statistical hw1

全金

2025-03-10

(1)

(a) 更好，大样本有助于训练复杂模型，避免过拟合。

(b) 光滑度低好，变量数大，样本数少时光滑模型不稳定，容易过拟合。

(c) 更好，非线性适合光滑模型。

(d) 光滑度低好，噪声大容易过拟合。

(3)

(a)

平方偏差：单调递减。
方差：单调递增。
训练误差：单调递减至接近零。
测试误差：先减后增。
贝叶斯误差：水平直线（常数）。

(b)

平方偏差：光滑度升高时，系统性拟合误差减小。
方差：高维拟合方法对噪声敏感。
训练误差：逐渐拟合直至完全过拟合。
测试误差：逐渐拟合直至完全过拟合。
贝叶斯误差：数据本身噪声决定，与模型无关。

(8)

(a)

```
library(ggplot2)
library(ISLR2)
set.seed(123)
```

```
college=read.csv("College.csv")
```

(b)

```
rownames(college)=college[,1]
#fix(college)
A=college
college=college[,-c(1,2)]
#fix(college)
```

(c)

```
### (i)

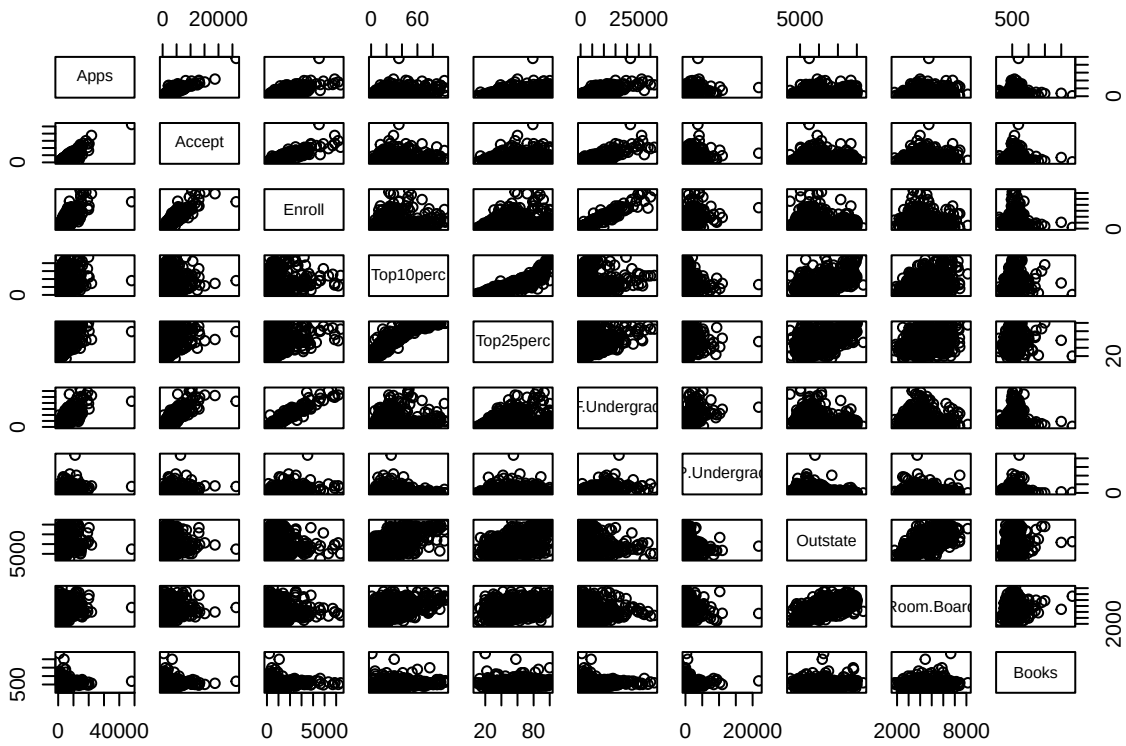

```r
summary(college)
```

```
##       Apps           Accept          Enroll       Top10perc       Top25perc
##  Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
##  1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
##  Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0
##  Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8
##  3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
##  Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0
##   F.Undergrad     P.Undergrad         Outstate       Room.Board
##  Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780
##  1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597
##  Median : 1707   Median :  353.0   Median : 9990   Median :4200
##  Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358
##  3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050
##  Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124
##      Books           Personal         PhD            Terminal
```
```

```
##  Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0
##  1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0
##  Median : 500.0   Median :1200   Median : 75.00   Median : 82.0
##  Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7
##  3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0
##  Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.0
##    S.F.Ratio      perc.alumni        Expend        Grad.Rate
##  Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
##  1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
##  Median :13.60   Median :21.00   Median : 8377   Median : 65.00
##  Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
##  3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
```

