Lecture 3: Polynomial and Spline regression

Wengang Mao (Marine Technology)
Department of Mechanics and Maritime Sciences,
Chalmers University of Technology,
Goteborg, Sweden

1

Contents of this lecture

- **Polynomial regression**
- **Spline regression/fitting (different from interpolation)**
- **Model estimation and evaluation of goodness**
  - **Cost/loss functions: penalties and regularization**
  - **Significant test and confidence interval (more complex model does not mean better model)**
  - **ANOVA, MSE and R2**
- **Advises when apply statistical and machine learning for modelling**

2 2023-04-11

2

Beyond linear models

- Let look at the Taylor expansion, for any complex function/model *f*(*x*), it can be expanded as

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots$$

- For multivariable functions, e.g., *f(x,y)*, it can be expanded as

$$f(x,y) = f(a,b) + f_x(a,b)(x-a) + f_y(a,b)(y-b) + \frac{1}{2!}\left[f_{xx}(a,b)(x-a)^2 + 2f_{xy}(a,b)(x-a)(y-b) + f_{yy}(y-b)^2\right] + \cdots$$

Statistical models in a general format

- Let the target "real" function/model is denoted by Y=*f*(*X*): Based on the Taylor expansion, the *f*(x)=*E*[Y|*X*] can be described by,

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$$

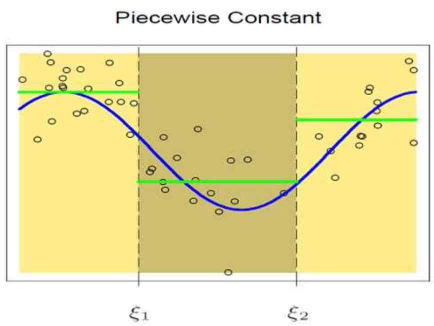
- The terms *h*(*X*) are called the basis function containing the effect of *X* to the prediction target *Y*
- For example, the basis function can be of the following format:

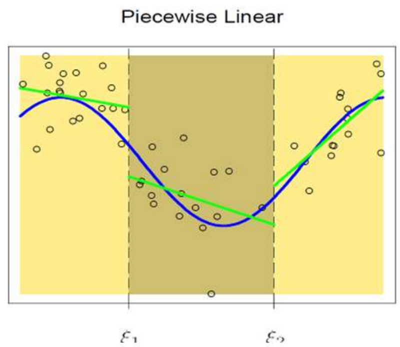
$h_m(X) = X_m \,, m = 1, \ldots, p$	Linear model
$h_m(X) = X_j^2$, or $h_m(X) = X_j X_k$	Polynominal model, can go to even higher orders
$h_m(X) = \log(X_j), \sqrt{X_j}, \ldots$	Other types of transformation
$h_m(X) = I(L_m \leq X_k \leq L_m)$	Indication function: used in e.g., spline function

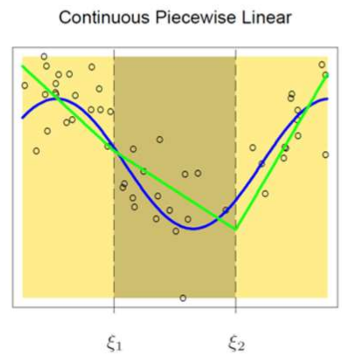
Piecewise linear basis function

- Let split data into various segments, e.g., ξ_1, ξ_2, for each segment, the model has the form,

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$$







Piecewise linear (additional) basis functions:

$$h_m(X) = I(\xi_{m-1} \leq X \leq \xi_m)$$

$$h_{m+3} = h_m(X)X, \; m = 1, \ldots, 3.$$

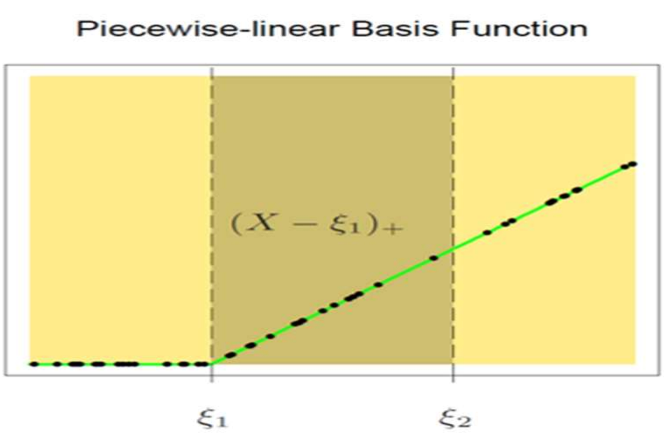
Piecewise constant basis functions:

$$h_1(X) = I(X < \xi_1), \quad h_2(X) = I(\xi_1 \leq X < \xi_2), \quad h_3(X) = I(\xi_2 \leq X).$$

