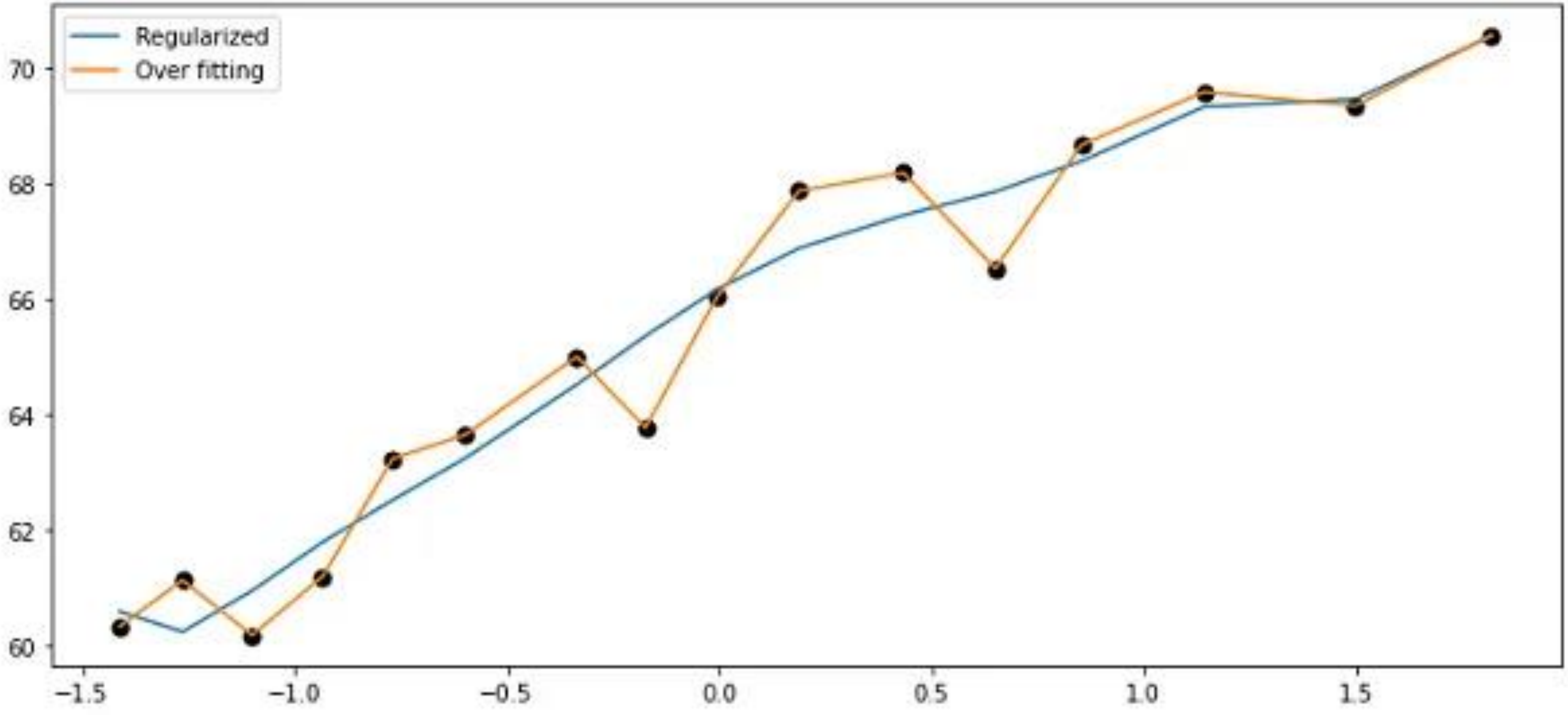
Regularized Linear Model



Regularized
Over fitting
70
68
66
64
62
60
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Intuitions on L1 and L2 Regularization

What's L1 and L2?