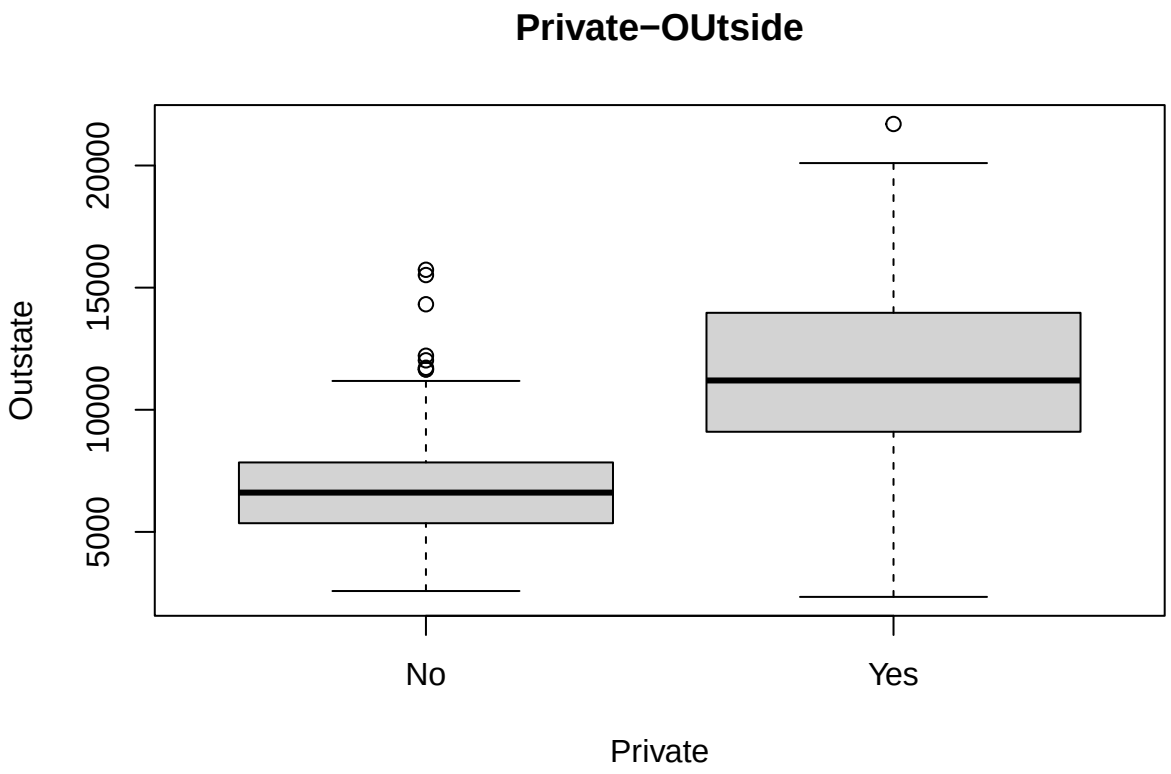
(ii)

```r
pairs(college[,1:10])
```



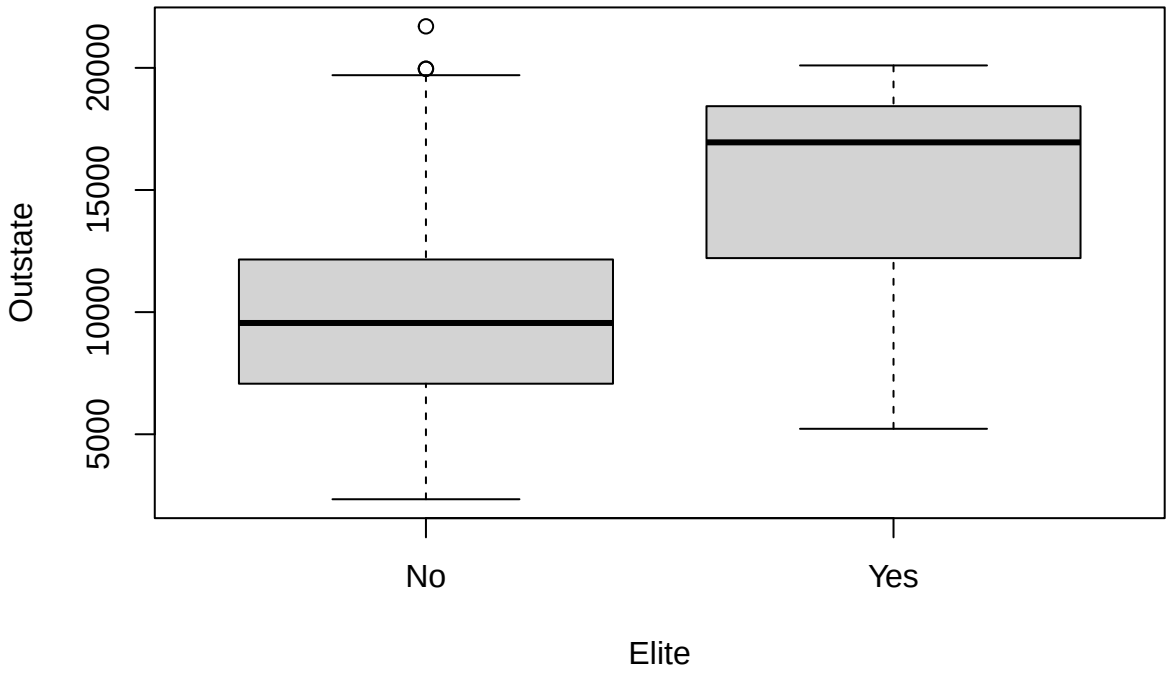
(iii)

```r
boxplot(Outstate~Private,data =A,xlab ="Private",ylab="Outstate",main="Private-OUtside")
```



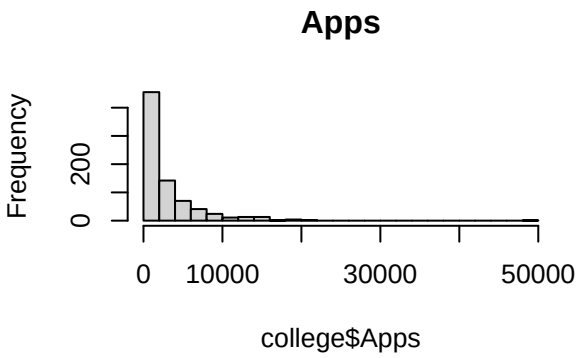

```r
### (iv)
```

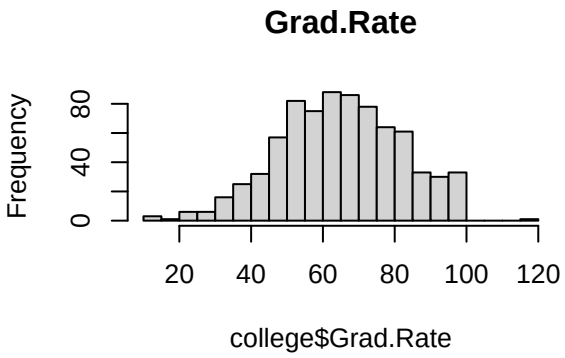
