MONASH University

ETC3250

Business Analytics

Week 6
Model assessment and selection

28 August 2017

Outline

Week	Topic	Chapter	Lecturer
1	Introduction to business analytics & R	1	Souhaib
2	Statistical learning	2	Souhaib
3	Regression for prediction	3,7	Tas & David
4	Classification	4	Souhaib
5	Classification	4, 9	Souhaib
6	Model selection and resampling methods	5	Souhaib
7	Dimension reduction	6,10	Souhaib
8	Advanced regression	6	Souhaib
9	Advanced learning methods	8	Souhaib
	Semester break		
10	Clustering	10	Souhaib
11	Visualization		Souhaib
12	Data wrangling		Souhaib

Model assessment and selection

- Which predictors should we choose?
- How do we choose the df for a spline?
- We need a way of comparing two competing models
- If there are a limited number of predictors, we can study **all possible models**.
- Otherwise we need a **search strategy** to explore some potential models.

- The process of evaluating a model's performance is known as model assessment
- The process of selecting the proper level of flexibility for a model is known as model selection

Training vs Test MSEs

Suppose we have a regression model $y = f(x) + \varepsilon$.

Estimate $\hat{f}$ from some **training data**, $Tr = \{x_i, y_i\}_1^n$.

Training Mean Squared Error

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}(x_i)]^2$$

Real accuracy using **test data** $Te = \{\tilde{x}_i, \tilde{y}_i\}_1^m$

Test Mean Squared Error

$$\frac{1}{m}\sum_{i=1}^{m}[\tilde{y}_i - \hat{f}(\tilde{x}_i)]^2$$

The bias-variance tradeoff

MSE decomposition

If $Y = f(x) + \varepsilon$ and $f(x) = \mathrm{E}[Y \mid X = x]$, then the expected **test** MSE for a new Y at x_0 will be equal to

$$\mathrm{E}[(Y - \hat{f}(x_0))^2] = [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\hat{f}(x_0)) + \mathrm{Var}(\varepsilon)$$

Test MSE = Bias^2 + Variance + Irreducible error

- The expectation averages over the variability of Y as well as the variability in the training data.
- As the flexibility of $\hat{f}$ increases, its variance increases and its bias decreases.
- Choosing the flexibility based on average test MSE amounts to a **bias-variance trade-off**.

Training vs Test MSEs

Let $\hat{\beta}$ denote the OLS estimate using the training data, $Tr = \{x_i, y_i\}_1^n$:

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

then

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\beta}' x_i]^2\right] \leq \mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}[\tilde{y}_i - \hat{\beta}' \tilde{x}_i]^2\right].$$

Training vs Test MSEs

1. $\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2\right]$
2. $A = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\beta}'x_i]^2$ and $B = \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \tilde{\beta}'\tilde{x}_i]^2$ where $\tilde{\beta}'$ is the OLS estimate using the test data, $\{\tilde{x}_i, \tilde{y}_i\}_1^n$.
    - A and B have the same distribution, so $\mathrm{E}[A] = \mathrm{E}[B]$.
3. $B = \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \tilde{\beta}'\tilde{x}_i]^2 \leq \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2$ by OLS property.
4. $\mathrm{E}[A] = \mathrm{E}[B] \leq \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2]$

More on Training vs Test MSEs

We want our model to have **small** expected test error for a **new random point** (x_0, y_0) where both x_0 and y_0 are random.

An easier task would be to produce predictions at the **same** values of the predictor variables as before, but with **different** noises. That is we fit our model using samples from

$$y = x\beta + \varepsilon,$$

and predict for samples from

$$y^{'} = x\beta + \varepsilon^{'}$$

where ε and $\varepsilon^{'}$ are independent but identically distributed.

We want to see if the coefficients estimated using (x_i, y_i) produce good predictions for $(x_i, y_i^{'})$.
$\rightarrow$ If the model can't predict well any more, it has just memorize the noise (only the noise has changed).

More on Training vs Test MSEs

Compare $\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i^{'} - \hat{m}_i]^2\right]$ with $\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{m}_i]^2\right]$ where $\hat{m}_i = x_i\hat{\beta}$.

