2013.5.15

第二十五讲. 统计学习 (I)：MSE

关联，而不是因果

统计学习或机器学习：通过分析数据 (训练,training)，建立关联关系模型用来预测、判别、分类。

不追求模型正确与否以及参数估计的无偏性，以预测精度为唯一标准。

1. 预测误差

样本：$Y_1,\ldots,Y_n$;

待预测变量：Y， 与$Y_1,\ldots,Y_n$独立

基于历史样本构造预测统计量：$\hat{Y}$ ($Y_1,\ldots,Y_n$的函数)

$$\hat{Y} \xrightarrow{\text{预测}} Y$$

预测误差：$\mathrm{E}(\hat{Y}-Y)^2$

记$\theta=\mathrm{E}(Y)$，

则预测误差分解为：$\mathrm{E}(\hat{Y}-Y)^2=\mathrm{E}(\hat{Y}-\theta+\theta-Y)^2$

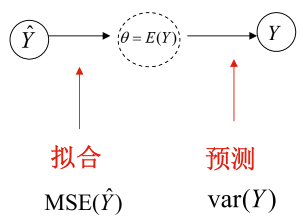
$=\mathrm{E}(\hat{Y}-\theta)^2+E(Y-\theta)^2 \triangleq \mathrm{MSE}(\hat{Y})+\mathrm{var}(Y)$

所以,预测误差由下面两部分构成：

(1) 均方误差$\mathrm{MSE}(\hat{Y})=\mathrm{E}(\hat{Y}-\theta)^2$： $\hat{Y}$ 估计参数θ的误差

(2) $\mathrm{var}(Y)$：被预测量Y在θ附近的波动程度(方差)

由该分解，预测过程可以理解为： 先拟合，后预测



而MSE可分解为 $\mathrm{MSE} = \mathrm{variance} + \mathrm{bias}^2$

设 $\hat{\theta}$ 是 θ 的一个估计，则

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2 = \mathrm{E}(\hat{\theta} - \mathrm{E}(\hat{\theta}) + \mathrm{E}(\hat{\theta}) - \theta)^2$$

$$= \mathrm{var}(\hat{\theta}) + [\mathrm{E}(\hat{\theta}) - \theta]^2$$

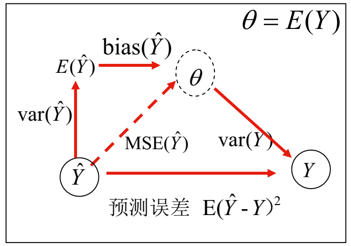
所以预测误差可以分解为三部分：

$$\mathrm{E}(\hat{Y} - Y)^2 = \mathrm{MSE}(\hat{Y}) + \mathrm{var}(Y) = \mathrm{var}(\hat{Y}) + \mathrm{bias}(\hat{Y})^2 + \mathrm{var}(Y)$$

由于被预测对象的方差无法控制，所以一个好的预测应该有尽量小的**MSE**，即尽量小的 $\mathrm{variance} + \mathrm{bias}^2$

2. 均方误差与有偏估计

预测误差 $\mathrm{E}(\hat{Y} - Y)^2 = \mathrm{MSE}(\hat{Y}) + \mathrm{var}(Y) = \mathrm{var}(\hat{Y}) + \mathrm{bias}(\hat{Y})^2 + \mathrm{var}(Y)$



为了减小**MSE**， 需综合平衡方差和偏差：

适度增大偏差，可能导致其方差大幅度下降，从而**MSE**大幅度减少。

如何减小方差？

(1) 压缩(shrinkage)：$X \to \lambda X$，$0 \le \lambda \le 1$，（Bayes方法、惩罚方法 ...）

(2) 截断(truncation)：$X \to X I_{(|X| \le c)}$

例1：设样本 $y_1, y_2, ..., y_n$ iid $\sim N(\theta, \sigma^2)$. 基于该样本预测 $y \sim N(\theta, \sigma^2)$.

一个自然的预测为 $\bar{y}$，其MSE： $\mathrm{E}(\bar{y} - \theta)^2 = \dfrac{\sigma^2}{n}$

假设预测变量取为 $\tilde{y} \equiv \lambda \bar{y}$，$0 \le \lambda \le 1$

其MSE：$\mathrm{E}(\tilde{y} - \theta)^2 = E(\lambda\bar{y} - \theta)^2 = \lambda^2\sigma^2/n + (1-\lambda)^2\theta^2$

如果 $|\theta| \le \dfrac{\sigma}{\sqrt{n}} \times \sqrt{\dfrac{1+\lambda}{1-\lambda}}$，则 $\mathrm{MSE}(\tilde{y}) \le \mathrm{MSE}(\bar{y})$

即，若 σ^2 很大或 $|\theta|$ 较小，则 $\tilde{y}$ 预测效果好于 $\bar{y}$

特别地，若 $|\theta| \le \dfrac{\sigma}{\sqrt{n}}$，则 $\mathrm{MSE}(\tilde{y} = 0) \le \mathrm{MSE}(\bar{y})$

例2. 设正确模型为： $y = a + bx + \varepsilon, \mathrm{E}(\varepsilon) = 0, \mathrm{var}(\varepsilon) = \sigma^2$

