生物信息学导论作业3

马家祺

January 17, 2016

目录

1　题目一

1.1　问题a

对于多元线性回归模型$y = x^T\beta + \epsilon$，ϵ服从$N(0, \sigma^2)$的多元线性模型，说明最大似然法与最小二乘法求解时是等价的。

1.1.1　最小二乘法

由$y = x^T\beta + \epsilon$，$\epsilon\ N(0, \sigma^2)$与$\hat{y} = x^T\beta$

得残差平方和为

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - x_i^T\beta)^2 = (Y - X^T\beta)^T(Y - X^T\beta) \tag{1}$$

其中$Y = (y_1, y_2, ..., y_n)^T, X = (x_1, x_2, ..., x_n)^T$。

最小二乘法将问题转为

$$\min_\beta (Y - X^T\beta)^T(Y - X^T\beta) \tag{2}$$

令目标函数对β的导数为0，

$$\begin{aligned} \frac{\partial(Y - X^T\beta)^T(Y - X^T\beta)}{\partial\beta} &= -2X(Y - X^T\beta) \\ &= -2XY + 2XX^T\beta = 0 \end{aligned} \tag{3}$$

解得，

$$\hat{\beta} = (XX^T)^{-1}XY \tag{4}$$

1.1.2　最大似然法

由残差服从$N(0, \sigma^2)$

得似然函数

$$\begin{aligned} L(\beta) &= \prod_i P(y_i|x_i, \beta) \\ &= C_1 e^{C_2 \sum_i (y_i - x_i^T\beta)^2} \\ &= C_1 e^{-C_2 (Y - X^T\beta)^T(Y - X^T\beta)} \end{aligned} \tag{5}$$

其中$C_1 > 0, C_2 > 0$为常数。

进而得对数似然函数

$$\ln L(\beta) = -C(Y - X^T\beta)^T(Y - X^T\beta) \tag{6}$$

其中$C > 0$为一常数。

最大似然法将问题转为

$$\max_\beta -C(Y - X^T\beta)^T(Y - X^T\beta) \tag{7}$$

该问题的解显然等于最小二乘法的解。

1.2 问题b

单变量回归模型中，对于x中的某一元x_j，有$y_j = x_j\beta_j + \epsilon$。

由最小二乘法优化如下问题

$$\min_{\beta_j} \sum_i (y_{ij} - x_{ij}\beta_j)^2 \tag{8}$$

可得

$$\beta_j^* = \frac{\sum_i x_{ij} y_{ij}}{\sum_i x_{ij}^2} \tag{9}$$

对比前面多元线性回归的解$\hat{\beta} = (XX^T)^{-1}XY$，可以看出，多元线性回归中，$\hat{\beta}$中的某一元$\hat{\beta}_j$是与数据的所有维元素相关的；而在单变量回归中，$\beta^*$中的每一元$\beta_j^*$仅和数据中的第$j$维元素有关。