第29讲 预测与均方误差

2020.6.3

Occam剃刀原则： 若无必要，勿增实体
Entities should not be multiplied unnecessarily

机器学习/统计学习

机器学习/统计学习：机器学习（machine learning）通过经验学习/分析数据，改善计算机算法，模仿实现人类学习行为。机器学习分为有监督的学习和无监督的学习（supervised、unsupervised learning）。

- 有监督学习（回归、分类）：

  以自变量（特征feature，预测变量）预测响应变量，当响应为类别时称为分类，当响应为连续变量时称为回归。

- 无监督学习（聚类、主成分等）：

  数据没有响应变量（即没有监督），目标是从数据(feature)中挖掘模式特征，比如聚类分析、主成分分析等。

预测：“估计”随机变量

基于历史数据，包括响应变量和自变量，建立模型描述变量之间的关联关系（未必因果，模型未必正确），并对仅含自变量的新数据的响应变量进行预测或分类。预测的一般原则为：

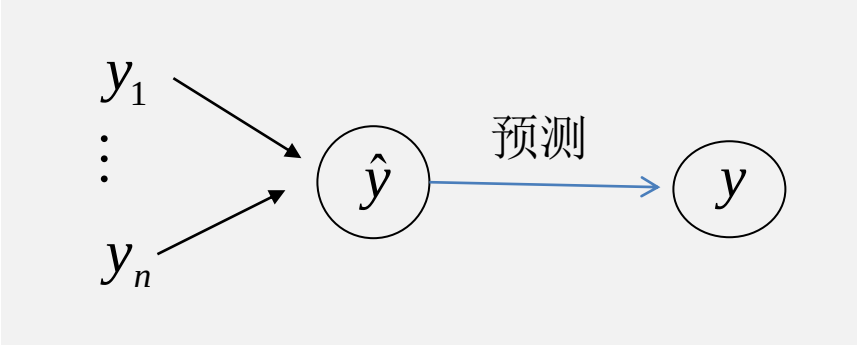
- 预测效果以预测精度为准。模型不必正确，预测量不必无偏。
- 可泛化（generalization）：可推广，适用于不同场景
- 简约原则（Occam’s Razor，Occam剃刀原则)：若无必要,勿增实体
  Entities should not be multiplied unnecessarily

预测误差依赖于均方误差！

历史数据/样本： $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$

待预测随机变量：对给定的自变量$\mathbf{x}$,预测对应的响应y

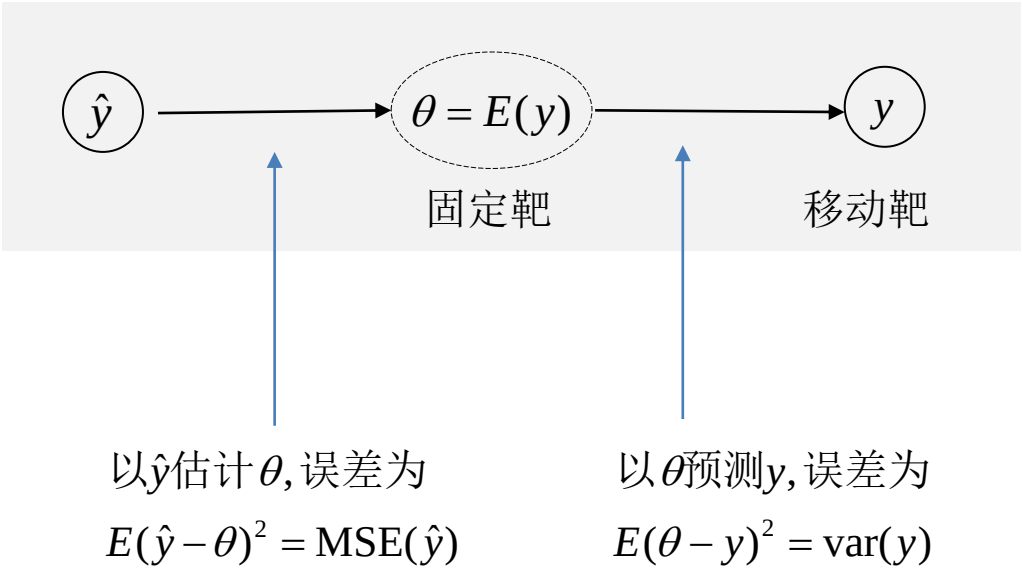
预测统计量： $\hat{y} = f(y_1, \ldots, y_n)$ -为简单计，略去自变量