基于历史数据 $(x_i, y_i), i = 1, .., n$，得LS估计 $\hat{a}$，$\hat{b}$。

需要 预测"新数据" x_0 对应的响应 y_0，假设 (x_0, y_0) 也满足上述模型：

$$y_0 = a + bx_0 + \varepsilon_0, \mathrm{E}(\varepsilon_0) = 0, \mathrm{var}(\varepsilon_0) = \sigma^2$$

则通常使用 $\hat{y}_0 = \hat{a} + \hat{b}x_0$ 进行预测，其MSE：

$$\mathrm{MSE}(\hat{y}_0) = \frac{1}{n}\sigma^2 + \frac{(x_0 - \bar{x})^2}{s_{xx}}\sigma^2.$$

若取b的估计为 $\tilde{b} \equiv 0$（有偏），以 $\tilde{y}_0 = \bar{y}$ 预测 y_0 效果如何？

$$\mathrm{MSE}(\tilde{y}_0) = \frac{\sigma^2}{n} + b^2(x_0 - \bar{x})^2$$

当 $|b| \le \dfrac{\sigma}{\sqrt{s_{xx}}}$（$b$ 较小，σ 较大，σ_x 较小）时，$\mathrm{MSE}(\tilde{y}_0) \le \mathrm{MSE}(\hat{y}_0)$，$\tilde{y}_0$ 预测效果更好.

3. 有偏统计与统计学习

- **James-Stein (1956,1961)**首次提出了正态分布均值向量的有偏估计 **– James-Stein**估计。
- **Hoerl and Kennard (1970)** 提出了岭估计**(ridge estimator).**
- 规则化/带惩罚的最小二乘：岭估计可以看作是**L2**惩罚下的最小二乘估计, **Tibshirani (1996)**提出了**L1** 惩罚下的最小二乘估计，即**LASSO**。
- **Vapnik** 认识到有偏统计对于拟合/预测/分类的重要性，认为统计分为 **Fisher**统计 **(**无偏统计**)** 和统计学习/机器学习 **(**有偏统计**).**

一般来说，统计学习不过分重视模型正确与否、有偏还是无偏，一切以预测方法在检验数据集上的预测精度为标准 **–** 这反映了统计学习的实用主义特性。

James-Stein 估计(1956，1961)

样本$\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_n$ iid $\sim N(\boldsymbol{\theta},\sigma^2 I_m)$. $m\geq 3$. 假设σ^2已知

最小二乘估计$\hat{\boldsymbol{\theta}}=\bar{\mathbf{y}}$ 是最小方差无偏估计

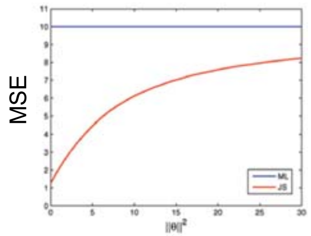
James和Stein定义了如下估计：

$$\hat{\boldsymbol{\theta}}_{JS}=\left(1-\frac{(m-2)\sigma^2}{\|\bar{\mathbf{y}}\|^2}\right)\bar{\mathbf{y}}$$

显然它是$\bar{\mathbf{y}}$的压缩估计，是有偏的。

James-Stein估计的MSE 小于$\hat{\boldsymbol{\theta}}=\bar{\mathbf{y}}$的MSE：

$$\mathrm{E}\|\hat{\boldsymbol{\theta}}_{JS}-\hat{\boldsymbol{\theta}}\|^2<\mathrm{E}\|\bar{\mathbf{y}}-\hat{\boldsymbol{\theta}}\|^2$$



统计学习的基本原则

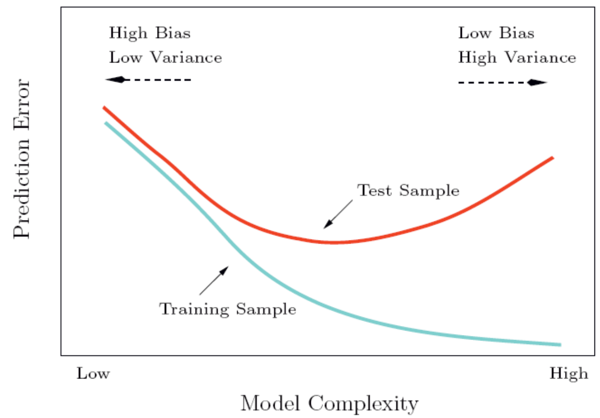
- **可泛化(Generalization):**

基于现有数据得到的预测模型需要有较好的泛化能力 **(generalization)**或预测能力。

一般说来，简单的方法才有强的泛化能力

- **Occam**剃刀原则 **(Occam's Razor, law of parsimony):**

若无必要，勿增实体 （简洁好于复杂）



模型/方法不要过于复杂, 避免过度拟合**(overfitting)**或者过度挖掘 **(data snooping, data dredging, data fishing)**；

4. 线性模型用于预测

- 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.

