机器学习导论 习题一

学号, 姓名, 邮箱

2024 年 3 月 13 日

作业提交注意事项

1. 作业所需的 LaTeX 及 Python 环境配置要求请参考: [Link];
2. 请在 LaTeX 模板中第一页填写个人的学号、姓名、邮箱;
3. 本次作业需提交的文件与对应的命名方式为:
   (a) 作答后的 LaTeX 代码 — HW1.tex;
   (b) 由 (a) 编译得到的 PDF 文件 — HW1.pdf;
   (c) 第三题代码 — Problem3.py;
   (d) 第五题代码 — Problem5.py;

   请将以上文件**打包为 学号 __ 姓名.zip** (例如 221300001_ 张三.zip) 后提交;
3. 若多次提交作业, 则在命名 .zip 文件时加上版本号, 例如 221300001_ 张三 _v1.zip" (批改时以版本号最高的文件为准);
4. 本次作业提交截止时间为 **3 月 29 日 23:59:59**. 未按照要求提交作业, 提交作业格式不正确, **作业命名不规范**, 将会被扣除部分作业分数; 除特殊原因 (如因病缓交, 需出示医院假条) 逾期未交作业, 本次作业记 0 分; **如发现抄袭, 抄袭和被抄袭双方成绩全部取消**;
5. 学习过程中, 允许参考 ChatGPT 等生成式语言模型的生成结果, 但必须在可信的信息源处核实信息的真实性; **不允许直接使用模型的生成结果作为作业的回答内容**, 否则将视为作业非本人完成并取消成绩;
6. 本次作业提交地址为 [Link], 请大家预留时间提前上交, 以防在临近截止日期时, 因网络等原因无法按时提交作业.

1 [25pts] Mathematical Foundations

(1) **[10pts]** (Derivatives of Matrices) 有 $\alpha \in \mathbb{R}$, $\boldsymbol{y} \in \mathbb{R}^{m\times 1}$, $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$ 且 $\mathbf{A} \in \mathbb{R}^{n\times n}$. 试完成如下题目, 并给出必要的计算过程:

(a) **[5pts]** 若 $\mathbf{b} \in \mathbb{R}^{n\times 1}$ 且 $\alpha = \boldsymbol{x}^\top \mathbf{A}\boldsymbol{x} + \mathbf{b}^\top \boldsymbol{x}$, 试求 $\frac{\partial \alpha}{\partial \boldsymbol{x}}$.

(b) **[5pts]** 若 $\mathbf{A}$ 可逆, $\mathbf{A}$ 为 α 的函数且 $\frac{\partial \mathbf{A}}{\partial \alpha}$ 已知, 试求 $\frac{\partial \mathbf{A}^{-1}}{\partial \alpha}$.

(2) **[15pts]** (Statistics) 有 $x_1, \cdots, x_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. 试完成如下题目, 并给出必要的过程:

(a) **[4pts]** 定义 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. 试证明 $\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

(b) **[7pts]** 定义 $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. 试证明 $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$ 且 s^2 独立于 $\bar{x}$.

(c) **[4pts]** 试证明 $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ 服从自由度为 $n-1$ 的学生 t 分布.

Solution. 此处用于写解答 (中英文均可)

(1a) $\frac{\partial \alpha}{\partial \boldsymbol{x}} = (\mathbf{A} + \mathbf{A}^\top)\boldsymbol{x} + \mathbf{b}$;

(1b) $\frac{\partial \mathbf{A}^{-1}}{\partial \alpha} = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial \alpha}\mathbf{A}^{-1}$. 关键在于使用 $\mathbf{I} = \mathbf{A}\mathbf{A}^{-1}$, $\frac{\partial \mathbf{I}}{\partial \alpha} = 0$.

