

机器学习导论

习题一

191250026, 丁云翔, 191250026@smail.nju.edu.cn

2021 年 4 月 2 日

学术诚信

本课程非常重视学术诚信规范，助教老师和助教同学将不遗余力地维护作业中的学术诚信规范的建立。希望所有选课学生能够对此予以重视。¹

- (1) 允许同学之间的相互讨论，但是**署你名字的工作必须由你完成**，不允许直接照搬任何已有的材料，必须独立完成作业的书写过程；
- (2) 在完成作业过程中，对他人工作（出版物、互联网资料）中文本的直接照搬（包括原文的直接复制粘贴及语句的简单修改等）都将视为剽窃，剽窃者成绩将被取消。**对于完成作业中有关键作用的公开资料，应予以明显引用；**
- (3) 如果发现作业之间高度相似将被判定为互相抄袭行为，**抄袭和被抄袭双方的成绩都将被取消**。因此请主动防止自己的作业被他人抄袭。

作业提交注意事项

- (1) 请在L^AT_EX模板中**第一页填写个人的姓名、学号、邮箱信息**；
- (2) 本次作业需提交该pdf文件、问题1,3可直接运行的源码(LinearRegression.py, PR.py, ROC.py, **不需要提交数据集**)，将以上三个文件压缩成zip文件后上传。zip文件格式为**学号.zip**，例如190000001.zip；pdf文件格式为**学号_姓名.pdf**，例如190000001_张三.pdf，**并通过教学立方提交**。
- (3) 未按照要求提交作业，或提交作业格式不正确，将会**被扣除部分作业分数**；
- (4) 本次作业提交截止时间为**4月2日23:55:00**。

¹参考尹一通老师高级算法课程中对学术诚信的说明。

1 [45pts] Linear Regression with a Regularization Term

给定数据集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$, 其中 $\mathbf{x}_i = (x_{i1}; x_{i2}; \dots; x_{id}) \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, 当我们采用线性回归模型求解时, 实际上是在求解下述优化问题:

$$\hat{\mathbf{w}}_{\text{LS}}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{X}\mathbf{w} + \mathbf{1}\mathbf{b}^\top - \mathbf{y}\|_2^2, \quad (1)$$

其中 $\mathbf{y} = [y_1, \dots, y_m]^\top \in \mathbb{R}^m$, $\mathbf{X} = [\mathbf{x}_1^\top; \mathbf{x}_2^\top; \dots; \mathbf{x}_m^\top] \in \mathbb{R}^{m \times d}$, $\mathbf{1}$ 为全1向量, 其维度可由其他元素推导而得。在实际问题中, 我们常常不会直接利用线性回归对数据进行拟合, 这是因为当样本特征很多, 而样本数相对较少时, 直接线性回归很容易陷入过拟合。为缓解过拟合问题, 常对公式(1)引入正则化项, 通常形式如下:

$$\hat{\mathbf{w}}_{\text{reg}}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{X}\mathbf{w} + \mathbf{1}\mathbf{b}^\top - \mathbf{y}\|_2^2 + \lambda \Omega(\mathbf{w}), \quad (2)$$

其中, $\lambda > 0$ 为正则化参数, $\Omega(\mathbf{w})$ 是正则化项, 根据模型偏好选择不同的 Ω 。

下面, 假设样本特征矩阵 \mathbf{X} 满足列满秩, 请回答下面的问题:

(1) [5pts] 考虑线性回归问题, 即对应于公式(1), 请给出最优解 $\hat{\mathbf{w}}_{\text{LS}}^*$ 和 \mathbf{b} 的闭式解表达式, 请使用矩阵形式表示;

(2) [10pts] 考虑岭回归(ridge regression)问题, 即对应于公式(2)中 $\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2$ 时, 请给出最优解 $\hat{\mathbf{w}}_{\text{Ridge}}^*$ 和 \mathbf{b} 的闭式解表达式, 请使用矩阵形式表示;

(3) [15pts] 请编程实现以上两种线性回归模型, 基于你求出的闭式解在训练集上构建模型。并汇报测试集上的 Mean Square Error (MSE)。

建议使用python语言实现, 本次采用波士顿房价预测数据, 数据集的获取依赖sklearn库, 你可以查阅相关资料进行安装。请参考作业中提供的LinearRegression.py进行模型的构造, 代码中已经完成了训练集和测试集的划分。对于线性回归模型, 你需要汇报测试集上的MSE, 对于岭回归问题, 你需要自行设置正则项 λ 的取值范围, 并观察训练集MSE, 测试集MSE和 λ 的取值的关系, 你有什么发现?

请注意, 除了示例代码中使用到的sklearn库函数以外, 你将不能使用其他的sklearn函数, 你需要基于numpy实现线性回归模型和MSE的计算。

(4) [5pts] 如果推广到分类问题, 应该如何设置 \mathbf{y} , 请谈谈你的看法;

(5) [10pts] 请证明对于任何矩阵 \mathbf{X} , 下式均成立

$$(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1} \mathbf{X} = \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda\mathbf{I})^{-1} \quad (3)$$

请思考, 上述的结论可以用在线性回归问题的什么情况中, 能带来怎样的帮助?

