Linear Regression

Dr. Mohamed Elshenawy

1

Previous session …

- Course Overview
- What is Learning?
- Why do we need machine learning?
- Types of learning
- Practice
  - Learn about the tools and libraries
  - Start playing with the data
    - Explore the dataset
    - Understand different data types
    - Pre- Processing

2

This session…

- Simple Linear Regression
- Steps to fit a linear regression model
- Multiple and Polynomial Regression
- Overfitting and Underfitting
- Regularization:
  - Ridge Regression
  - Lasso Regression
  - Elastic-Net Regression

3

Example

- Given a history of sold houses in the last 5 years within a certain neighborhood, your client wants you to build a system that can estimate the fair price of a house given its characteristics (size, number of rooms, etc.)
- What is the output of the model?
- Is it quantitative or qualitative?
- What are the input features?
- Are these features quantitative or qualitative?

4

Regression Analysis

- Regression: a measure of the relationship between the mean value of one variable (output of your model) and corresponding values of other variables (input features).

Let's assume the following Dataset

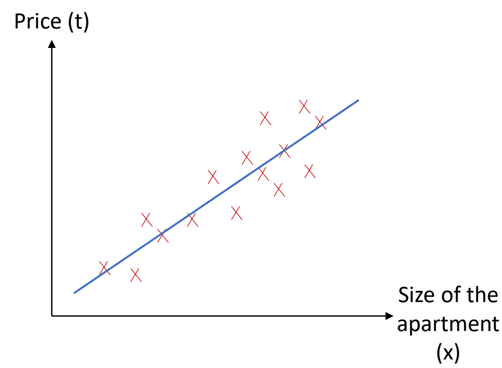
Data set

- *n* cases *i= 1,2,…….n*
- *1 target variable (price)*
    - t_i , $i = 1,2,\ldots\ldots.n$
    - $t_1=290$, $t_2=405$, $t_3=200$,….
- *1 input variable (size)*
    - x_i , $i = 1,2,\ldots\ldots.n$
    - $x_1=1320$, $x_2=1900$, $x_3=900$,….

Size in feet^2	Price in thousands
1320	290
1900	405
900	200
1600	340
…	…

Linear Regression

- Assumes a linear relationship between the mean of the *target/ output* variable (the factor you are trying to predict (also known as the dependent variable) and the *input* variables (features that are expected to affect the target variable (also known as "predictor/ explanatory" variables)
- $t_i = \theta_1 * x_i + \theta_2 + \epsilon_i$



7

What shall the model learn in that case?

- We need to learn θ_0 and θ_1 (model parameters)
- If I can estimate θ_0 and θ_1 correctly, I have an estimation of the output given the input attributes with some error (ϵ_i).
- Prediction error (residual) is defined as $\epsilon_i = t_i - y_i$
  - t_i: observed output value for dataset record *i (true price value in the previous example)*.
  - y_i: our estimate of the target variable for dataset record *I* ($y_i = \theta_1 * x_i + \theta_2$).
- How can we estimate θ_0 and θ_1?

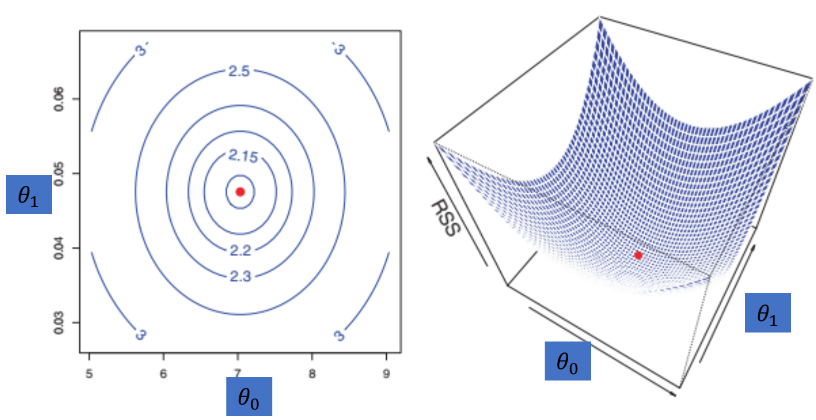
8

Least Squares Criterion

- The most popular estimation method.
- **Least Squares Criterion:** *"minimize the sum of the squared prediction errors."*
- **Residual Sum of Squares** $(RSS) = e_1^2 + e_2^2 + \cdots . + e_n^2$
- find the values θ_0 and θ_1 that make the sum of the squared prediction errors the smallest it can be.
- How? Using the **Ordinary least squares (OLS) method**