L1 and L2 regularisation owes its name to L1 and L2 norm of a vector *w* respectively. Here's a primer on norms:

$$||\mathbf{w}||_1 = |w_1| + |w_2| + \ldots + |w_N|$$

1-norm (also known as L1 norm)

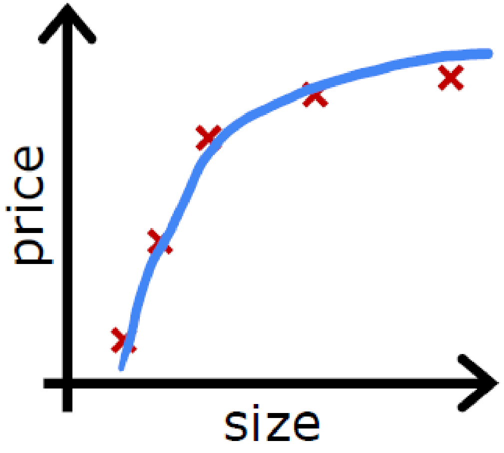
$$\|\mathbf{w}\|_2 = \left(|w_1|^2 + |w_2|^2 + \ldots + |w_N|^2\right)^{\frac{1}{2}}$$

2-norm (also known as L2 norm or Euclidean norm)

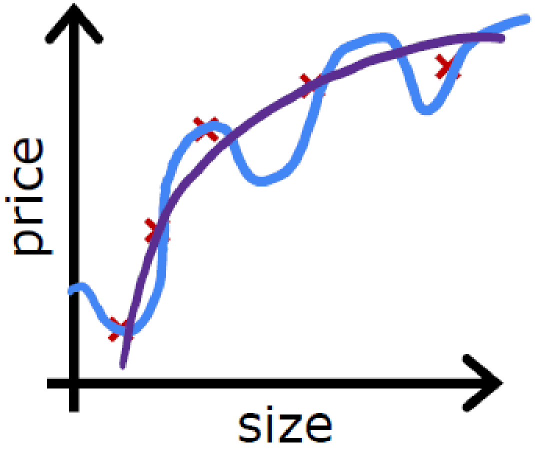
$$\|\mathbf{w}\|_p = \left(|w_1|^p + |w_2|^p + \ldots + |w_N|^p\right)^{\frac{1}{p}}$$

p-norm

Intuition



$w_1x + w_2x^2 + b$



$w_1x + w_2x^2 + \cancel{\underbrace{w_3}_{\approx 0}x^3} + \cancel{\underbrace{w_4}_{\approx 0}x^4} + b$

make w_3, w_4 really small (≈ 0)

$$\min_{\vec{w},b} \frac{1}{2m}\sum_{i=1}^{m}\left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}\right)^2 \cancel{+ 1000\,\underbrace{w_3^2}_{0.001}} \cancel{+ 1000\,\underbrace{w_4^2}_{0.002}}$$

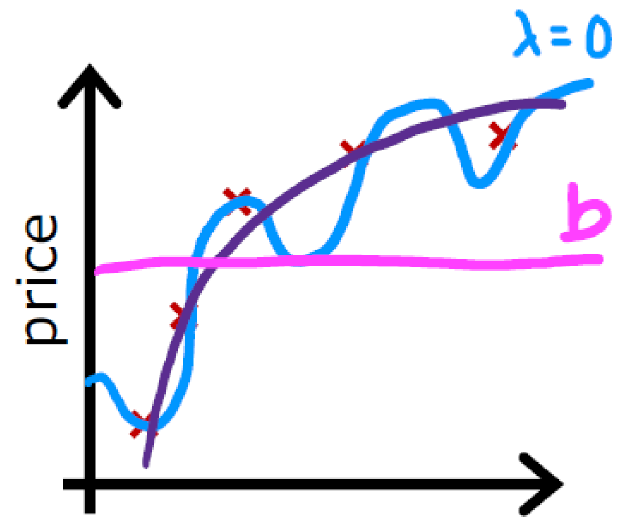
Regularization

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

fit data

Keep w_j small

λ balances both goals



choose $\lambda = 10^{10}$

$$f_{\vec{w},b}(\vec{x}) = \underset{\approx 0}{\cancel{w_1 x}} + \underset{\approx 0}{\cancel{w_2 x^2}} + \underset{\approx 0}{\cancel{w_3 x^3}} + \underset{\approx 0}{\cancel{w_4 x^4}} + b$$

$f(x) = b$

Choose λ

Case study on L2 Regression

Diabetes dataset

	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6	TARGET
0	0.038076	0.050680	0.061696	0.021872	-0.044223	-0.034821	-0.043401	-0.002592	0.019908	-0.017646	151.0
1	-0.001882	-0.044642	-0.051474	-0.026328	-0.008449	-0.019163	0.074412	-0.039493	-0.068330	-0.092204	75.0
2	0.085299	0.050680	0.044451	-0.005671	-0.045599	-0.034194	-0.032356	-0.002592	0.002864	-0.025930	141.0
3	-0.089063	-0.044642	-0.011595	-0.036656	0.012191	0.024991	-0.036038	0.034309	0.022692	-0.009362	206.0
4	0.005383	-0.044642	-0.036385	0.021872	0.003935	0.015596	0.008142	-0.002592	-0.031991	-0.046641	135.0

- Multivariate Linear Regression Model
- How many parameters ?