```r
Elite=rep("No",nrow(college))
Elite[college$Top10perc>50]="Yes"
Elite=as.factor(Elite)
college=data.frame(college ,Elite)
boxplot(Outstate ~ Elite, data = college,xlab = "Elite", ylab = "Outstate",main="Elite-Outside")
```

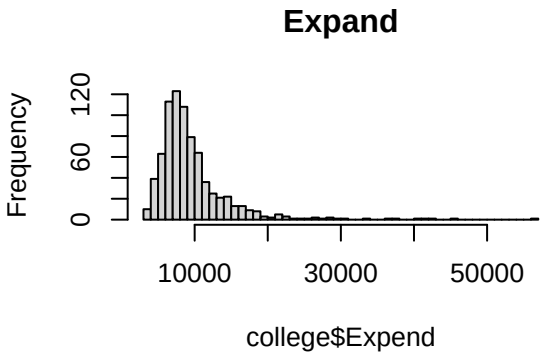


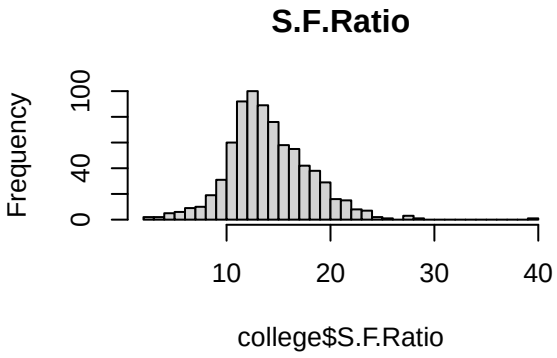

```
### (v)