(2a) 作为高斯分布的线性组合, $\bar{x}$ 也为高斯分布. 其均值与方差可计算如下:

$$\mathbb{E}[\bar{x}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[x_i\right] = \frac{1}{n}\cdot n\mu = \mu$$

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(x_i\right) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}$$

(2b) 记 $z_i = \frac{x_i-\mu}{\sigma} \sim \mathcal{N}(0,1)$. 容易推得:

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n}(\frac{x_i-\mu}{\sigma} - \frac{\bar{x}-\mu}{\sigma})^2 = \sum_{i=1}^{n}(z_i - \bar{z})^2 = \sum_{i=1}^{n} z_i^2 - n\bar{z}^2$$

构造一个第一行行向量为 $(\frac{1}{\sqrt{n}}, \cdots, \frac{1}{\sqrt{n}})$ 的 n 维正交矩阵 $\mathbf{P}$, 并记 $\boldsymbol{y} = \mathbf{P}\mathbf{z}$, 成立:

$$\sum_{i=1}^{n} z_i^2 = \mathbf{z}^\top\mathbf{z} = \mathbf{z}^\top\mathbf{P}^\top\mathbf{P}\mathbf{z} = \boldsymbol{y}^\top\boldsymbol{y} = \sum_{i=1}^{n} y_i^2; \quad y_1^2 = \frac{1}{n}\cdot(n\bar{z})^2 = n\bar{z}^2$$

易知 $\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, 因此 $\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n} z_i^2 - n\bar{z}^2 = \sum_{i=2}^{n} y_i^2 \sim \chi^2_{n-1}$ 且 $y_1 \perp\!\!\!\perp \sum_{i=2}^{n} y_i^2$. 由此可以推出 s^2 与 $\bar{x}$ 独立.

(2c) 已知 $\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \Rightarrow \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$ 且两者独立. 因此成立:

$$\frac{\sqrt{n}(\bar{x}-\mu)/\sigma}{\sqrt{\frac{(n-1)s^2}{\sigma^2}/(n-1)}} \sim t_{n-1}$$

化简即得 $\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$. 这一经典统计量常被用来对 μ 做假设检验.

2 [10pts] Performance Measure

性能度量是衡量模型泛化能力的评价标准, 在对比不同模型的能力时, 使用不同的性能度量往往会导致不同的评判结果. 请仔细阅读《机器学习》第二章 2.3.2 节. 在书中, 我们学习并计算了模型的二分类性能度量. 下面我们给出一个多分类 (三分类) 的例子, 请根据学习器的具体表现, 回答如下问题.

表 1: 类别的真实标记与预测标记

真实类别 \ 预测类别	第一类	第二类	第三类
第一类	9	0	2
第二类	2	8	1
第三类	0	0	7

(1) **[3pts]** 如表 1 所示, 请计算该学习器的错误率及精度.

(2) **[3pts]** 请分别计算宏查准率, 宏查全率, 微查准率, 微查全率.

(3) **[4pts]** 分别使用宏查准率, 宏查全率, 微查准率, 微查全率计算宏 $F1$ 度量, 微 $F1$ 度量.

Solution. 此处用于写解答 (中英文均可)

精度是该学习器在所有样例中正确预测的比例; 错误率是该学习器在所有样例中错误预测的比例. 在实际计算中, 正确预测的样例是混淆矩阵中对角线上的元素, 因此精度等于对角线上的元素和除以矩阵中的总元素和.

$$\text{Accuracy} = \frac{9+8+7}{9+0+2+2+8+1+0+0+7} = \frac{24}{29}.$$

$$\text{ErrorRate} = 1 - \text{Accuracy} = \frac{5}{29}$$

查准率和查全率的计算需要将多分类混淆矩阵改写为二分类混淆矩阵. 分别以第一类, 第二类, 第三类作为正例, 以其他类作为负例, 可以得到以下三个二分类混淆矩阵:

真实类别 \ 预测类别	正类	反类
正类	9	2
反类	2	16

真实类别 \ 预测类别	正类	反类
正类	8	3
反类	0	18

真实类别 \ 预测类别	正类	反类
正类	7	0
反类	3	19

依照教材公式, 可以计算以下指标:

$$\begin{aligned}
\text{Macro-P} &= \frac{1}{3}\sum_{i=1}^{3} P_i = \frac{1}{3} \times \left(\frac{9}{11} + \frac{8}{8} + \frac{7}{10}\right) \approx 0.839 \\
\text{Micro-P} &= \frac{\overline{TP}}{\overline{TP} + \overline{FP}} = \frac{9+8+7}{9+8+7+2+0+3} \approx 0.828 \\
\text{Macro-R} &= \frac{1}{3}\sum_{i=1}^{3} R_i = \frac{1}{3} \times \left(\frac{9}{11} + \frac{8}{11} + \frac{7}{7}\right) \approx 0.848 \\
\text{Micro-R} &= \frac{\overline{TP}}{\overline{TP} + \overline{FN}} = \frac{9+8+7}{9+8+7+2+3+0} \approx 0.828
\end{aligned}$$

依照教材公式, 可以计算以下指标:

$$\begin{aligned}
\text{Macro-F1} &= \frac{2 \times \text{ Macro-P } \times \text{ Macro-R}}{\text{Macro-P } + \text{ Macro-R}} \approx 0.844 \\
\text{Micro-F1} &= \frac{2 \times \text{ Micro-P } \times \text{ Micro-R}}{\text{Micro-P } + \text{ Micro-R}} \approx 0.828
\end{aligned}$$

3 [20pts] Cross Validation & Model Selection

机器学习常涉及两类参数: 一类是算法的参数, 亦称 "超参数", 如对数几率回归模型训练的迭代总次数; 另一类是模型的参数, 如对数几率回归模型的 $\mathbf{w}$ 与 b. 大多数学习算法的性能都会受到超参数设置的影响. 在《机器学习》第二章 2.2.2 节中介绍了一种模型评估方法 — 交叉验证, 它也经常被用于算法的参数调节. 下面, 我们尝试通过交叉验证, 寻找在所给数据集上最适合岭回归分类器 (RidgeClassifier) 的超参数 α. 请仔细阅读代码框架 Problem3.py, 补全空缺的代码片段, 实现以下的功能并回答相关问题.

(1) **[6pts]** 补全空缺代码, 实现 k 折交叉验证方法.

(2) **[4pts]** 通过单次 10 折交叉验证, 评估不同 α 值对分类器的性能影响. 请将生成的 cross_validation.png 图表放置在解答区域.

(3) **[5pts]** 基于上一题的结果选取最优的 α 值, 并计算模型在测试集上的分类精度. 请汇报选取的最优超参数 α 的取值与对应的分类精度.

(4) **[5pts]** 基于上述实验, 阅读《机器学习》2.2.4 节的内容, 简要谈谈在评估学习算法的泛化性能时, 数据集划分与超参数调节的大致流程.

Solution. 此处用于写解答 (中英文均可)