Piecewise linear basis function

- Let split data into various segments, e.g., ξ_1, ξ_2, for each segment, the model has the form,

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$$



Or alternatively, for the piecewise linear basis,

$$h_1(X) = 1, \quad h_2(X) = X,$$

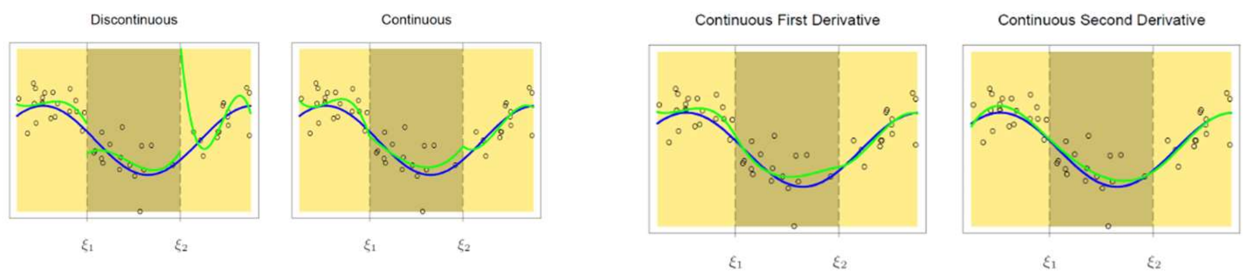
$$h_3(X) = (X - \xi_1)_+, \quad h_4(X) = (X - \xi_2)_+,$$

Piecewise Polynomial and Spline

- Let split data into various segments, e.g., ξ_1, ξ_2, for each segment, the model has the form,

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$$

Spline: the M-order basis functions continuous in their *(M-2)*-th derivatives at the knots are known as the spline curve



Piecewise cubic basis functions:

$$h_1(X) = 1, \quad h_3(X) = X^2, \quad h_5(X) = (X - \xi_1)_+^3,$$
$$h_2(X) = X, \quad h_4(X) = X^3, \quad h_6(X) = (X - \xi_2)_+^3.$$

The number of basis functions is:

$$\#(h_m) = \#(regions) * PolyOrder - \#(knots) * (PolyOrder - 1)$$

2023-04-11

7

Two spline curves (basis function)

- Natural Spline (data outside boundaries be modelled by linear function, reduce 2*2 df): K knots with K basis functions as

$$N_1(X) = 1, \quad N_2(X) = X, \quad N_{k+2}(X) = d_k(X) - d_{K-1}(X),$$

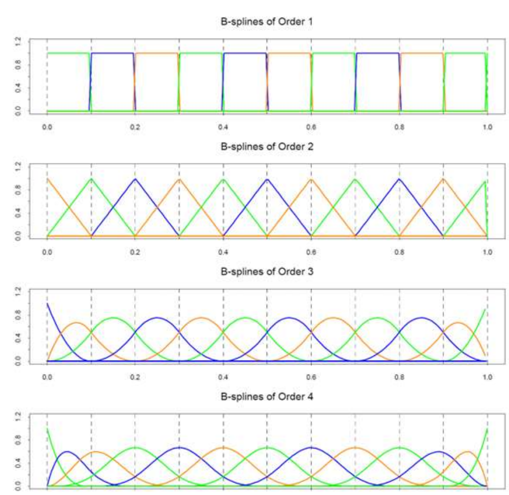
$$d_k(X) = \frac{(X - \xi_k)_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi_k}$$

- B-Spline (with order *m*, *K* knots, *K+2M-m* basis functions)

  - $\tau_1 \le \tau_2 \le \cdots \le \tau_M \le \xi_0$;
  - $\tau_{j+M} = \xi_j, \ j = 1, \cdots, K$;
  - $\xi_{K+1} \le \tau_{K+M+1} \le \tau_{K+M+2} \le \cdots \le \tau_{K+2M}$.

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x)$$

for $i = 1, \ldots, K + 2M - m$.



8

8

We have the models (mathematical formulas) and also the data, but how to estimate the parameters within the models?

Model estimation (1): regression

- **The goal of a regression is to find these coefficients that minimize the cost function**
- **Let a model write as**

$$f(x) = \sum_{k=1}^{K} \alpha_k T_k(x),$$

- **Three cost functions are often used (regularization), to solve the parameters** α_k

$$\min_{\alpha} \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{K} \alpha_k T_k(x_i) \right)^2 + \lambda \cdot J(\alpha) \right\}$$

  - *Ordinary least squares method* $\lambda=0$
  - *Ridge regression method* $J(\alpha) = \sum_{k=1}^{K} |\alpha_k|^2$
  - *Lasso regression method* $J(\alpha) = \sum_{k=1}^{K} |\alpha_k|$