```r
par(mfrow=c(2,2))
hist(college$Apps, breaks=30, main="Apps")
hist(college$Grad.Rate, breaks=20, main="Grad.Rate")
hist(college$Expend, breaks=50, main="Expand")
hist(college$S.F.Ratio, breaks=40, main="S.F.Ratio")
```
```










```
### (vi)
```

通过 (iii) 可看出，私立学校的学费普遍比公立学校高出 $5000 左右。
通过 (iv) 可看出，Top10% 学校的学费也普遍比公立学校高 $7000 左右。
通过 (v) 可看出，学校的申请数量绝大多数在 10000 以内，极少超过 20000 份，
毕业率大致都在 50%-80% 左右，大学教育支出大部分都在 $10000 左右，师生率在 13% 左右。

(10)

(a)

```
library(MASS)
```

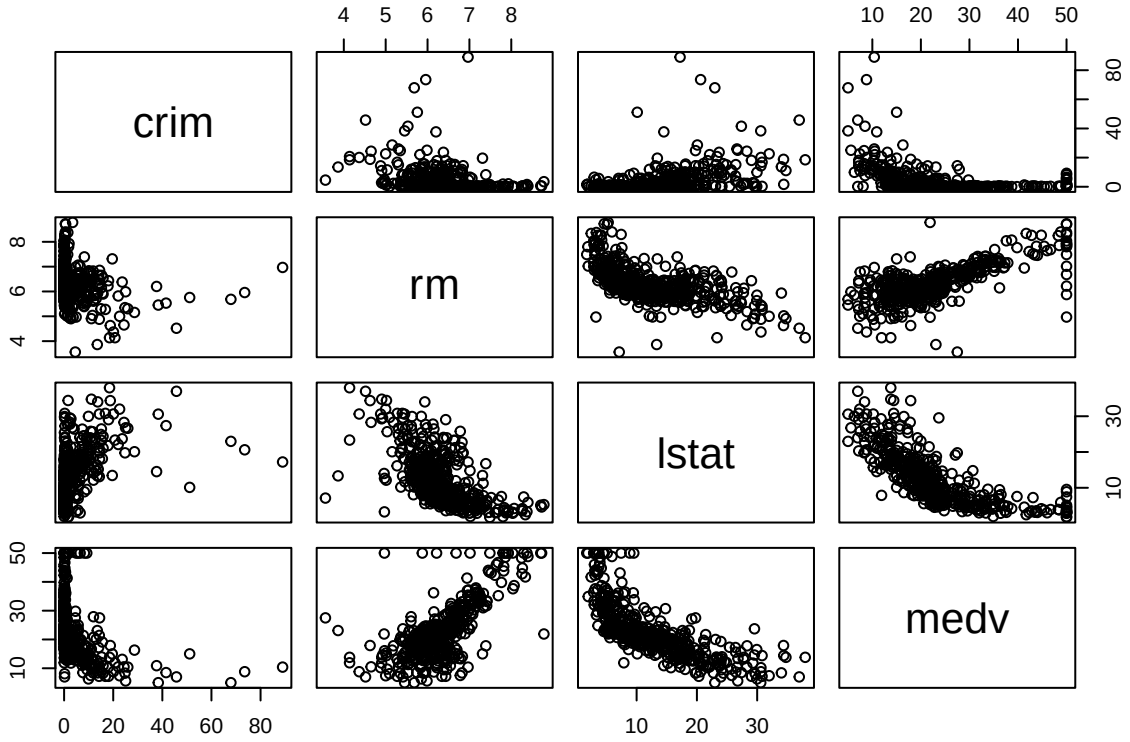
```
##
## Attaching package: 'MASS'
## The following object is masked from 'package:ISLR2':
##
##     Boston
```

```
#Boston
#?Boston
```

