统计机器学习

（小班研讨）

主讲教师：刘峤

回归算法

An introduction to regression

Three interpretations of regression

- **Linear regression**

$$\hat{y} = \mathbf{w} \cdot \mathbf{x}$$

- **Probabilistic (Maximum Likelihood Estimation, MLE)**

$$y \sim \mathcal{N}(\mathbf{w} \cdot \mathbf{x}, \sigma^2)$$

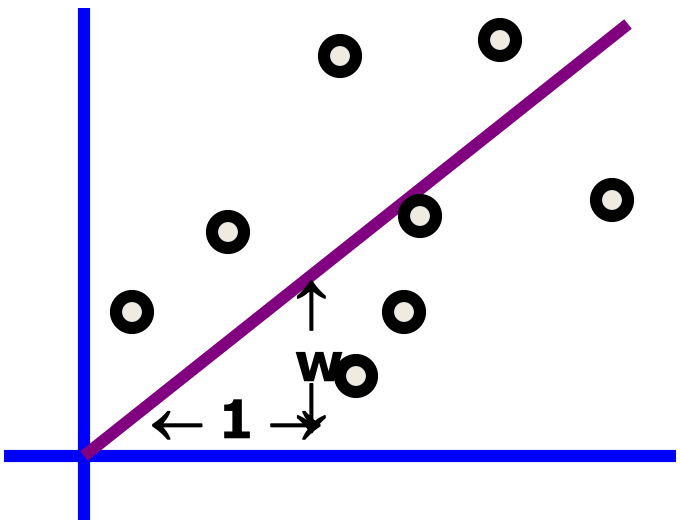
$$\underset{\mathbf{w}}{\operatorname{argmax}}\, p(y|\mathbf{w}, \mathbf{x})$$

- **Bayesian (Maximum Posterior Probability Estimation, MAP)**

$$\underset{\mathbf{w}}{\operatorname{argmax}}\, p(y|\mathbf{w}, \mathbf{x})p(\mathbf{w}, \mathbf{x})$$

Single-Parameter Linear Regression

Linear Regression



DATASET

inputs	outputs
$x_1 = 1$	$y_1 = 1$
$x_2 = 3$	$y_2 = 2.2$
$x_3 = 2$	$y_3 = 2$
$x_4 = 1.5$	$y_4 = 1.9$
$x_5 = 4$	$y_5 = 3.1$

- Linear regression assumes that the expected value of the output given an input, *E[y|x]*, is linear.
- Simplest case: Out(x) = wx for some unknown w.
- Given the data, we can estimate w.

1-parameter linear regression

Assume that the data is formed by

$$y_i = wx_i + \text{noise}_i$$

where…

- the noise signals are independent
- the noise has a normal distribution with mean 0 and unknown variance σ^2

Means?

- $p(y|w, x)$ has a normal distribution with
  - mean wx
  - variance σ^2

Bayesian Linear Regression

$$p(y|w,x) = \text{Normal (mean } wx\text{, var } \sigma^2)$$

- We have a set of datapoints (x_1,y_1) (x_2,y_2) … (x_n,y_n) which are EVIDENCE about w.

- We want to infer w from the data.

$$p(w|x_1, x_2, x_3,\ldots x_n, y_1, y_2\ldots y_n)$$

- **You can use BAYES rule to work out a posterior distribution for *w* given the data.**
- **Or you could do Maximum Likelihood Estimation**

Maximum likelihood estimation of *w*

MLE asks: For which value of *w* is this data most likely to have happened?

<=> For what *w* is

$$p(y_1, y_2 \ldots y_n \mid x_1, x_2, x_3, \ldots x_n, w) \text{ maximized?}$$

<=> For what *w* is

$$\prod_{i=1}^{n} p(y_i \mid w, x_i) \text{ maximized?}$$

For what w is

$$\prod_{i=1}^{n} p(y_i | w, x_i) \text{ maximized?}$$

For what w is

$$\prod_{i=1}^{n} \exp\left(-\frac{1}{2}\left(\frac{y_i - wx_i}{\sigma}\right)^2\right) \text{ maximized?}$$

For what w is

$$\sum_{i=1}^{n} -\frac{1}{2}\left(\frac{y_i - wx_i}{\sigma}\right)^2 \text{ maximized?}$$