提示1: 你可以参考 The Matrix Cookbook 获取矩阵求导的一些知识。

Solution. 此处用于写证明(中英文均可)

(1) 对于

$$\hat{\mathbf{w}}_{\text{LS}}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{X}\mathbf{w} + \mathbf{1}\mathbf{b}^\top - \mathbf{y}\|_2^2,$$

令 $\mathbf{E}_{\mathbf{w}, \mathbf{b}} = \frac{1}{2} \|\mathbf{X}\mathbf{w} + \mathbf{1}\mathbf{b}^\top - \mathbf{y}\|_2^2$, 又 $\|\mathbf{X}\mathbf{w} + \mathbf{1}\mathbf{b}^\top - \mathbf{y}\|_2^2 = (\mathbf{X}\mathbf{w} + \mathbf{1}\mathbf{b}^\top - \mathbf{y})^\top (\mathbf{X}\mathbf{w} + \mathbf{1}\mathbf{b}^\top - \mathbf{y})$, 化简得:

$$\mathbf{E}_{\mathbf{w}, \mathbf{b}} = \frac{1}{2} (\mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w} + 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{1}\mathbf{b}^\top - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{b}^\top \mathbf{1}\mathbf{b}^\top - 2\mathbf{y}^\top \mathbf{1}\mathbf{b}^\top + \mathbf{y}^\top \mathbf{y}),$$

对 $\hat{\mathbf{w}}$ 求偏导得:

$$\frac{\partial \mathbf{E}_{\hat{\mathbf{w}}, \mathbf{b}}}{\partial \hat{\mathbf{w}}} = \mathbf{X}^\top \mathbf{X} \mathbf{w} + \mathbf{X}^\top \mathbf{1} \mathbf{b}^\top - \mathbf{X}^\top \mathbf{y} \quad (4)$$

$$\frac{\partial \mathbf{E}_{\hat{\mathbf{w}}, \mathbf{b}}}{\partial \mathbf{b}} = \mathbf{w}^\top \mathbf{X}^\top \mathbf{1} + \mathbf{b} \mathbf{1}^\top \mathbf{1} - \mathbf{y}^\top \mathbf{1} \quad (5)$$

然后令(4)和(5)式为0可得到如下方程:

$$\mathbf{X}^\top \mathbf{X} \mathbf{w} = \mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{1} \mathbf{b}^\top \quad (6)$$

$$\mathbf{w}^\top \mathbf{X}^\top \mathbf{1} = \mathbf{y}^\top \mathbf{1} - \mathbf{b} \mathbf{1}^\top \mathbf{1} \quad (7)$$

联立(6)(7)两个方程可得:

$$\mathbf{w} = (\mathbf{X}^\top \mathbf{X} - \frac{1}{\mathbf{m}} \mathbf{X}^\top \mathbf{1} \cdot \mathbf{1}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{y} - \frac{1}{\mathbf{m}} \mathbf{X}^\top \mathbf{1} \cdot \mathbf{1}^\top \mathbf{y}) \quad (8)$$

$$\mathbf{b} = \frac{1}{\mathbf{m}} (\mathbf{y}^\top \mathbf{1} - (\mathbf{y}^\top \mathbf{X} - \frac{1}{\mathbf{m}} \mathbf{y}^\top \mathbf{1} \cdot \mathbf{1}^\top \mathbf{X}) (\mathbf{X}^\top \mathbf{X} - \frac{1}{\mathbf{m}} \mathbf{X}^\top \mathbf{1} \cdot \mathbf{1}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{1}) \quad (9)$$

由(8)和(9)的解可得: 最优解 $\hat{\mathbf{w}}_{\text{LS}}^*$ 和 \mathbf{b} 的闭式解表达式:

$$\hat{\mathbf{w}}_{\text{LS}}^* = (\mathbf{X}^\top \mathbf{X} - \frac{1}{\mathbf{m}} \mathbf{X}^\top \mathbf{1} \cdot \mathbf{1}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{y} - \frac{1}{\mathbf{m}} \mathbf{X}^\top \mathbf{1} \cdot \mathbf{1}^\top \mathbf{y})$$

$$\mathbf{b} = \frac{1}{\mathbf{m}} (\mathbf{y}^\top \mathbf{1} - (\mathbf{y}^\top \mathbf{X} - \frac{1}{\mathbf{m}} \mathbf{y}^\top \mathbf{1} \cdot \mathbf{1}^\top \mathbf{X}) (\mathbf{X}^\top \mathbf{X} - \frac{1}{\mathbf{m}} \mathbf{X}^\top \mathbf{1} \cdot \mathbf{1}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{1})$$

(2) 对于

$$\hat{\mathbf{w}}_{\text{reg}}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{X} \mathbf{w} + \mathbf{1} \mathbf{b}^\top - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2,$$

令 $\mathbf{E}_{\hat{\mathbf{w}}, \mathbf{b}} = \frac{1}{2} \|\mathbf{X} \mathbf{w} + \mathbf{1} \mathbf{b}^\top - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$, 又 $\|\mathbf{X} \mathbf{w} + \mathbf{1} \mathbf{b}^\top - \mathbf{y}\|_2^2 = (\mathbf{X} \mathbf{w} + \mathbf{1} \mathbf{b}^\top - \mathbf{y})^\top (\mathbf{X} \mathbf{w} + \mathbf{1} \mathbf{b}^\top - \mathbf{y})$, $\|\mathbf{w}\|_2^2 = \mathbf{w}^\top \mathbf{w}$ 化简得:

$$\mathbf{E}_{\hat{\mathbf{w}}, \mathbf{b}} = \frac{1}{2} (\mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w} + 2 \mathbf{w}^\top \mathbf{X}^\top \mathbf{1} \mathbf{b}^\top - 2 \mathbf{w}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{b} \mathbf{1}^\top \mathbf{1} \mathbf{b}^\top - 2 \mathbf{y}^\top \mathbf{1} \mathbf{b}^\top + \mathbf{y}^\top \mathbf{y}) + \lambda \mathbf{w}^\top \mathbf{w},$$