$Y = X\beta + \varepsilon$

重点是$\hat{Y}$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 $Y_{n\times 1} = X_{n\times p}\beta + \varepsilon, \varepsilon \sim (0, \sigma^2 I_n)$

预测新数据 x_0所对应的y_0：$y_0 = x_0'\beta + \varepsilon_0, \varepsilon_0 \sim (0, \sigma^2)$

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$\tilde{y}_0 = x_0'\tilde{\beta}$预测$y_0$，其预测误差

$$\begin{aligned}
E(\tilde{y}_0 - y_0)^2 &= E(\tilde{y}_0 - E(y_0) + E(y_0) - y_0)^2 = E(\tilde{y}_0 - E(y_0))^2 + \sigma^2 \\
&= E\left(x_0'(\tilde{\beta} - \beta)\right)^2 + \sigma^2 = E\left(x_0'(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'x_0\right) + \sigma^2 \\
&= x_0'M(\tilde{\beta})x_0 + \sigma^2
\end{aligned}$$

其中$\tilde{\beta}$的MSE定义为(向量形式)：

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

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：$\tilde{\beta} = \hat{\beta}_{LS}/2$, $\tilde{\beta} = \hat{\beta}_{LS} 1_{(|\hat{\beta}_{LS}|<c)}$
- 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。

基于部分模型的预测误差

全模型(真模型):

$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$，$\varepsilon \sim (0, \sigma^2 I_n)$

$\Rightarrow LS$估计$\hat{\beta}_1 = \left(X_1^{\perp}{}'X_1^{\perp}\right)^{-1} X_1^{\perp}{}'Y$，$X_1^{\perp} = X_1 - P_{X_2}X_1$

部分模型：$Y = X_1\beta_1 + \delta$，$\delta \sim (0, \tau^2 I_n)$

$\Rightarrow LS$估计$\tilde{\beta}_1 = \left(X_1'X_1\right)^{-1} X_1'Y$

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 $\mathrm{MSE}(\tilde{\beta})\le\mathrm{MSE}(\hat{\beta})$，即基于部分模型的预测误差较小。

注: 条件 $\|\mathrm{X}_2^{\perp}\beta_2\|\le\sigma\overset{近似地}{\Leftrightarrow}\|\mathrm{X}_2^{\perp}\hat{\beta}_2\|^2\le\hat{\sigma}^2$

$\Leftrightarrow \mathrm{H}_0:\beta_2=0$ 的F检验统计量

$$F=\frac{\hat{\beta}_2'\mathrm{X}_2^{\perp}\mathrm{X}_2^{\perp}{}'\hat{\beta}_2}{k\hat{\sigma}^2}\le\frac{1}{k},\ k=\beta_2\text{的长度。}$$

证明：

(1) 给定自变量条件下，

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

除非 $\mathrm{X}_1'\mathrm{X}_2=0$

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

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

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

(3) 给定自变量条件下，

由(1), $\mathrm{E}(\tilde{\beta}_1)=\beta_1+(\mathrm{X}_1'\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{X}_2\beta_2\mathrel{\hat=}\beta_1+A\beta_2$

bias = $\mathrm{E}(\tilde{\beta}_1)-\beta_1=A\beta_2$

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

$$\Rightarrow\mathrm{MSE}(\tilde{\beta})=E\left(\begin{pmatrix}\tilde{\beta}_1-\beta_1\\0-\beta_2\end{pmatrix}\left((\tilde{\beta}_1-\beta_1)',(0-\beta_2)'\right)\right)$$
$$=\begin{pmatrix}E(\tilde{\beta}_1-\beta_1)(\tilde{\beta}_1-\beta_1)' & -E(\tilde{\beta}_1-\beta_1)\beta_2'\\ \beta_2E(\tilde{\beta}_1-\beta_1)' & \beta_2\beta_2'\end{pmatrix}=\begin{pmatrix}\sigma^2(\mathrm{X}_1'\mathrm{X}_1)^{-1}+A\beta_2\beta_2'A' & -A\beta_2\beta_2'\\ -\beta_2\beta_2'A' & \beta_2\beta_2'\end{pmatrix}$$
$$=\begin{pmatrix}\sigma^2(\mathrm{X}_1'\mathrm{X}_1)^{-1} & 0\\ 0 & 0'\end{pmatrix}+\begin{pmatrix}A\beta_2\beta_2'A' & -A\beta_2\beta_2'\\ -\beta_2\beta_2'A' & \beta_2\beta_2'\end{pmatrix}$$

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

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

所以

$$\mathrm{MSE}(\tilde{\beta})=\begin{pmatrix}\sigma^2(\mathrm{X}_1'\mathrm{X}_1)^{-1} & 0\\ 0 & 0'\end{pmatrix}+\begin{pmatrix}A\beta_2\beta_2'A' & -A\beta_2\beta_2'\\ -\beta_2\beta_2'A' & \beta_2\beta_2'\end{pmatrix}$$
$$\le\sigma^2\begin{pmatrix}(\mathrm{X}_1'\mathrm{X}_1)^{-1}+ABA' & -AB\\ -BA' & B\end{pmatrix}=\sigma^2(\mathrm{X'X})^{-1}=\mathrm{MSE}(\hat{\beta})$$