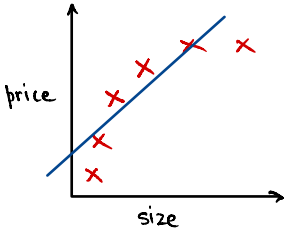
overfitting ⟺ underfitting

We don't need to be overly accurate while choosing our parameters, but at the same time we must not disregard too many data points.

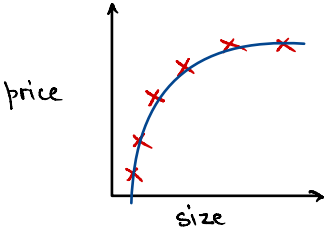


$\vec{w}\vec{x} + b$

underfitting (same thing) highly biased

↓

not taking many data points in regard

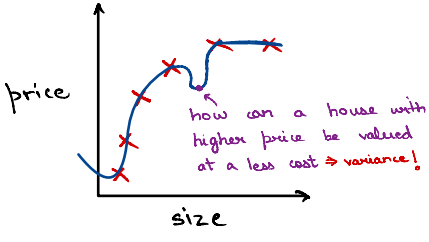


$w_1\vec{x} + w_2x^2 + b$

generalized

↓

best! doesn't perfectly fit, but does a pretty good job



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

overfitting

high variance

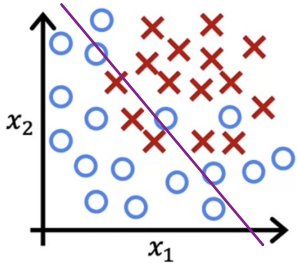
↓

fits example really well, but as soon as new predictions are made it will fail and vary.

Always, the generalized fit will give consistent predictions, even for new examples.

Overfitting may cause the shape of the graph to change after gradient descent when a new example is added, leading to previous points not matching.

Same thing happens for classification

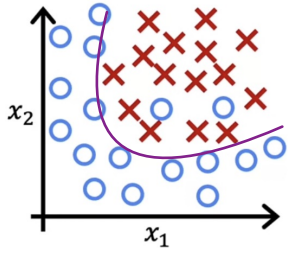


$z = w_1 x + w_2 x + b$

$f_{\vec{w},b}(\vec{x}) = g(z)$

g is the sigmoid function

underfit or high bias

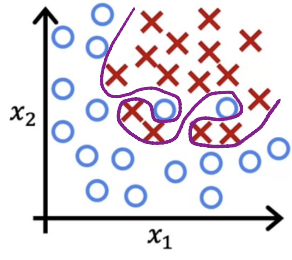


$z = w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + w_5 x_1 x_2 + b$

just right

will generalize

too confusing !



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

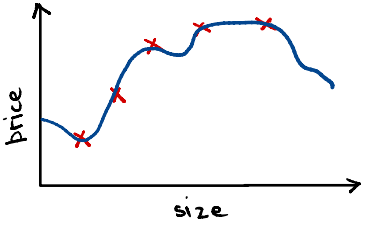
overfitting or high variance

Addressing overfitting

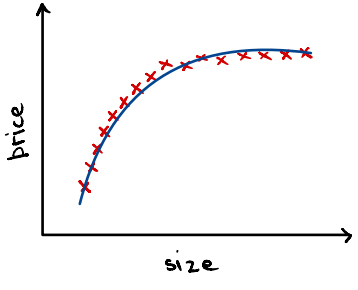
If we've identified that an algorithm is overfitting then there are a couple steps we can take to reduce this :-

- Collect more data
- Select certain features and excluding useless ones.
- Apply "regularization" - reducing size of the parameters

1. Collect More Training Data



→ Maybe not enough data! With a larger data set a training set, the learning algorithm will learn to fit a less wiggly and more accurate function.



getting more data isn't always an easy option!

2. Select features to include / exclude

size	bedrooms	floors	age	avg income	...	distance to coffee shop	price
x_1	x_2	x_3	x_4	x_5	...	x_n	y

maybe only some features are adequate and other features aren't useful or don't have enough training data.

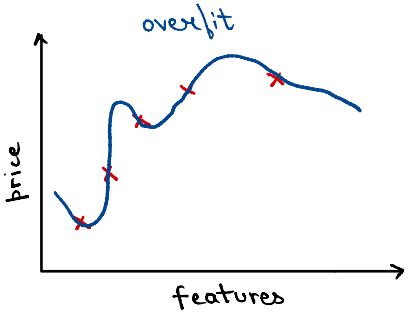
all features
+
insufficient data

= overfit

selected features
size
bedrooms
age
just right
(feature selection)

3. Regularization

Regularization is used when all features are important, but the model is overfit.



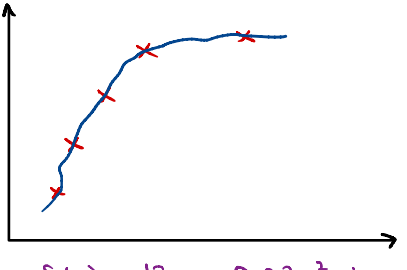
$f(x) = 28x - 385x^2 + 39x^3 - 174x^4 + 100$

large values

can reduce this

don't want to eliminate this feature by setting w to 0.

Regularization refers to as reducing size of the parameters w_j.



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

small values of w_j

w_1 to w_j can be regularized but you may or may not regularize 'b'.