/ /

problem 2: conceptual questions

[ISL book] chapter 2

1) (+ definition: our estimating model → may be too far from true f
(because of simplicity)

⇓

to solve this problem

↓

by choosing flexible models

⇓

- that can fit many different possible functional forms for f.
- these more complex models can lead to a phenomenon known as overfitting the data.

$\boxed{\uparrow \text{Complexity} \longrightarrow \text{flexibility} \uparrow}$

(my study from ISL-chap2 _ 2.1.2))

a) + sample size (n) → extremely large
+ p (number of predictors) → small
→ the more number of sample size lead to more complex model that can fit better to the data.

↓

in this case flexible model is better

/ /

b) the number of predictors (p) → extremely large

the number of observations (n) → small

↳ Since the number of predictors is high ($p \uparrow$) → complexity ↑

estimating model may overfit

⇓

+ flexible model is not good.

+ and simple (inflexible) model in this situation is better.

c) the relationship between the predictors and response is highly non_linear.

⇓

in this case complex and flexible models is better.

d) $\sigma^2 = Var(\varepsilon)$ → is extremely high ?

↓ by flexible models maybe overfitting occurs,

so inflexible models is better.

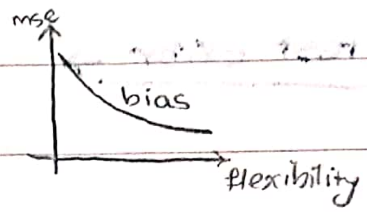
/ /

3) revisit the bias-variance decomposition

a) single plot for less flexible learning to more flexible approaches.

b) explain why each of five curves has this shape?

bias → inflexible models $\xrightarrow{\text{bias}(\hat{f})\downarrow}$ flexible models



⇓

because → $\text{biase}(\hat{f}) = \underbrace{E(\hat{f}) - f}$

↓

when $\hat{f}$ is more complex

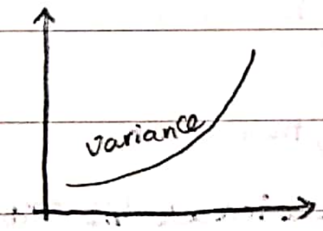
⇓

has better fit to the data → so $\text{bias}(\hat{f})\downarrow$

variance → inflexible models $\xrightarrow{\text{variance}\uparrow}$ flexible models

⇓

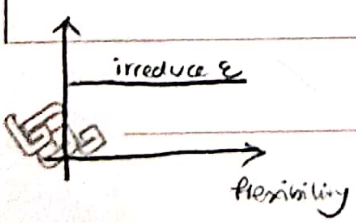
(like lecture slides) if we only suppose 2 points for training model, in complex models with more parametrs $(\beta_0, \beta_1, \dots \beta_d)$, changing 2 data points can change estimating model more (than linear models).



⇓

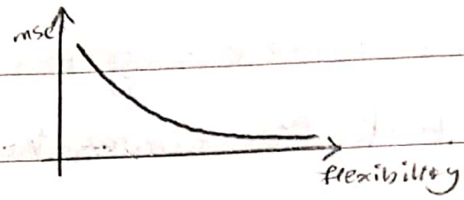
so flexible models have more variance

irreducable error → (due to unmissured variables) It is constant. + we can not reduce this error by changing model from inflexible to flexible.

