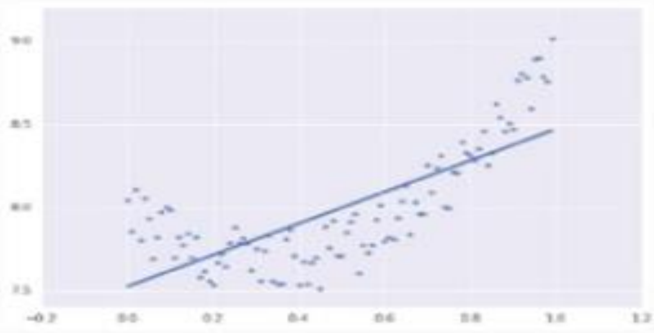
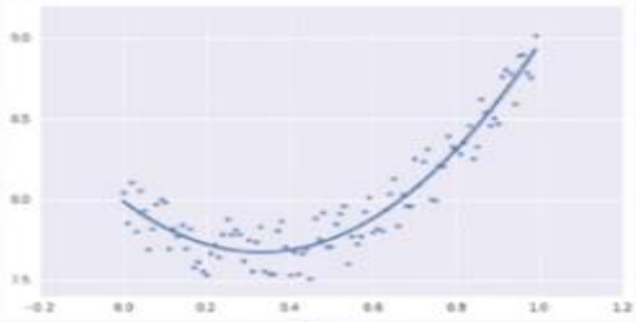
Agenda for today

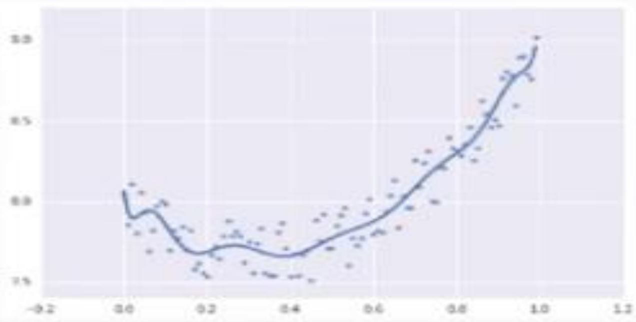
- High Bias/Underfitting and High Variance/Overfitting
- Ridge regression/ L2 regularization
- Lasso Regression / L1 regularization
- Broadcasting using Numpy



Underfit



Fit



Overfit

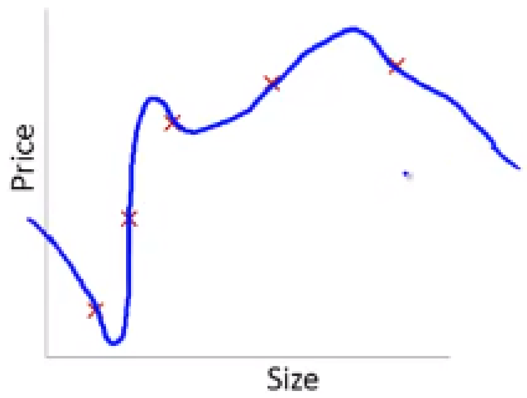
Overfitting

- If we have too many features, the learned hypothesis may fit the training set very well, but fail to generalise to new examples
- It memorises the data

Usually occurs due to high number of features

Addressing overfitting:

$x_1 =$ size of house
$x_2 =$ no. of bedrooms
$x_3 =$ no. of floors
$x_4 =$ age of house
$x_5 =$ average income in neighborhood
$x_6 =$ kitchen size
$\vdots$
x_{100}