Algorithm 16.1 *Forward Stagewise Linear Regression.*

1. Initialize $\check{\alpha}_k = 0,\ k = 1, \ldots, K$. Set $\varepsilon > 0$ to some small constant, and M large.
2. For $m = 1$ to M:
   (a) $(\beta^*, k^*) = \arg\min_{\beta,k} \sum_{i=1}^{N} \left(y_i - \sum_{l=1}^{K} \check{\alpha}_l T_l(x_i) - \beta T_k(x_i) \right)^2$.
   (b) $\check{\alpha}_{k^*} \leftarrow \check{\alpha}_{k^*} + \varepsilon \cdot \operatorname{sign}(\beta^*)$.
3. Output $f_M(x) = \sum_{k=1}^{K} \check{\alpha}_k T_k(x)$.

Model estimation (2): Spline

- **Coefficients of Spline basis functions: their properties are associated with the coefficients (# of basis functions)**
  - Number of knots
  - Spline order (e.g., linear, cubic)
  - Cost function in the coefficient estimation/optimization

- **Cost function: Penalized residual sum of squares (RSS):**

$$\mathrm{RSS}(f, \lambda) = \sum_{i=1}^{N} \{y_i - f(x_i)\}^2 + \lambda \int \{f''(t)\}^2 dt,$$

$\lambda = 0:$ f can be any function that interpolates the data.

$\lambda = \infty:$ the simple least squares line fit, since no second derivative can be tolerated.

Model estimation (3): Spline

- How to estimate the parameters/coefficients within the model
- In the Spline regression, we assume the number of knots and spline order are predefined.
  1. For the Natural Spline regression:

$$\hat{\theta} = (\mathbf{N}^T\mathbf{N} + \lambda\mathbf{\Omega}_N)^{-1}\mathbf{N}^T\mathbf{y}$$

$$\{\mathbf{N}\}_{ij} = N_j(x_i) \text{ and } \{\mathbf{\Omega}_N\}_{jk} = \int N_j''(t)N_k''(t)dt.$$

$$\hat{f}(x) = \sum_{j=1}^{N} N_j(x)\hat{\theta}_j$$

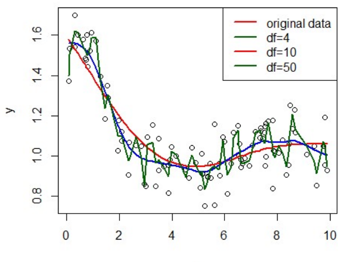
  2. For the B-Spline regression

$$\hat{\gamma} = (\mathbf{B}^T\mathbf{B} + \lambda\mathbf{\Omega}_B)^{-1}\mathbf{B}^T\mathbf{y}$$

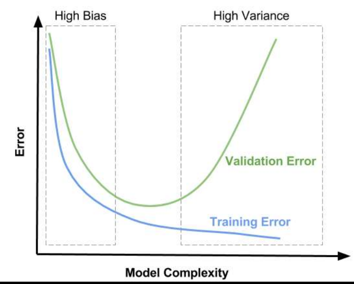
$$f(x) = \sum_1^{N+4} \gamma_j B_j(x)$$

Model evaluation (1): e.g., Spline case

- **Should we use more complex model or simple model? How complex should it be?**
- **Where is the point of trade-off between biased and variance?**
  - How to choose the penalized parameter λ in the cost function?
  - How many knots should we use for the Spline regression?
  - Which Spline model should we use for the Spline regression?




```
Call:
lm(formula = y ~ bs(x, knots = c(2, 4, 6, 8)), data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.250265 -0.064692  0.000633  0.063775  0.212458

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.48211    0.06149  24.103  < 2e-16 ***
bs(x, knots = c(2, 4, 6, 8))1  0.27932    0.11686   2.390  0.01887 *
bs(x, knots = c(2, 4, 6, 8))2 -0.60465    0.08513  -7.103 2.52e-10 ***
bs(x, knots = c(2, 4, 6, 8))3 -0.48877    0.09435  -5.181 1.30e-06 ***
bs(x, knots = c(2, 4, 6, 8))4 -0.59253    0.07819  -7.578 2.69e-11 ***
bs(x, knots = c(2, 4, 6, 8))5 -0.29603    0.09226  -3.209  0.00184 **
bs(x, knots = c(2, 4, 6, 8))6 -0.48531    0.10356  -4.686 9.60e-06 ***
bs(x, knots = c(2, 4, 6, 8))7 -0.47054    0.08314  -5.660 1.70e-07 ***
```



-04-11

13

Model evaluation (2): trade-off under/over-fitting

- Criteria for model evaluation and selection: RSS → *f*(Bias) + *f*(Variance) [under/over-fitting]
- Let a regressed model denote by $\hat{f}(x)$. For a new input $X = x_0$, the squared error loss of the new prediction is:

$$\begin{aligned}
\mathrm{Err}(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + [\mathrm{E}\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)) \\
&= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance.}
\end{aligned}$$