For what w is

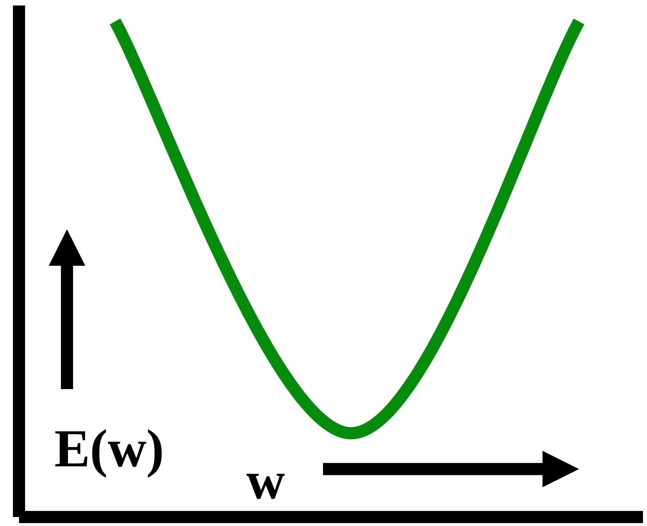
$$\sum_{i=1}^{n} \left(y_i - wx_i\right)^2 \text{ minimized?}$$

First result

- **MLE with Gaussian noise is the same as minimizing the L_2 error**

$$\arg\min \sum_{i=1}^{n} (y_i - w x_i)^2$$

Linear Regression



The maximum likelihood *w* is the one that minimizes sum-of-squares of <u>residuals</u>

$$E = \sum_i \left(y_i - wx_i\right)^2$$

$$= \sum_i y_i^{\,2} - \left(2\sum x_i y_i\right)w + \left(\sum x_i^{\,2}\right)w^2$$

We want to minimize a quadratic function of *w*.

Linear Regression

Easy to show the sum of squares is minimized when

$$w = \frac{\sum x_i y_i}{\sum x_i^{\,2}}$$

The maximum likelihood model is

$$\mathrm{Out}(x) = wx$$

We can use it for prediction

But what about MAP?

- **MLE**

$$\underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{n} p(y_i|\mathbf{w}, x_i)$$

- **MAP**

$$\underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{n} p(y_i|\mathbf{w}, x_i)p(\mathbf{w})$$

But what about MAP?

- MAP

$$\underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{n} p(y_i|\mathbf{w}, x_i)p(\mathbf{w})$$

- We assumed

$$y \sim \mathcal{N}(\mathbf{w} \cdot \mathbf{x}, \sigma^2)$$

- Now add a prior that assumption that

$$\mathbf{w} \sim \mathcal{N}(0, \gamma^2)$$

For what *w* is

$$\prod_{i=1}^{n} p(y_i|\mathbf{w}, x_i)p(\mathbf{w}) \quad \text{maximized?}$$

For what *w* is

$$\prod_{i=1}^{n} \exp\left(-\tfrac{1}{2}(\tfrac{y_i-\mathbf{w}x_i}{\sigma})^2\right) \exp\left(-\tfrac{1}{2}(\tfrac{\mathbf{w}}{\gamma})^2\right) \text{ maximized?}$$

For what *w* is

$$\sum_{i=1}^{n} -\tfrac{1}{2}(\tfrac{y_i-\mathbf{w}x_i}{\sigma})^2 - \tfrac{1}{2}(\tfrac{\mathbf{w}}{\gamma})^2 \quad \text{maximized?}$$

For what *w* is

$$\sum_{i=1}^{n}(y_i - \mathbf{w}x_i)^2 + (\tfrac{\sigma\mathbf{w}}{\gamma})^2 \text{ minimized?}$$

Second result

- MAP with Gaussian prior on *w* is the same as minimizing the L_2 error plus an L_2 penalty on w

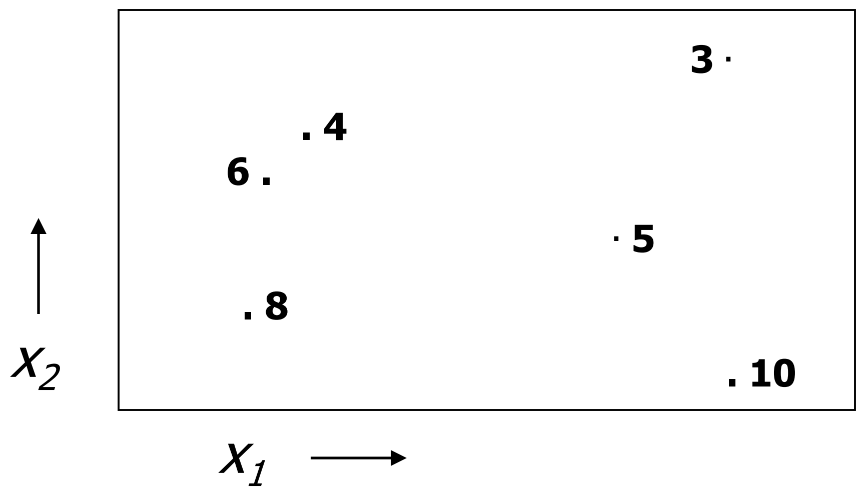
$$\underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mathbf{w}x_i)^2 + \lambda \mathbf{w}^2$$

- This is called: Regularization
  - Ridge regression
  - Shrinkage

Multivariate Linear Regression

Multivariate Regression

What if the inputs are vectors?