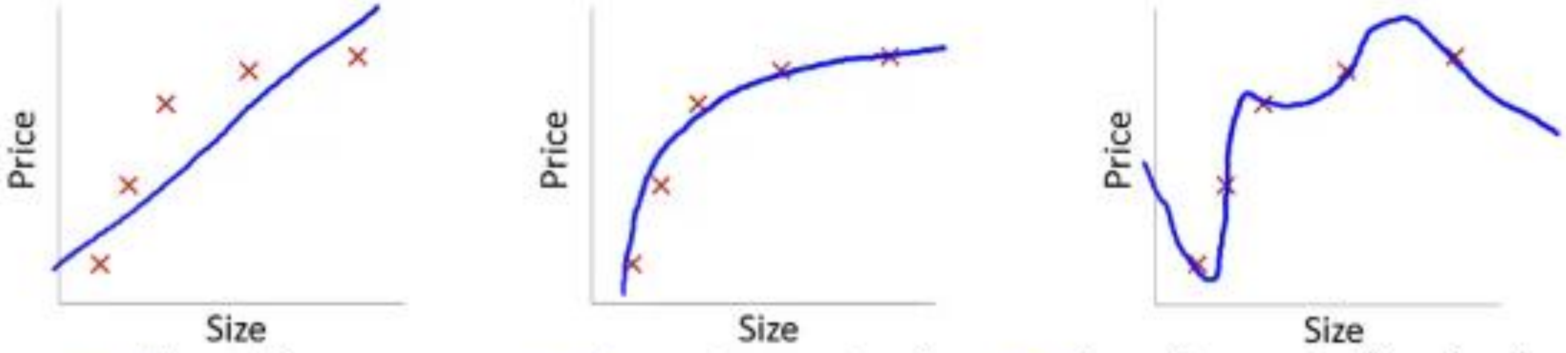


Regression example

Example: Linear regression (housing prices)

$\theta_0 + \theta_1 x$ $\qquad$ $\theta_0 + \theta_1 x + \theta_2 x^2$ $\qquad$ $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

Example: Linear regression (housing prices)



$\rightarrow \theta_0 + \theta_1 x$

"Underfit" "High bias"

$\rightarrow \theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

$\rightarrow \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

"Overfit" "High variance"


NOT SURE IF GOOD MODEL...
...OR JUST OVERFITTING
memegenerator.net

Is Overfitting really bad??

How to avoid overfitting??

Option1

Reduce number of features:

✓Usually not advisable.

✓Difficult to handpick few features

Option2

Increase the data with variance:

✓Not easy to get more data

Option3

Regularization:

✓Keep all the features, but reduce magnitude/values of parameters

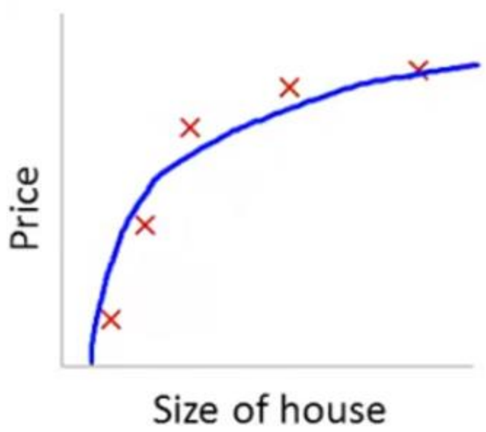
✓Works well when we have a lot of features, each of which contributes a bit to predicting output

Others:

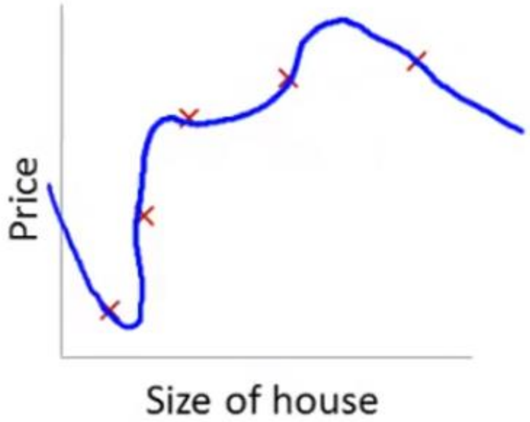
- Batch Normalization, Dropout, Early stopping, etc
- Usually used in Neural Networks

Regularization and cost function

Intuition



$\theta_0 + \theta_1 x + \theta_2 x^2$



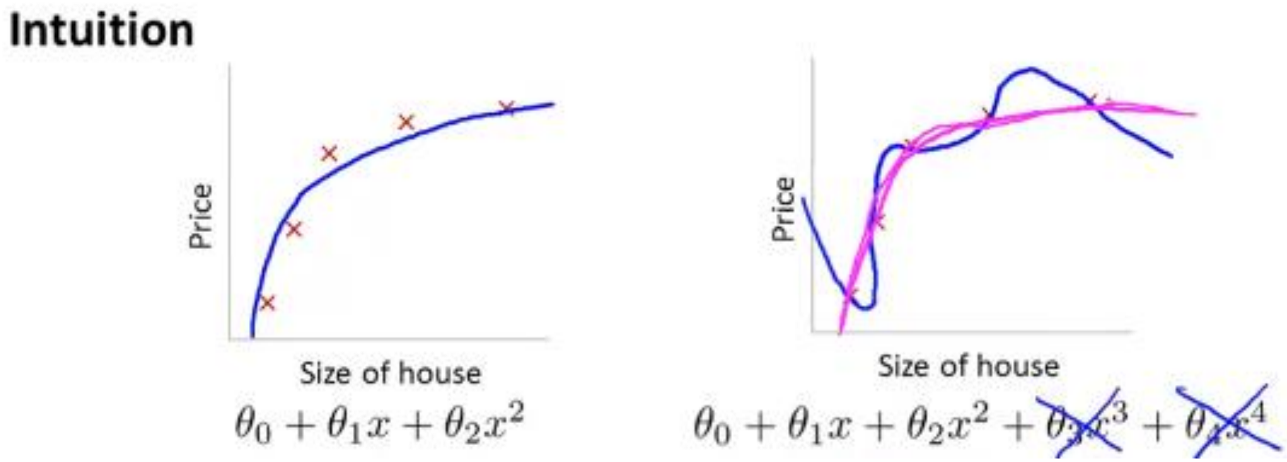
$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