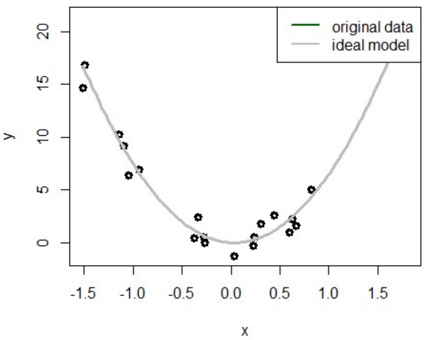
NB: for a new input x_0, its prediction $\hat{f}(x_0)$ is also a random variable since the parameters of the fitted model $\hat{f}$ are also random variable. We often get one prediction for the inputs and that is the mean/expected prediction.

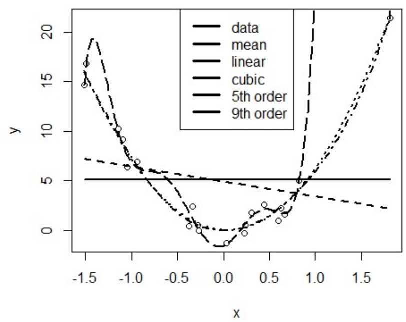
14 023-04-11

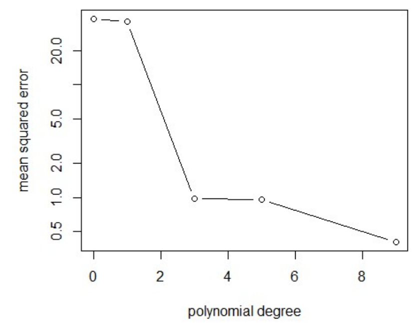
14

Model evaluation (3): An example

- Let a known model denote by $\boldsymbol{Y = 7X^2 - 0.5X}$. We get a series of data from this quadratic model, of course with observation uncertainties (noise).
- Our target is to find a proper model (no underfitting, no overfitting) from the data.
- The more complex the model, the smaller of the MSE. But smaller MSE not necessarily better!







2023-04-11

15

Model evaluation (4): Criteria

- Criteria for model evaluation and selection
  - RSS $\min_{\alpha}\left\{\sum_{i=1}^{N}\left(y_i - \sum_{k=1}^{K}\alpha_k T_k(x_i)\right)^2 + \lambda \cdot J(\alpha)\right\}$
  - The coefficient of determination R^2 is a measure of the amount of variability in the data accounted for by the regression model, i.e. ***strength of the relationship***.

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{= \sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  - In general the higher the value of R^2, the ***better*** the model fits the data.
    - $R^2 = 1$: Perfect match between the line and the data points.
    - $R^2 = 0$: There are no linear relationship between x and y.

16 2023-04-11

16

Model evaluation (5): methods

- **Significant test ANOVA (T-test, F-test)**
  - This method is to check if one or several features/variables/effects could be deleted but will not affect the accuracy of the model, i.e., the significant test.
  - If one variable is checked, the T-distribution should be used.
  - If several variables are checked at the same time, use F-distribution.
  - The significant level is often set to be 5%. Or the probably of accept the hypothesis ”delete the concerned variable” can be given.
- **Bootstrap methods**
- **Cross validation**

ANOVA (1): T- & F- distribution

- Let Z_1, Z_2, Z_3,…, Z_n denote independent, standard normal RVs
- Then we construct the following statistics:

$$Q(k) = \sum_{i=1}^{k} Z_i^2 \qquad t(k) = \frac{Z}{\sqrt{Q(k)/k}} \qquad F(m,n) = \frac{Q(m)/m}{Q(n)/n}$$

  - *Q* follows the **chi-square distribution** (also **chi-squared** or **χ^2-distribution**) with *k* df as the parameter
  - *t* follows the student’s *t* distribution with *k* df as the parameter
  - F follows the F distribution with *m*, *n* df as parameters

ANOVA (2): T- & F- distribution

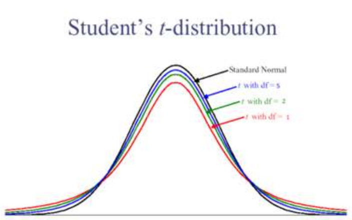
- **For the hypothesis test (ANOVA), we need to define the hypothesis for the modelling**
  - For example, a complex model as: $Y = \beta_0 + \beta_1 f(X_1) + +\beta_2 f(X_2)$, we would like to see if the effect/variable X_1 is necessary in the model.
  - We define the test of the hypothesis condition as: Hypothesis: $\beta_1 = 0$
  - Then, we should construct and estimate the t, or F statistics based on the data
- **The formula for the t and F statistics are related to the following variables:**
  - **sums of squares within** (SSwithin) indicates the total amount of dispersion within groups;
  - **degrees of freedom within** (DFwithin) is (n - k) for n observations and k groups and
  - **mean squares within** (MSwithin) -basically the variance within groups is SSwithin / DFwithin.