2-d input example

Dataset has form

$$\begin{array}{ll} \mathbf{x}_1 & y_1 \\ \mathbf{x}_2 & y_2 \\ \mathbf{x}_3 & y_3 \\ \vdots & \vdots \\ \mathbf{x}_R & y_R \end{array}$$

Multivariate Regression

Write matrix X and Y thus:

$$\mathbf{x} = \begin{bmatrix} \ldots\ldots\mathbf{x}_1\ldots\ldots \\ \ldots\ldots\mathbf{x}_2\ldots\ldots \\ \vdots \\ \ldots\ldots\mathbf{x}_R\ldots\ldots \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ & & \vdots & \\ x_{R1} & x_{R2} & \cdots & x_{Rm} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_R \end{bmatrix}$$

(there are *R* datapoints. Each input has *m* components)

The linear regression model assumes a vector ***w*** such that

$$\text{Out}(\boldsymbol{x}) = \boldsymbol{x}\,\boldsymbol{w} = w_1 x[1] + w_2 x[2] + \ldots . w_m x[D]$$

The max. likelihood ***w*** is $\boldsymbol{w} = (X^T X)^{-1}(X^T Y)$

Multivariate Regression (con't)

The max. likelihood ***w*** is ***w*** = $(X^{\mathrm{T}}X)^{-1}(X^{\mathrm{T}}Y)$

$X^{\mathrm{T}}X$ is an *m* x *m* matrix: i,j'th elt is $\sum_{k=1}^{R} x_{ki}x_{kj}$

$X^{\mathrm{T}}Y$ is an *m*-element vector: i'th elt $\sum_{k=1}^{R} x_{ki}y_{k}$

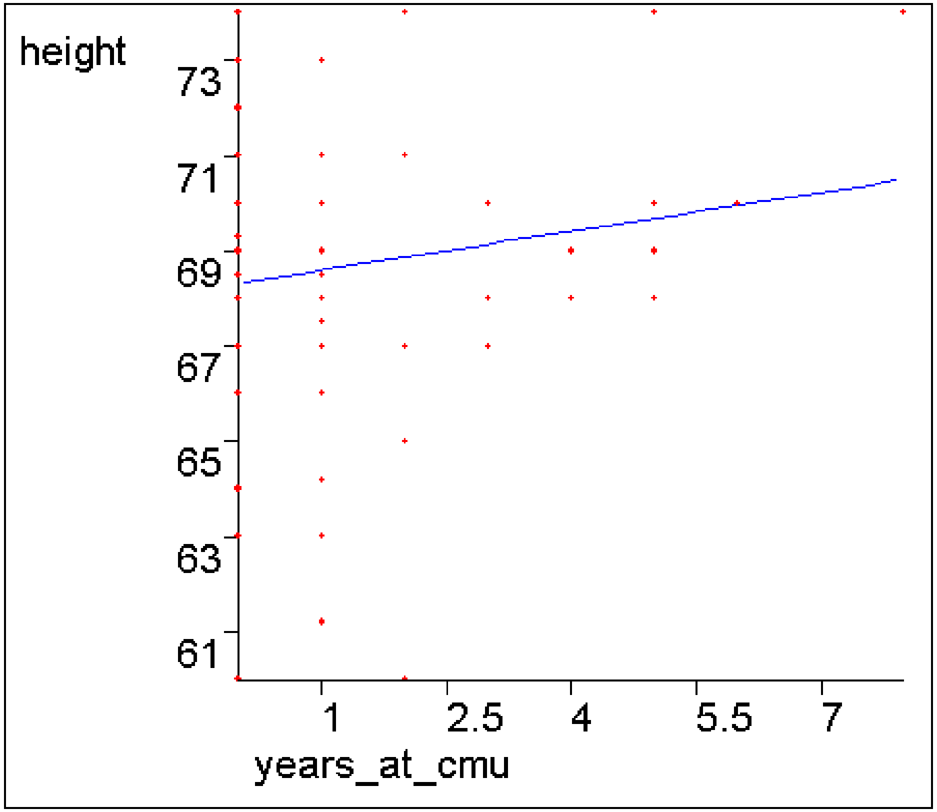
Constant Term in Linear Regression

What about a constant term?