对 $\hat{\mathbf{w}}$ 求偏导得:

$$\frac{\partial \mathbf{E}_{\hat{\mathbf{w}}, \mathbf{b}}}{\partial \hat{\mathbf{w}}} = \mathbf{X}^\top \mathbf{X} \mathbf{w} + \mathbf{X}^\top \mathbf{1} \mathbf{b}^\top - \mathbf{X}^\top \mathbf{y} + 2 \lambda \mathbf{w} \quad (10)$$

$$\frac{\partial \mathbf{E}_{\hat{\mathbf{w}}, \mathbf{b}}}{\partial \mathbf{b}} = \mathbf{w}^\top \mathbf{X}^\top \mathbf{1} + \mathbf{b} \mathbf{1}^\top \mathbf{1} - \mathbf{y}^\top \mathbf{1} \quad (11)$$

然后令(10)和(11)式为0可得到如下方程:

$$(\mathbf{X}^\top \mathbf{X} + 2 \lambda \mathbf{I}) \mathbf{w} = \mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{1} \mathbf{b}^\top \quad (12)$$

$$\mathbf{w}^\top \mathbf{X}^\top \mathbf{1} = \mathbf{y}^\top \mathbf{1} - \mathbf{b} \mathbf{1}^\top \mathbf{1} \quad (13)$$

联立(12)(13)两个方程可得:

$$\mathbf{w} = (\mathbf{X}^\top \mathbf{X} + 2 \lambda \mathbf{I} - \frac{1}{\mathbf{m}} \mathbf{X}^\top \mathbf{1} \cdot \mathbf{1}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{y} - \frac{1}{\mathbf{m}} \mathbf{X}^\top \mathbf{1} \cdot \mathbf{1}^\top \mathbf{y}) \quad (14)$$

$$\mathbf{b} = \frac{1}{\mathbf{m}} (\mathbf{y}^\top \mathbf{1} - (\mathbf{y}^\top \mathbf{X} - \frac{1}{\mathbf{m}} \mathbf{y}^\top \mathbf{1} \cdot \mathbf{1}^\top \mathbf{X}) (\mathbf{X}^\top \mathbf{X} + 2 \lambda \mathbf{I} - \frac{1}{\mathbf{m}} \mathbf{X}^\top \mathbf{1} \cdot \mathbf{1}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{1}) \quad (15)$$

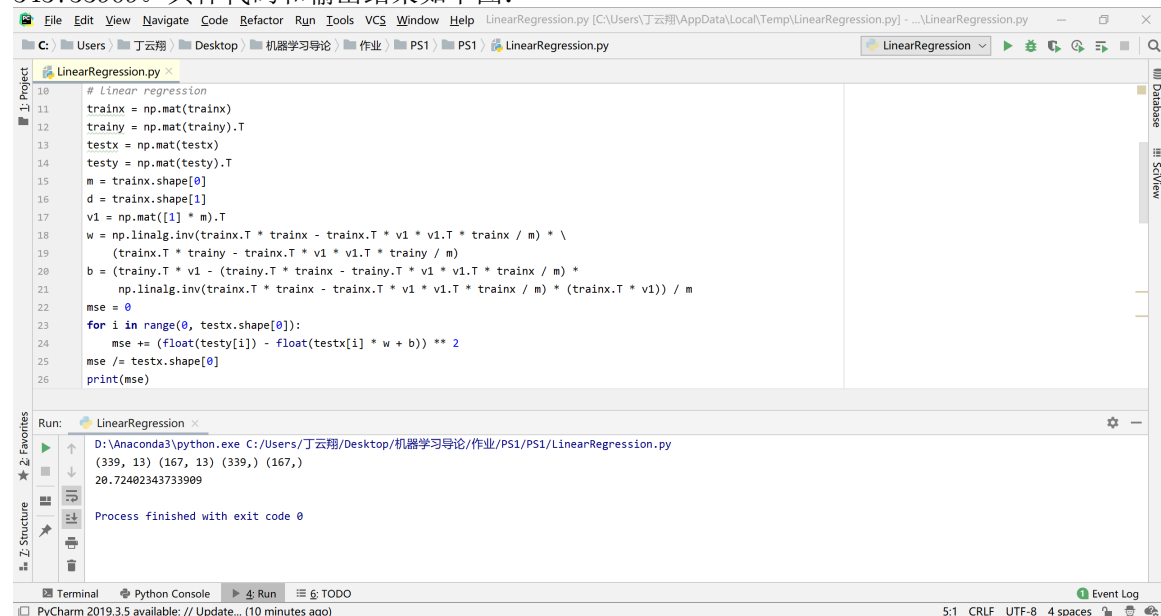
由(14)和(15)的解可得：最优解 $\hat{\mathbf{w}}_{\text{Ridge}}^*$ 和 \mathbf{b} 的闭式解表达式：

$$\hat{\mathbf{w}}_{\text{Ridge}}^* = (\mathbf{X}^T \mathbf{X} + 2\lambda \mathbf{I} - \frac{1}{m} \mathbf{X}^T \mathbf{1} \cdot \mathbf{1}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y} - \frac{1}{m} \mathbf{X}^T \mathbf{1} \cdot \mathbf{1}^T \mathbf{y})$$

$$\mathbf{b} = \frac{1}{m} (\mathbf{y}^T \mathbf{1} - (\mathbf{y}^T \mathbf{X} - \frac{1}{m} \mathbf{y}^T \mathbf{1} \cdot \mathbf{1}^T \mathbf{X}) (\mathbf{X}^T \mathbf{X} + 2\lambda \mathbf{I} - \frac{1}{m} \mathbf{X}^T \mathbf{1} \cdot \mathbf{1}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{1})$$