9

Optimization Problem: find θ_0 and θ_1 to minimize RSS



Contour and three-dimensional plots of RSS along with the model parameters

Book: An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0.

10

Solving the optimization problem (mathematically)

$$J(\theta) = \sum_{i=1}^{n} (t_i - y_i)^2 \,, y_i = \theta_0 + \theta_1 x_i$$

$$J(\theta) = \sum_{i=1}^{n} (t_i - (\theta_0 + \theta_1 x_i))^2$$

To minimize $J(\theta) =$, take the derivative with respect to θ_0 and θ_1, set to 0, and the model parameters

11

Solution

$$\theta_0 = \bar{t} - \theta_1 \bar{x},$$

$$\bar{t} = \frac{\sum_{i=1}^{n} t_i}{n}, \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

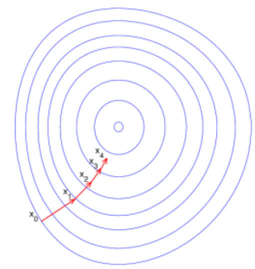
$$\theta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(t_i - \bar{t})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

12

We can also solve it using Gradient Descent (iterative learning method)

- One of the generic algorithms used to solve optimization problems

1. Initialize θ randomly.
2. repeatedly update θ based on the gradient



$$\theta \leftarrow \theta - \lambda \frac{\partial J(\theta)}{\partial \theta}$$

For a single training case i:

$$\frac{\partial J_i(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}(t_i - y_i)^2$$

13

Gradient Descent (cont)

$$y_i = \theta_0 + \theta_1 x_i$$

$$J_i(\theta) = \left(t_i - (\theta_0 + \theta_1 x_i)\right)^2$$

$$\frac{\partial J_i(\theta)}{\partial \theta_0} = 2(t_i - y_i)(\text{-}1)$$

$$\theta_0 \leftarrow \theta_0 - 2\lambda(t_i - y_i)(\text{-}1)$$

$$\frac{\partial J_i(\theta)}{\partial \theta_1} = 2(t_i - y_i)(\text{-}\,x_i)$$

$$\theta_1 \leftarrow \theta_1 - 2\lambda(t_i - y_i)(\text{-}\,x_i)$$

14

Gradient Descent (cont.)

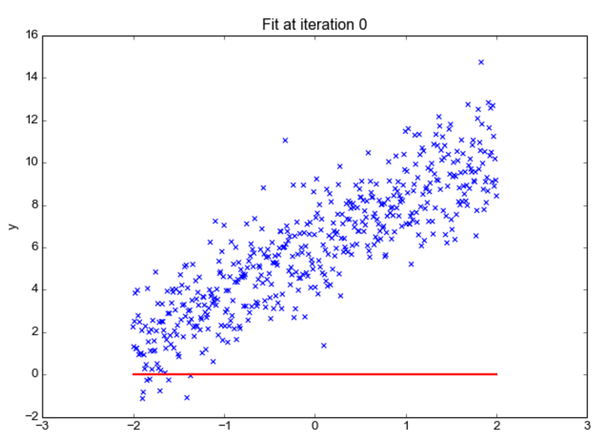
- For all points in the data set
- Batch update
  - Sum or average updates across every example n, update for all samples

$$\theta_0 \leftarrow \theta_0 + \frac{2\lambda}{n}\sum_{i=1}^{n}(t_i - y_i)\,;\ \theta_1 \leftarrow \theta_1 + \frac{2\lambda}{n}\sum_{i=1}^{n}(t_i - y_i)x_i$$

- Stochastic/online updates: update the parameters for each training case in turn, according to its own gradients

