Ve406 Lecture 13

Jing Liu

UM-SJTU Joint Institute

June 27, 2018

- Suppose we have a large number of potential predictor variables, e.g.

$$\underbrace{50{,}000{,}000}_{queries} \times \underbrace{154}_{locations} \times \underbrace{182}_{weeks} = 1{,}393.7 \text{ billion data points} \approx 11{,}149.60\text{GB}$$

  that we could use to build a regression model.

Q: How do we decide which variables should enter the model?

Q: What can go wrong?

- If we introduce too few, we are underfitting, and if we introduce too many, we are overfitting, both which lead to unpleasant consequences.
- The type of consequences that we would try to avoid depends on the reason for building a regression model, there are two reasons for building a model:

2. To predict
1. To explain

Q: Why is it unlikely to find a model that fulfil both reasons?

- There is a bias-variance trade-off,

$$\underbrace{\mathbb{E}\left[\left(Y_{n+1}-\hat{Y}_{n+1}\right)^2 \mid \cdot\right]}_{\text{mean squared error}} = \sigma^2 + \underbrace{\mathrm{Var}\left[\hat{Y}_{n+1} \mid \cdot\right]}_{\text{variance}} + \underbrace{\mathbb{E}\left[\hat{Y}_{n+1}-\mu_{n+1} \mid \cdot\right]^2}_{\text{bias}}$$

  however, the trade-off is not necessarily one-for-one.

- The effect of overfitting is to increase the variance, whereas the effect of underfitting is to increase the bias.
- If the aim is to predict, then we try minimising MSE while keeping the variance down, i.e., overfitting is worse than underfitting in general.
- If the aim is to explain, then we try minimising MSE while keeping the bias down, i.e. underfitting is worse than overfitting in general.

- In both cases, we would like to select a model that has a small MSE.

Q: How can we get some idea on MSE?

- Data splitting
- If we have a lot of data, we can divide the data into two parts, the training set and the test set. We use the training set to build models and test set to estimate the MSE of prediction.
- Cross-validation
- If there is not sufficient data, we split the data into several parts. Treating one of them as the test set and estimate MSE, and the rest as the training set. Repeat by treating each of other parts as the test set and the rest as the training set. Take the mean of all of those estimates.
- Bootstrap
- Sample the data with replacement to create a training set, and use the original data set as the test set and estimate the MSE. Repeat this many times and take the mean of all of those estimates.

- The last two are computationally intense, thus were avoided in the old days.
- Traditionally, followings are used
- Residual sum of squares
- This simply uses the residual sum of squares as the estimate for MSE. This is usually too optimistic and underestimates the prediction error.
- If we are simply trying to decide whether a submodel is appropriate, then
- t-test
- Adjusted R^2
- Akaike information criterion

$$AIC = 2k - 2\ln(L_m)$$

where L_m denotes the maximum value of the likelihood function, and k is the number of parameters estimated. Small values of AIC means better.

- Bayesian information criterion

$$AIC = k \ln n - 2 \ln (L_m)$$

Similar to AIC, but penalises the number of parameters more severely, since n is the number of observations. Small values of BIC means better.

- Mallow's Cp
- Before cross-validation and the bootstrap were realistic, it was used as an estimation of MSE. Small values of Cp means better models.
- With modern computing power, we could consider all possible regression models for a given set of variables/transformation of variables.

```
> # install.packages("leaps")
> source("~/Desktop/allpossregs.R")
> allpossregs(yield~conversion+flow+ratio,
+             data = chem_pro.df)
```

```
> allpossregs(yield~conversion+flow+ratio,
+                 data = chem_pro.df)
```

	rssk	sigma2	adjRsq	Cp	AIC	BIC	CV	conversion	flow	ratio
1	426.892	10.164	0.393	3.906	47.906	51.474	41.695	1	0	0
2	388.917	9.486	0.434	2.000	46.000	51.353	41.110	1	1	0
3	388.917	9.723	0.420	4.000	48.000	55.137	44.535	1	1	1

- It produces the best of models of size 1, 2, 3 according to cross-validation.

```
> allpossregs(yield~conversion+flow+I(1/ratio),
+                 data = chem_pro.df, best = 2)
```

	rssk	sigma2	adjRsq	Cp	AIC	BIC	CV	conversion	flow	I(1/ratio)
1	426.892	10.164	0.393	9.290	53.290	56.858	41.968	1	0	0
1	634.515	15.108	0.098	33.263	77.263	80.831	64.039	0	0	1
2	388.917	9.486	0.434	6.905	50.905	56.258	41.109	1	1	0
2	421.494	10.280	0.386	10.667	54.667	60.019	70.830	1	0	1
3	346.433	8.661	0.483	4.000	48.000	55.137	51.532	1	1	1

- If we are fitting the model for the purposes of prediction, we would use the prediction error criteria CV/BOOT.
- If we are to explain, then AIC or BIC is more appropriate.