(3)

对于线性回归模型，利用训练集和第(1)问中的闭式解，求得 \mathbf{w} 与 \mathbf{b} ，对应模型的MSE为20.72402343733909。具体代码和输出结果如下图：



```

10 # Linear regression
11 trainx = np.mat(trainx)
12 trainy = np.mat(trainy).T
13 testx = np.mat(testx)
14 testy = np.mat(testy).T
15 m = trainx.shape[0]
16 d = trainx.shape[1]
17 v1 = np.mat([1] * m).T
18 w = np.linalg.inv(trainx.T * trainx - trainx.T * v1 * v1.T * trainx / m) * \
19     (trainx.T * trainy - trainx.T * v1 * v1.T * trainy / m)
20 b = (trainy.T * v1 - (trainy.T * trainx - trainy.T * v1 * v1.T * trainx / m) *
21     np.linalg.inv(trainx.T * trainx - trainx.T * v1 * v1.T * trainx / m) * (trainx.T * v1)) / m
22 mse = 0
23 for i in range(0, testx.shape[0]):
24     mse += (float(testy[i]) - float(testx[i] * w + b)) ** 2
25 mse /= testx.shape[0]
26 print(mse)

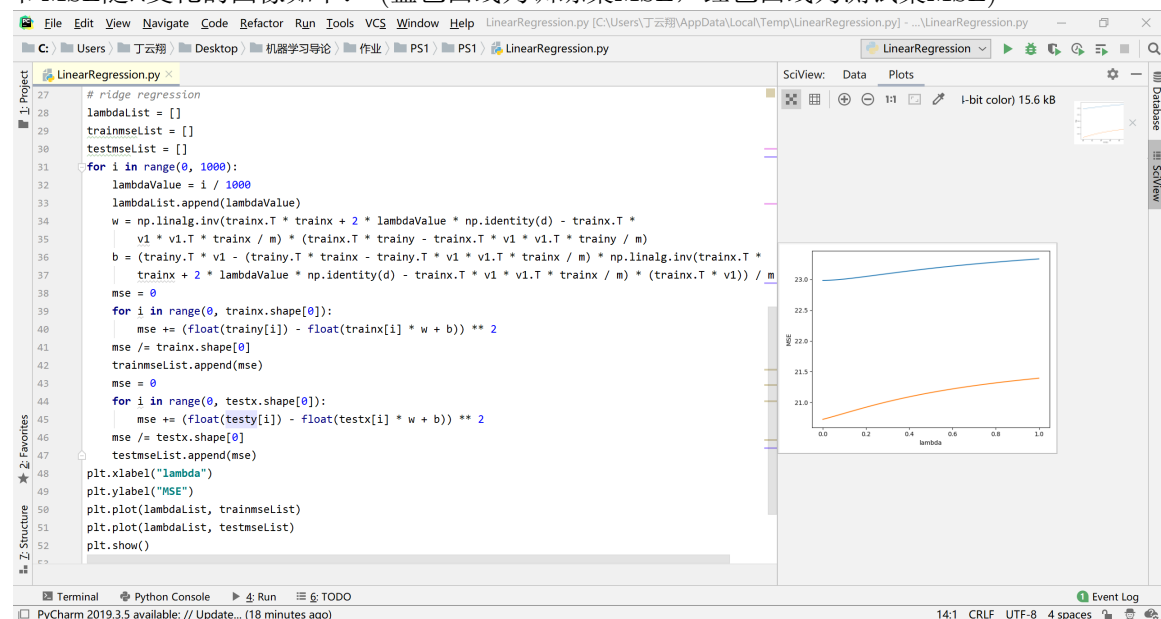
```

