2013.5.20

第二十六讲. 统计学习 (II)：变量选择

Occam剃刀原则：若无必要，勿增实体

4. 线性模型用于预测 （续上节课）

- 预测与通常的统计推断问题不同,预测关心的是对响应Y的“估计”而不是回归系数.

$\mathrm{Y} = \mathrm{X}\beta + \varepsilon$

重点是$\hat{\mathrm{Y}}$而不是$\hat{\beta}$，即使$\hat{\beta}$是有偏的。

容许回归系数估计有偏,可能会大幅度地降低预测统计量的方差,从而提高预测精度.

预测问题框架

训练数据集 (X, Y)：假设模型 $\mathrm{Y}_{n\times 1} = \mathrm{X}_{n\times p}\beta + \varepsilon, \varepsilon \sim (0, \sigma^2 I_n)$

预测新数据 x_0所对应的 y_0(与Y独立)：$\mathrm{y}_0 = \mathrm{x}_0'\beta + \varepsilon_0, \varepsilon_0 \sim (0, \sigma^2)$

注意：假设了训练数据 和待预测数据服从同样 模型。

预测误差与MSE

设$\tilde{\beta}$为β的任一估计(可能有偏), 以$\tilde{\mathrm{y}}_0 = \mathrm{x}_0'\tilde{\beta}$预测$\mathrm{y}_0$，其预测误差

$$\begin{aligned}
&\mathrm{E}(\tilde{\mathrm{y}}_0 - \mathrm{y}_0)^2 = \mathrm{E}\left(\tilde{\mathrm{y}}_0 - \mathrm{E}(\mathrm{y}_0) + \mathrm{E}(\mathrm{y}_0) - \mathrm{y}_0\right)^2 = \mathrm{E}\left(\tilde{\mathrm{y}}_0 - \mathrm{E}(\mathrm{y}_0)\right)^2 + \sigma^2 \\
&= \mathrm{E}\left(\mathrm{x}_0'(\tilde{\beta} - \beta)\right)^2 + \sigma^2 = \mathrm{E}\left(\mathrm{x}_0'(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'\mathrm{x}_0\right) + \sigma^2 \\
&= \mathrm{x}_0'\mathrm{M}(\tilde{\beta})\mathrm{x}_0 + \sigma^2
\end{aligned}$$

其中$\tilde{\beta}$的MSE定义为:

向量形式：$\mathrm{M}(\tilde{\beta}) = \mathrm{E}\left((\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'\right) = \mathrm{var}(\tilde{\beta}) + (E\tilde{\beta} - \beta)(E\tilde{\beta} - \beta)'$

$\mathrm{MSE}(\tilde{\beta}) = \mathrm{E}\left((\tilde{\beta} - \beta)'(\tilde{\beta} - \beta)\right) = tr\,\mathrm{var}(\tilde{\beta}) + \| E\tilde{\beta} - \beta \|^2$

为了改进LS估计的预测精度 / MSE，通常采用有偏估计，
即以牺牲无偏性为代价，换取方差的减少. 常用方法有：

- 变量选择方法(variable selection)：选取部分变量,减少估计的方差;
- 压缩估计方法(shrinkage)：比如：$\tilde{\beta} = \hat{\beta}_{\mathrm{LS}}/2$，$\tilde{\beta} = \hat{\beta}_{\mathrm{LS}} 1_{(|\hat{\beta}_{\mathrm{LS}}| < c)}$
- 规则化方法(regularization) /惩罚最小二乘:对回归系数进行限制或约束
- Bayes方法：通过假设参数的随机性，实现平滑、压缩、减少参数的效果。

注：这些方法界限不一定分明，比如
如果压缩估计把某些估计压缩为0, 则达到了选择变量的效果;
规则化方法可理解为Bayes方法, 也是压缩方法。

基于部分模型（选变量模型）的预测误差

全模型(真模型):

$$\mathrm{Y} = \mathrm{X}\beta + \varepsilon = \mathrm{X}_1\beta_1 + \mathrm{X}_2\beta_2 + \varepsilon,\quad \varepsilon \sim (0, \sigma^2 I_n)$$

$$\Rightarrow LS\text{估计}\hat{\beta}_1 = \left(\mathrm{X}_1^{\perp}{}'\mathrm{X}_1^{\perp}\right)^{-1}\mathrm{X}_1^{\perp}{}'\mathrm{Y},\quad \mathrm{X}_1^{\perp} = \mathrm{X}_1 - \mathrm{P}_{\mathrm{X}_2}\mathrm{X}_1$$