Suppose we penalize and make θ_3, θ_4 really small.

$$\min_{\theta} \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

$$\min_{\theta} \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + 1000\,\theta_3^2 + 1000\,\theta_4^2$$

- Here $\theta 3$ *and* $\theta 4$ will have to be reduced to almost zero in order to minimise this function



Cost function

Regularized linear regression

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

$$\min_\theta J(\theta)$$

Gradient descent

$$\theta_j := \theta_j(1 - \alpha\frac{\lambda}{m}) - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

What if λ is too high??

Hint: Bias - Variance trade-off

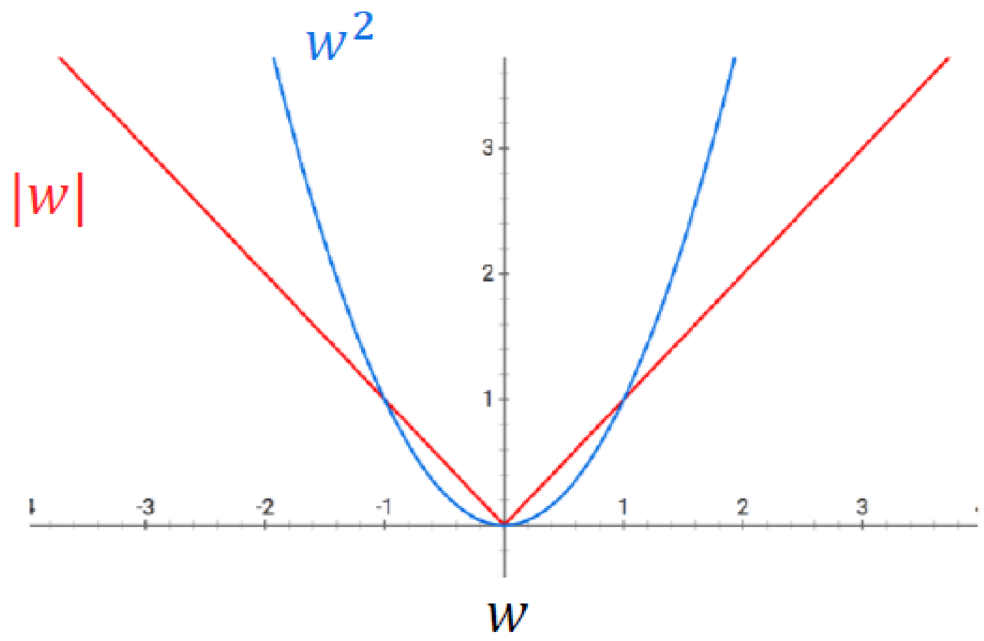
Tips to resolve overfitting

- Getting more training examples
- Try smaller set of features
- Try increasing λ

L1 regularization / Lasso regression

Visualizing of each type of regularizer in 1D

$$\min_{w} \|Xw - y\|_2^2 + \lambda w^2 \quad \text{vs} \quad \min_{w} \|Xw - y\|_2^2 + \lambda |w|$$



- For the L2 norm, the gradient gets smaller and smaller, so at some point, the pressure from that gradient will let up (counter-balanced by the data fit term).
- For the L1 norm, the pressure is constant, so keeps pushing strongly toward $w = 0$.