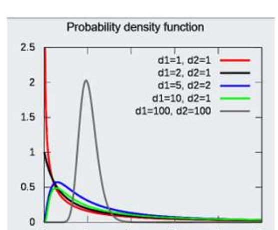
ANOVA (3): T- & F- distribution

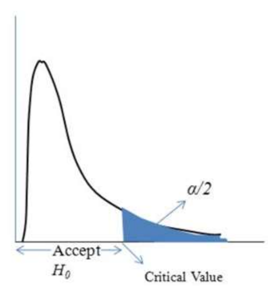
- For example,

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

- Finally, one should check the calculated statistics with the critical value for a significant level





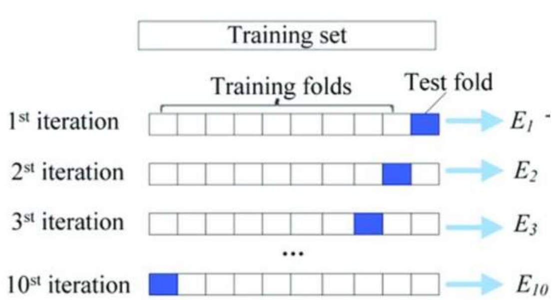


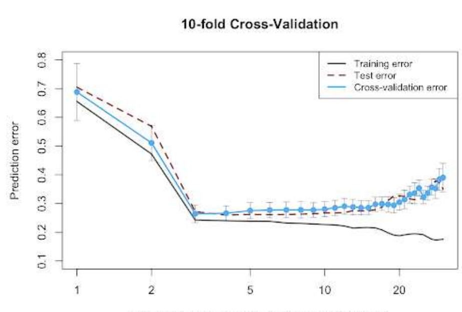
ANOVA (4): T- distribution (example)

```
Call:
lm(formula = y ~ bs(x, knots = c(2, 4, 6, 8)), data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.250265 -0.064692  0.000633  0.063775  0.212458

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.48211    0.06149  24.103  < 2e-16 ***
bs(x, knots = c(2, 4, 6, 8))1  0.27932    0.11686   2.390  0.01887 *
bs(x, knots = c(2, 4, 6, 8))2 -0.60465    0.08513  -7.103 2.52e-10 ***
bs(x, knots = c(2, 4, 6, 8))3 -0.48877    0.09435  -5.181 1.30e-06 ***
bs(x, knots = c(2, 4, 6, 8))4 -0.59253    0.07819  -7.578 2.69e-11 ***
bs(x, knots = c(2, 4, 6, 8))5 -0.29603    0.09226  -3.209  0.00184 **
bs(x, knots = c(2, 4, 6, 8))6 -0.48531    0.10356  -4.686 9.60e-06 ***
bs(x, knots = c(2, 4, 6, 8))7 -0.47054    0.08314  -5.660 1.70e-07 ***
```

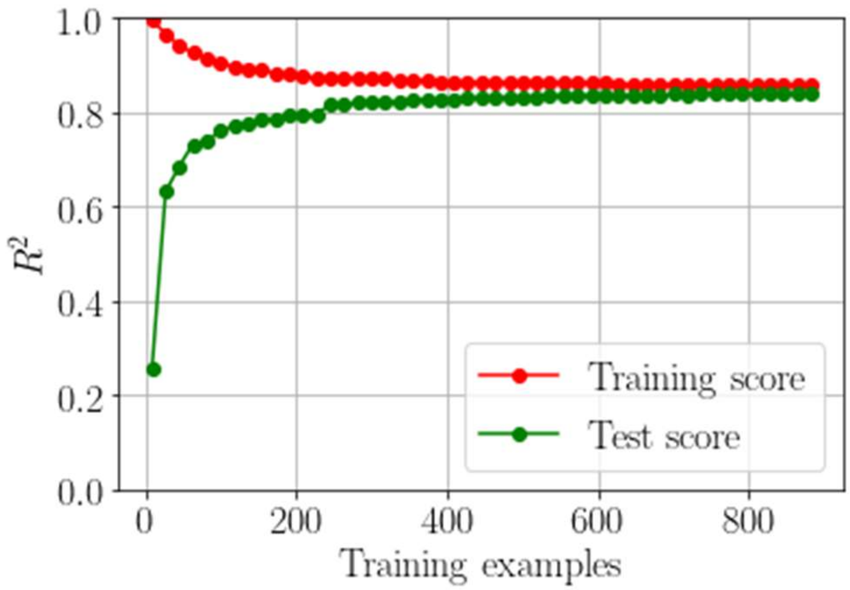

Model evaluation(6): crossing validation

- **Ideas of crossing validation**
  - Split the data into training and test dataset
  - Training data is now used for model evaluation and selection
  - Because all models contain uncertainties, in order to build the “best” model from the training data, we split the training data into two part: training and validation.
  - The cross validation can be formulated as: k-fold and leave-one-out cross validation