部分模型：$\mathrm{Y} = \mathrm{X}_1\beta_1 + \delta,\quad \delta \sim (0, \tau^2 I_n)$

$$\Rightarrow LS\text{估计}\tilde{\beta}_1 = \left(\mathrm{X}_1{}'\mathrm{X}_1\right)^{-1}\mathrm{X}_1{}'\mathrm{Y}$$

即在全模型中, β的估计取为 $\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_1 \\ 0 \end{pmatrix}$

定理1: 真模型成立的条件下：

(1) $\mathrm{E}(\tilde{\beta}_1) \neq \beta_1$, 除非$\mathrm{X}_1{}'\mathrm{X}_2 = 0$或$\beta_2 = 0$;

(2) $\mathrm{var}(\tilde{\beta}_1) \le \mathrm{var}(\hat{\beta}_1)$

(3) 若$\| \mathrm{X}_2^{\perp}\beta_2 \| \le \sigma$，其中$\mathrm{X}_2^{\perp} = \mathrm{X}_2 - \mathrm{P}_{\mathrm{X}_1}\mathrm{X}_2$，则$\mathrm{M}(\tilde{\beta}) \le \mathrm{M}(\hat{\beta})$。

(4)若$\| X_2^{\perp}\beta_2 \| \le \sigma\sqrt{p-q}$, 则$\mathrm{MSE}(\tilde{Y}) \le \mathrm{MSE}(\hat{Y})$, 其中$\tilde{Y} = X_1\tilde{\beta}_1$

注: 条件$\| \mathrm{X}_2^{\perp}\beta_2 \| \le \sqrt{p-q}\,\sigma \;\overset{\text{近似地}}{\Leftrightarrow}\; \| \mathrm{X}_2^{\perp}\hat{\beta}_2 \|^2 \le \hat{\sigma}^2 (p-q)$

$$\Leftrightarrow \mathrm{H}_0: \beta_2 = 0\text{的F检验统计量}F = \frac{\hat{\beta}_2{}'\mathrm{X}_2^{\perp}\mathrm{X}_2^{\perp}{}'\hat{\beta}_2}{(p-q)\hat{\sigma}^2} \le 1$$

引理：实数$a > 0$, 矩阵$A_{n\times n} > 0$，向量$\mathrm{x} \in R^n$, 则$aA \ge \mathrm{xx}' \Leftrightarrow \mathrm{x}'A^{-1}\mathrm{x} \le a$

证明: (1) 给定自变量条件下，

$$\mathrm{E}(\tilde{\beta}_1) = \mathrm{E}\left(\mathrm{X}_1{}'\mathrm{X}_1\right)^{-1}\mathrm{X}_1{}'\mathrm{Y} = \left(\mathrm{X}_1{}'\mathrm{X}_1\right)^{-1}\mathrm{X}_1{}'(\mathrm{X}_1\beta_1 + \mathrm{X}_2\beta_2)$$

$$= \beta_1 + \left(\mathrm{X}_1{}'\mathrm{X}_1\right)^{-1}\mathrm{X}_1{}'\mathrm{X}_2\beta_2 \triangleq \beta_1 + A\beta_2 \neq \beta_1,\quad \text{除非}\mathrm{X}_1{}'\mathrm{X}_2 = 0$$

(2) $\mathrm{var}(\tilde{\beta}_1) = \mathrm{var}\left(\left(\mathrm{X}_1{}'\mathrm{X}_1\right)^{-1}\mathrm{X}_1{}'\mathrm{Y}\right) = \sigma^2\left(\mathrm{X}_1{}'\mathrm{X}_1\right)^{-1} \le \sigma^2\left(\mathrm{X}_1^{\perp}{}'\mathrm{X}_1^{\perp}\right)^{-1} = \mathrm{var}(\hat{\beta}_1)$

(4) $\tilde{Y} = X_1\tilde{\beta}_1$, $\mathrm{var}(\tilde{Y}) = \mathrm{var}(X_1(X_1{}'X_1)^{-1}_{\sigma^2}X_1{}'Y) = \sigma^2 P_{X_1}$

$\mathrm{bias} = \mathrm{E}(\tilde{Y}) - X\beta = P_{X_1}X\beta - X\beta = -X_2^{\perp}\beta_2$

$\mathrm{MSE}(\tilde{Y}) = \mathrm{E}\| \tilde{Y} - X\beta \|^2 = \mathrm{E}\| \tilde{Y} - E(\tilde{Y}) \|^2 + \| E(\tilde{Y}) - X\beta \|^2 = q\sigma^2 + \beta_2{}'X_2^{\perp}{}'X_2^{\perp}\beta_2$