We may expect linear data that does not go through the origin.

Statisticians and Neural Net Folks all agree on a simple obvious hack.

Can you guess??



The constant term

- The trick is to create a fake input "X_0" that always takes the value 1

X_1	X_2	Y
2	4	16
3	4	17
5	5	20

Before:

$Y=w_1X_1+ w_2X_2$

...has to be a poor model

X_0	X_1	X_2	Y
1	2	4	16
1	3	4	17
1	5	5	20

After:

$Y= w_0X_0+w_1X_1+ w_2X_2$

$= w_0+w_1X_1+ w_2X_2$

...has a fine constant term

In this example, You should be able to see the MLE w_0 , w_1 and w_2 by inspection

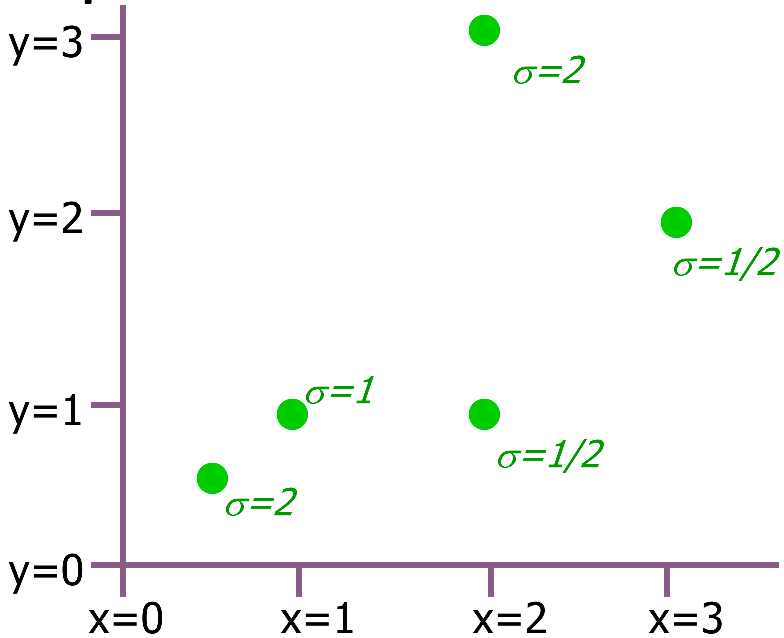
Heteroscedasticity…

Linear Regression with varying noise

Regression with varying noise

- Suppose you know the variance of the noise that was added to each data point.

x_i	y_i	σ_i^2
½	½	4
1	1	1
2	1	1/4
2	3	4
3	2	1/4



Assume $$y_i \sim N(wx_i, \sigma_i^2)$$

What's the MLE estimate of w?

MLE estimation with varying noise

$$\underset{w}{\operatorname{argmax}} \log p(y_1, y_2, \ldots, y_R \mid x_1, x_2, \ldots, x_R, \sigma_1^2, \sigma_2^2, \ldots, \sigma_R^2, w) =$$

$$\underset{w}{\operatorname{argmin}} \sum_{i=1}^{R} \frac{(y_i - wx_i)^2}{\sigma_i^2} =$$

Assuming independence among noise and then plugging in equation for Gaussian and simplifying.

$$\left(w \text{ such that} \sum_{i=1}^{R} \frac{x_i(y_i - wx_i)}{\sigma_i^2} = 0 \right) =$$