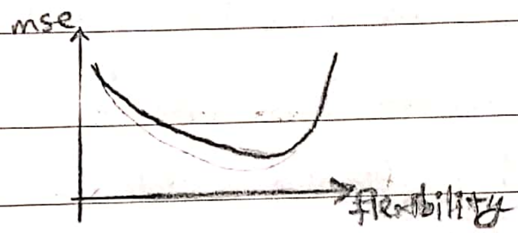


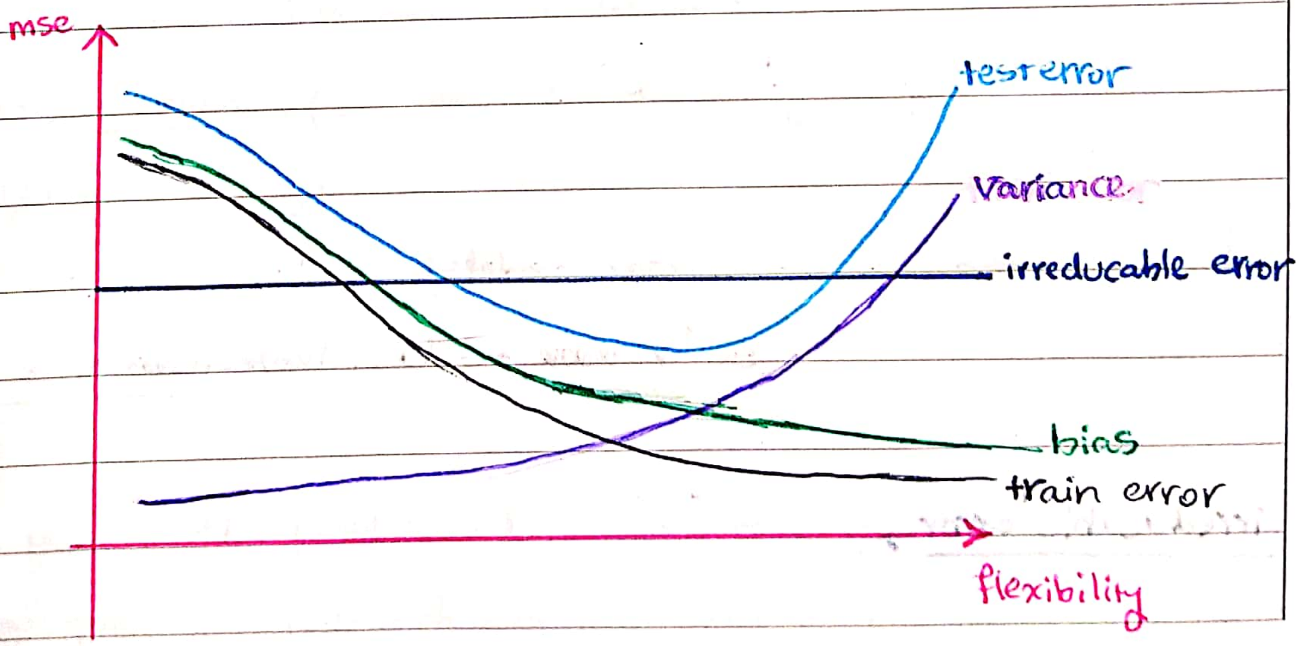
/ /

train error → more complex models can fit better to training data, so mse alway reduce from inflexible models towards flexible models.



test error → flexible models can reduce mse in test data, but overestimating in model complexity can increase mse in test data, and call over fitting.





5) - very flexible model → + can fit very well to data (training)

+ but can overfit and increase error on test data.

- when there is a large sample size → more flexible models can be useful.

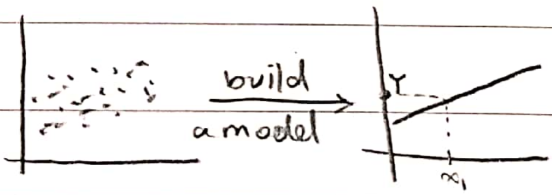
6) differences between parametric & non-parametric statistical learning.

[advantages & disadvantages of parametric model]

+ In a parametric model → first we use all data and build a model.

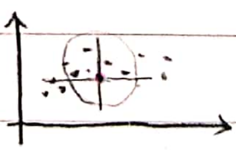
then we can use this model for estimating

output of another datapoints.



+ In nonparametric model → we use directly all data to find output of

another data points (like KNN or NW)



* we need to estimate $\hat{Y}$ of x according to all data, because we don't have any model or parameter, but if we want to have best estimation, we need large size of data.

/ /

7) d] since Bayes decision boundry is nonlinear, small k is better.

[ISL book] (chap3)

4) a) training RSS for cubic regression is lower than linear regression. because cubic regression have more complexity and fit better to training data.

b) probably, cubic regression RSS is high for test data. (because overfitting)

c) cubic regression fit better and has lower RSS.

d) since, we know the relationship between X & Y is nonlinear, may be cubic regression has better & lower RSS on test data.

[ISL book - chap 6]

4) a) iii → λ↑ → model simpilicity ↑ → fit to data ↓ → $\boxed{\text{training RSS} \uparrow}$

b) ii → λ↑ → $\boxed{\text{test RSS} \downarrow \Rightarrow \text{test RSS} \uparrow}$

c) iv → λ↑ → complexity of model ↓ → $\boxed{\text{Variance} \downarrow}$

[chap6-ISL book] 4)

d) $\lambda\uparrow \to$ complexity $\downarrow \to$ bias $\uparrow$ iii

e) v $\to$ irreducable error Remain constant.

problem2) a] repeat (94-ISL book-chap6) for K in KNN.