而 $\mathrm{MSE}(\hat{Y}) = E\| \hat{Y} - X\beta \|^2 = tr\,\mathrm{var}(\hat{Y}) = p\sigma^2$

故若$\| X_2^{\perp}\beta_2 \| \le \sigma\sqrt{p-q}$, 则$\mathrm{MSE}(\tilde{Y}) \le \mathrm{MSE}(\hat{Y})$

(3) 由(1), $\mathrm{E}(\tilde{\beta}_1) = \beta_1 + \left(\mathrm{X}_1{}'\mathrm{X}_1\right)^{-1}\mathrm{X}_1{}'\mathrm{X}_2\beta_2 \triangleq \beta_1 + A\beta_2$ bias = $\mathrm{E}(\tilde{\beta}_1) - \beta_1 = A\beta_2$

$$\Rightarrow \mathrm{M}(\tilde{\beta}_1) = \mathrm{var}(\tilde{\beta}_1) + (\mathrm{E}(\tilde{\beta}_1) - \beta_1)(\mathrm{E}(\tilde{\beta}_1) - \beta_1)' = \sigma^2\left(\mathrm{X}_1{}'\mathrm{X}_1\right)^{-1} + A\beta_2\beta_2{}'A'$$

$$\Rightarrow \mathrm{M}(\tilde{\beta}) = E\left(\begin{pmatrix} \tilde{\beta}_1 - \beta_1 \\ 0 - \beta_2 \end{pmatrix}\left((\tilde{\beta}_1 - \beta_1)', (0 - \beta_2)'\right)\right) = \begin{pmatrix} E(\tilde{\beta}_1 - \beta_1)(\tilde{\beta}_1 - \beta_1)' & -E(\tilde{\beta}_1 - \beta_1)\beta_2{}' \\ \beta_2 E(\tilde{\beta}_1 - \beta_1)' & \beta_2\beta_2{}' \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2\left(\mathrm{X}_1{}'\mathrm{X}_1\right)^{-1} & 0 \\ 0 & 0' \end{pmatrix} + \begin{pmatrix} A\beta_2\beta_2{}'A' & -A\beta_2\beta_2{}' \\ -\beta_2\beta_2{}'A' & \beta_2\beta_2{}' \end{pmatrix}$$

条件$\| \mathrm{X}_2^{\perp}\beta_2 \| \le \sigma \Leftrightarrow \beta_2{}'\mathrm{X}_2^{\perp}{}'\mathrm{X}_2^{\perp}\beta_2 \le \sigma^2 \;\overset{\text{由引理}}{\Leftrightarrow}\; \beta_2\beta_2{}' \le \sigma^2(\mathrm{X}_2^{\perp}{}'\mathrm{X}_2^{\perp})^{-1} \triangleq \sigma^2\mathrm{B}$

$$\Rightarrow \begin{pmatrix} A\beta_2\beta_2{}'A' & -A\beta_2\beta_2{}' \\ -\beta_2\beta_2{}'A' & \beta_2\beta_2{}' \end{pmatrix} = \begin{pmatrix} A \\ -I \end{pmatrix}\beta_2\beta_2{}'(A', -I) \le \sigma^2\begin{pmatrix} A \\ -I \end{pmatrix}B(A', -I) = \sigma^2\begin{pmatrix} ABA' & -AB \\ -BA' & B \end{pmatrix}$$

$$\text{所以}\mathrm{M}(\tilde{\beta}) \le \sigma^2\begin{pmatrix} \left(\mathrm{X}_1{}'\mathrm{X}_1\right)^{-1} + ABA' & -AB \\ -BA' & B \end{pmatrix} = \sigma^2(\mathrm{X}'\mathrm{X})^{-1} = \mathrm{M}(\hat{\beta})$$

5. 变量选择准则

全模型(p个变量)：$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$，$\varepsilon \sim (0, \sigma^2 I_n)$
选变量模型 / 部分模型(q个变量)：$Y = X_1\beta_1 + \tilde{\varepsilon}$，$\tilde{\varepsilon} \sim (0, \tau^2 I_n)$

自变量个数q越多，RSS越小，R^2越大,这是因为:

$$RSS_q = \min_{\beta_1} \| Y - X_1\beta_1 \|^2 = \min_{\beta_1,\ \beta_2=0} \| Y - X_1\beta_1 - X_2\beta_2 \|^2 \geq \min_{\beta_1,\ \beta_2} \| Y - X_1\beta_1 - X_2\beta_2 \|^2 = RSS_p$$

故若以RSS或R^2作为变量选择标准,则将选择所有变量.