What is your observation?

alpha	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6
0.0000	-9.160885	-205.462260	516.684624	340.627341	-895.543609	561.214533	153.884786	126.734316	861.121400	52.419828
0.0001	-9.118336	-205.337133	516.880570	340.556792	-883.415291	551.553259	148.578680	125.355917	856.480254	52.467627
0.0010	-8.763583	-204.321125	518.371729	339.975385	-787.690766	475.274718	106.786540	114.632063	819.739542	52.872100
0.0100	-6.401088	-198.669767	522.048548	336.348363	-383.709187	152.663678	-66.060583	75.611090	659.869402	55.828128
0.1000	6.642753	-172.242166	485.523872	314.682122	-72.939323	-80.590053	-174.466515	83.616653	484.363285	73.584154
1.0000	42.242217	-57.305508	282.170831	198.061386	14.363544	-22.551274	-136.930053	102.023193	260.104308	98.552274
10.0000	21.174004	1.659796	63.659772	48.493240	18.421492	12.875448	-38.915435	38.842464	61.612405	35.505355
100.0000	2.858979	0.629452	7.540604	5.849997	2.710879	2.142134	-4.834047	5.108223	7.448466	4.576129
1000.0000	0.295726	0.069290	0.769004	0.597829	0.282900	0.225936	-0.495607	0.527031	0.761497	0.471029
10000.0000	0.029674	0.006995	0.077054	0.059915	0.028412	0.022715	-0.049686	0.052870	0.076321	0.047241

Case study on L1 Regression (LASSO)

Diabetes dataset

	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6	TARGET
0	0.038076	0.050680	0.061696	0.021872	-0.044223	-0.034821	-0.043401	-0.002592	0.019908	-0.017646	151.0
1	-0.001882	-0.044642	-0.051474	-0.026328	-0.008449	-0.019163	0.074412	-0.039493	-0.068330	-0.092204	75.0
2	0.085299	0.050680	0.044451	-0.005671	-0.045599	-0.034194	-0.032356	-0.002592	0.002864	-0.025930	141.0
3	-0.089063	-0.044642	-0.011595	-0.036656	0.012191	0.024991	-0.036038	0.034309	0.022692	-0.009362	206.0
4	0.005383	-0.044642	-0.036385	0.021872	0.003935	0.015596	0.008142	-0.002592	-0.031991	-0.046641	135.0

- Multivariate Linear Regression Model
- How many parameters ?

What is your observation?

alpha	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6
0.0000	-9.160885	-205.462260	516.684624	340.627341	-895.543596	561.214523	153.884780	126.734314	861.121395	52.419828
0.0001	-9.071288	-205.337332	516.780313	340.539730	-888.652320	555.952271	150.585260	125.453044	858.639860	52.379002
0.0010	-8.264924	-204.213177	517.641106	339.751339	-826.653342	508.609613	120.899583	113.924518	836.314382	52.011583
0.0100	-1.361404	-192.944226	526.348511	332.649058	-430.205495	191.277876	-44.048113	68.990747	688.384976	47.939528
0.1000	0.000000	-113.976046	526.737112	292.635423	-82.691928	-0.000000	-152.691332	0.000000	551.077200	7.169852
1.0000	0.000000	0.000000	363.882636	27.278420	0.000000	0.000000	-0.000000	0.000000	336.135971	0.000000
10.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000
100.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000
1000.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000
10000.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000

When to use which?

- Ridge Regression is suitable if you want to keep all the features and avoid that the model becomes over sensitive to the noise/fluctuations in the training data.
- If you think that only few features are useful, then it would be better to use Lasso Regression or Elastic Net as they sets the weights of less-important features to zero. But keep in mind that Lasso gives a higher level of sparsity i.e. most of the coefficients are set to zero.
- In general, Elastic Net is preferred over Lasso because Lasso may behave in an unpredictable way when the number of features is greater than number of training samples or when many features are highly correlated.

Early Stopping