Setting dLL/dw equal to zero

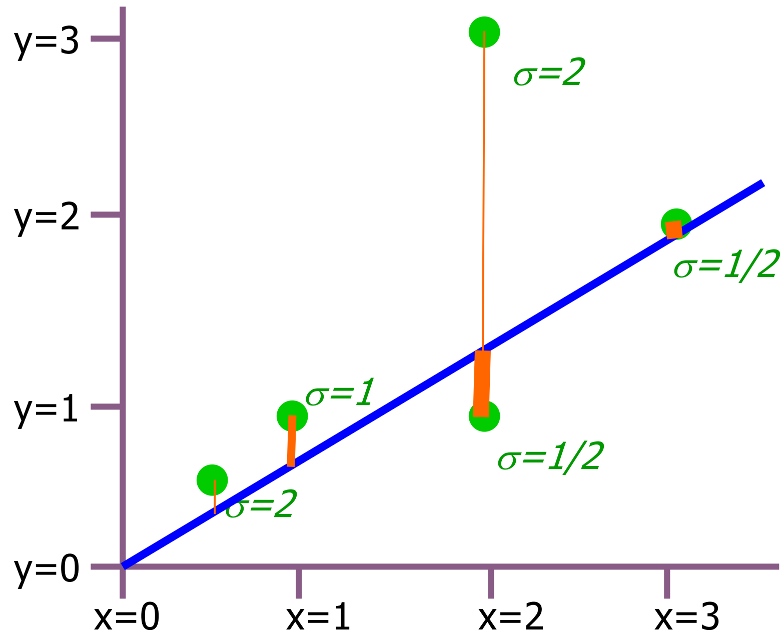
$$\frac{\left(\displaystyle\sum_{i=1}^{R} \frac{x_i y_i}{\sigma_i^2} \right)}{\left(\displaystyle\sum_{i=1}^{R} \frac{x_i^2}{\sigma_i^2} \right)}$$

Trivial algebra

This is Weighted Regression

- We are asking to minimize the weighted sum of squares

$$\underset{w}{\operatorname{argmin}} \sum_{i=1}^{R} \frac{(y_i - wx_i)^2}{\sigma_i^2}$$



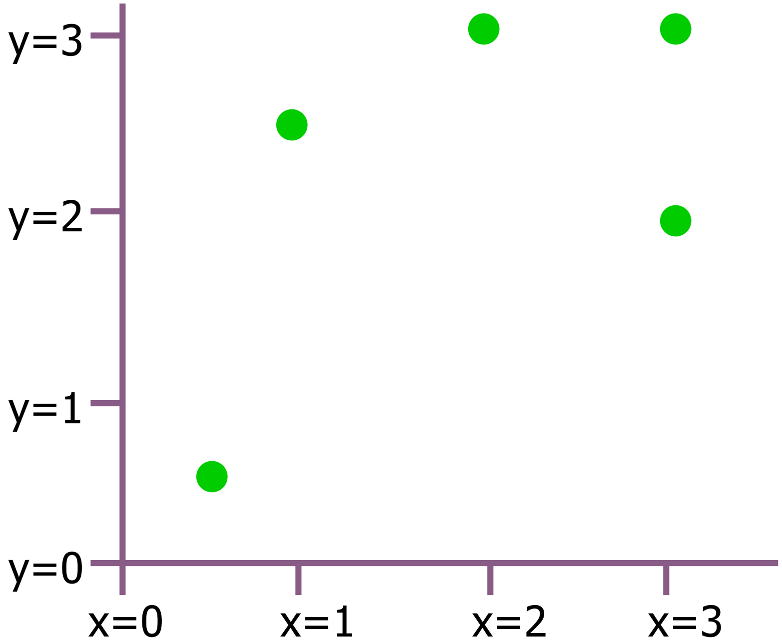
where weight for i'th datapoint is $\frac{1}{\sigma_i^2}$

Non-linear Regression

Non-linear Regression

- Suppose you know that y is related to a function of x in such a way that the predicted values have a non-linear dependence on w, e.g:

x_i	y_i
½	½
1	2.5
2	3
3	2
3	3



Assume $$y_i \sim N(\sqrt{w + x_i}, \sigma^2)$$

What's the MLE estimate of w?

Non-linear MLE estimation

$$\underset{w}{\operatorname{argmax}} \log p(y_1, y_2, \ldots, y_R \mid x_1, x_2, \ldots, x_R, \sigma, w) =$$

$$\underset{w}{\operatorname{argmin}} \sum_{i=1}^{R} \left(y_i - \sqrt{w + x_i} \right)^2 =$$

Assuming i.i.d. and then plugging in equation for Gaussian and simplifying.

$$\left(w \text{ such that} \sum_{i=1}^{R} \frac{y_i - \sqrt{w + x_i}}{\sqrt{w + x_i}} = 0 \right) =$$

Setting dLL/dw equal to zero

So guess what we do?