定义. 预测误差 $e(\hat{y}) = E(\hat{y} - y)^2$

向量情形的预测误差 $e(\hat{\mathbf{y}}) = E \| \hat{\mathbf{y}} - \mathbf{y} \|^2 = \mathrm{E}(\hat{\mathbf{y}} - \mathbf{y})^{\mathrm{T}} (\hat{\mathbf{y}} - \mathbf{y})$

预测随机目标 y (移动靶), 实质在于估计其期望位置$E(y)=\theta$ (固定靶)



均方误差：预测统计量与被预测变量的均值的期望距离平方，
它决定了预测误差：$\mathbf{e}(\hat{y}) = E(\hat{y}-y)^2 = MSE(\hat{y}) + var(y)$

命题1：假设 $y_1,...,y_n,y$ 为独立$r.v.'s$，$\theta = E(y)$, $\hat{y} = f(y_1,...,y_n)$ 的预测误差

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

其中$\mathrm{MSE}(\hat{y}) = \mathrm{MSE}(\hat{y};\theta) = E(\hat{y}-\theta)^2$是$\hat{y}$作为$\theta$的估计的均方误差.

证：$\theta = E(y)$，由$\hat{y}$与y独立，

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

注：

被预测对象y的方差$\mathrm{var}(y)$不可控,为了减小预测误差，我们应该减小$\mathrm{MSE}(\hat{y})$。

而MSE又可分解为方差于偏差平方之和（命题2）

方差与偏差的权衡：MSE=Variance+Bias^2

命题2 (The bias - variance trade - off/dilemma/decompostion).

设 $\hat{y}$ 是参数 θ 的一个估计，$\mu = E(\hat{y})$，$\hat{y}$ 的偏差为 $\text{Bias}(\hat{y}) = E(\hat{y}) - \theta = \mu - \theta$，则均方误差可分解为：

$$\text{MSE}(\hat{y}) = \text{var}(\hat{y}) + [\text{Bias}(\hat{y})]^2 .$$

证明：$\text{MSE}(\hat{y}) = \text{MSE}(\hat{y};\theta) = E(\hat{y} - \theta)^2 = E\left(\hat{y} - \mu + \mu - \theta\right)^2$

$$= E\left(\hat{y} - \mu\right)^2 + \left[\mu - \theta\right]^2 = \text{var}(\hat{y}) + \left[\text{Bias}(\hat{y})\right]^2$$

注意：$\mu = E(\hat{y})$是预测统计量$\hat{y}$的均值，$\theta = E(y)$是被预测量y的均值.

而$\text{Bias}(\hat{y}) = \mu - \theta$是预测量偏离被预测量的期望差异。

为了使MSE较小，应该在偏差和方差之间进行平衡（tradeoff）。显然方差和偏差平方最小都是0， 其中之一为0不足以保证MSE最小。

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

(1) 以$\bar{y}$预测y, 无偏，均方误差$\mathrm{MSE}(\bar{y}) = \mathrm{var}(\bar{y}) = \dfrac{\sigma^2}{n}$,