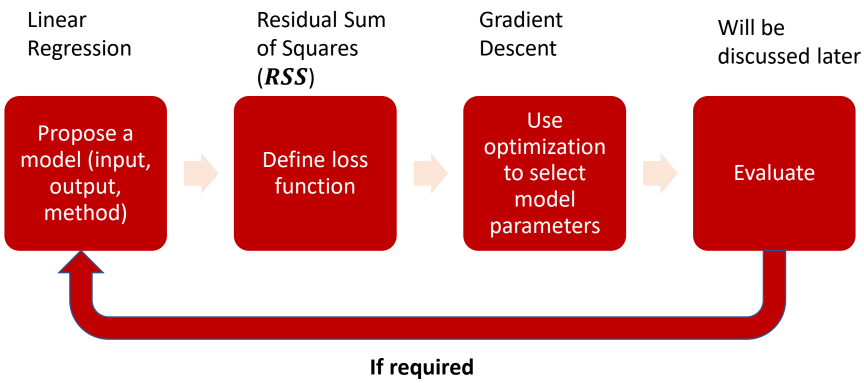
15

Fitting a model



16

Steps to train a regression model



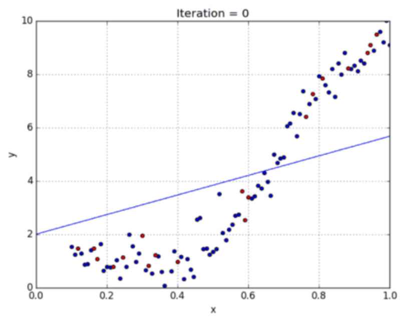
17

How about if we have non-linear relationship ?

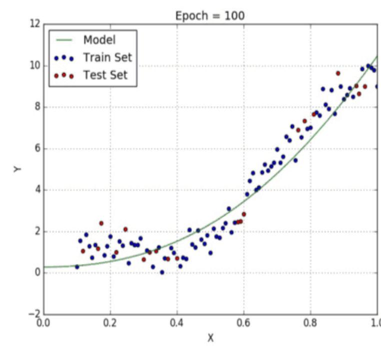
18

Polynomial Regression

$$T = \theta_0 + \theta_1 X_1 + \theta_2 X_1^2 + \theta_3 X_1^3 + \cdots + \theta_m X_1^m + \mathrm{E}$$



$$Y = \theta_0 + \theta_1 X_1$$

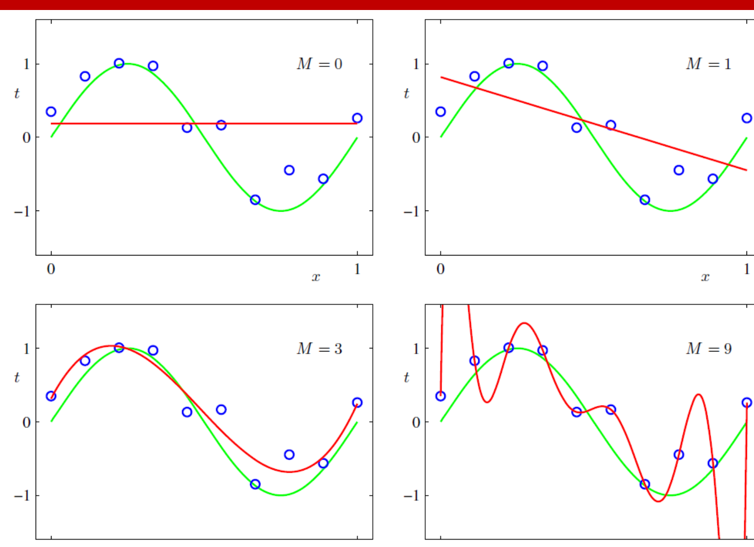


$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_1^2$$

19

How many parameters shall we use?

Plots of polynomials having various degrees, shown as red curves, Which fit is the best?



20

Model Capacity

- A model's capacity is its ability to fit a wide variety of functions.