- y_i and $\hat{m}_i$ are dependent random variables since $\hat{m}_i$ depends notably on y_i
- $y_i^{'}$ and $\hat{m}_i$ are independent random variables

More on Training vs Test MSEs

$$\begin{aligned}\mathrm{E}\left[(y_i-\hat{m}_i)^2\right] &= \mathrm{Var}(y_i-\hat{m}_i)+(\mathrm{E}[y_i-\hat{m}_i])^2\\ &= \mathrm{Var}(y_i)+\mathrm{Var}(\hat{m}_i)-2\mathrm{Cov}(y_i,\hat{m}_i)+(\mathrm{E}[y_i]-\mathrm{E}[\hat{m}_i])^2\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left[(y_i^{'}-\hat{m}_i)^2\right] &= \mathrm{Var}(y_i^{'}-\hat{m}_i)+(\mathrm{E}[y_i^{'}-\hat{m}_i])^2\\ &= \mathrm{Var}(y_i^{'})+\mathrm{Var}(\hat{m}_i)-2\mathrm{Cov}(y_i^{'},\hat{m}_i)+(\mathrm{E}[y_i^{'}]-\mathrm{E}[\hat{m}_i])^2\end{aligned}$$

- y_i is independent of $y_i^{'}$ but has the same distribution: $\mathrm{E}[y_i]=\mathrm{E}[y_i^{'}]$ and $\mathrm{Var}(y_i)=\mathrm{Var}(y_i^{'})$
- $\mathrm{Cov}(y_i^{'},\hat{m}_i)=0$

$$\begin{aligned}\mathrm{E}\left[(y_i^{'}-\hat{m}_i)^2\right] &= \mathrm{Var}(y_i)+\mathrm{Var}(\hat{m}_i)+(\mathrm{E}[y_i]-\mathrm{E}[\hat{m}_i])^2\\ &= \mathrm{E}\left[(y_i-\hat{m}_i)^2\right]+2\mathrm{Cov}(y_i,\hat{m}_i)\end{aligned}$$

More on Training vs Test MSEs

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i' - \hat{m}_i]^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{m}_i]^2\right] + \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(y_i, \hat{m}_i)$$

$$\text{linear model: } \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(y_i, \hat{m}_i) = \frac{2}{n}\sigma^2 \ \mathrm{tr}(H) = \frac{2}{n}\sigma^2(p+1)$$

since $\mathrm{Cov}(y_i, \hat{m}_i) = \sigma^2 H_{ii}$ and $\mathrm{tr}(H) = p + 1$.

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i' - \hat{m}_i]^2\right] \approx \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{m}_i]^2 + \frac{2}{n}\sigma^2(p+1)$$

- The **optimism** of a model is the amount by which its training MSE systematically under-estimates its true expected squared error
- It grows with σ^2 and p. It shrinks with n.

Model assessment

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i' - \hat{m}_i]^2\right] \approx \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{m}_i]^2 + \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(y_i, \hat{m}_i)$$

- One way to estimate test error is to **estimate the optimism** and then **add it to the training error**.
  - Examples for estimates that are linear in their parameters include AIC and BIC.
- Another way is to **directly estimate the test error** using resampling methods.
  - Cross-validation and bootstrap

Assessing regression models

Sum of squared errors

$$\text{SSE} = \sum_{i=1}^{n} e_i^2$$

Minimizing SSE will always choose the model with the most predictors.

Estimated residual variance

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n - k - 1}$$

where $k =$ no. predictors.

Minimizing $\hat{\sigma}^2$ works quite well for choosing predictors (but better methods to follow).

Assessing regression models

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

However . . .

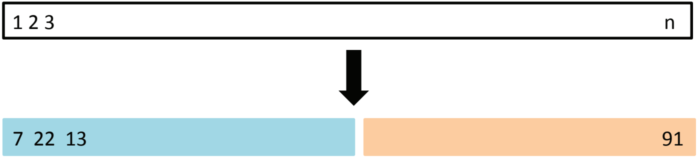
- R^2 does not allow for “degrees of freedom”.
- Adding *any* variable tends to increase the value of R^2, even if that variable is irrelevant.

