机器学习导论
习题一

Solution

2021 年 4 月 12 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LaTeX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$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$, 其中$\mathbf{x}_i = (x_{i1}; x_{i2}; \cdots; x_{id}) \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, 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：

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

其中$\mathbf{y} = [y_1, \cdots, y_m]^{\mathrm{T}} \in \mathbb{R}^m, \mathbf{X} = [\mathbf{x}_1^{\mathrm{T}}; \mathbf{x}_2^{\mathrm{T}}; \cdots; \mathbf{x}_m^{\mathrm{T}}] \in \mathbb{R}^{m\times d}$, $\mathbf{1}$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：

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

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$\hat{\mathbf{w}}^*_{\mathbf{LS}}$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$\hat{\mathbf{w}}^*_{\mathbf{Ridge}}$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X，下式均成立

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

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) 定义目标函数：

$$f(\mathbf{w}, \mathbf{b}) := \frac{1}{2}\left(\mathbf{X}\mathbf{w} + 1\mathbf{b}^\top - \mathbf{y}\right)^\top\left(\mathbf{X}\mathbf{w} + \mathbf{1}\mathbf{b}^\top - \mathbf{y}\right)$$

求微分：

$$\frac{\partial}{\partial\mathbf{w}} f(\mathbf{w}, \mathbf{b}) = \mathbf{X}^\top\left(\mathbf{X}\mathbf{w} + 1\mathbf{b}^\top - \mathbf{y}\right) = \left(\mathbf{X}^\top\mathbf{X}\right)\mathbf{w} + \left(\mathbf{X}^\top\mathbf{1}\right)(\mathbf{b}) - \mathbf{X}^\top\mathbf{y}$$

$$\frac{\partial}{\partial \mathbf{b}} f(\mathbf{w}, \mathbf{b}) = \mathbf{1}^\top \left(\mathbf{X}\mathbf{w} + 1\mathbf{b}^\top - \mathbf{y}\right) = \left(\mathbf{1}^\top \mathbf{X}\right)\mathbf{w} + \left(\mathbf{1}^\top \mathbf{1}\right)\mathbf{b} - \mathbf{1}^\top \mathbf{y}$$

由于目标函数对于$\mathbf{w}, \mathbf{b}$是凸函数，联立以下方程组：

$$\begin{cases} \left(\mathbf{X}^\top \mathbf{X}\right)\mathbf{w} + \left(\mathbf{X}^\top \mathbf{1}\right)(\mathbf{b}) - \mathbf{X}^\top \mathbf{y} = 0 \\ \left(\mathbf{1}^\top \mathbf{X}\right)\mathbf{w} + m\mathbf{b} - \mathbf{1}^\top \mathbf{y} = 0 \end{cases} \tag{4}$$

得到：

$$\left(\mathbf{X}^\top \mathbf{X} - \frac{1}{m}\mathbf{X}^\top \mathbf{1}\mathbf{1}^\top \mathbf{X}\right)\mathbf{w} - \left(\mathbf{X}^\top - \frac{1}{m}\mathbf{X}^\top \mathbf{1}\mathbf{1}^\top\right)\mathbf{y} = 0$$

所求闭式解为：

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

$$\mathbf{b}^*_{\mathrm{LS}} = \frac{1}{m}\left(\mathbf{1}^\top \mathbf{y} - \mathbf{1}^\top \mathbf{X}\mathbf{w}^*_{\mathbf{LS}}\right)$$

(2) 类似上一问，定义目标函数：

$$g(\mathbf{w}, \mathbf{b}) := \frac{1}{2}\left\|\mathbf{X}\mathbf{w} + 1\mathbf{b}^\top - \mathbf{y}\right\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

求微分：

$$\left(\mathbf{X}^\top \mathbf{X} - \frac{1}{m}\mathbf{X}^\top \mathbf{1}\mathbf{1}^\top \mathbf{X} + 2\lambda \mathbf{I}_d\right)\mathbf{w} - \left(\mathbf{X}^\top - \frac{1}{m}\mathbf{X}^\top \mathbf{1}\mathbf{1}^\top\right)\mathbf{y} = 0$$

同样可以得到：

$$\hat{\mathbf{w}}^*_{\mathbf{Ridge}} = \left(\mathbf{X}^\top \mathbf{X} - \frac{1}{m}\mathbf{X}^\top \mathbf{1}\mathbf{1}^\top \mathbf{X} + 2\lambda \mathbf{I}_d\right)^{-1}\left(\mathbf{X}^\top - \frac{1}{m}\mathbf{X}^\top \mathbf{1}\mathbf{1}^\top\right)\mathbf{y}$$

$$\mathbf{b}^*_{\mathbf{Ridge}} = \frac{1}{m}\left(\mathbf{1}^\top \mathbf{y} - \mathbf{1}^\top \mathbf{X}\hat{\mathbf{w}}^*_{\mathbf{Ridge}}\right)$$

(4) 考虑二分类问题，可以将两个类别的标记分别定为(0,1)或(-1,1)，再利用相应的映射函数（如Sigmoid）将输出限制到指定范围内。对于多分类问题，由于真实场景下的应用很少存在标记“序”的关系，因此可以设计one-hot编码来建模分类任务。

