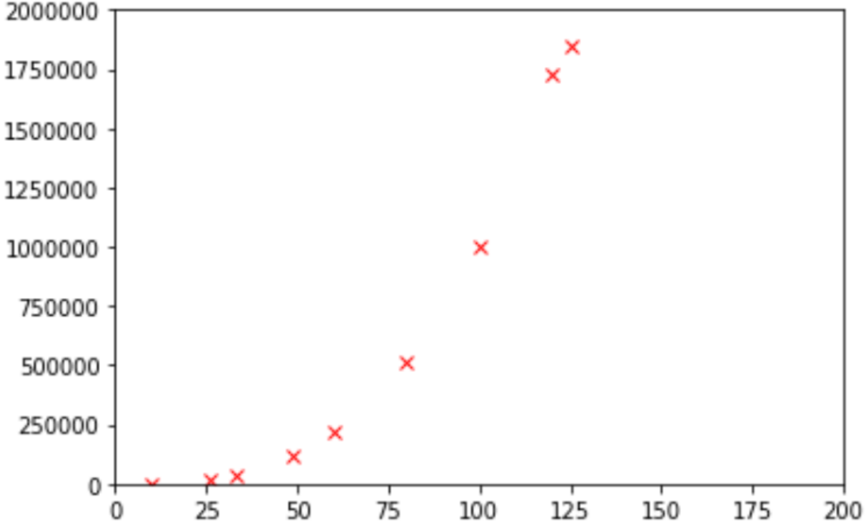
Chapter 7

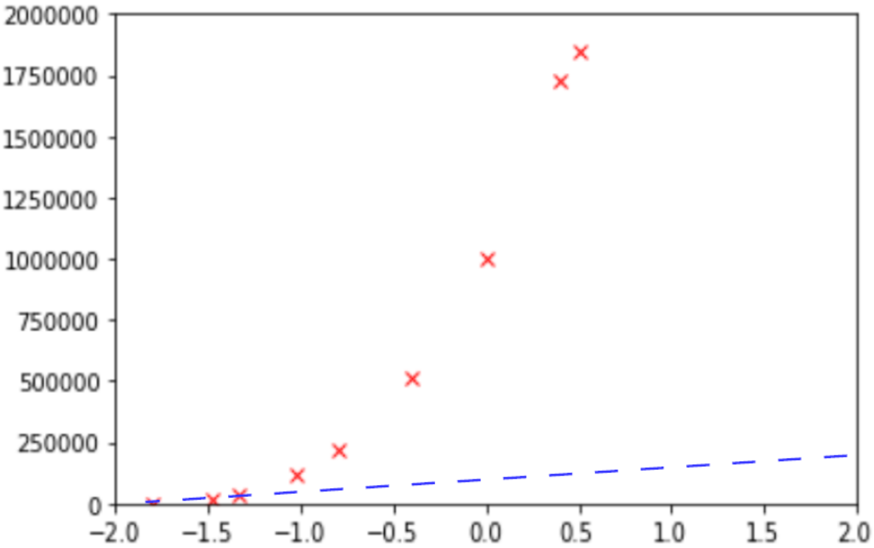
Polynomial Regression

Motivation

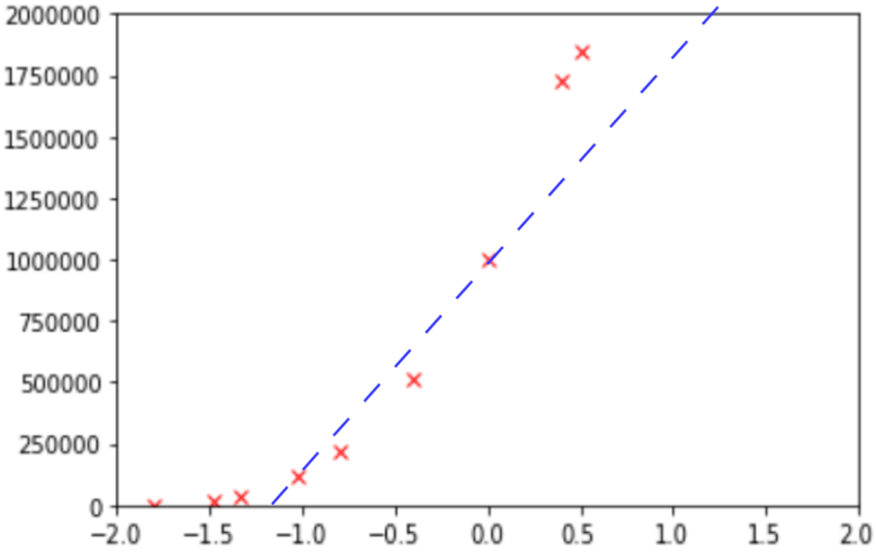
- Linear regression can't fit all kind of data
- Sometimes the pattern between variables can only be described with polynomial equations