R2 from CV



Next step

- We have discussed the procedure to choose more "optimal" model to describe the data
- For the Spline regression, there are some explicit formulas to establish the model, i.e., to estimate the parameters in the Spline model, such as optimal knots arrangement, optimal choice of smoothness parameter λ and the coefficients of the basis functions in the Spline model
- Questions: For other general ML regression, how can we estimate the parameters in the model? To answer this question, we need to understand:
  - What is the investigated model (Linear, Polynomial, Spline, GAM, GLM) for the ML regression?
  - What is the "optimization" objective for the ML regression?
  - Based on the data, what mathematical algorithms can help to solve the above problems?

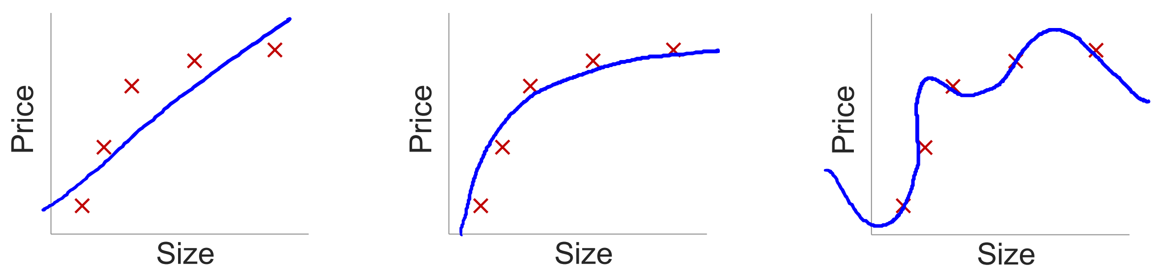
Advices for applying ML methods

1, Diagnosing bias vs. variance

2, Regularization and bias/variance

3, Learning curve

Bias/variance



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

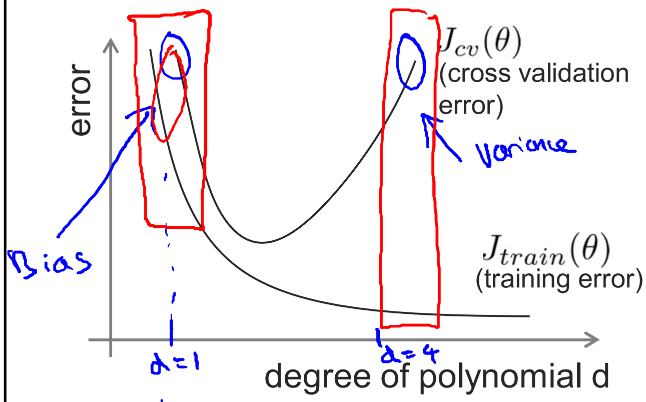
High bias (underfit)
$d=1$

"Just right"
$d=2$

High variance (overfit)
$d=4$

Bias vs. variance

Suppose your learning algorithm is performing less well than you were hoping. ($J_{cv}(\theta)$ or $J_{test}(\theta)$ is high.) Is it a bias problem or a variance problem?



Bias (underfit):

→ $J_{train}(\theta)$ will be high

$J_{cv}(\theta) \approx J_{train}(\theta)$

Variance (overfit):

→ $J_{train}(\theta)$ will be low

$J_{cv}(\theta) \gg J_{train}(\theta)$

$\gg$

2023-04-11

27

2, Regularization and bias/variance

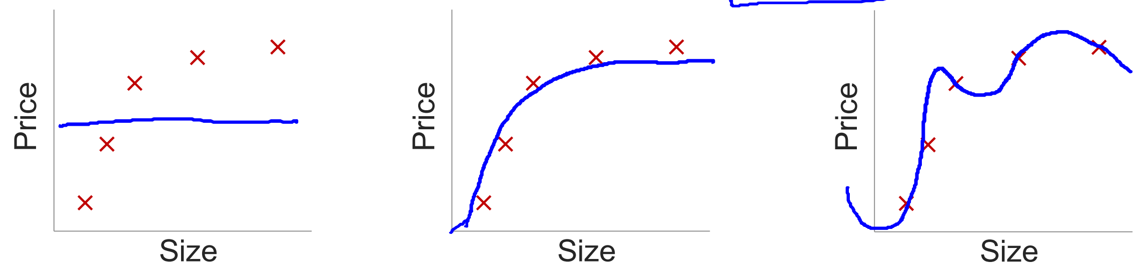
2023-04-11

28

Linear regression with regularization

Model: $\boxed{h_\theta(x) = \theta_0} + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

