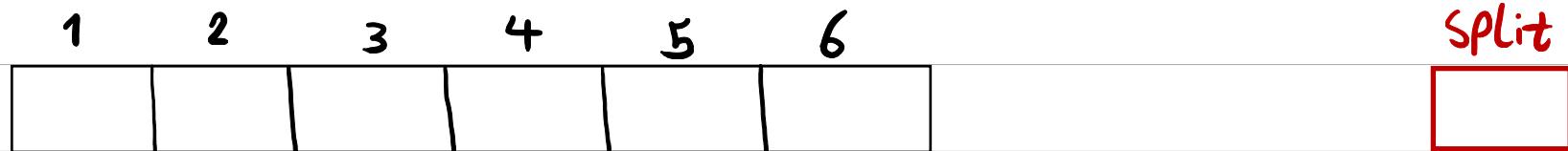
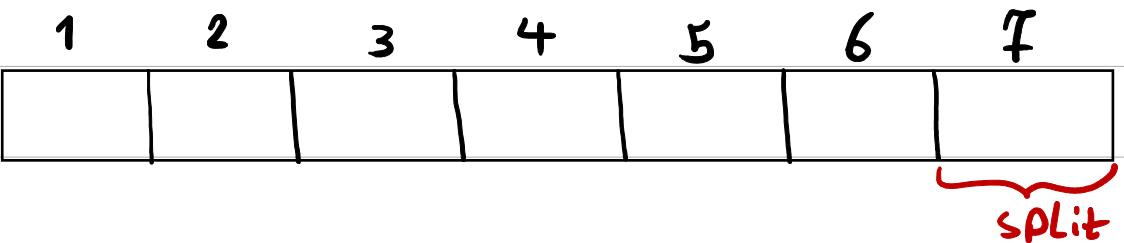


k -fold:



$k=7$, $CV=2$





train

validation

test

زهرا امینی
@zahraamini_ai