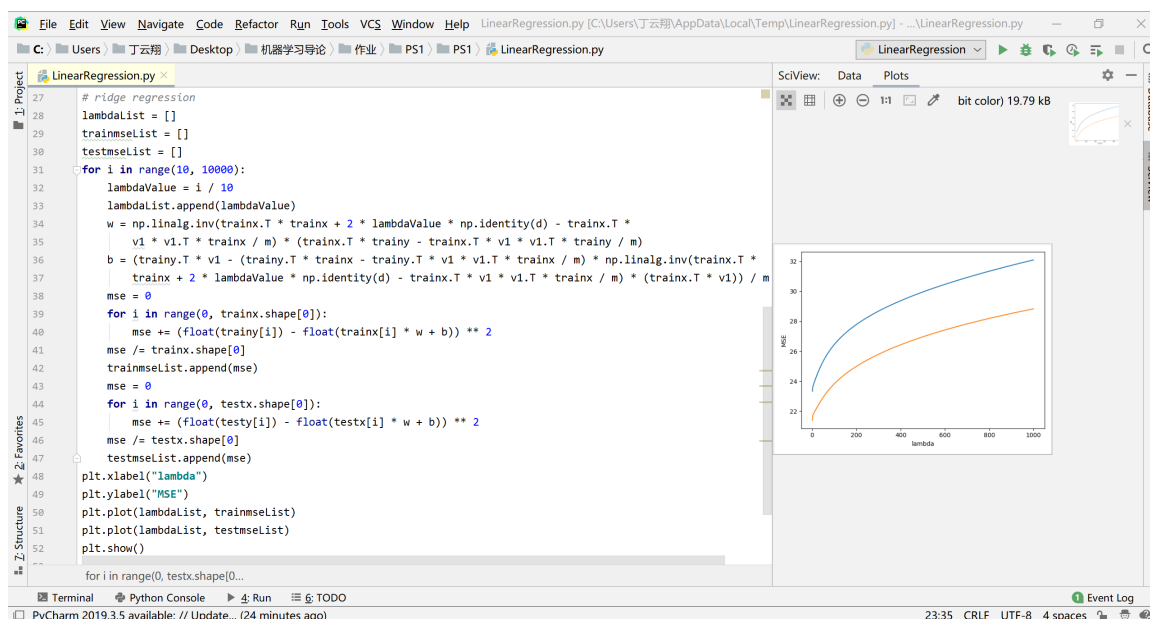
Run: LinearRegression

D:\Anaconda3\python.exe C:/Users/丁云翔/Desktop/机器学习导论/作业/PS1/PS1/LinearRegression.py
 (339, 13) (167, 13) (339,) (167,)
 20.72402343733909

Process finished with exit code 0

对于岭回归模型，分别在 $[0,1]$ 和 $[1,1000]$ 中对 λ 进行取值、训练对应的模型，并计算模型在训练集和测试集上的MSE，可以看出随着 λ 值的增大，训练集MSE和测试集MSE均增大。对应代码和MSE随 λ 变化的图像如下：（蓝色曲线为训练集MSE，红色曲线为测试集MSE）





(4)

对于分类问题，我认为应该从以下几种情况来设置：对二分类任务，可以将两种类别转化为特定值，如0和1，然后可以依次以数据集中样本的类别属性对应的数值作为各分量组成一个向量 y 。对于多分类问题，若不同类别属性间存在序的关系，则可以按序关系将其转换为特定的数值，如在 $[0,1]$ 间按类别的序取值作为类别属性的值，然后可以依次以数据集中样本的类别属性对应的数值作为各分量组成一个向量 y 。若不同类别属性间不存在序关系，则每个样本的类别属性都可以用一个 k 维向量表示，所以 y 可以设置为数据集中各样本的类别属性转换成向量后依次组成的矩阵。

(5)

对于等式两侧的表达式 $(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{X}$ 和 $\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}$ ，均分别左乘 $(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})$ 和右乘 $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})$ ，即分别计算

$$(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})$$

和

$$(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})$$

化简可得两式的结果均为

$$\mathbf{X}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\mathbf{X}$$

故等式左右两边相等，原等式对任何矩阵 \mathbf{X} 均成立。

此结论可以应用于线性回归中求解 w 与 b 中，若求解的表达式中出现形如等式左边或等式右边的表达式时，可以进行互相转换以提高运算效率。因为 $\mathbf{X}\mathbf{X}^\top$ 为一个 $m \times m$ 维矩阵，而 $\mathbf{X}^\top\mathbf{X}$ 为一个 $d \times d$ 维的矩阵，同理， $(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}$ 为一个 $m \times m$ 维矩阵，而 $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}$ 为一个 $d \times d$ 维的矩阵。在 m 大于 d 时，可将 $(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{X}$ 转化为 $\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}$ 来计算；同理，当 m 小于 d 时，可将 $\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}$ 转化为 $(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{X}$ 来计算，这样可以减少运算量，提高运算速度。当 m 与 d 相差较大时，这样的转换带来的性能提升会很明显。

2 [25+5pts] Multi-Class Logistic Regression

教材的章节3.3介绍了对数几率回归解决二分类问题的具体做法。假定现在的任务不再是二分类问题，而是多分类问题，其中 $y \in \{1, 2, \dots, K\}$ 。请将对数几率回归算法拓展到该多分类问题。

- (1) [15pts] 给出该对率回归模型的“对数似然”(log-likelihood);
- (2) [10pts] 请仿照课本公式3.30，计算该“对数似然”的梯度;
- (3) [Bonus 5pts] 对于样本类别分布不平衡的问题，基于以上的推导会出现怎样的问题，应该进行怎样的应对？谈谈你的看法。

提示1：假设该多分类问题满足如下 $K-1$ 个对数几率，

$$\begin{aligned} \ln \frac{p(y=1|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_1^T \mathbf{x} + b_1 \\ \ln \frac{p(y=2|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_2^T \mathbf{x} + b_2 \\ &\dots \\ \ln \frac{p(y=K-1|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_{K-1}^T \mathbf{x} + b_{K-1} \end{aligned}$$

提示2：定义指示函数 $\mathbb{I}(\cdot)$,

$$\mathbb{I}(y=j) = \begin{cases} 1 & \text{若 } y \text{ 等于 } j \\ 0 & \text{若 } y \text{ 不等于 } j \end{cases}$$

Solution. 此处用于写解答(中英文均可)

(1)

由西瓜书公式(3.22)到(3.23)和(3.24)的变换和提示1中的多分类问题对数几率可得：

$$\begin{aligned} p(y=1|x) &= \frac{e^{\mathbf{w}_1^T \mathbf{x} + b_1}}{1 + \sum_{j=1}^{K-1} e^{\mathbf{w}_j^T \mathbf{x} + b_j}} \\ p(y=2|x) &= \frac{e^{\mathbf{w}_2^T \mathbf{x} + b_2}}{1 + \sum_{j=1}^{K-1} e^{\mathbf{w}_j^T \mathbf{x} + b_j}} \\ &\dots \\ p(y=K-1|x) &= \frac{e^{\mathbf{w}_{K-1}^T \mathbf{x} + b_{K-1}}}{1 + \sum_{j=1}^{K-1} e^{\mathbf{w}_j^T \mathbf{x} + b_j}} \\ p(y=K|x) &= \frac{1}{1 + \sum_{j=1}^{K-1} e^{\mathbf{w}_j^T \mathbf{x} + b_j}} \end{aligned}$$