a) $K\uparrow \to$ $\boxed{\text{training RSS}\uparrow}$ iii

b) $K\uparrow \to$ tess RSS $\to$ depends on test data ($K \to$ very big $\to$ $\boxed{\text{tess RSS}\uparrow}$))

c) $K\uparrow \to$ variance $\downarrow$ iv

d) $K\uparrow \to$ bias $\downarrow$ iv

e) $K\uparrow \to$ irreducable error (constant) (v)

Scanned by CamScanner

/ /

b] Repeat for λ of NW kernel regression.

a) $\lambda \uparrow \rightarrow$ training RSS $\uparrow$

b) $\lambda \uparrow \rightarrow$ test RSS $\uparrow$

c) $\lambda \uparrow \rightarrow$ Variance $\uparrow$

d) $\lambda \uparrow \rightarrow$ bias $\uparrow$

e) irreducible error $\rightarrow$ constant

/ /

problem 3 _ Review _ Column space

$$A=\begin{pmatrix}1&1&4\\2&3&9\\2&1&7\end{pmatrix} \xrightarrow{\text{span}} a\begin{bmatrix}1\\2\\2\end{bmatrix}+b\begin{bmatrix}1\\3\\1\end{bmatrix}+c\begin{bmatrix}4\\9\\7\end{bmatrix}=\begin{bmatrix}5\\7\\3\end{bmatrix}$$

$$\Downarrow$$

$$\left(\begin{array}{ccc|c}1&1&4&5\\2&3&9&7\\2&1&7&3\end{array}\right) \xrightarrow[R_3\to R_3-2R_1]{R_2\to R_2-2R_1} \left(\begin{array}{ccc|c}1&1&4&5\\0&1&1&-3\\0&-1&-1&-7\end{array}\right)$$

$$\xrightarrow{R_3\to R_3+R_2} \left(\begin{array}{ccc|c}1&1&4&5\\0&1&1&-3\\0&0&0&-10\end{array}\right)$$

$$a(0)+b(0)+c(0)=-10$$

$$\boxed{0\neq-10}$$

چون معادله برقرار نیست پس وکتور w ، در column space ماتریس A نیست.

/ /

Problem 4: Feature selection & cross validation

(ESL-chap 7-10.2)

The wrong & Right way to do cross validation

- we suppose to have a large number of predictors. (e.g 5000 predictor, n=50) and the sample size is small. ($p \gg n$)

⇓

for building a model, we can use data (obs) and then calculate Cross validation error in usual way includes 3 steps:

(1) find a "good predictor" subset (100) that have strong correlation with the class labels (2) using selection predictors → build a multivariate classifier (3) divide obs to k-fold and estimate prediction error of final model (CV)

but → suppose ($n=50$, $p=5000$) → $\boxed{\text{test (true) error of any classifier} \to 50\%}$

(selection 100 of p) → with high cor → $\boxed{\text{CV} \to \text{error} = 3\%}$

problem: since these predictors "have already seen" the left out samples → we have unfair predictors which show 3% error in CV (far lower than True error (50%))

optional

solution: [(1) select 1000 predictors with highest variance across whole predictors in 50 samples] (2) divide all samples into k-fold (randomly)

(3) [a] select a subset (100) of "good predictors" that show strong correlation with the class labels, using all samples except fold k. [b] build a multivariate classifier by this subset of p (by using all samples except fold k)

[c] use classifier for prediction fold k class label ⇒ (4) estimate CV error

/ /

problem 5: orthogonal projection

H (Hat matrix) is an orthogonal projection $\xrightarrow{\text{if}} \begin{cases} H^2 = H \\ H = H^T \end{cases}$

$$H = x(x^T x)^{-1} x^T \rightarrow H^2 = (x(x^T x)^{-1} x^T)^2 = (x(x^T x)^{-1} \underbrace{x^T x (x^T x)^{-1}}_{I} x^T)$$

$$= \boxed{x(x^T x)^{-1} x^T} \; H$$

$$\Longrightarrow \underline{H^2 = H}$$

$$H = H^T \rightarrow H^T = (x(x^T x)^{-1} x^T)^T$$

$(ABC)^T = C^T B^T A^T$

$$\Longrightarrow x\left[(x^T x)^{-1}\right]^T x^T$$

$(AB)^T = B^T A^T$

$(A^T)^T = A$

$$\Longrightarrow = x\left[(x^T x)^T\right]^{-1} x^T$$

$$= \underbrace{\boxed{x(x^T x)^{-1} x^T}}_{H} \Longrightarrow H^T = H$$

problem6 : K-nearest neighbor

decision boundaries by the 1-NN algorithm. ⇒ K=1 → class label of each point is similar to nearest neighbor by K=1.