Non-linear MLE estimation

$$\left(w \,\text{such that} \sum_{i=1}^{R} \frac{y_i - \sqrt{w + x_i}}{\sqrt{w + x_i}} = 0 \right) =$$

Common (but not only) approach:

Numerical Solutions:

- Line Search
- **Simulated Annealing**
- **Gradient Descent**
- Conjugate Gradient
- Levenberg Marquart
- **Newton's Method**

Also, special purpose statistical-optimization-specific tricks such as E.M. (See Gaussian Mixtures lecture for introduction)

Polynomial Regression

Polynomial Regression

So far we've mainly been dealing with linear regression

X_1	X_2	Y
3	2	7
1	1	3
:	:	:

X=

3	2
1	1
:	:

y=

7
3
:

$y_1=7..$

Z=

1	3	2
1	1	1
:		:

y=

7
3
:

$\mathbf{z}_1=(1,3,2)..$ $\quad y_1=7..$

$\mathbf{z}_k=(1,x_{k1},x_{k2})$

$$\beta=(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{y})$$

$$y^{est} = \beta_0+ \beta_1 x_1+ \beta_2 x_2$$

Quadratic Regression

It's trivial to do linear fits of fixed nonlinear basis functions

X_1	X_2	Y
3	2	7
1	1	3
:	:	:

X=

3	2
1	1
:	:

y=

7
3
:

y_1=7..

z=

1	3	2	9	6	4
1	1	1	1	1	1
:					:

y=

7
3
:

$\mathbf{z}=(1,\ x_1,\ x_2,\ x_1^2,\ x_1x_2,\ x_2^2)$

$$\boldsymbol{\beta}=(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{y})$$

$$y^{est}=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1^2+\beta_4x_1x_2+\beta_5x_2^2$$

Quadratic Regression

- $\boldsymbol{z}=(1,\ x_1,\ x_2,\ x_1^2, x_1x_2, x_2^2)$
  - Each component of a z vector is called a term.
  - Each column of the Z matrix is called a term column
- How many terms in a quadratic regression with m inputs?
  - 1 constant term
  - m linear terms
  - (m+1)-choose-2 = m(m+1)/2 quadratic terms
  - (m+2)-choose-2 terms in total $= O(m^2)$.
- Note that solving $\beta=(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}(\boldsymbol{Z}^T\boldsymbol{y})$ is thus $O(m^6)$

Q^{th}-degree polynomial Regression

X_1	X_2	Y
3	2	7
1	1	3
:	:	:

X=

3	2
1	1
:	:

y=

7
3
:

Z=

1	3	2	9	6	…
1	1	1	1	1	…
:					…

y=

7
3
:

z=*(all products of powers of inputs in which sum of powers is q or less)*

$$\boldsymbol{\beta}=(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{y})$$

$$y^{est} = \beta_0 + \beta_1 x_1 + \ldots$$

m inputs, degree Q: how many terms?

= the number of unique terms of the form

$$x_1^{q_1} x_2^{q_2} \ldots x_m^{q_m} \text{ where} \sum_{i=1}^{m} q_i \le Q$$

= the number of unique terms of the form

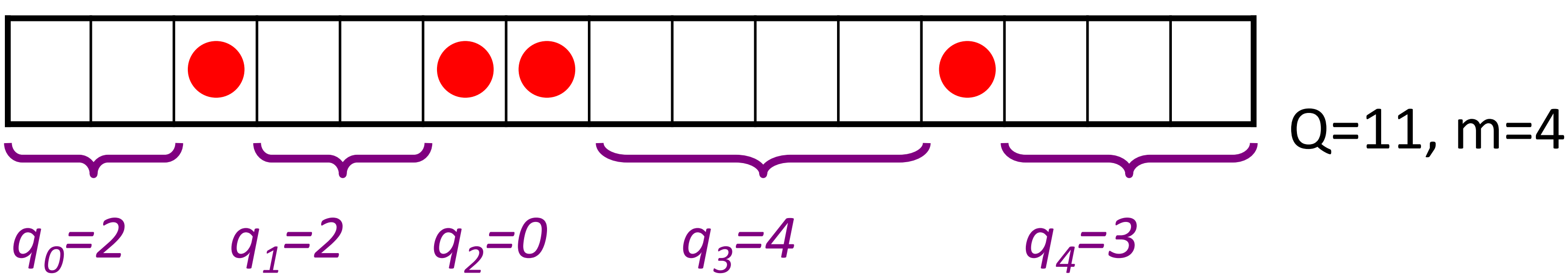
$$1^{q_0}\, x_1^{q_1} x_2^{q_2} \ldots x_m^{q_m} \text{ where} \sum_{i=0}^{m} q_i = Q$$