Motivation



2000000
1750000
1500000
1250000
1000000
750000
500000
250000
0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0



2000000
1750000
1500000
1250000
1000000
750000
500000
250000
0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

Polynomial Regression - Single Variable

- p is the degree of the polynomial

x	y
x_1	y_1
x_2	y_2
…	…
x_i	y_i
…	…
x_n	y_n

→

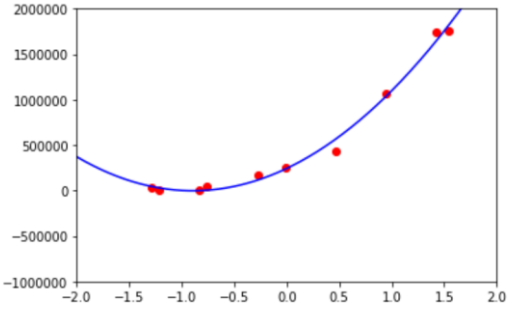
x	x^2	…	x^p	y
x_1	$x_1^{\,2}$	…	$x_1^{\,p}$	y_1
x_2	$x_2^{\,2}$	…	$x_2^{\,p}$	y_2
…	…	…	…	…
x_i	$x_i^{\,2}$	…	$x_i^{\,p}$	y_i
…	…	…	…	…
x_n	$x_n^{\,2}$	…	$x_n^{\,p}$	y_n

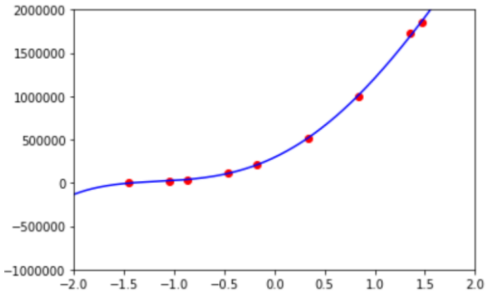
$$\hat{y} = w^{(0)} + w^{(1)}x + w^{(2)}x^2 + \ldots + w^{(p)}x^p =$$
$$= w^{(0)}x^0 + w^{(1)}x + w^{(2)}x^2 + \ldots + w^{(p)}x^p$$
$$= \sum_{j=0}^{p} w^{(j)}x^p$$

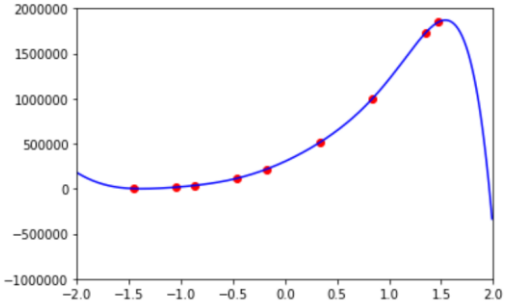
- We introduce new predictor features for each polynomial component and use linear regression on the extended data