Capacity

- A linear regression model (M1) with two parameters (A polynomial of degree one)

$$y_{M1} = \theta_0 + \theta_1 x$$

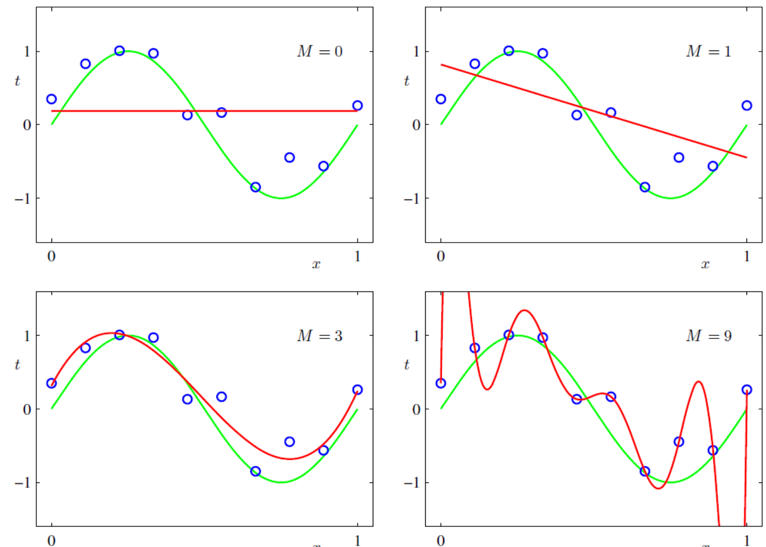
- The model can fit several functions. For example $y = 4;\ y = 2x; y = 3 + 1.5x; y = 0.5 + 0.25x$
- By adding one more parameter and use x^2(A polynomial of degree 2)

$$y_{M2} = \theta_0 + \theta_1 x + \theta_2 x^2$$

- The model (M2) can fit several more functions that cannot be represented by the first mode. For example $y = 2x^2; y = 2 + 1.5x^2; y = 0.5x + 0.2x^2$. It can also fit all functions of M1.
- We say that M2 has more capacity than M1

Now back to our example

- Does having more capacity mean better performance?
- How can we evaluate the performance in that case?

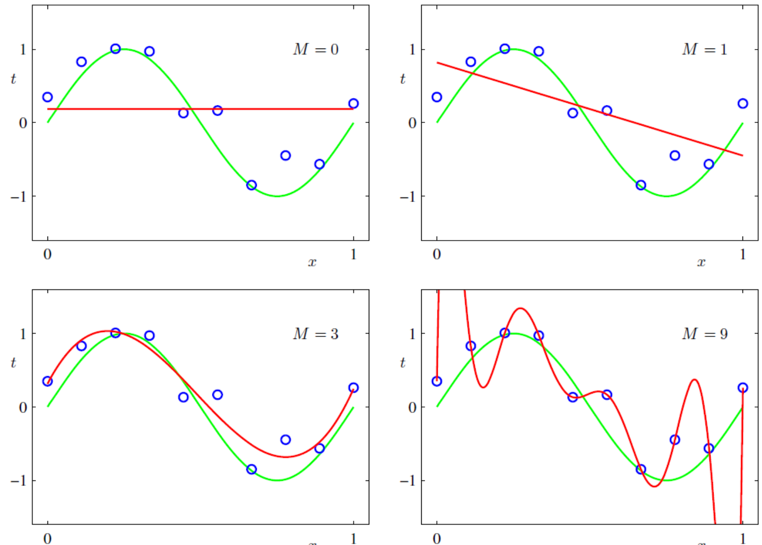


Training and test data

- We split the data into two segments:
  - Training set, used to learn the model parameters. The error calculated using this set is called the training error.
  - Test set, used to assess the generalization of the model (Model's ability to perform well on unseen data). Why is this important?