大多变量选择准则是RSS的修正 (基于当前样本对预测误差的估计).
对各个子模型计算该准则，选择使其达到最小的模型。

变量选择准则有多种 ：

In - sample criteria (使用建立预测模型的数 据评判预测误差）

- 修正的R^2： $\overline{R}_q^2 = 1 - \frac{n-1}{n-q}(1 - R_q^2)$
- 平均残差平方和(Residual mean squares):
  $RMS_q = RSS_q/(n\text{-}q)$

思路是对的，即惩罚大的模型(大的p)，但已很少使用.

- Mallow's C_p
- Akaike's Information Criterion (AIC)
  Schwarz's Bayesian Information Criterion (BIC)

AIC或BIC最为常用。Cp准则接近于AIC，其推导简单易于理解

Out - of - sample criteria (使用另外的数据对预测模型进行评判)

- 交叉验证 (Cross - Validation, CV). 将数据集划分为两部分：
  - training data (learning set)
  - testing data (validation set)

  训练数据集用来构建预 测模型,检验数据集用来评价预 测效果。

最简单的CV方法是 Leave - one - out Cross Validation，即测试集只有一个点，由于测试集太小（与训练集相差太大），预测误差的估计有偏差，故该方法已不常用。但优点是，平方预测误差 PRESS (Prediction sum of squares)有简洁的显式表达:

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{i(-i)})^2 = \sum_{i=1}^{n} \frac{e_i^2}{(1-h_{ii})^2},$$

C_p准则

Mallow's C_p准则： $C_q = \frac{RSS_q}{\hat{\sigma}^2} - n + 2q,$
其中RSS_q为部分模型(q个变量)下的RSS，
$\hat{\sigma}^2$为全模型下的误差方差估计。

全 / 真模型(p个变量)：$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$，$\varepsilon \sim (0, \sigma^2 I_n)$
选变量模型 / 部分模型(q个变量)：$Y = X_1\beta_1 + \tilde{\varepsilon}$，$\tilde{\varepsilon} \sim (0, \tau^2 I_n)$

推导C_q:
考虑in - sample预测(用现有数据(X, Y)得到参数估计，并预测/拟合Y）

以$\tilde{Y} = X_1\tilde{\beta}_1$预测$Y$,其中$\tilde{\beta}_1$为部分模型的LS估计,其MSE:

我们已经证明了：$\text{MSE}(\tilde{Y}) = q\sigma^2 + \beta_2' X_2^{\perp}{}' X_2^{\perp} \beta_2$

考虑到方差/单位，我们以MSE/σ^2作为预测效果的评价：

$$\mathrm{M}_q = \mathrm{MSE}(\tilde{Y})/\sigma^2 = q + \| X_2^{\perp}\beta_2 \|^2 / \sigma^2$$

但M_q不是一个可计算出的量（其中含有未知参数,需要对其进行估计）

其中的$\| X_2^{\perp}\beta_2 \|^2$可用$\mathrm{RSS}_q = \| Y - \tilde{Y} \|^2$估计，其中$\tilde{Y} = \mathrm{P}_{\mathrm{X}_1} Y$，理由如下：

注意到全模型可以写为$\mathrm{Y} = \mathrm{X}_1^* \beta_1 + \mathrm{X}_2^{\perp}\beta_2 + \varepsilon$

故我们可以$\| Y - X_1^* \beta_1 \|^2$ 替代("估计") $\| X_2^{\perp}\beta_2 \|^2$,

其中的参数β_1可使用部分模型进行估计(注意:目标是评价部分模型)

所以可用$\mathrm{RSS}_q = \| Y - X_1^* \tilde{\beta}_1 \|^2 = \| Y - \tilde{Y} \|^2$ 估计$\| X_2^{\perp}\beta_2 \|^2$

但使用 $\mathrm{RSS}_q = (\tilde{Y} - Y)'(\tilde{Y} - Y) = Y'(I - P_{X_1})Y$ 估计$\| X_2^{\perp}\beta_2 \|^2$ 有偏差,这是因为由二次型的期望公式(期望都是在真模型下计算）：

$$E(\mathrm{RSS}_q) = \| X_2^{\perp}\beta_2 \|^2 + \sigma^2 tr(I - P_{X_1}) = \| X_2^{\perp}\beta_2 \|^2 + (n-q)\sigma^2 \qquad (*)$$

