第十章（特征选择与稀疏学习）作业

10.1 在本题中，我们将通过理论计算推导线性回归模型的偏差与方差。考虑如下线性回归模型：

$$y_i = \boldsymbol{x_i^T}\boldsymbol{\beta}^* + \epsilon_i, \qquad i = 1, \dots, n$$

其中$\boldsymbol{x_1}, \dots, \boldsymbol{x_n} \in \mathbb{R}^p$是确定的样本值，$\boldsymbol{\beta}^* \in \mathbb{R}^p$是未知的系数向量，$\epsilon_1, \dots, \epsilon_n$是误差，它们独立同$\mathcal{N}(0, \sigma^2)$分布。记$\boldsymbol{x} = \left(\boldsymbol{x_1^T}; \dots; \boldsymbol{x_n^T}\right) \in \mathbb{R}^{n \times p}$，$x$的每一列互相独立，$\boldsymbol{y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$为输出向量。利用最小二乘法可拟合得到系数的估计：

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x^T x})^{-1} \boldsymbol{x^T y}$$

同时，记模型在任意样本点$\boldsymbol{x_0}$处的回归输出为$\hat{f}(\boldsymbol{x_0}) = \boldsymbol{x_0^T}\widehat{\boldsymbol{\beta}}$。

（1） 回忆模型偏差的定义，在任意样本点$\boldsymbol{x_0}$处，模型的偏差为：

$$Bias\left(\hat{f}(\boldsymbol{x_0})\right) = \mathbb{E}\left[\hat{f}(\boldsymbol{x_0})\right] - y(\boldsymbol{x_0})$$

其中$y(\boldsymbol{x_0}) = \boldsymbol{x_0^T}\boldsymbol{\beta}^*$为真实回归值。请证明$Bias\left(\hat{f}(\boldsymbol{x_0})\right) = 0$，从而所有样本线性回归的平均偏差也为 0，即：

$$\frac{1}{n}\sum_{i=1}^{n} Bias\left(\hat{f}(\boldsymbol{x_i})\right) = 0$$

（2） 现在我们考虑回归输出的方差。试证明：

$$\frac{1}{n}\sum_{i=1}^{n} Var\left(\hat{f}(\boldsymbol{x_i})\right) = \frac{\sigma^2 p}{n}$$

（3） 根据课上所讲均方误差与偏差和方差的关系，计算模型的期望测试误差$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i' - \hat{f}(\boldsymbol{x_i})\right)^2\right]$。

其中，$y_1', \dots, y_n'$是独立的测试集。

提示：

（2）中将等式左边化为等价的矩阵形式，其中包括某个$n \times n$矩阵的迹，可以简化计算。

答案：

（1） 依题意：

$$Bias\left(\hat{r}(\boldsymbol{x_0})\right) = \mathbb{E}[\hat{r}(\boldsymbol{x_0})] - r(\boldsymbol{x_0}) = \mathbb{E}\left(\boldsymbol{x_0^T}\widehat{\boldsymbol{\beta}}\right) - \boldsymbol{x_0^T}\boldsymbol{\beta}^* = \boldsymbol{x_0^T}\left(\mathbb{E}\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)$$

而：

$$\mathbb{E}\widehat{\boldsymbol{\beta}} = \mathbb{E}((\boldsymbol{x^T x})^{-1}\boldsymbol{x^T y}) = (\boldsymbol{x^T x})^{-1}\boldsymbol{x^T}\mathbb{E}\boldsymbol{y}$$

由于$\mathbb{E}y_i = \boldsymbol{x_i^T}\boldsymbol{\beta}^*$，从而$\mathbb{E}\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta}^*$，代入上式，得到：

$$\mathbb{E}\widehat{\boldsymbol{\beta}} = (\boldsymbol{x^T x})^{-1}\boldsymbol{x^T x}\boldsymbol{\beta}^* = \boldsymbol{\beta}^*$$

从而$Bias\big(\hat{r}(\boldsymbol{x_0})\big)=0$。

（2） 依题意：

$$\frac{1}{n}\sum_{i=1}^{n}Var\big(\hat{r}(\boldsymbol{x_i})\big)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\hat{r}(\boldsymbol{x_i})^2)-\mathbb{E}\big(\hat{r}(\boldsymbol{x_i})\big)^2$$

