

机器学习导论 习题一

221300066, 季千^①, qkjia@smail.nju.edu.cn

2024 年 3 月 27 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 **4 月 2 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; 除特殊情况 (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 $\alpha \in \mathbb{R}$, $\mathbf{y} \in \mathbb{R}^{m \times 1}$, $\mathbf{x} \in \mathbb{R}^{n \times 1}$ 且 $\mathbf{A} \in \mathbb{R}^{n \times n}$. 试完成如下题目(请给出必要的计算步骤, 否则不予计分):

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

(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 $x_1, \dots, x_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. 试完成如下题目:

(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 $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$ 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) (1)

(a) 求偏导数 $\frac{\partial \alpha}{\partial \mathbf{x}}$, 即对向量 \mathbf{x} 的每一个分量求导:

对于第一项 $\mathbf{x}^\top \mathbf{A} \mathbf{x}$ (一个关于 \mathbf{x} 的二次型), 因为 \mathbf{A} 是对称的 ($\mathbf{A} = \mathbf{A}^\top$), 根据矩阵微积分的规则有, $\frac{\partial (\mathbf{x}^\top \mathbf{A} \mathbf{x})}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top) \mathbf{x}$.

对于第二项 $\mathbf{b}^\top \mathbf{x}$, 这是一个线性的项, 其导数就是常数向量 \mathbf{b} .

所以 $\frac{\partial \alpha}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top) \mathbf{x} + \mathbf{b}$

(b) 根据链式法则有:

$$\frac{\partial (\mathbf{A}^{-1})}{\partial \alpha} = \frac{d(\mathbf{A}^{-1})}{d\alpha} \cdot \frac{\partial \mathbf{A}}{\partial \alpha}$$

因为 \mathbf{A}^{-1} 是 \mathbf{A} 的逆, 有:

$$\mathbf{A} \mathbf{A}^{-1} = \mathbf{I}$$

其中 \mathbf{I} 是单位矩阵。对上式两边关于 α 取导数, 可得:

$$\frac{d(\mathbf{A} \mathbf{A}^{-1})}{d\alpha} = \frac{d(\mathbf{I})}{d\alpha}$$

再使用乘积法则, 有:

$$\mathbf{A} \frac{d(\mathbf{A}^{-1})}{d\alpha} + \mathbf{A}^{-1} \frac{d(\mathbf{A})}{d\alpha} = 0$$

因为 $\frac{d(\mathbf{I})}{d\alpha} = 0$ (单位矩阵不随 α 变化)。只需解下面的方程来找到 $\frac{d(\mathbf{A}^{-1})}{d\alpha}$:

$$\mathbf{A} \frac{d(\mathbf{A}^{-1})}{d\alpha} = -\mathbf{A}^{-1} \frac{d(\mathbf{A})}{d\alpha}$$

$$\frac{d(\mathbf{A}^{-1})}{d\alpha} = -\mathbf{A}^{-1} \frac{d(\mathbf{A})}{d\alpha} \mathbf{A}^{-1}$$

将 $\frac{d(\mathbf{A})}{d\alpha}$ 表示为 $\frac{\partial \mathbf{A}}{\partial \alpha}$, 即得:

$$\frac{\partial \mathbf{A}^{-1}}{\partial \alpha} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \alpha} \mathbf{A}^{-1}$$

(2)

(a) 因为 $x_i \sim N(\mu, \sigma^2)$, 根据正态分布的性质, x_1, \dots, x_n 的和也是正态分布的。

使所以, $\sum_{i=1}^n x_i \sim N(n\mu, n\sigma^2)$ 。

题中定义 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, 则 \bar{x} 的期望值 $E[\bar{x}] = \frac{1}{n} E[\sum_{i=1}^n x_i] = \frac{1}{n}(n\mu) = \mu$ 。

\bar{x} 的方差 $Var[\bar{x}] = \frac{1}{n^2} Var[\sum_{i=1}^n x_i] = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$ 。

所以 $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ 。

(b) 因为 $x_i - \mu \sim N(0, \sigma^2)$, 所以 $(x_i - \mu)^2/\sigma^2$ 服从自由度为 1 的卡方分布 χ_1^2 。