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



Large λ | Intermediate λ | Small λ

High bias (underfit) | "Just right" | High variance (overfit)

$\lambda = 10000.\ \theta_1 \approx 0, \theta_2 \approx 0, \ldots$

$h_\theta(x) \approx \theta_0$

$\lambda = 0$

2023-04-11

29

Choosing the regularization parameter λ

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

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

$J(\theta)$

$$J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$J_{cv}(\theta) = \frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}\left(h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)}\right)^2$$

$$J_{test}(\theta) = \frac{1}{2m_{test}}\sum_{i=1}^{m_{test}}\left(h_\theta(x_{test}^{(i)}) - y_{test}^{(i)}\right)^2$$

J_{train}

J_{cv}

J_{test}

2023-04-11

30

Choosing the regularization parameter λ

Model: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

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

1. Try $\lambda = 0$ → $\min_\theta J(\theta)$ → $\theta^{(1)}$ → $J_{cv}(\theta^{(1)})$
2. Try $\lambda = 0.01$ → $\min_\theta J(\theta)$ → $\theta^{(2)}$ → $J_{cv}(\theta^{(2)})$
3. Try $\lambda = 0.02$ → $\theta^{(3)}$ → $J_{cv}(\theta^{(3)})$
4. Try $\lambda = 0.04$
5. Try $\lambda = 0.08$ → $\theta^{(5)}$ $J_{cv}(\theta^{(5)})$

   ⋮

12. Try $\lambda = 10$ → $\theta^{(12)}$ → $J_{cv}(\theta^{(12)})$

10.24

Pick (say) $\theta^{(5)}$. Test error: $J_{test}(\theta^{(5)})$

2023-04-11

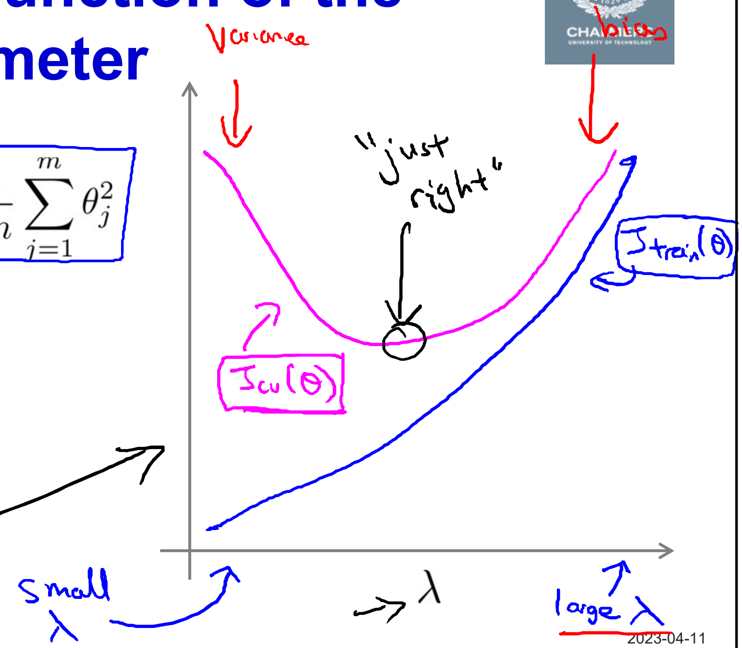
31

Bias/variance as a function of the regularization parameter λ

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

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

$$\boxed{J_{cv}(\theta)} = \frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}(h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$