预测误差： $e(\bar{y}) = E(\bar{y} - y)^2 = \dfrac{\sigma^2}{n} + \sigma^2.$

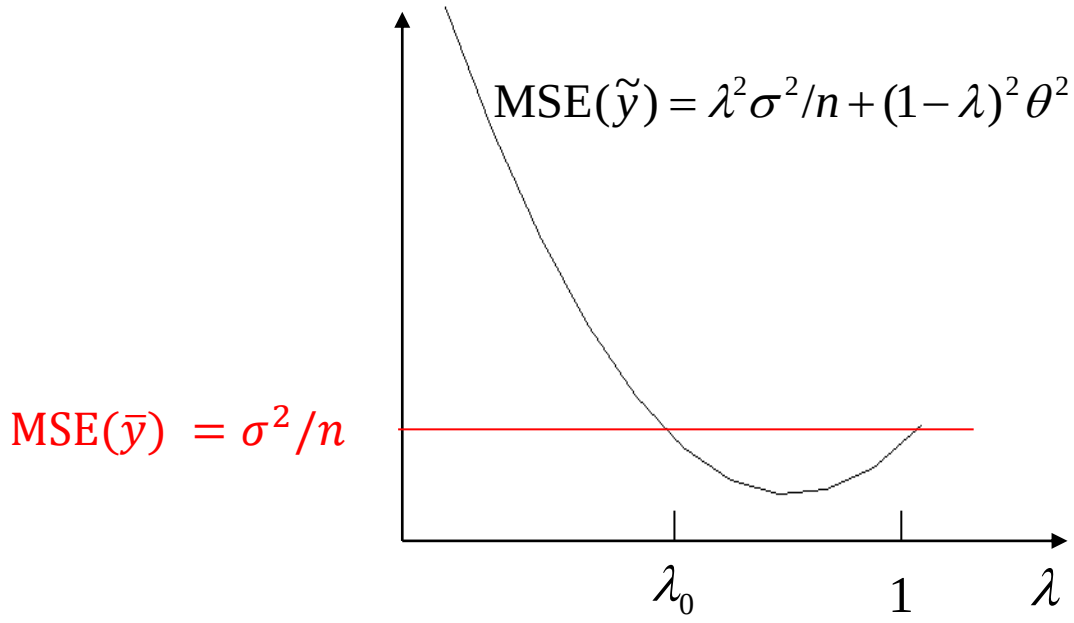
(2) 取 $\tilde{y} = \lambda \bar{y}$, $\mathrm{Bias}(\tilde{y}) = (\lambda - 1)\theta$，有偏，

$\mathrm{MSE}(\tilde{y}) = \mathrm{var}(\tilde{y}) + (\mathrm{Bias}(\tilde{y}))^2 = \lambda^2 \sigma^2 / n + (1-\lambda)^2 \theta^2$

预测误差： $e(\tilde{y}) = E(\tilde{y} - y)^2 = \mathrm{MSE}(\tilde{y}) + \sigma^2$

何时预测误差$e(\tilde{y}) = MSE(\tilde{y}) + \sigma^2 < e(\bar{y}) = \mathrm{MSE}(\bar{y}) + \sigma^2$?

即，何时$\mathrm{MSE}(\tilde{y}) < \mathrm{MSE}(\bar{y})$?



$$\lambda_0 = \frac{\theta^2 - \sigma^2/n}{\theta^2 + \sigma^2/n}$$

当$\lambda_0 < \lambda < 1$时，$MSE(\tilde{y}) < MSE(\bar{y})$

特别地，何时常数预测$\tilde{y} = 0$优于$\bar{y}$？

若$|\theta| \le \sigma/\sqrt{n}$时($\theta$较小或偏大$\sigma$)，则$MSE(\tilde{y} = 0) < MSE(\bar{y})$

经典统计与统计学习

- 经典统计在无偏的条件下最小化方差;
- 统计学习牺牲无偏性，在允许有偏的前提下尽量减少方差

（大致上，方差越小，模型越简单，方差为0时是最简单的常数预测）

适度增大预测量的偏差, 可能导致其方差大幅度下降，从而降低 MSE。

如何减小方差？最常用的也是最直观的方法是压缩(shrinkage)：

$$X \to \lambda X,\ 0 \le \lambda \le 1$$

Bayes方法、约束/惩罚*LS*方法都是压缩方法。

截断(truncation)：$X \to XI_{(|X|\le c)}$也能减小方差, 但不常用。

例如Bayes估计是一种压缩估计。

设样本$y_1, y_2, ..., y_n$ iid $\sim N(\theta, \sigma^2)$, 假设θ具有先验分布$N(\mu_0, \tau^2)$, 则后验分布

$$\theta \mid y\text{'s} \sim N\left(\frac{\tau^2 \bar{y} + \sigma^2 \mu_0}{\tau^2 + \sigma^2}, \frac{\tau^2 \sigma^2}{\tau^2 + \sigma^2} \right)$$

