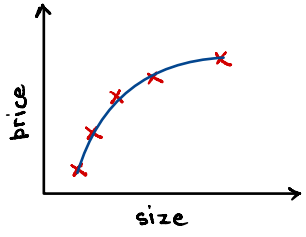
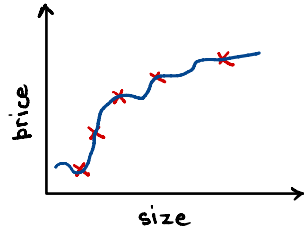


$w_1x + w_2x^2 + b$



$w_1x + w_2x^2 + w_3x^5 + w_4x^4 + b$

We can make the function fit better by making some parameters' value very small (such as w_3, w_4 in this case).

Cost Function after applying regularization :-

"1000" is just an example

$$\min_{\vec{w},b} \frac{1}{2m} \sum_{i=1}^{m} \left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}\right)^2 + 1000\, w_3^2 + 1000\, w_4^2$$

taking big constants infront of square of parameters

Since the function "minimizes" by nature, taking big constants and squaring the parameters with result in w_3 and w_4 being very small (≈ 0.001 or something)

If such parameters like w_3 and w_4 result in something like 0.001 or 0.002 they become effectively 0 and only have a tiny contribution, therefore resulting in a model that is better fitting.

Regularization

small values $w_1, w_2, \cdots, w_n, b$

simpler model

less likely to overfit

$w_3 \approx 0$

$w_4 \approx 0$

size	bedrooms	floors	age	avg income	...	distance to coffee shop	price
x_1	x_2	x_3	x_4	x_5		x_{100}	y

$w_1, w_1, w_2, \cdots, w_{100}, b$

n features

When there are a lot of parameters then we have no choice but to penalize all w_j parameters.

Custom cost function will be:-

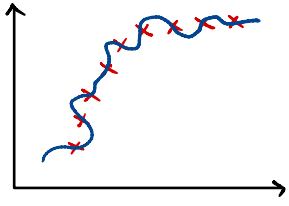
$$J(\vec{w}, b) = \frac{1}{2m} \sum_{i=1}^{m} \left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)} \right)^2 + \underbrace{\frac{\lambda}{2m} \sum_{j=1}^{n} w_j^2}_{\text{regularization term}}$$

We are dividing $\lambda \sum_{j=1}^{m} w_j$ by 2m so that it is in correspondence with the former part and it is also easy to choose the value of λ. with.

λ is a similar value as α learning rate. Choosing a small value such as 0 will disregard the regularization term and thus the model will be overfit.

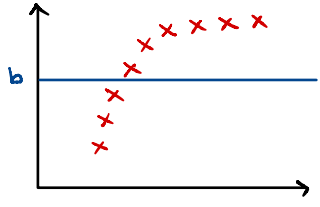
Choosing a big value like 10^{10} will result in all w_j's to be 0, because you know when the constant is large then the minimizing function results in the w's to be 0 $\Rightarrow$ $f_{\vec{w},b}(\vec{x}^{(i)}) = w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + b$ ($w_1 \approx 0$, $w_2 \approx 0$, $w_3 \approx 0$, $w_4 \approx 0$)

$\lambda = 0$:



overfit!

$\lambda = 1$:



underfit!

$$J(\vec{w},b) = \frac{1}{2m}\sum_{i=1}^{m}\overbrace{\left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}\right)^2}^{\text{mean squared error}} + \underbrace{\frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2}_{\text{regularization term}}$$

Choose the best possible λ for optimum result.

Our Goals:-

1. Fit Data
2. Keep w_j small

λ balances both goals ($0 < \lambda < 1$)

We can also regularize b, but conventionally it is not needed:-

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