Again, back to our example

- Which of the four models has low training error and high test error?
- We call this overfitting

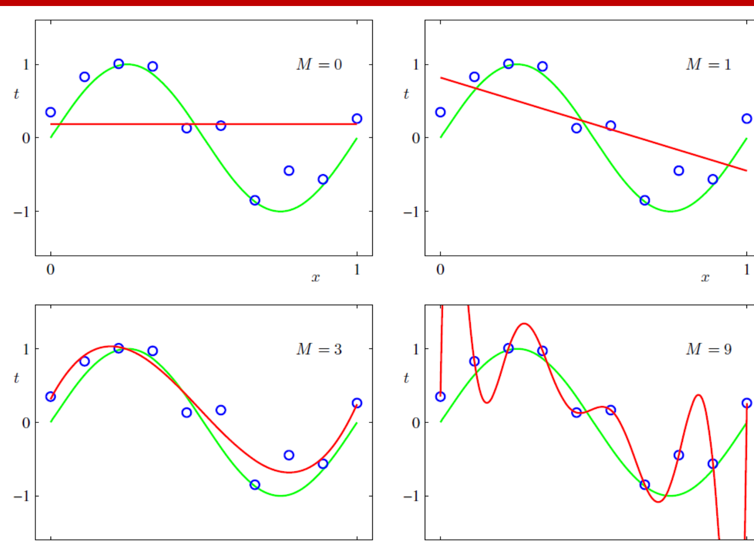


25

Again, back to our example

- Which of the four models will has high training error?
- We call this underfitting

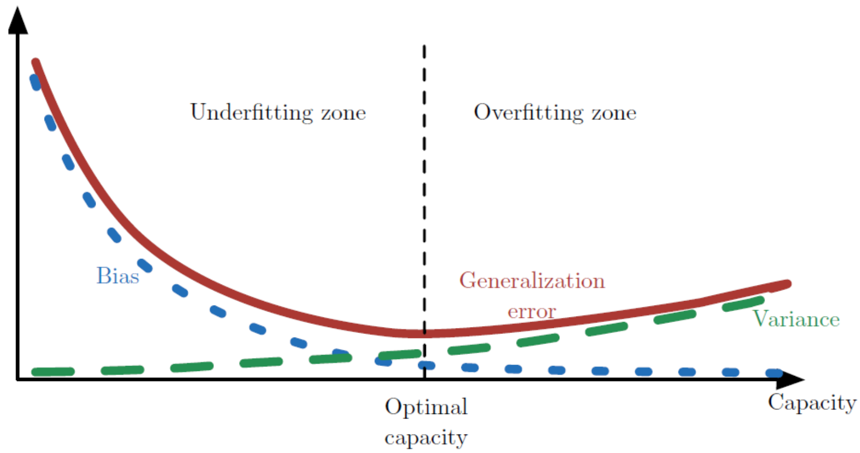
From Bishop's textbook section (1.1)



26

Overfitting, underfitting, and model capacity

We will discuss bias-variance tradeoff in the upcoming lecture



Section 5.4 from the book:Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. Vol. 1. Cambridge: MIT press, 2016.

27

How can we prevent overfitting

- Best way is to train the model using more data
- Sometimes this is difficult. We can also
  - Reduce model complexity.
  - Use regularization techniques

28

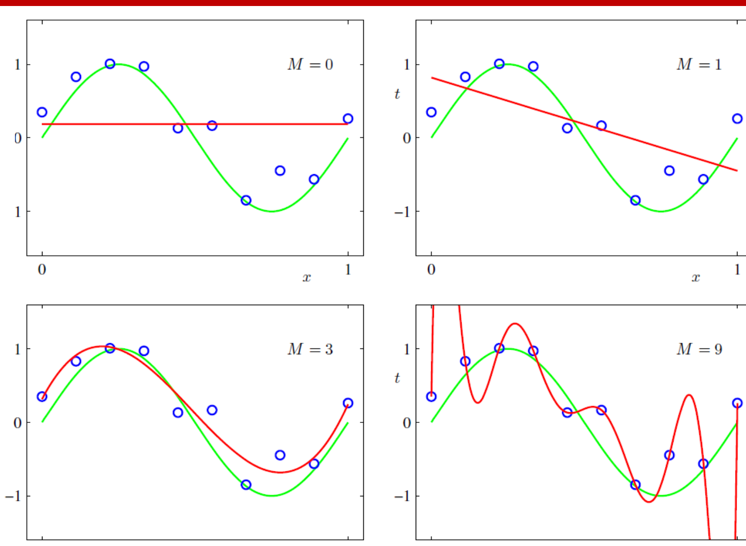
Regularization

- Any modification we make to a learning algorithm that is intended to reduce its generalization/test error, but not its training error.