θ的后验估计 $\tilde{\theta}_{\text{Bayes}} = \dfrac{\bar{y}/\sigma^2 + \mu_0/\tau^2}{1/\tau^2 + 1/\sigma^2}$。

特别地当已知θ较小, $\mu_0 = 0$, $\tilde{\theta}_{\text{Bayes}} = \dfrac{\tau^2}{\tau^2 + \sigma^2} \bar{y}$

约束或惩罚LS:

设样本$y_1, y_2, ..., y_n$ iid $\sim N(\theta, \sigma^2)$, 假设已知$|\theta| < c$,

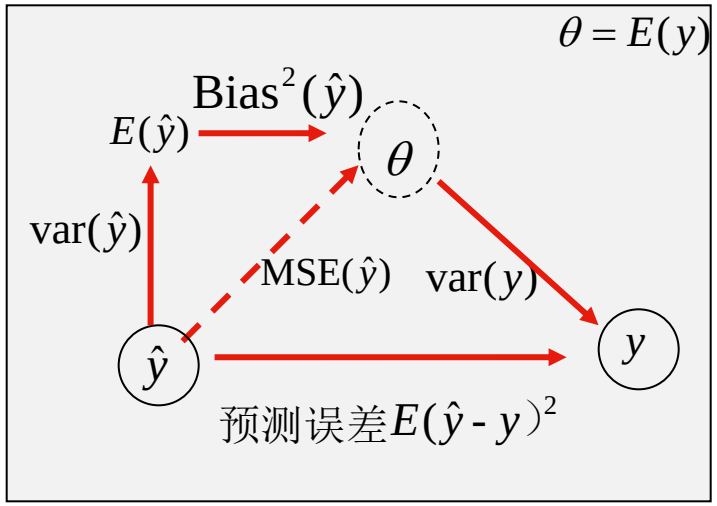
$$\min \sum (y_i - \theta)^2 = \min \sum (y_i - \bar{y})^2 + n(\bar{y} - \theta)^2, s.t. |\theta| < c,$$

最优解 $\tilde{\theta}_{\text{c}} = \begin{cases} \bar{y} & |\bar{y}| < c \\ c & \bar{y} \geq c \\ -c & \bar{y} \leq -c \end{cases}$

综上，预测误差(也称为expected generalization error) 的分解

$$\text{预测误差} = \text{方差} + \text{偏差}^2 + \text{被预测量的方差}$$

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



预测误差与均方误差：向量情形

定义（预测误差）. 以向量$\hat{\mathbf{y}}$预测$\mathbf{y}$，记$\boldsymbol{\theta} = E(\mathbf{y})$,

预测误差： $e(\hat{\mathbf{y}}) = E \| \hat{\mathbf{y}} - \mathbf{y} \|^2 = E(\hat{\mathbf{y}} - \mathbf{y})^{\mathsf{T}}(\hat{\mathbf{y}} - \mathbf{y})$

定义（均方误差）. 若参数$\boldsymbol{\theta}$、统计量$\hat{\mathbf{y}}$是向量，定义

- MSE矩阵： $M(\hat{\mathbf{y}}) = E\left((\hat{\mathbf{y}} - \boldsymbol{\theta})(\hat{\mathbf{y}} - \boldsymbol{\theta})^{\mathsf{T}}\right) = \mathrm{var}(\hat{\mathbf{y}}) + \mathbf{b}\mathbf{b}^{\mathsf{T}}$,
  其中 $\mathbf{b} = Bias(\hat{\mathbf{y}}) = E\hat{\mathbf{y}} - \boldsymbol{\theta}$.
- MSE： $m(\hat{\mathbf{y}}) = E \| \hat{\mathbf{y}} - \boldsymbol{\theta} \|^2 = \mathrm{tr}(M(\hat{\mathbf{y}})) = \mathrm{tr}\left(\mathrm{var}(\hat{\mathbf{y}})\right) + \mathbf{b}^{\mathsf{T}}\mathbf{b}$