则该多分类问题的对率回归模型的“对数似然”可以写为：

$$\mathbf{L}(w, b) = \sum_{i=1}^m \ln p(\mathbf{y}_i | \mathbf{x}_i; w, b) \quad (16)$$

同样的，为便于讨论，令 $\beta_j = (\mathbf{w}_j; \mathbf{b}_j)$ ， $\hat{x} = (\mathbf{x}; 1)$ ，则 $\mathbf{w}_j^T \mathbf{x} + \mathbf{b}_j$ 可简写为 $\beta_j^T \hat{x}$ ，再令 $\mathbf{p}_j(\hat{x}; \beta) = p(\mathbf{y} = j | \hat{x}; \beta)$ ($1 \leq j \leq K-1$)， $\mathbf{p}_K(\hat{x}; \beta) = p(\mathbf{y} = K | \hat{x}; \beta) = 1 - \sum_{j=1}^{K-1} \mathbf{p}_j(\hat{x}; \beta)$ ，则式(16)中

的似然项可重写为：

$$p(\mathbf{y}_i | \mathbf{x}_i; w, b) = \sum_{j=1}^K \mathbb{I}(\mathbf{y}_i = j) \mathbf{p}_j(\hat{x}_i; \beta) \quad (17)$$

将式(17)代入(16)，“对数似然”式可改写为：

$$\mathbf{L}(\beta) = \sum_{i=1}^m \ln \sum_{j=1}^K \mathbb{I}(\mathbf{y}_i = j) \mathbf{p}_j(\hat{x}_i; \beta) \quad (18)$$

又 $\sum_{j=1}^K \mathbb{I}(\mathbf{y}_i = j) \ln \mathbf{p}_j(\hat{x}_i; \beta)$ 与 $\ln \sum_{j=1}^K \mathbb{I}(\mathbf{y}_i = j) \mathbf{p}_j(\hat{x}_i; \beta)$ 单调性相同，故最大化式(18)即最大化：

$$\mathbf{L}(\beta) = \sum_{i=1}^m \sum_{j=1}^K \mathbb{I}(\mathbf{y}_i = j) \ln \mathbf{p}_j(\hat{x}_i; \beta) \quad (19)$$

将 $\mathbf{p}_j(\hat{x}_i; \beta)$ 的表达式代入式(19)得：

$$\mathbf{L}(\beta) = \sum_{i=1}^m \left[\sum_{j=1}^{K-1} \mathbb{I}(\mathbf{y}_i = j) \beta_j^T \hat{x}_i - \sum_{j=1}^K \mathbb{I}(\mathbf{y}_i = j) \ln \left(1 + \sum_{q=1}^{K-1} e^{\beta_q^T \hat{x}_i} \right) \right] \quad (20)$$

又对于确定的 i ，有且只有一个 j 使得 $\mathbb{I}(\mathbf{y}_i = j) = 1$ ，故 $\sum_{j=1}^K \mathbb{I}(\mathbf{y}_i = j) \ln \left(1 + \sum_{q=1}^{K-1} e^{\beta_q^T \hat{x}_i} \right) = \ln \left(1 + \sum_{q=1}^{K-1} e^{\beta_q^T \hat{x}_i} \right)$ ，式(20)可化简为：

$$\mathbf{L}(\beta) = \sum_{i=1}^m \left[\sum_{j=1}^{K-1} \mathbb{I}(\mathbf{y}_i = j) \beta_j^T \hat{x}_i - \ln \left(1 + \sum_{q=1}^{K-1} e^{\beta_q^T \hat{x}_i} \right) \right] \quad (21)$$

上式即为该对率回归模型的“对数似然”。

(2)

令式(21)对 β_t 求偏导，可得：

$$\frac{\partial \mathbf{L}(\beta)}{\partial \beta_t} = \sum_{i=1}^m \hat{x}_i \left[\mathbb{I}(\mathbf{y}_i = t) - \frac{e^{\beta_t^T \hat{x}_i}}{1 + \sum_{q=1}^{K-1} e^{\beta_q^T \hat{x}_i}} \right]$$

则该“对数似然”的梯度为

$$\nabla \mathbf{L}(\beta) = \left(\frac{\partial \mathbf{L}(\beta)}{\partial \beta_1}, \frac{\partial \mathbf{L}(\beta)}{\partial \beta_2}, \dots, \frac{\partial \mathbf{L}(\beta)}{\partial \beta_{K-1}} \right)$$

(3)

上述推导都是基于样本类别分布平衡的情况，所以产生的分类器阈值也是对应样本类别分布平衡时的阈值。若样本类别分布不平衡，则上述推导出的分类器的阈值就不适用了。故需要在推导前依照各类别样本数和基准类别样本数对每个类别的原对数几率函数中的几率进行再缩放，用过缩放的对数几率函数进行推导。