= the number of lists of non-negative integers $[q_0, q_1, q_2, .. q_m]$

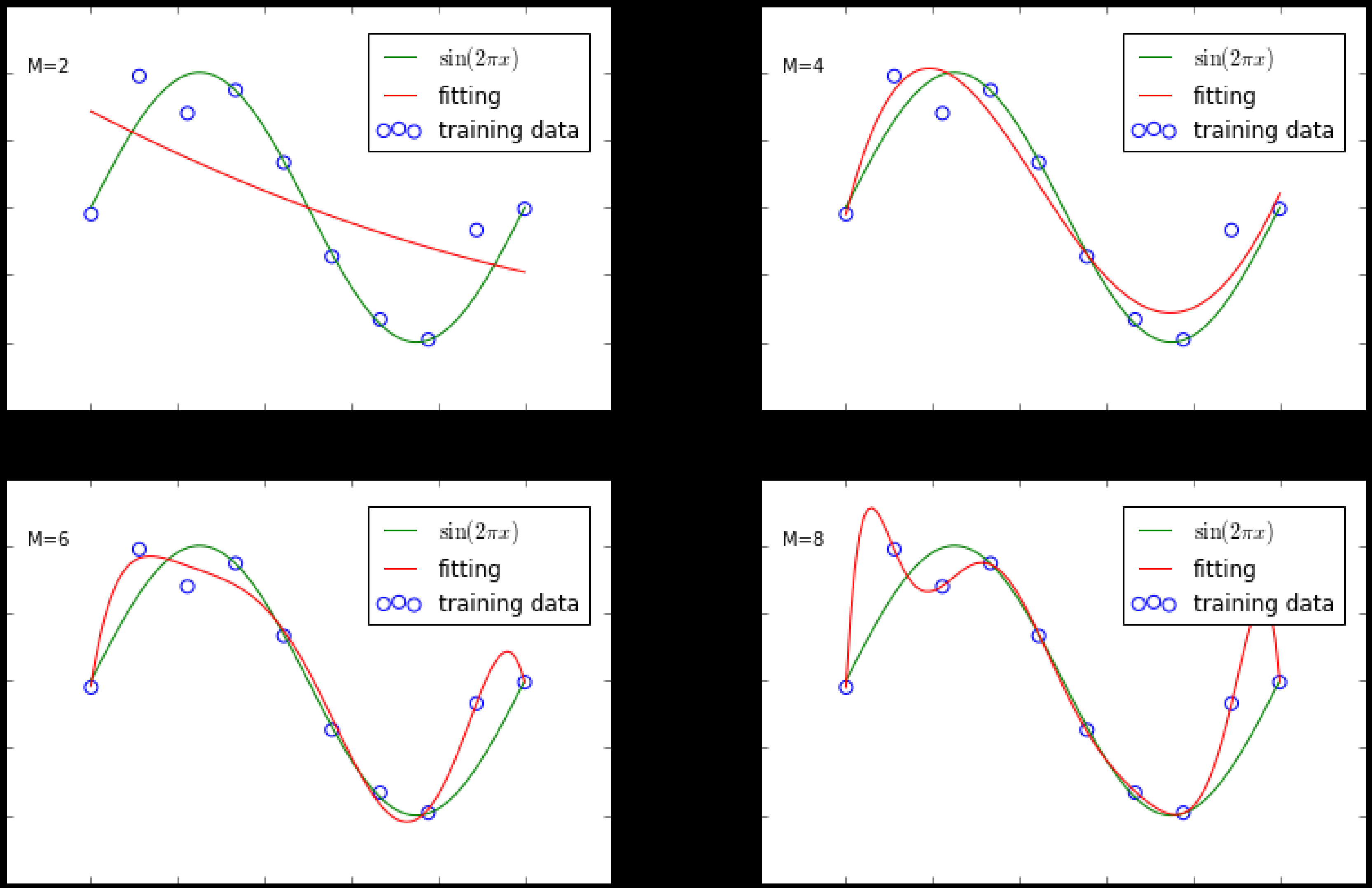
in which $\sum q_i = Q$

= the number of ways of placing Q red disks on a row of squares of

length Q+m = (Q+m)-choose-Q



Polynomial Regression



Multivariate vs Ridge Regression

- The linear regression model assumes a vector ***w*** such that

$$\text{Out}(\boldsymbol{x}) = \boldsymbol{x}\,\boldsymbol{w} = w_1x_1 + w_2x_2 + \ldots. w_mx_m$$