To overcome this problem, we can use *adjusted* R^2:

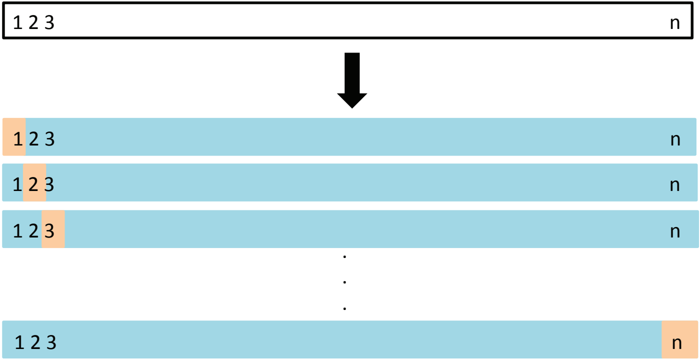
$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Maximizing $\bar{R}^2$ is equivalent to minimizing $\hat{\sigma}^2$.

The validation set



Leave one-out cross-validation



LOO Cross-validation

Leave-one-out cross-validation (LOOCV) for regression can be carried out using the following steps.

- Remove observation i from the data set, and fit the model using the remaining data. Then compute the error ($e_i^* = y_i - \hat{y}_i$) for the omitted observation.
- Repeat step 1 for $i = 1, \ldots, n$.
- Compute the MSE from $\{e_1^*, \ldots, e_n^*\}$. We shall call this the CV.

The best model is the one with the smallest value of CV.

LOOCV vs validation sets

- LOOCV has less bias
  - We repeatedly fit the statistical learning method using training data that contains $n-1$ obs., i.e. almost all the data set is used
- LOOCV produces a less variable MSE
  - The validation approach produces different MSE when applied repeatedly due to randomness in the splitting process, while performing LOOCV multiple times will always yield the same results, because we split based on 1 obs. each time
- LOOCV is (usually) computationally intensive
  - We fit each model n times!

LOOCV for linear models

Fitted values

$$\hat{\boldsymbol{Y}} = \mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the "hat matrix".

Leave-one-out residuals

Let $h_1, \dots, h_n$ be the diagonal values of $\boldsymbol{H}$, then the cross-validation statistic is

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}[e_i/(1-h_i)]^2,$$

where e_i is the residual obtained from fitting the model to all n observations.

Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2(k+1)$$

where L is the likelihood and k is the number of predictors in the model.

- This is a *penalized likelihood* approach.
- *Minimizing* the AIC gives the best model for prediction.
- AIC penalizes terms more heavily than $\bar{R}^2$.
- Minimizing the AIC is asymptotically equivalent to minimizing MSE via leave-one-out cross-validation.

Corrected AIC

For small values of n, the AIC tends to select too many predictors, and so a bias-corrected version of the AIC has been developed.

$$\mathrm{AIC_C} = \mathrm{AIC} + \frac{2(k+2)(k+3)}{n-k-1}$$

As with the AIC, the $\mathrm{AIC_C}$ should be minimized.

Schwartz Bayesian Information Criterion

$$\text{BIC} = -2\log(L) + (k+1)\log(n)$$

where L is the likelihood and k is the number of predictors in the model.

- *Minimizing* the BIC gives the best model for prediction.
- BIC penalizes terms more heavily than AIC
- Also called SBIC and SC.
- Minimizing BIC is asymptotically equivalent to leave-v-out cross-validation when $v = n[1 - 1/(log(n) - 1)]$.

Selecting regression variables

Different strategies to search for the best model: Best subsets, Forward stepwise, Backwards stepwise, ...

Best subsets regression

- Fit all possible regression models using one or more of the predictors.
- Choose the best model based on CV (or an asymptotic equivalent: AIC, AICc).

Warning!

- If there are a large number of predictors, this is not possible.

  For example, 44 predictors leads to 18 trillion possible models!

Selecting regression variables

Backwards stepwise regression

- Start with a model containing all variables.
- Try subtracting one variable at a time. Keep the model if it has lower CV, AICc or AIC.
- Iterate until no further improvement.

Notes:

- Stepwise regression is not guaranteed to lead to the best possible model.
- If you are trying several different models, use the CV, AICc or AIC value to select between them.