有 506 行，14 列，每一行代表一个房子，每一列代表房子的一个指标。

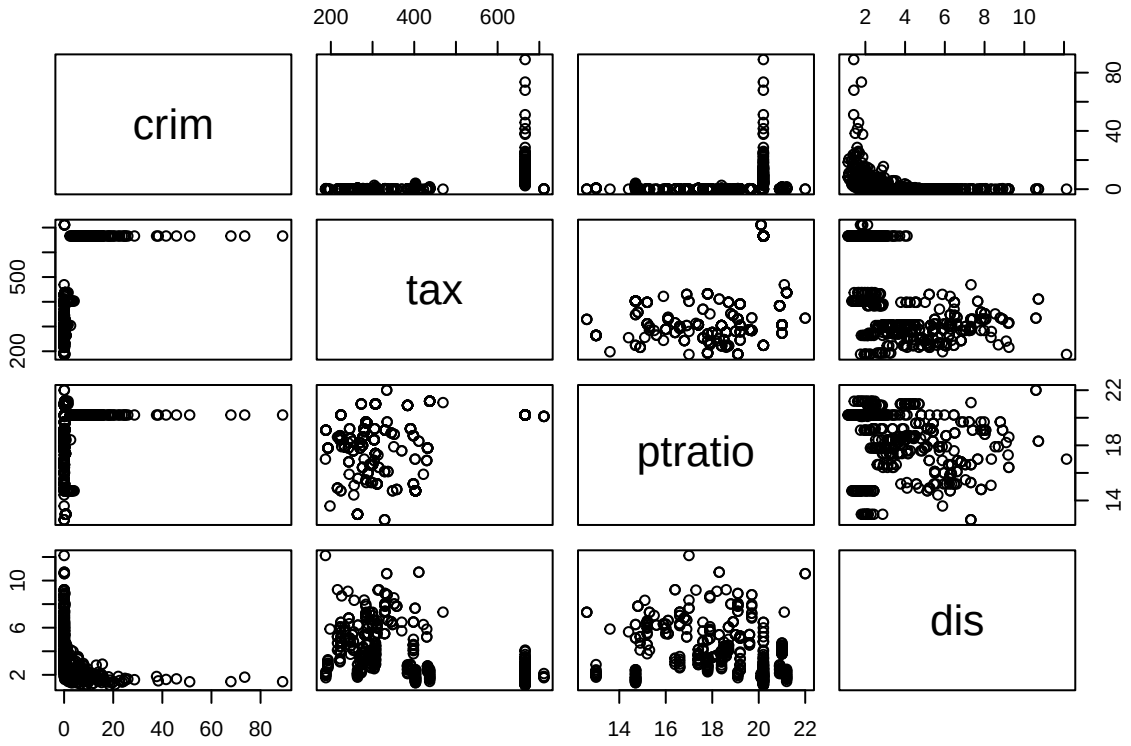
(b)

```
pairs(Boston[, c("crim", "rm", "lstat", "medv")])
```



medv 和 lsat 呈负相关，rm 和 medv 呈正相关。## (c) 犯罪率和交通便利程度，税收比例呈正相关，和地方房价呈负相关。## (d)

```
pairs(Boston[, c("crim", "tax", "ptratio", "dis")])
```



Boston 郊外犯罪率不会特别高，反而是越靠近市中心越高；税率和师生比与郊区与否关系不大。## (e)

```
sum(Boston$chas)
```

```
## [1] 35
```

(f)

```
median(Boston$ptratio)
```

```
## [1] 19.05
```

(g)

```
lowestmedv=Boston[which.min(Boston$medv), ]
lowestmedv
```

```
##         crim zn indus chas   nox    rm age    dis rad tax ptratio black lstat
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.9 30.59
##     medv
## 399    5
```

该地区犯罪率和低收入人口比例都明显偏高，说明该地区就业情况糟糕，许多人为了谋生只能铤而走险。

(h)

```
cat(" 房间数 >7 的区域数: ", sum(Boston$rm > 7), "\n")
```

```
## <U+623F><U+95F4><U+6570> >7<U+7684><U+533A><U+57DF><U+6570><U+FF1A> 64
```

```
cat(" 房间数 >8 的区域数: ", sum(Boston$rm > 8), "\n")
```

```
## <U+623F><U+95F4><U+6570> >8<U+7684><U+533A><U+57DF><U+6570><U+FF1A> 13
```