14) Bayesian Approach for Model Assessment and Selection

Vitor Kamada

February 2018

Reference

Tables, Graphics, and Figures from

The Elements of Statistical Learning

Hastie et al. (2017): Chapters: 7.1 to 7.9

Model Selection vs Model Assessment

Analytically (AIC and BIC)

vs

Efficient Sample Re-Use
(Cross-Validation and Bootstrap)

Loss Function

$$L(Y,\hat{f}(X)) = \begin{cases} (Y-\hat{f}(X))^2 \\ |Y-\hat{f}(X)| \end{cases}$$

$$Err_{\mathcal{T}} = E[L(Y,\hat{f}(X))|\mathcal{T}]$$

$$Err = E[L(Y,\hat{f}(X))] = E[Err_{\mathcal{T}}]$$

$$\bar{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i,\hat{f}(x_i))$$

The Bias–Variance Decomposition

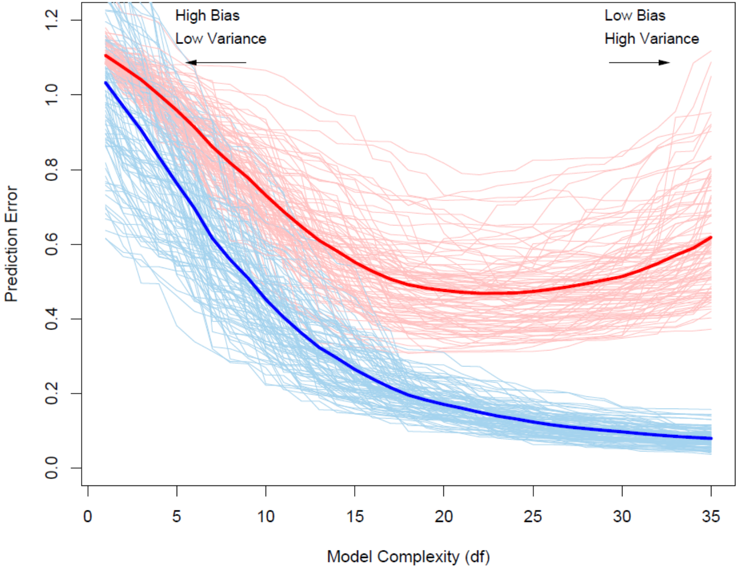
$$Err(x_0) = E[(Y - \hat{f}(x_0))^2 | X = x_0]$$

$$\sigma_\epsilon^2 + [E\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2$$

$$\sigma_\epsilon^2 + \frac{1}{N}\sum_{i=1}^{N}[f(x_i) - E\hat{f}(x_i)]^2 + \frac{p}{N}\sigma_\epsilon^2$$

$$\sigma_\epsilon^2 + [f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})]^2 + \frac{\sigma_\epsilon^2}{k}$$

Test Sample and Training Sample Error



Categorical Response G

$$L(G, \hat{G}(X)) = I(G \neq \hat{G}(X))$$

$$L(G, \hat{p}(X)) = -2 \sum_{k=1}^{K} I(G = k) log \hat{p}_k(X)$$

$$= -2 log \hat{p}_G(X) = \textbf{-2(log-likelihood)}$$

$$L(Y, \theta(X)) = -2 log Pr_{\theta(X)}(Y)$$

$$\bar{err} = -\frac{2}{N} \sum_{i=1}^{N} log \hat{p}_{g_i}(x_i)$$

Optimism of the Training Error Rate

$$op = Err_{in} - \bar{err}$$

$$\frac{1}{N}\sum_{i=1}^{N} E_{Y^0}[L(Y_i^0, \hat{f}(x_i))|\tau] - \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

$$E_y(op) \cong \frac{2}{N}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i)$$

$$E_y(Err_{in}) = E_y(\bar{err}) + 2\frac{d}{N}\sigma_\epsilon^2$$

$$E_y(Err_{in}) = E_y(\bar{err}) + 2\frac{d}{N}\sigma_\epsilon^2$$

For Squared Error Loss:

$$C_p = \hat{Err}_{in} = \bar{err} + 2\frac{d}{N}\hat{\sigma}_\epsilon^2$$

Akaike Information Criterion (AIC)

$$AIC(\alpha) = \bar{err}(\alpha) + 2\frac{d(\alpha)}{N}\hat{\sigma}_{\epsilon}^{2}$$

If $N \to \infty$, then

$$-2E[logPr_{\hat{\theta}}(Y)]$$

$$\approx -\frac{2}{N}E[loglik] + 2\frac{d}{N}$$

Bayesian Approach to Model Selection

$$Pr(M_m|Z) \propto Pr(M_m)Pr(Z|M_m)$$

$$\frac{Pr(M_m|Z)}{Pr(M_l|Z)} = \frac{Pr(M_m)Pr(Z|M_m)}{Pr(M_l)Pr(Z|M_l)}$$

$Pr(M_m)$ is constant if prior over models is uniform

$$logPr(Z|M_m)$$

$$= logPr(Z|\hat{\theta}_m, M_m) - \frac{d_m}{2} logN + O(1)$$

Bayesian Information Criterion (BIC)

$$BIC = -2loglik + (logN)d$$

$$AIC(\alpha) = \bar{err}(\alpha) + 2\frac{d(\alpha)}{N}\hat{\sigma}_\epsilon^2$$

As $N \to \infty$, BIC selects the correct model, but AIC tends to choose too complex model

library(ISLR); library(stargazer); stargazer(Hitters)