Polynomial Regression - Single Variable

Slightly different training samples can cause very different learned parameters



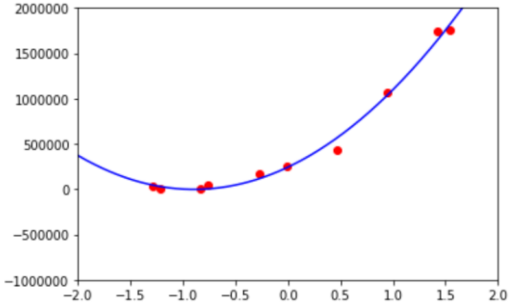


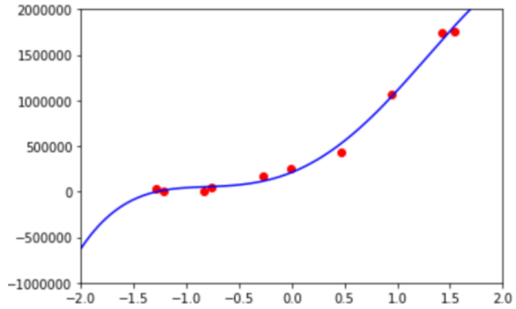


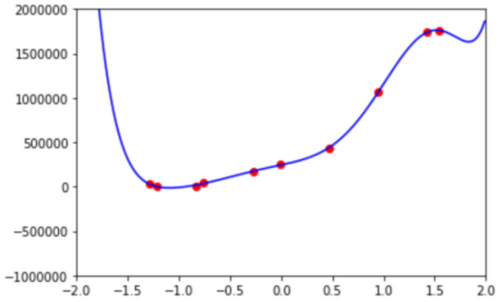
d = 2

d = 4

d = 8







Regularization methods

Regularization methods help reduce variance (sensitivity to training samples) by introducing cost function on the learned weights. This is very useful when we have many variables and we are not sure if we need all of them.

Ridge regression

$$argmin_{\hat{\beta}_0,\ldots,\hat{\beta}_p}\left\{\sum_{i=1}^{n}\left[y_i-\left(\underbrace{\hat{\beta}_0+\sum_{j=1}^{p}\hat{\beta}_j\times x_{ij}}_{\hat{y}_i}\right)\right]^2+\lambda\times\sum_{j=1}^{p}\hat{\beta}_j^2\right\}$$

- Deals better with correlated attributes than ordinary least squares regression
- L2 regularization

Lasso regression

$$argmin_{\hat{\beta}_0,\ldots,\hat{\beta}_p}\left\{\sum_{i=1}^{n}\left[y_i-\left(\underbrace{\hat{\beta}_0+\sum_{j=1}^{p}\hat{\beta}_j\times x_{ij}}_{\hat{y}_i}\right)\right]^2+\lambda\times\sum_{j=1}^{p}|\hat{\beta}_j|\right\}$$

- Deals better with correlated attributes than ridge regression
- produces simpler models than ordinary least squares or ridge regression
- Automatically discounts irrelevant attributes
- L1 regularization

Polynomial Regression - Multiple Variables

$$\hat{y} = w^{(0)} + w^{(1,1)}x^{(1)} + w^{(1,2)}x^{(2)} + \ldots +$$
$$+ w^{(2,1)}(x^{(1)})^2 + w^{(2,2)}(x^{(2)})^2 + w^{(2,3)}(x^{(1)}x^{(2)})$$

$x^{(1)}$	$x^{(2)}$	…	$x^{(m)}$	y
$x_1^{(1)}$	$x_1^{(2)}$	…	$x_1^{(m)}$	y_1
$x_2^{(1)}$	$x_2^{(2)}$	…	$x_2^{(m)}$	y_2
…	…	…	…	…
$x_i^{(1)}$	$x_i^{(2)}$	…	$x_i^{(m)}$	y_i
…	…	…	…	…
$x_n^{(1)}$	$x_n^{(2)}$	…	$x_n^{(m)}$	y_n

$x^{(0)}$	$x^{(1)}$	$x^{(2)}$	…	$x^{(m)}$	$(x^{(1)})^2$	$(x^{(2)})^2$	$x^{(1)}x^{(2)}$	…	y
1	$x_1^{(1)}$	$x_1^{(2)}$	…	$x_1^{(m)}$	$(x_1^{(1)})^2$	$(x_1^{(2)})^2$	$x_1^{(1)}x_1^{(2)}$	…	y_1
1	$x_2^{(1)}$	$x_2^{(2)}$	…	$x_2^{(m)}$	$(x_2^{(1)})^2$	$(x_2^{(2)})^2$	$x_2^{(1)}x_2^{(2)}$	…	y_2
…	…	…	…	…	…	…	…	…	…
1	$x_i^{(1)}$	$x_i^{(2)}$	…	$x_i^{(m)}$	$(x_i^{(1)})^2$	$(x_i^{(2)})^2$	$x_i^{(1)}x_i^{(2)}$	…	y_i
…	…	…	…	…	…	…	…	…	…
1	$x_n^{(1)}$	$x_n^{(2)}$	…	$x_n^{(m)}$	$(x_n^{(1)})^2$	$(x_n^{(2)})^2$	$x_n^{(1)}x_n^{(2)}$		y_n

Literature

- https://towardsdatascience.com/polynomial-regression-bbe8b9d97491
- Polynomial features explained https://datascience.stackexchange.com/a/71942/53849
- http://biointelligence.hu/pdf/02-from-linear-regression-to-deep-learning.pdf (Special thanks to Krisztian Buza for allowing me to use materials from this presentation)

Questions?