(5) 由于

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

即

$$\mathbf{X}\left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right) = \left(\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}\right)\mathbf{X}$$

左右两边各乘上相同的矩阵：

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

即：

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

这个公式主要区别在于改变了求逆矩阵的规模。由于二者维度不同，$\mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{m\times m}$，$\mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{d\times d}$，其中d为样本特征维度，而m为训练集样本数目。在很多情况下，训练集样本数目是远大于样本特征维度的，因此使用这个公式可以显著地降低矩阵求逆开销。

助教注： 前两问的目的是让大家自己练一下求闭式解，有的同学解出来的闭式解没有联立，在w中含有b，而b中含有w，这样的结果被扣除了一定分数。第五问的证明比较简单，但是很多同学只回答了一半，没有回答“它有什么用”，这一问只要提到能够在样本数较多的情况下降低矩阵求逆开销的均给了满分。

在实际批改过程中，有部分同学标量和向量混用，加粗和不加粗混用，虽然没有扣分，但是请大家之后的作业中注意数学公式书写。

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$y \in \{1, 2 \ldots, K\}$。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，

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

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：

$$\begin{aligned}
\ln \frac{p(y=1 \mid \mathbf{x})}{p(y=K \mid \mathbf{x})} &= \mathbf{w}_1^{\mathrm{T}}\mathbf{x} + b_1 = \boldsymbol{\beta}_1^{\mathrm{T}}\hat{\mathbf{x}} \\
\ln \frac{p(y=2 \mid \mathbf{x})}{p(y=K \mid \mathbf{x})} &= \mathbf{w}_2^{\mathrm{T}}\mathbf{x} + b_2 = \boldsymbol{\beta}_2^{\mathrm{T}}\hat{\mathbf{x}} \\
&\cdots \\
\ln \frac{p(y=K-1 \mid \mathbf{x})}{p(y=K \mid \mathbf{x})} &= \mathbf{w}_{K-1}^{\mathrm{T}}\mathbf{x} + b_{K-1} = \boldsymbol{\beta}_{k-1}^{\mathrm{T}}\hat{\mathbf{x}}
\end{aligned}$$

为了归一化后验概率，$\sum_{i=1}^{K} p(y=i \mid \mathbf{x}) = 1$，得到：

$$p(y=i \mid \mathbf{x}) = \frac{e^{\boldsymbol{\beta}_i^{\mathrm{T}}\hat{\mathbf{x}}}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k^{\mathrm{T}}\hat{\mathbf{x}}}}, i = 1, 2, \ldots K-1$$

$$p(y=K \mid \mathbf{x})=\frac{1}{1+\sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k^T \hat{x}}}$$

考虑对数似然：

$$\ell(\beta)=\sum_{i=1}^{m}\sum_{k=1}^{K}\mathbb{I}\left(y_i=k\right)\ln\left(p\left(y_i=k \mid x_i\right)\right) \tag{5}$$

(2) 仿照书上的过程，可以求Eq. 5梯度：

$$\begin{aligned}\frac{\partial \ell\beta}{\partial \beta_j}&=\sum_{i=1}^{m}\sum_{k=1}^{K}\left(\mathbb{I}\left(y_i=j\right)\ln\left(e^{\boldsymbol{\beta}_i^{\mathrm{T}}\hat{\mathbf{x}}}\right)-\mathbb{I}\left(y_i=j\right)\ln\left(1+\sum_{k=1}^{K}e^{\boldsymbol{\beta}_k^{\mathrm{T}}\hat{\mathbf{x}}}\right)\right)\\&=\sum_{i=1}^{m}\sum_{k=1}^{K}\left(\mathbb{I}\left(y_i=j\right)\hat{\mathbf{x}}_i-\mathbb{I}\left(y_i=j\right)\ln\left(1+\sum_{k=1}^{K}e^{\boldsymbol{\beta}_k^{\mathrm{T}}\hat{\mathbf{x}}}\right)\right)\\&=\sum_{i=1}^{m}\left(\sum_{k=1}^{K}\mathbb{I}\left(y_i=j\right)\hat{\mathbf{x}}_i-p\left(y_i=j \mid x_i\right)\hat{\mathbf{x}}_i\right)\\&=\sum_{i=1}^{m}\hat{\mathbf{x}}_i\left(\mathbb{I}\left(y_i=j\right)-p\left(y_i=j \mid \hat{\mathbf{x}}_i\right)\right)\end{aligned} \tag{6}$$

(3) 类别分布不平衡会导致模型倾向于预测训练集中出现较多的类。可以利用课本3.6节提供的一些思路进行解决。

助教注： 这题比较简单，借助提示基本都能做出来，得分率较高。附加题的设计和目前计算机视觉领域比较热门的长尾分布问题也有一定关系；只要用书上提到的解决类别不平衡问题的方法均给满分。

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)

助教注： 这题实现的代码应该不超过50行，但是还是有同学来问能不能直接调包做。部分同学没有写对应的分析，或者没有贴图，均扣除了一定比例的分数。最普遍的扣分点在于没有给出横轴和纵轴的名称。