样本方差 s^2 是这些独立卡方分布变量的平均值, 但分母是 $n-1$ 而不是 n 。

即 $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ 。

因为每个 $(x_i - \mu)^2/\sigma^2$ 都是独立的卡方分布变量, 它们的和除以自由度数也应当服从卡方分布。

所以 $(n-1)s^2/\sigma^2 = \sum_{i=1}^n ((x_i - \mu)^2/\sigma^2)$ 服从自由度为 $n-1$ 的卡方分布 χ_{n-1}^2 。

另外, 由于 \bar{x} 只涉及 x_i 的线性组合而与平方项无关, s^2 是基于 x_i 平方项的一个度量。根据 Lemma, 如果两个随机变量 U 和 V 独立, 那么它们的任意函数 $g(U)$ 和 $g(V)$ 也独立。因此 \bar{x} 和 s^2 是独立的。证毕。

(c) 由 (a) 可知 $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$, 所以 $\sqrt{n}(\bar{x} - \mu) \sim N(0, \sigma^2)$ 。

由 (b) 可知 $(n-1)s^2/\sigma^2 \sim \chi_{n-1}^2$, 所以 s^2 是 σ^2 的无偏估计。

因为学生 t 分布定义为一个正态分布和一个独立的卡方分布变量的比值的分布, 所以有:

$\frac{\sqrt{n}(\bar{x} - \mu)}{\sqrt{\frac{\sigma^2}{n-1}}} = \frac{\sqrt{n}(\bar{x} - \mu)}{s/\sqrt{n}}$ 服从自由度为 $n-1$ 的学生 t 分布。

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)

(1) 错误率是衡量该学习器在所有样例中预测失误的比例; 精度是衡量该学习器在所有样例中正确预测的比例. 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

$$Acc = \frac{9+8+7}{9+2+2+8+1+7} = \frac{24}{29}, \text{ 错误率为 } 1 - Acc = \frac{5}{29}$$

(2) 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
	7	0
正类	7	0
反类	3	19

根据教材公式有：

$$Macro_P = \frac{1}{3} \sum_{i=1}^3 P_i = \frac{1}{3} \left(\frac{9}{11} + \frac{8}{8} + \frac{7}{10} \right) = \frac{277}{330}$$

$$Macro_R = \frac{1}{3} \sum_{i=1}^3 P_i = \frac{1}{3} \left(\frac{9}{11} + \frac{8}{11} + \frac{7}{7} \right) = \frac{28}{33}$$

$$Micro_P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}} = \frac{24}{29}$$

$$Micro_R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} = \frac{24}{29}$$

(3) 同样根据教材公式有：

$$MacroF1 = \frac{2 \times Macro_P \times Macro_R}{Macro_P + Macro_R} \approx 0.8843$$

- MicroF1:

$$MicroF1 = \frac{2 \times Micro_P \times Micro_R}{Micro_P + Micro_R} = \frac{24}{29}$$

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)

- (1) 实现 k 折交叉验证方法需补全的代码如下：

```
# Split the dataset into training and validation set
start = i * fold_size
end = (i + 1) * fold_size if i < k - 1 else N
X_train = np.concatenate((X[:start], X[end:]), axis=0)
y_train = np.concatenate((y[:start], y[end:]), axis=0)
X_val = X[start:end]
y_val = y[start:end]

# Train and evaluate the accuracy of the model using the training and validation set
model = train_model(X_train, y_train)
accuracy = evaluate_accuracy(model, X_val, y_val)
```

- (2)