Statistic	N	Mean	St. Dev.	Min	Max
AtBat	263	403.643	147.307	19	687
Hits	263	107.829	45.125	1	238
HmRun	263	11.620	8.757	0	40
Runs	263	54.745	25.540	0	130
RBI	263	51.487	25.883	0	121
Walks	263	41.114	21.718	0	105
Years	263	7.312	4.794	1	24
CAtBat	263	2,657.544	2,286.583	19	14,053
CHits	263	722.186	648.200	4	4,256
CHmRun	263	69.240	82.198	0	548
CRuns	263	361.221	331.199	2	2,165
CRBI	263	330.418	323.368	3	1,659
CWalks	263	260.266	264.056	1	1,566
PutOuts	263	290.711	279.935	0	1,377
Assists	263	118.760	145.081	0	492
Errors	263	8.593	6.607	0	32
Salary	263	535.926	451.119	67.500	2,460.000

library(leaps); regfit.full=regsubsets(Salary~.,Hitters)

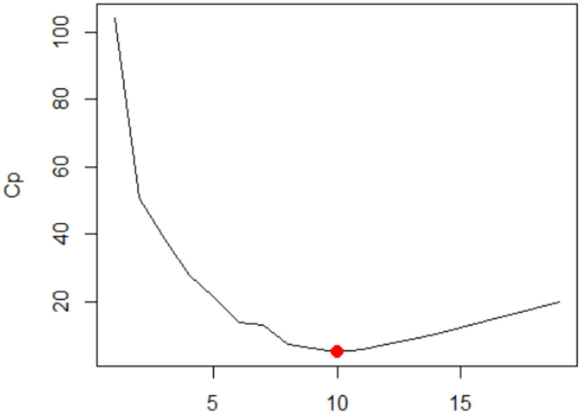
summary(regfit.full)

```
         AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns
1  ( 1 ) " "   " "  " "   " "  " " " "   " "   " "    " "   " "    " "  
2  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "  
3  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "  
4  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "  
5  ( 1 ) "*"   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "  
6  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    " "  
7  ( 1 ) " "   "*"  " "   " "  " " "*"   " "   "*"    "*"   "*"    " "  
8  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   "*"    "*"  
         CRBI CWalks LeagueN DivisionW PutOuts Assists Errors NewLeagueN
1  ( 1 ) "*"  " "    " "     " "       " "     " "     " "    " "       
2  ( 1 ) "*"  " "    " "     " "       " "     " "     " "    " "       
3  ( 1 ) "*"  " "    " "     " "       "*"     " "     " "    " "       
4  ( 1 ) "*"  " "    " "     "*"       "*"     " "     " "    " "       
5  ( 1 ) "*"  " "    " "     "*"       "*"     " "     " "    " "       
6  ( 1 ) "*"  " "    " "     "*"       "*"     " "     " "    " "       
7  ( 1 ) " "  " "    " "     "*"       "*"     " "     " "    " "       
8  ( 1 ) " "  "*"    " "     "*"       "*"     " "     " "    " "       
```

plot(reg.summary$cp,xlab="Number of Variables", ylab="Cp",type='l')

which.min(reg.summary$cp)

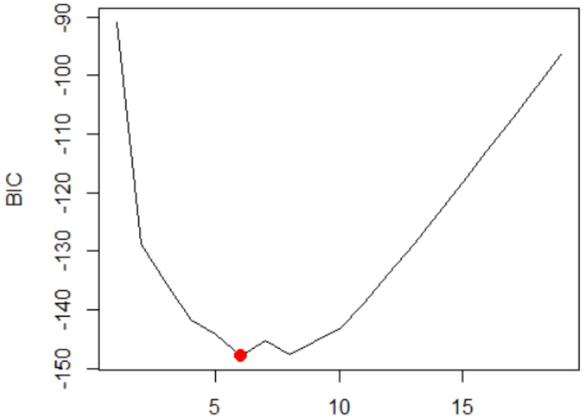
points(10,reg.summary$cp[10],col="red",cex=2,pch=20)



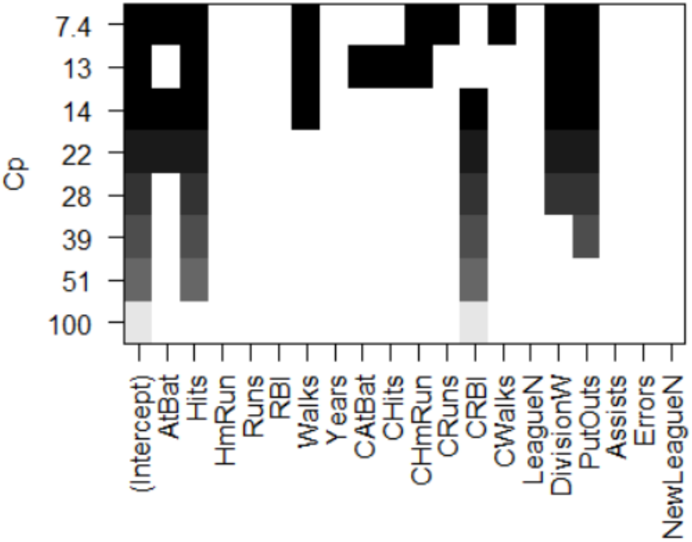
plot(reg.summary$bic,xlab="Number of Variables", ylab="BIC",type='l')

which.min(reg.summary$bic)

points(6,reg.summary$bic[6],col="red",cex=2,pch=20)



plot(regfit.full,scale="Cp")



plot(regfit.full,scale="bic")