所以我们应该使用 $\mathrm{RSS}_q - (n-q)\sigma^2$ “估计” $\| X_2^{\perp}\beta_2 \|^2$

以$q + (\mathrm{RSS}_q - (n-q)\sigma^2)/\sigma^2 = \dfrac{\mathrm{RSS}_q}{\sigma^2} - (n-2q)$ "估计"$\mathrm{M}_q = q + \| X_2^{\perp}\beta_2 \|^2 / \sigma^2$

最后，σ^2以全模型下的$\hat{\sigma}^2 = \mathrm{RSS}_p/(n-p)$代替，即得到$\mathrm{C}_q$

$$\mathrm{C}_q = \hat{\mathrm{M}}_q = \frac{\mathrm{RSS}_q}{\hat{\sigma}^2} - (n-2q)$$

因此为了最小化 $\mathrm{MSE}(\tilde{Y})/\sigma^2$，我们可以近似最小化 C_q

C_p推导中的关键是认识到：
在给定的模型下最小二乘法或极大似然法估计回归系数(或位置参数),
之后再用它们估计残差平方和(二次型)的期望，会出现偏差，
偏差与位置参数个数有关(上页(*)式）。

若$\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$，$\mathrm{var}(\mathbf{x}) = \sigma^2 I_n$，
则$\mathbf{x}$是$\boldsymbol{\mu}$的无偏估计，但$\mathbf{x}'\mathrm{A}\mathbf{x}$不是$\boldsymbol{\mu}'\mathrm{A}\boldsymbol{\mu}$的无偏估计
这是因为(二次型的期望公式):

$$\mathrm{E}(\mathbf{x}'\mathrm{A}\mathbf{x}) = \boldsymbol{\mu}'\mathrm{A}\boldsymbol{\mu} + tr\mathrm{A}\mathrm{var}(\mathbf{x}) = \boldsymbol{\mu}'\mathrm{A}\boldsymbol{\mu} + \sigma^2 \mathrm{tr}(\mathrm{A})$$

回顾经典问题：假设$y_1,\ldots,y_n$ iid $\sim N(\mu,\sigma^2)$,

μ的极大似然估计或最小二乘估计 $\hat{\mu} = \bar{y}$; σ^2的极大似然估计$\hat{\sigma}^2 = \sum (y_i - \bar{y})^2/n$。

$\hat{\sigma}^2$有偏，$\mathrm{E}(\hat{\sigma}^2) = \dfrac{n-1}{n}\sigma^2$，但偏差几乎可以忽略不计。

估计σ^2时使用了μ的估计，导致了偏差（μ已知时$\sum (y_i - \mu)^2/\mathrm{n}$ 为σ的无偏估计）

但在考虑平方和的期望 时，偏差就不是可以忽 略的了：

$$\mathrm{E}\left(\sum (y_1 - \bar{y})^2\right) = (n-1)\sigma^2 = n\sigma^2 - \sigma^2$$
$$\mathrm{E}\left(\sum (y_i - \mu)^2\right) = n\sigma^2$$

同样地，在回归问题中，假设$y_1,\ldots,y_n$ 独立，$y_i \sim N(x_i'\beta, \sigma^2)$,

β长度为p,其最小二乘估计$\hat{\beta}$

$$\mathrm{E}\left(\sum (y_i - x_i'\hat{\beta})^2\right) = (n-p)\sigma^2 = n\sigma^2 - p\sigma^2$$
$$\mathrm{E}\left(\sum (y_i - x_i'\beta)^2\right) = n\sigma^2$$

AIC准则 以及BIC准则

Akaike Information Criterion (AIC, Hirotugu Akaike,1974)：

$\mathrm{AIC} = -2\log L(\hat{\theta}) + 2q$

其中$L(\hat{\theta})$为似然函数极大值，q为参数个数

- AIC是一般的模型选择(包括变量选择)准则,适用于一般概率模型。
- 在正态回归情况下，AIC与C_p类似。
- AIC的推导与C_p类似，但使用Kullback - Leibler距离度量预测误差

BIC (Bayesian Information Criterion, Schwarz 1978)是一个类似的准则，

$\mathrm{BIC} = -2\log L(\hat{\theta}) + (\log n)q$

其中n是样本量，q是参数个数. BIC相比于AIC倾向于选择更小的q

对于正态回归模型

$\mathrm{AIC} = n\log(\mathrm{RSS}_q) + 2q + $常数项; $\mathrm{BIC} = n\log(\mathrm{RSS}_q) + (\log n)q + $常数项