2023-04-11

32

3, Learning curves

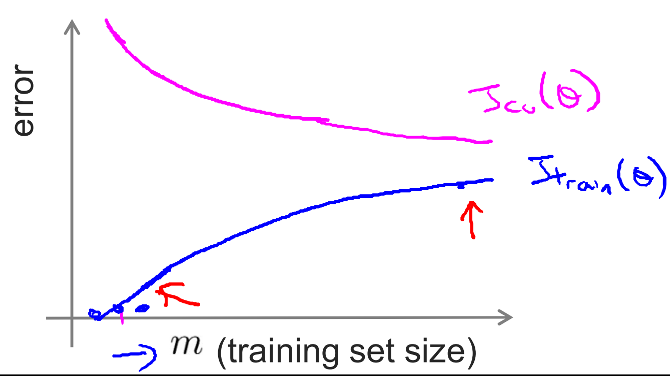
33

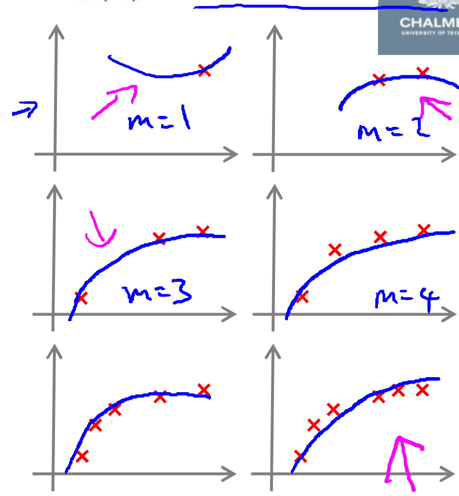
Learning curves

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

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

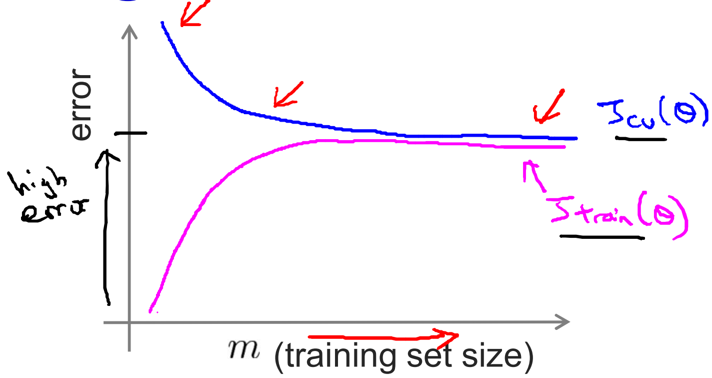
$$J_{cv}(\theta) = \frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}(h_\theta(x^{(i)}_{cv}) - y^{(i)}_{cv})^2$$



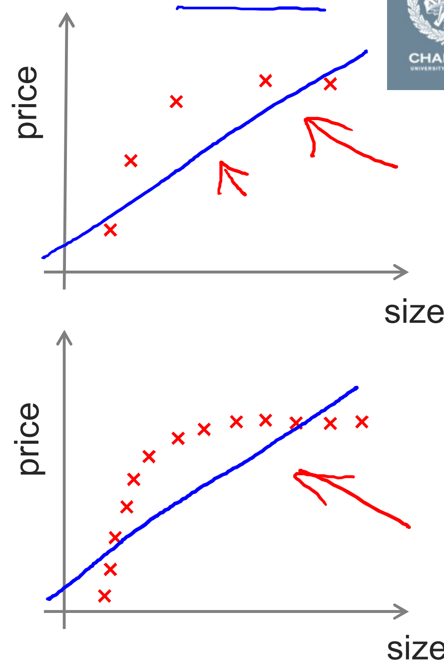


34

High bias

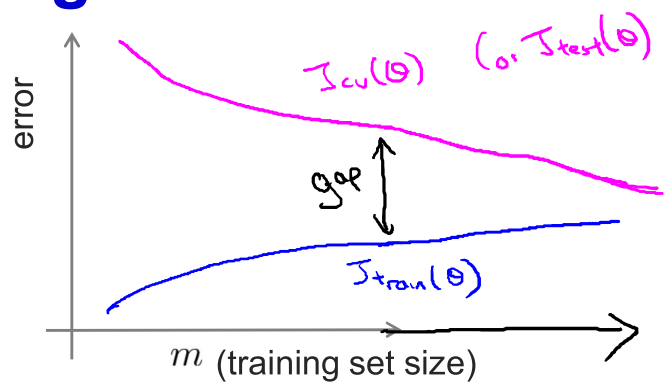


$h_\theta(x) = \theta_0 + \theta_1 x$

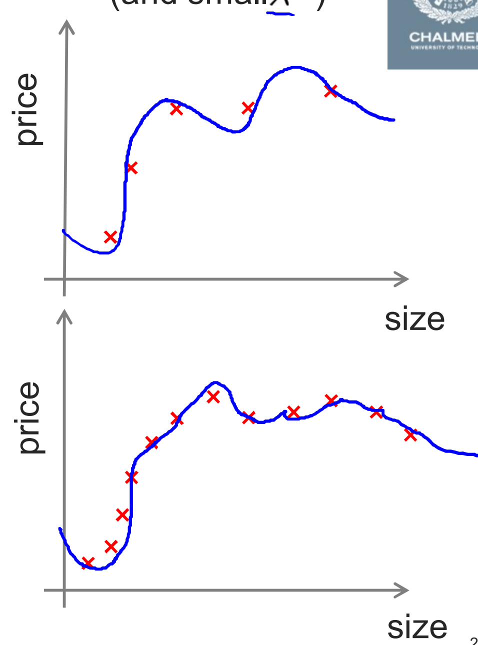


If a learning algorithm is suffering from high bias, getting more training data will not (by itself) help much.

High variance



$h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{100} x^{100}$
(and small λ)



If a learning algorithm is suffering from high variance, getting more training data is likely to help.

CHALMERS
UNIVERSITY OF TECHNOLOGY

37