- The maximum likelihood ***w*** is ***w*** $= (X^TX)^{-1}(X^TY)$

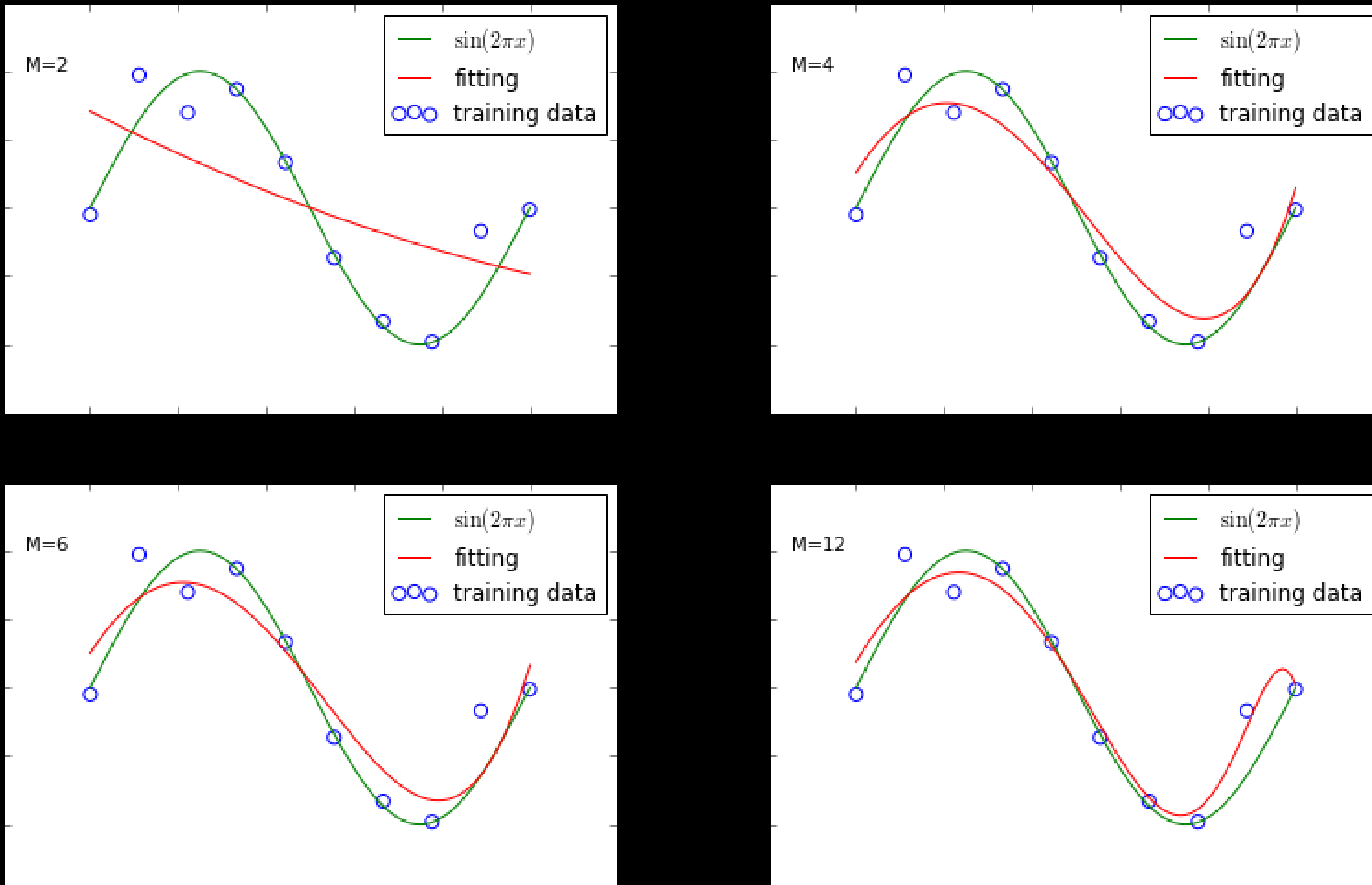
- **MAP with Gaussian prior on w**

$$\underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{n} p(y_i|\mathbf{w}, x_i)p(\mathbf{w})$$

$$y \sim \mathcal{N}(\mathbf{w}\cdot\mathbf{x}, \sigma^2) \qquad \mathbf{w} \sim \mathcal{N}(0, \gamma^2)$$

$$\underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mathbf{w}x_i)^2 + \lambda \mathbf{w}^2$$

Ridge Regression



What we have seen

- MLE with Gaussian noise is the same as minimizing the L_2 error
  - Other noise models will give other loss functions
- MLE with a Gaussian prior adds a penalty to the L_2 error, giving Ridge regression
  - Other priors will give different penalties
- One can make nonlinear relations linear by transforming the features
  - Polynomial regression
  - Radial Basis Functions (RBF) – will be covered laterx

电子科技大学

UESTC

1956