由（1）知，$\mathbb{E}\big(\hat{r}(\boldsymbol{x_i})\big)=\boldsymbol{x_i^T\beta^*}$，将上式化为矩阵形式：

$$\sum_{i=1}^{n}\mathbb{E}(\hat{r}(\boldsymbol{x_i})^2)=\sum_{i=1}^{n}\mathbb{E}\big(\boldsymbol{x_i^T\widehat{\beta}\widehat{\beta}^T x_i}\big)=\mathbb{E}\left(tr\big(\boldsymbol{x\widehat{\beta}\widehat{\beta}^T x^T}\big)\right)=tr\big(\boldsymbol{x}\mathbb{E}(\boldsymbol{\widehat{\beta}\widehat{\beta}^T})\boldsymbol{x^T}\big)$$

又：

$$\begin{aligned}\boldsymbol{x^T}\mathbb{E}\big(\boldsymbol{\widehat{\beta}\widehat{\beta}^T}\big)\boldsymbol{x}&=\boldsymbol{x}\mathbb{E}[(\boldsymbol{x^T x})^{-1}\boldsymbol{x^T y y^T x}(\boldsymbol{x^T x})^{-1}]\boldsymbol{x^T}\\&=\boldsymbol{x}(\boldsymbol{x^T x})^{-1}\boldsymbol{x^T}\mathbb{E}(\boldsymbol{yy^T})\boldsymbol{x}(\boldsymbol{x^T x})^{-1}\boldsymbol{x^T}\end{aligned}$$

所以$\sum_{i=1}^{n}\mathbb{E}(\hat{r}(\boldsymbol{x_i})^2)=tr(\boldsymbol{x}(\boldsymbol{x^T x})^{-1}\boldsymbol{x^T}\mathbb{E}(\boldsymbol{yy^T})\boldsymbol{x}(\boldsymbol{x^T x})^{-1}\boldsymbol{x^T})$。

同样的：

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}\big(\hat{r}(\boldsymbol{x_i})\big)^2&=\sum_{i=1}^{n}\boldsymbol{x_i^T\beta^*\beta^{*T}x_i}=tr\big(\boldsymbol{x\beta^*\beta^{*T}x^T}\big)\\&=tr(\boldsymbol{x}(\boldsymbol{x^T x})^{-1}\boldsymbol{x^T}\mathbb{E}\boldsymbol{y}*\mathbb{E}\boldsymbol{y^T x}(\boldsymbol{x^T x})^{-1}\boldsymbol{x^T})\end{aligned}$$

于是：

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}Var\big(\hat{r}(\boldsymbol{x_i})\big)\\&=\frac{1}{n}\left[tr\left(\boldsymbol{x}(\boldsymbol{x^T x})^{-1}\boldsymbol{x^T}\Big(\mathbb{E}(\boldsymbol{yy^T})-(\mathbb{E}\boldsymbol{y}*\mathbb{E}\boldsymbol{y^T})\Big)\boldsymbol{x}(\boldsymbol{x^T x})^{-1}\boldsymbol{x^T}\right)\right]\\&=\frac{1}{n}tr(\sigma^2\boldsymbol{x}(\boldsymbol{x^T x})^{-1}\boldsymbol{x^T x}(\boldsymbol{x^T x})^{-1}\boldsymbol{x^T})=\frac{\sigma^2}{n}tr(\boldsymbol{x}(\boldsymbol{x^T x})^{-1}\boldsymbol{x^T})\\&=\frac{p\sigma^2}{n}\end{aligned}$$

（3） 由课件公式：

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\big(y_i'-\hat{r}(\boldsymbol{x_i})\big)^2\right]\\&=\frac{1}{n}\sum_{i=1}^{n}Bias\big(\hat{r}(\boldsymbol{x_i})\big)+\frac{1}{n}\sum_{i=1}^{n}Var\big(\hat{r}(\boldsymbol{x_i})\big)+\frac{1}{n}\sum_{i=1}^{n}Var(\epsilon_i)\\&=0+\frac{p\sigma^2}{n}+\sigma^2=\sigma^2(\frac{p}{n}+1)\end{aligned}$$