Let's have a closer look at the previous example

	$M = 0$	$M = 1$	$M = 6$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

What do you notice about the values of the model parameters?



Regularization

- A possible way to avoid overfitting is to add extra terms in the objective function that can be thought as corresponding to a soft constraint on the parameter values.
- How?

31

Review: Norm

- A measure the size of a vector (maps a vector to a non-negative value representing its size).
- L^p norm is given by

$$L^p \text{ norm } of\ some\ vector\ x = \|x\|_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$$

For $p \geq 1$

32

L^1 and L^2 Norm

- For p=2 (L^2 norm)

$$\|x\|_2 = \sqrt{\sum_i |x_i|^2}$$

- L^2 norm is simply the Euclidean distance from the origin to the point identified by x
- The squared L^2($\|x\|_2^2$) norm is typically more convenient to work with mathematically and computationally than the L2 norm itself.
- For p=1 (L^1 norm)

$$\|x\|_1 = \sum_i |x_i|$$

33

Using the norm to constrain parameter values

- Basic idea: add a parameter norm penalty $\Omega(\theta)$ to the objective function J

$$\tilde{J}(\theta; x, y) = J(\theta; x, y) + \lambda\Omega(\theta)$$

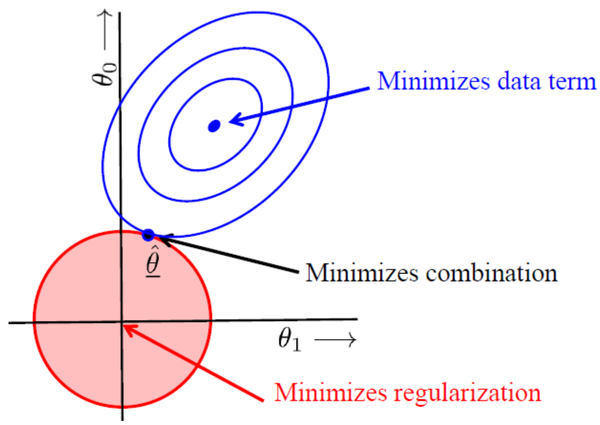
- Where $\lambda \in [0, \infty)$is a hyper-parameter that weights the relative contribution of the norm penalty term $\Omega(\theta)$

34

Ridge Regression(L^2 Parameter Regularization)

- Basic idea:

  Add regularization term
  $\Omega(\theta)=\|\theta\|_2^2$



35

Ridge Regression

$$\hat{\theta}^{ridge} = \min\left\{\sum_{i=1}^{n}\left(t_i - \theta_0 - \sum_{j=1}^{p} x_{ij}\,\theta_j\right)^2 + \lambda\sum_{j=0}^{p}\theta_j^2\right\}$$

$\lambda \geq 0$: a complexity parameter that controls the amount of shrinkage

Leads to changing the update rule for gradient descent`

$$\frac{\partial}{\partial\theta_0}\lambda\sum_{j=0}^{p}\theta_j^2 = \lambda\frac{\partial}{\partial\theta_0}[\theta_0^2 + \theta_1^2 + .. + \theta_p^2] = 2\lambda\theta_0$$

Similarly, you can find $\frac{\partial}{\partial\theta_1}$, ..., $\frac{\partial}{\partial\theta_p}$

$$\theta_0 \leftarrow \theta_0 + \frac{2\lambda}{n}\left[\sum_{i=1}^{n}(t_i - y_i) - 2\alpha\theta_0\right];\ \theta_k \leftarrow \theta_k + \frac{2\lambda}{n}\left[\sum_{i=1}^{n}(t_i - y_i)x_{ik} - 2\alpha\theta_k\right]$$