所以，预测误差分解为

$$e(\hat{\mathbf{y}}) = E \| \hat{\mathbf{y}} - \mathbf{y} \|^2 = \mathrm{tr}M(\hat{\mathbf{y}}) + \mathrm{tr}(\mathrm{var}(\mathbf{y})) = \mathrm{tr}(\mathrm{var}(\hat{\mathbf{y}})) + \| bias(\hat{\mathbf{y}}) \|^2 + \mathrm{tr}\left(\mathrm{var}(\mathbf{y})\right)$$

例2. 假设数据$(x_i, y_i), i = 1,.., n,$ 满足模型：

$$y_i = a + bx_i + \varepsilon_i, \quad \varepsilon_i \sim (0, \sigma^2),$$

训练数据：得LS估计$\hat{a}$，$\hat{b}$。

预测目标：对于"新"数据x_0, 预测其相应的y_0, 假设(x_0, y_0)满足同样的模型：

$$y_0 = a + bx_0 + \varepsilon_0, \varepsilon_0 \sim (0, \sigma^2), \ \varepsilon_0 \text{与} \varepsilon_1, ..., \varepsilon_n \text{独立}$$

通常使用$\hat{y}_0 = \hat{a} + \hat{b}x_0$ 预测y_0， y_0的期望：

$$\theta = E(y_0) = a + bx_0,$$

$E(\hat{y}_0) = E(\hat{a} + \hat{b}x_0) = a + bx_0 = \theta,$ 其MSE：

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

$$\text{var}\begin{pmatrix}\hat{a}\\ \hat{b}\end{pmatrix} = \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} & -\frac{\bar{x}^2}{s_{xx}} \\ -\frac{\bar{x}^2}{s_{xx}} & \frac{1}{s_{xx}} \end{pmatrix}$$

预测误差：$e(\hat{y}_0) = E(\hat{y}_0 - y_0)^2 = m(\hat{y}_0) + \sigma^2 = \frac{1}{n}\sigma^2 + \frac{(x_0 - \bar{x})^2}{s_{xx}}\sigma^2 + \sigma^2$

若取b的估计为$\tilde{b} \equiv 0$，即认为模型为$y_0 = a + \varepsilon_0$,

以$\tilde{y}_0 = \bar{y}$ 预测y_0 :

$\text{bias}(\tilde{y}_0) = E(\bar{y}) - a - bx_0 = (a + b\bar{x}) - a - bx_0 = b(\bar{x} - x_0)$, 有偏

$$\text{var}(\tilde{y}_0) = \text{var}(\bar{y}) = \frac{\sigma^2}{n},$$

$$m(\tilde{y}_0) = \text{var}(\tilde{y}_0) + bias(\tilde{y}_0)^2 = \frac{\sigma^2}{n} + b^2(x_0 - \bar{x})^2$$

预测误差$e(\tilde{y}_0) = E(\tilde{y}_0 - y_0)^2 = \frac{\sigma^2}{n} + b^2(x_0 - \bar{x})^2 + \sigma^2$

当$|b| \leq \frac{\sigma}{\sqrt{s_{xx}}}$ (b 较小, σ 较大, s_{xx}较小) 时,

$m(\tilde{y}_0) \leq m(\hat{y}_0)$， 预测误差$e(\tilde{y}_0) \leq e(\hat{y}_0)$

$\tilde{y}_0$预测效果更好.

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, LASSO， 贝叶斯方法..
- Vapnik 统计学习/机器学习理论

James-Stein 估计(1956，1961)

样本$\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$ iid ~ $N(\boldsymbol{\theta}, \sigma^2 I_p)$. $p \geq 3$. 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-Stein估计：

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

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：

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