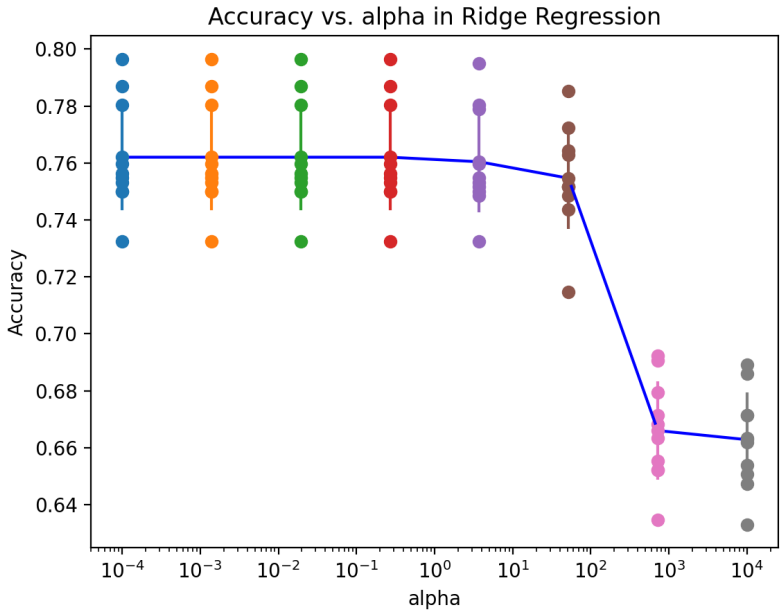

图 1: alpha 取值与验证集精度的关系

(1) 样例解答代码请参考附件: Problem3-sol.py.

(2) 生成的 cross_validation.png 图表如图 1 所示.

(3) 当 $\alpha \in \{0.0001, 0.0014, 0.0193, 0.2683\}$ 时, 模型在 10 折交叉验证中均取得最高的平均精度 0.7620. 选取任一最优取值 $\alpha = 0.0001$, 在完整的训练集上重新训练模型后, 模型的测试集分类精度为 0.7827.

(4) 在评估学习算法的泛化性能时, 往往将完整数据集按一定的比例划分训练数据与测试数据 (例如, 所给的代码框架中采用了 8:2), 训练数据用于模型的训练与选择, 测试数据用于评估模型的泛化性能. 此外, 由于模型的性能受到超参数的影响, 为了找到最优的超参数设置, 可以进一步将训练数据划分为训练集与验证集, 评估每组超参数在训练集上训练后的验证集表现. 最终, 选用验证集上最优的超参数在完整训练数据上训练, 再评估模型在测试数据上的泛化性能. 在资源允许的条件下, 这一数据集划分与超参数选择的过程也应当选用不同的随机种子重复多次.

4 [25pts] ROC & AUC

ROC 曲线与其对应的 AUC 值可以反应分类器在一般情况下泛化性能的好坏. 请仔细阅读《机器学习》第二章 2.3.3 节, 并完成本题 (请按定义给出必要的计算步骤, 否则不予计分).

表 2: 样例的真实标记与预测

样例	x_1	x_2	x_3	x_4	x_5	x_6	x_7
标记	0	1	0	1	0	1	0
分类器输出值	0.32	0.89	0.63	0.32	0.25	0.66	0.48

(1) **[5pts]** 如表 2 所示, 第二行为样例的真实标记, 第三行为某分类器对样例的预测结果. 请根据上述结果, 绘制分类器在该样例集合上的 ROC 曲线, 并计算其对应的 AUC 值.

(2) **[6pts]** 除表 2 外另有负类样本 x_8, 预测值为 0.8. 请绘制此时的 ROC 曲线, 并计算其对应的 AUC 值. 试分析增加一个预测值高的负类样本对 AUC 带来的影响及原因.

(3) **[6pts]** 除表 2 外另有正类样本 x_8, 预测值为 0.8. 请绘制此时的 ROC 曲线, 并计算其对应的 AUC 值. 试分析增加一个预测值高的正类样本对 AUC 带来的影响及原因, 并相比上问, 分析这两种情况下 AUC 值的变化幅度差异.

(4) **[8pts]** 试证明对有限样例成立 (请给出详尽的证明过程, 直接使用书中结论不计分):

$$\mathrm{AUC} = \frac{1}{m^+ m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left(\mathbb{I}\left\{ f(x^+) > f(x^-) \right\} + \frac{1}{2} \mathbb{I}\left\{ f(x^+) = f(x^-) \right\} \right). \quad (4.1)$$

Solution. 此处用于写解答 (中英文均可)

(1) 如图3所示，可以计算此时的 AUC 值如下所示：

$$\mathrm{AUC} = \frac{2}{3} \cdot \frac{1}{2} + \frac{1}{2} \cdot \left(\frac{2}{3} + 1 \right) \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = \frac{19}{24} \approx 0.79167$$