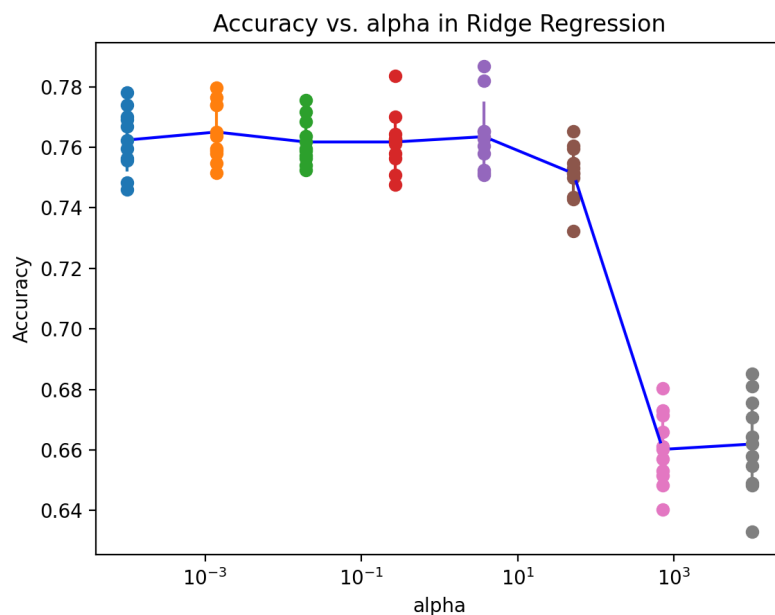


图 1: cross_validation

(3)

```

Shape of X: (7800, 26)
Shape of y: (7800,)
Shape of X_train: (6240, 26)
Shape of y_train: (6240,)
Shape of X_test: (1560, 26)
Shape of y_test: (1560,)

Candidate alpha values: [1.00000000e-04 1.38949549e-03 1.93069773e-02 2.68269580e-01
 3.72759372e+00 5.17947468e+01 7.19685673e+02 1.00000000e+04]

alpha = 0.0001, average accuracy = 0.7624
alpha = 0.0014, average accuracy = 0.7651
alpha = 0.0193, average accuracy = 0.7618
alpha = 0.2683, average accuracy = 0.7618
alpha = 3.7276, average accuracy = 0.7635
alpha = 51.7947, average accuracy = 0.7514
alpha = 719.6857, average accuracy = 0.6602
alpha = 10000.0000, average accuracy = 0.6619

Plot is saved as cross_validation.png

Best alpha = 0.0014, accuracy on test set = 0.7827

```

图 2: 运行结果

由图 2 可知：选取的最优超参数 $\alpha = 0.0014$, 对应的分类精度 $acc = 0.7827$

(4) 大致流程如下：

1. 数据预处理：首先对原始数据进行预处理，包括数据清洗、特征选择、缺失值处理等，以

确保数据的质量和可用性。

2. 数据集划分：将预处理后的数据集划分为训练集和测试集。通常情况下使用分层抽样或随机抽样的方法进行划分，以确保训练集和测试集的数据分布相似。

3. 模型训练：使用训练集数据训练机器学习模型。根据具体的算法和问题，选择合适的模型进行训练。

4. 超参数调节：对于一些需要设置超参数的模型（如正则化参数、学习率等），需要通过交叉验证等方法来选择最优的超参数。可以使用网格搜索、随机搜索或者贝叶斯优化等方法来进行超参数调节。

5. 模型评估：使用测试集数据对训练好的模型进行评估，计算模型的泛化性能指标，如准确率、召回率、F1 值等。

6. 分析优化：根据模型的评估结果，分析模型在不同类别、不同特征上的表现，找出可能存在的问题并进行改进，不断调整模型结构、超参数等，直至达到满意的泛化性能。

7. 模型选择：根据模型评估的结果，选择优化后表现最好的模型作为最终的模型。

8. 模型应用：将最终选择的模型应用于实际问题中，进行预测或分类等任务。

为了确保评估结果的可靠性，还可以将数据集划分为多个子集，进行多次交叉验证，然后取平均值作为最终的评估结果。此外，还可以使用不同的评估指标来综合考虑模型的性能，以及结合实际的业务需求进行模型选择。

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; 本题涉及的 ROC 曲线手绘或编程绘制均可).

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):

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

Solution. 此处用于写解答 (中英文均可)