对类别 t ：

$$\frac{y'}{1 - y'} = \frac{y}{1 - y} \times \frac{m^K}{m^t}$$

3 [30pts] P-R Curve & ROC Curve

现有500个测试样例，其对应的真实标记和学习器的输出值如表1所示（完整数据见data.csv文件）。该任务是一个二分类任务，1表示正例，0表示负例。学习器的输出越接近1表明学习器认为该样例越可能是正例，越接近0表明学习器认为该样例越可能是负例。

表 1: 测试样例表

样本	x_1	x_2	x_3	x_4	x_5	...	x_{496}	x_{497}	x_{498}	x_{499}	x_{500}
标记	1	1	0	0	0	...	0	1	0	1	1
输出值	0.206	0.662	0.219	0.126	0.450	...	0.184	0.505	0.445	0.994	0.602

(1) [10pts]请编程绘制P-R曲线；

(2) [15pts]请编程绘制ROC曲线，并计算AUC；

(3) [5pts] 需结合关键代码说明思路，并附最终绘制的曲线。建议使用python编程实现。实验报告需要有层次和条理性，能让读者仅通过实验报告便能了解实验的目的，过程和结果。

提示1: 需要注意数据中存在输出值相同的样例。

提示2: 在python中，数值计算通常使用numpy, 表格数据操作通常使用pandas, 画图可以使用matplotlib, 可以通过上网查找相关资料学习使用这些工具。未来大家会接触到更多的python扩展库，如集成了众多机器学习方法的sklearn, 深度学习工具包pytorch等。

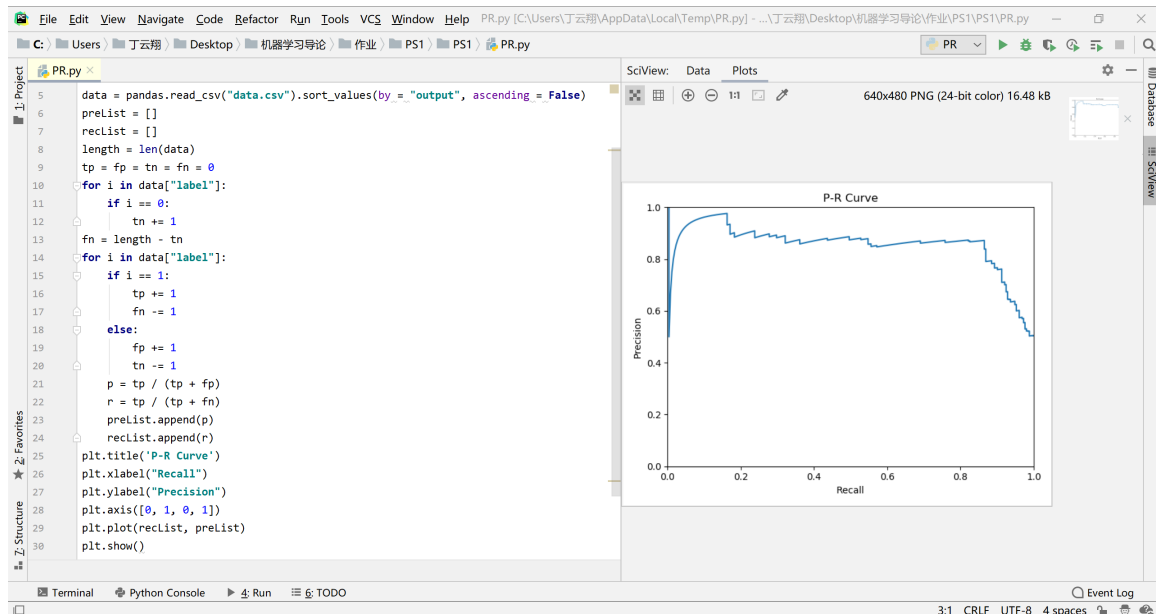
Solution. 此处用于写解答(中英文均可)

(1)

实验目的：根据学习器的测试样例表数据绘制该学习器的P-R曲线

实验过程：使用Python编程，先将data.csv中的数据读入，再按“output”的值从大到小排序，然后计算出500个样本中的负例数，并按全预测为负例的情况初始化TP、FP、TN、FN，然后按排序的顺序逐个把样本作为正例进行预测，并检查该样本是否为正例，选择更新TP、FN或FP、TN，然后依照定义分别计算出对应的查全率和查准率并存储起来。最后以查准率为纵轴、查全率为横轴作图，绘制出P-R曲线。

实验结果：绘制出的P-R曲线和绘制的代码如下图：



(2)