(2) 如图3所示，可以计算此时的 AUC 值如下所示：

$$\mathrm{AUC} = \frac{1}{3} \cdot \frac{1}{5} + \frac{2}{3} \cdot \frac{2}{5} + \frac{1}{2} \cdot \left(\frac{2}{3} + 1 \right) \cdot \frac{1}{5} + 1 \cdot \frac{1}{5} = \frac{7}{10} = 0.7$$

增加一个预测值高的负类样本, 使得 AUC 下降很多. 因为在排序损失中, 所有预测值小于该负类样本的正类样本, 都会累加一份损失, 因此会导致 AUC 值明显下降.

(3) 如图3所示，可以计算此时的 AUC 值如下所示：

$$\mathrm{AUC} = \frac{3}{4} \cdot \frac{1}{2} + \frac{1}{2} \cdot \left(\frac{3}{4} + 1 \right) \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = \frac{27}{32} = 0.84375$$

增加一个预测值高的正类样本, AUC 会增加. 因为预测值高的正类样本 (此时) 不会带来新的错误排序对, 因此不会增加额外的排序损失; 同时因为增加了一个正例数目, 使得每个错

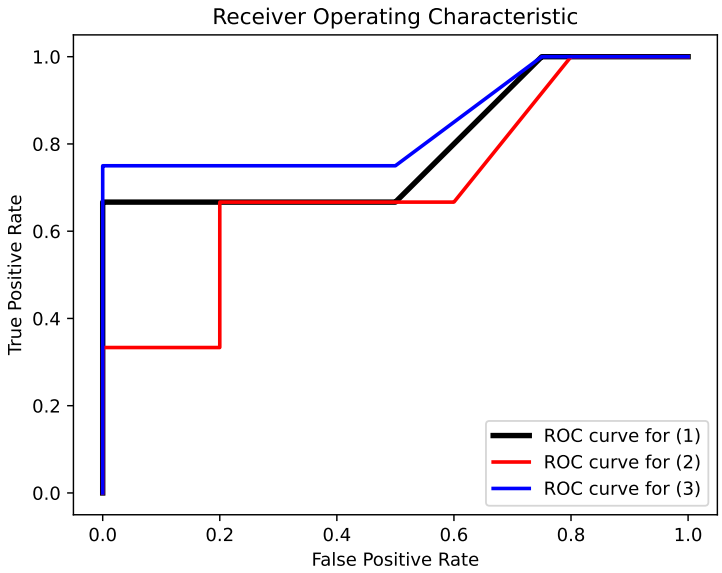

图 2: 分类器对应 ROC 曲线

误排序对的损失相对变小 ($\frac{1}{m^+ + 1}\frac{1}{m^-} < \frac{1}{m^+}\frac{1}{m^-}$), 导致 ℓ_{rank} 相对变小, AUC 变大. 在这两问中, 增加一个预测值高的正类样本导致 AUC 增加的值, 要比增加一个预测值高的负类样本导致 AUC 减小的值小. 造成这种差异的原因是导致 AUC 升高和减小的原因存在差异.

(4) 证明过程如下:
考虑 ROC 曲线的绘制过程，设前一个样例在 ROC 曲线上的坐标为 (x, y),

1) 若当前样例为真正例，则对应在 ROC 曲线上的坐标为 $(x, y + \frac{1}{m^+})$;

2) 若当前样例为假正例，则对应在 ROC 曲线上的坐标为 $(x + \frac{1}{m^-}, y)$。

由此可知，考虑任何一对正例和负例对,

1) 若其中正例预测值小于反例，则 x 先增加，y 后增加，曲线下方的面积 (即 AUC) 将不会因此而增加;

2) 若其中正例预测值大于反例，则 y 值会先增加，x 后增加，曲线下方的面积 (即 AUC) 将增加一个矩形格子，其面积为 $\frac{1}{m^+ m^-}$;

3) 若一个正例预测值等于反例，对应标记点 x, y 坐标值同时增加，曲线下方的面积 (即 AUC) 将增加一个三角形，其面积为 $\frac{1}{2}\frac{1}{m^+ m^-}$.