10.2 计算机小实验 1：线性回归、岭回归和 LASSO 回归

请编写代码生成以下仿真数据，探索线性回归、岭回归和 LASSO 模型对共线性问题的表现。

$$y = 3x_1 + 2 + \varepsilon_1, x_1 = 1, \dots, 20$$
$$x_2 = 0.05x_1 + \varepsilon_2$$
$$\varepsilon_1 \in \mathrm{N}(0, 2.5), \varepsilon_2 \in \mathrm{N}(0, 0.5)$$

若我们将与x_1有强相关关系的噪声x_2误认为是一维特征（即输入特征变为了$[x_1, x_2]$），请同学们尝试使用上述三种模对y进行回归，并回答以下问题。

(1) 请给出x_1, x_2的相关系数。

(2) 请多次生成数据，观察正则化系数为 1 情况下三种模型拟合参数的稳定性。

答案：

(1) 多次实验平均，0.38 左右

(2) 由于共线性变量的存在，普通的线性回归出现了不稳定，x_1和x_2的系数都会波动，x_2的波动范围更大

在正则化系数为 1 情况下， 岭回归也有一定波动，但是x_1的波动较小。

在正则化系数为 1 情况下， LASSO 非常稳定，x_2的系数直接为 0。

10.3 计算机小实验 2：特征选择

附件 feature_selection_X.txt 中给出了 400 个组织样本数据，每一行是一维样本，每一列代表一维特征，feature_selection_Y.txt 中给出了样本对应的标签（1 代表肿瘤组织，0 代表正常组织）。请随机抽取 300 个样本作为训练集，100 个样本作为测试集。使用特征选择算法，挑选出区分不同组织的特征，利用分类器进行分类：

(1) 分别用**类内类间距离和最大信息系数(互信息的另一种度量方式)的判据**选择 1, 5, 10, 20, 50, 100 个特征，用 Logistic 回归进行分类，并比较与不做特征选择时候的模型预测效果；除此之外，请比较两种方法在这些特征个数时挑选出的特征子集有多少特征是相同的；

(2) 请简述前向算法的流程，使用**前向算法**进行特征选择,并使用逻辑回归作为分类器。并比较与（1）中选出特征的异同。

(3) 决策树算法在学习过程中会自动选择特征。请使用决策树对数据进行分类，并观察比较决策树中用到的特征与（1）和（2）中选出的特征的重合程度。

提示：

最大信息系数的计算可以调用 minepy 库中的 MINE 模块；逻辑回归和决策树可以调用 sklearn 库。

答案：

(1)

特征个数	1	5	10	20	50	100	未做筛选
类内类间距离	0.84	0.95	0.93	0.92	0.93	0.89	0.89
最大信息系数	0.84	0.95	0.94	0.89	0.9	0.88	
相同个数	1	5	5	5	9	18	

由于直接应用互信息进行特征选择效果不是很好，使用将互信息转换成一种度量方式的最大信息系数作为判据。

从结果上可以看出，未做特征筛选时，分类准确率为 0.89，通过特征的筛选，分类的准确率有了一定的提高，基本可以在 0.9 以上。除此之外，两种不同的判据在较少特征个数时相同的特征较多，当个数增多时，相同的特征并没有随着个数增多而增高，可能是由于数据集中很多特征具有一定的共线性，因此使用不同的特征组合也可以实现较高的分类准确率。

(2)

由于最优特征的组合搜索运行时间比较长，目前没有对最优特征进行搜索，只是依次叠特征，使用前 151 维特征效果最好，分类准确率为 0.78 左右。在类内类间距离找到的 100 个特征中，有 20 个是在前 150 维的；在 MIC 找到的 100 个特征中，有 13 个是在前 150 维的；

(3)

使用决策树进行数据分类，准确率为 0.9，使用到的特征有 8 个，索引为 4，47,48，74，219,354，677,916，(1)中与之最相近的特征个数为 10，在所挑选的 10 个特征中，类内类间距离挑选出的特征索引为 47,916,219,415,4,427,299，467,91,294，最大信息系数挑选出的特征索引为 47,916,4,219,415,747,334,828,787,835。可以看出，特征集合的重复性较高，三种方法都挑选出了 4,47,219,916 这 4 个位置的特征，说明这 4 个特征对于分类是很重要的。