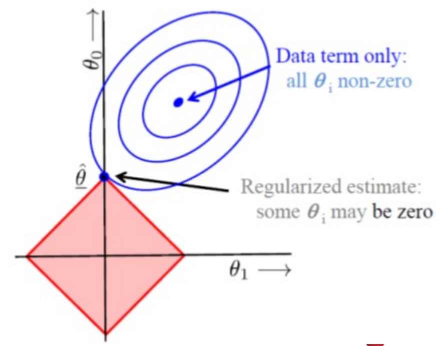
36

Lasso Regression(L^1Parameter Regularization)

- Basic idea:

  Add regularization term $\Omega(\theta)=\|\theta\|_1$
- Used as a method of feature selection

$$\hat{\theta}^{lasso} = \min\left\{\sum_{i=1}^{n}\left(t_i - \theta_0 - \sum_{j=1}^{p} x_{ij}\,\theta_j\right)^2 + \lambda\sum_{j=0}^{p}|\theta_j|\right\}$$



37

L^1 and L^2 regularization

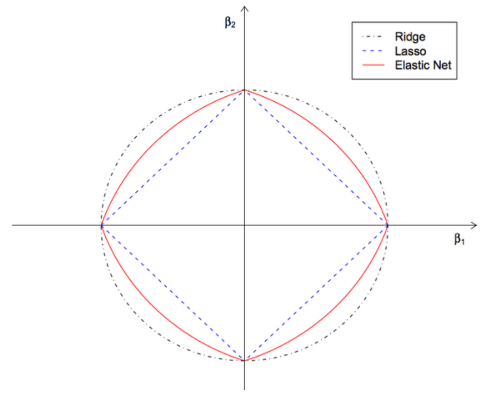
- L^1regularization, minimizes the sum of the absolute values
- L^2 regularization minimizes the sum of squares.

- Choosing L^1 or L^2 That depends on the specific problem
- L^1 regularization has an important advantage: it tends to produce a sparse model. That is, it often sets many weights to zero, effectively declaring the corresponding attributes to be irrelevant.

38

Elastic Net Regularization

- Emerged as a result of critique on lasso
- Combines $\boldsymbol{L^1}$ and $\boldsymbol{L^2}$ penalties to get the best of both
- $\tilde{J}(\theta; x, y) = J(\theta; x, y) + \lambda_1 \|\theta\|_1 + \lambda_2 \|\theta\|_2^2$



https://medium.com/Fmlearning-ai/elasticnet-regression-fundamentals-and-modeling-in-python

39

How about if we have multiple inputs?

40

Multiple Linear Regression

- Given multiple inputs $[X_1, X_2, \dots, X_p]$, the linear regression model has the form

$$Y = \theta_0 + \sum_{j=1}^{p} \theta_j X_j$$

- $X_1 = [x_{11}, x_{21}, x_{31}, \dots \dots, x_{n1}], X_2 = [x_{12}, x_{22}, x_{32}, \dots \dots, x_{n2}], \dots . X_p = [x_{1p}, x_{2p}, x_{3p}, \dots \dots, x_{np}]$
- Y= $[y_1, y_2, y_3, \dots \dots, y_n]$

- What do we need to learn in this case?
- $\theta_0, \theta_1, \theta_2, \dots \dots, \theta_p$
- We use OLS in a way similar to we discussed before

41

Solution

- $\min_{\theta} (t - X\theta)^T (t - X\theta)$
- Closed form solution
- $\theta = (X^T X)^{-1} X^T \mathrm{t}$

42

Multiple Linear Regression

- X_j can be:
  - Quantitative input;
  - Numeric or "dummy" coding of the levels of **qualitative** inputs (e.g. 1 for category adult and 0 for category child).
  - Transformations of quantitative inputs, such as log, square-root or square;
  - Basis expansions, such as:
    - for sample i, $x_{i2} = x_{i1}^2, x_{i3} = x_{i1}^3$ leading to a polynomial representation;
  - Interactions between variables, for example, for sample I, $x_{i3} = x_{i1}.x_{i2}$

43

mmelshenawy@gmail.com

44