考虑所有正例和负例对，AUC 的面积即为曲线下方的面积，根据上述情况进行累加，则有

$$\mathrm{AUC} = \frac{1}{m^+ m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left(\mathbb{I}(f(x^+) > f(x^-)) + \frac{1}{2}\mathbb{I}(f(x^+) = f(x^-)) \right)$$

5 [20pts] Logistic Regression

对数几率回归 (Logistic Regression) 是常用的分类学习算法, 通常使用 AUC 值评估其分类性能. 下面, 我们利用 Python 实现二分类的对数几率回归模型, 并采用牛顿法进行模型的优化求解. 请仔细阅读代码框架 Problem5.py, 补全空缺的代码片段, 实现以下的功能并回答相关问题.

(1) **[5pts]** 实现 $\ell(\beta)$ 关于 β 的二阶导数的计算. (即书中公式 3.31)
提示: 可以参考框架代码中 $\ell(\beta)$ 关于 β 的一阶导数的计算方法.

(2) **[5pts]** 实现牛顿法的迭代步骤. (即书中公式 3.29)

(3) **[5pts]** 实现基于参数 β, 计算 $\mathbf{X}$ 对应的类别概率的方法.

(4) **[5pts]** 绘制训练后的模型在测试集上的 ROC 曲线图, 并汇报对应的 AUC 数值 (保留四位小数). 请将生成的 roc.png 图片放置在解答区域.
提示: 若你未能完成 (1-3) 题, 你可以使用 sklearn 的对数几率回归模型作为替代.

Solution. 此处用于写解答 (中英文均可)

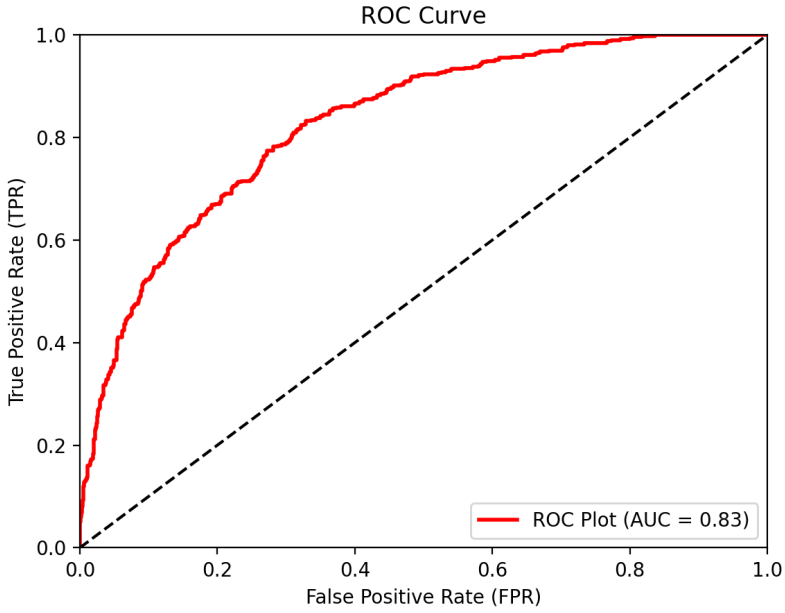

图 3: 测试集 ROC 曲线图

(1-3) 样例解答代码请参考附件: Problem5-sol.py.

(4) 绘制得到的 ROC 曲线图如图 3 所示. 采用代码框架实现的 MyLogisticRegression 时, 对应的测试集 AUC 数值为 0.8292; 采用 sklearn.linear_model.LogisticRegression 时, 对应的测试集 AUC 数值为 0.8298.