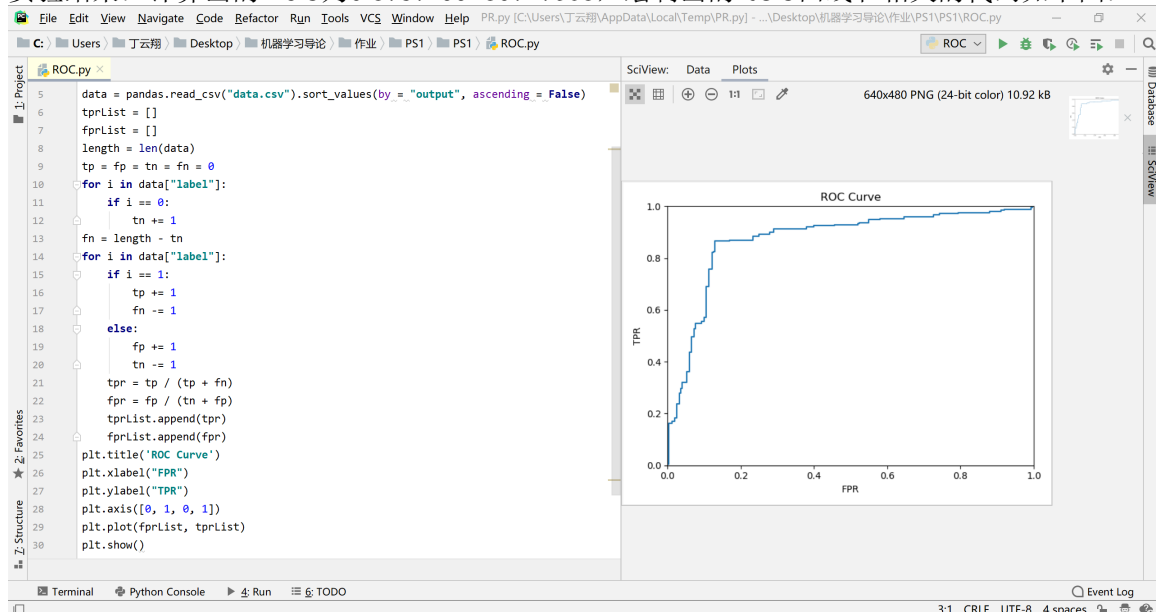
实验目的：根据学习器的测试样例表数据绘制该学习器的ROC曲线，并计算AUC

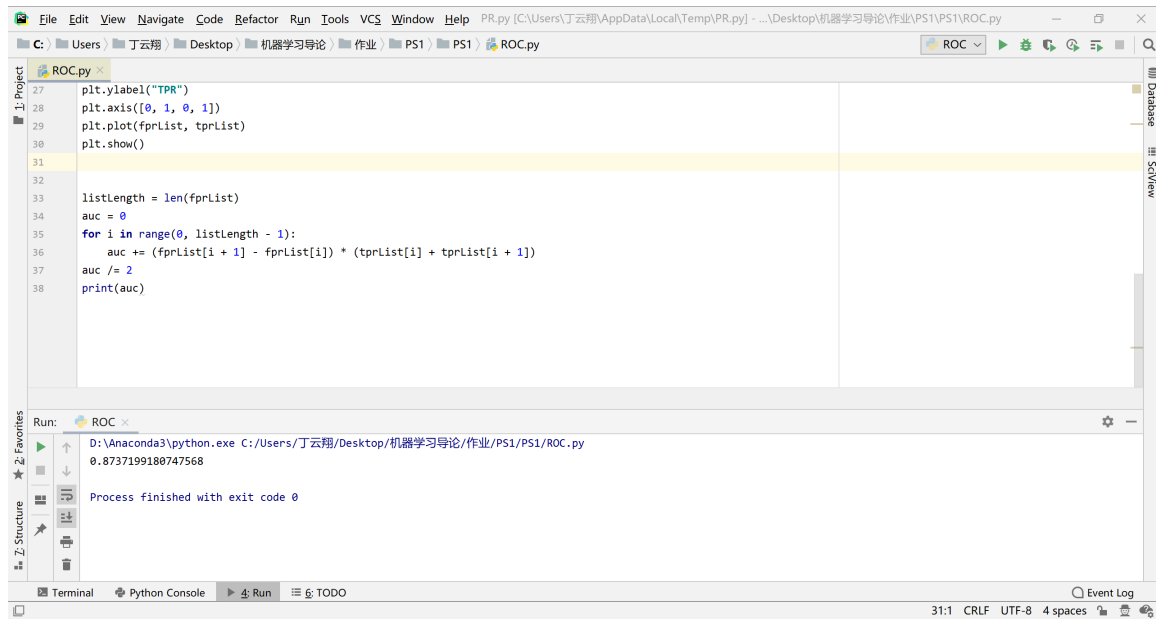
实验过程：使用Python编程，先将data.csv中的数据读入，再按“output”的值从大到小排序，然后计算出500个样本中的负例数，并按全预测为负例的情况初始化TP、FP、TN、FN，然后按排序的顺序逐个把样本作为正例进行预测，并检查该样本是否为正例，选择更新TP、FN或FP、TN，然后依照定义分别计算出对应的真正例率和假正例率并存储起来。最后以真正例率为纵轴、假正例率为横轴作图，绘制出ROC曲线。然后，由西瓜书中的公式(2.20)

$$AUC = \frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} - x_i)(y_i + y_{i+1})$$

利用Python和存储了真正例率、假正例率的列表计算出AUC的值。

实验结果：计算出的AUC为0.8737199180747568，绘制出的ROC曲线和相关的代码如下图：





The screenshot shows an IDE window with a Python file named `ROC.py`. The code calculates the Area Under the Curve (AUC) for a binary classifier. It starts by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). The AUC is then calculated using the trapezoidal rule.

```
27 plt.ylabel("TPR")
28 plt.axis([0, 1, 0, 1])
29 plt.plot(fprList, tprList)
30 plt.show()
31
32
33 listLength = len(fprList)
34 auc = 0
35 for i in range(0, listLength - 1):
36     auc += (fprList[i + 1] - fprList[i]) * (tprList[i] + tprList[i + 1])
37 auc /= 2
38 print(auc)
```

The Run window shows the output of the script:

```
Run: ROC
D:\Anaconda3\python.exe C:/Users/丁云翔/Desktop/机器学习导论/作业/PS1/ROC.py
0.8737199180747568
Process finished with exit code 0
```

The status bar at the bottom indicates the file encoding is UTF-8 with 4 spaces, and the cursor is at line 31, column 1.

(3)

如前两问所述。