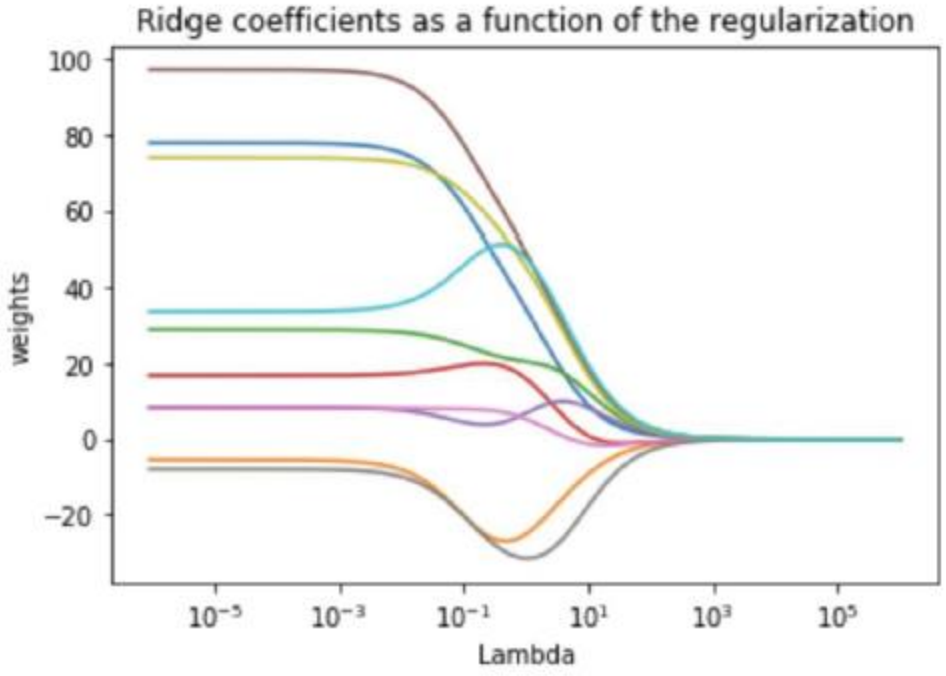
Points to remember for Lasso

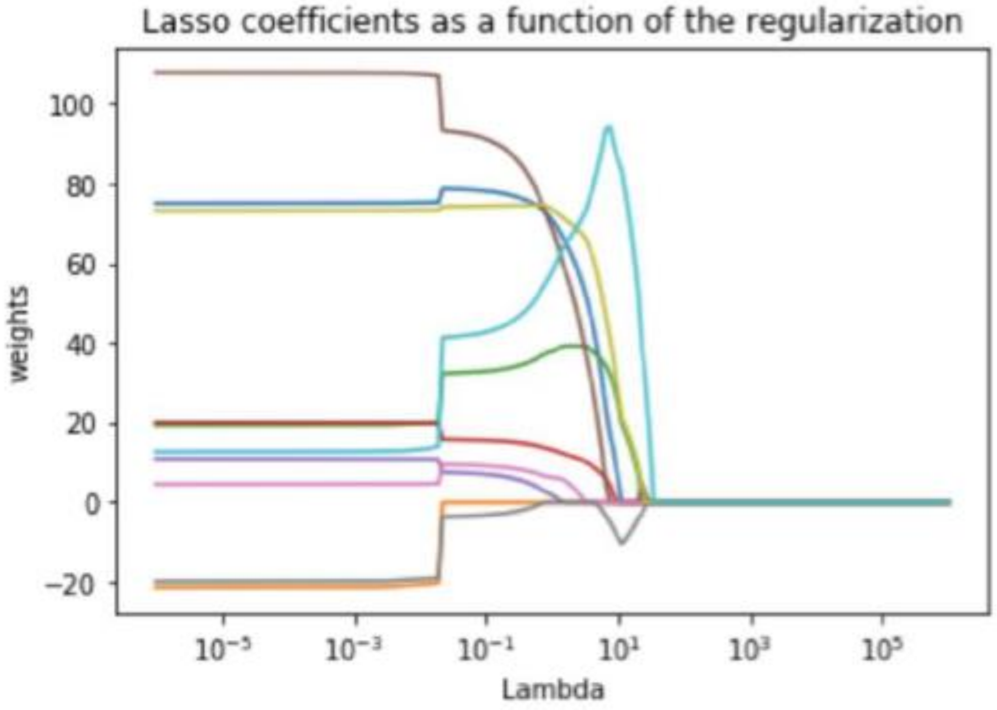
- The gradient of the regularizer term is constant λ always, unlike ridge where it is dependent on the weight as well
- So the pressure to reduce the weight values is constant in L1. Thus, weights will tend to go to zero
- However, in L2, pressure to reduce weight decreases as w decreases. Consequently, w will not tend to go to zero, they will just get small

Summarising L1 regularization

- As we increase the L1 (aka “Lasso”) penalty, λ, the weights shrink, as in ridge, but in a way such that more and more become precisely zero, unlike in ridge.
- Used for feature selection

Comparing and contrasting Ridge and Lasso





Quiz

RIDGE : GAUSSIAN :: LASSO : ?

RIDGE : GAUSSIAN :: LASSO : LAPLACIAN !!

Is model still linear after L2?

ONLY LOSS FUNCTION CHANGES!!

Why not L3 norm??


CONVEXITY
IS A HELL OF A DRUG