(1) $\text{AUC} \approx 0.7916667$

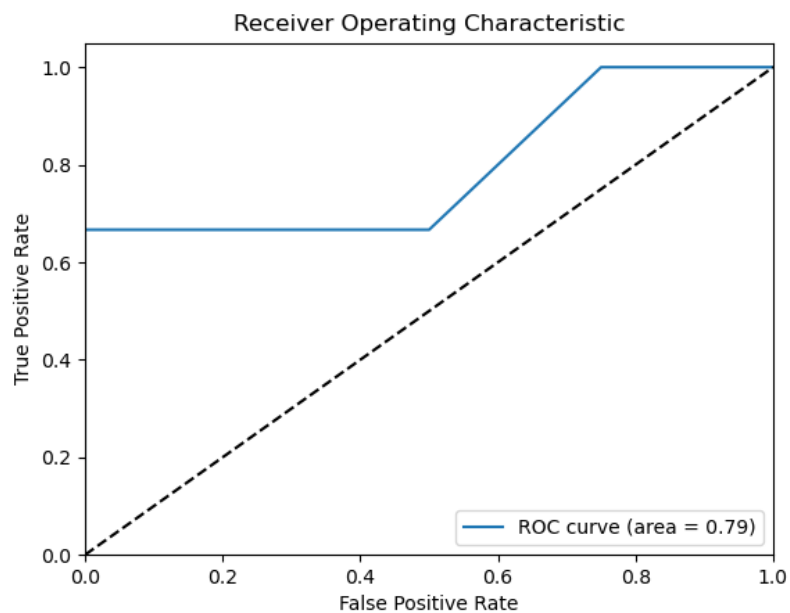


图 3: ROC Curve

(2) $AUC = 0.7$

当增加一个预测值高的负类样本（即题目中实际标签为 0，预测值为 0.8）时，影响了分类器的假阳性率（FPR）和真阳性率（TPR）的计算。因为该样本的预测值较高，而它实际上是负类样本，所以在较低的阈值下，它会被错误地分类为正类。这将导致在较低的阈值下 FPR 增加，而不会显著影响 TPR，因为真正的正类样本的预测值通常更高。这个效果会使得 ROC 曲线向左移动，由于 AUC 是 ROC 曲线下的面积，所以 AUC 的值可能会略微下降。

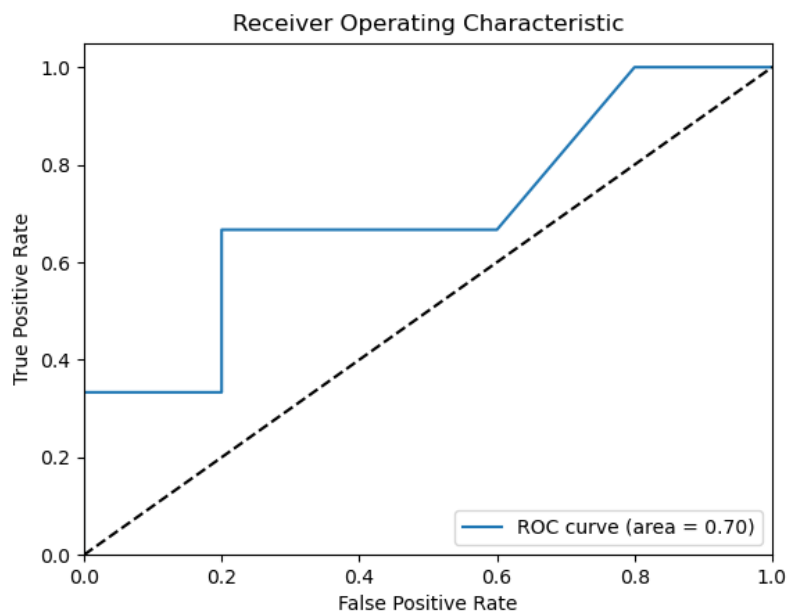


图 4: ROC Curve

(3) $AUC = 0.84375$

当增加一个预测值高的正类样本（即题目中实际标签为 1，预测值为 0.8）时，同样会影响分类器的 FPR 和 TPR。由于该样本的预测值较高，它将正确地被分类为正类，即分类器能够更准确地识别正类样本。这将增加 TPR 而不会显著影响 FPR。则会使得 ROC 曲线向右移动，AUC 值增加。

相比上问可知，变化幅度的差异取决于新样本的预测值与现有样本的预测值分布。如果新样本的预测值接近于现有样本的阈值，则意味着更多的样本将被错误分类或正确分类，那么对 AUC 的影响会更大。如果新样本的预测值远高于或远低于现有样本的阈值，那么对 AUC 的影响会更小。

总结来说，增加预测值高的正类样本通常会对 AUC 产生更积极的影响，而增加预测值高的负类样本可能会对 AUC 产生轻微的负面影响。这些变化的具体幅度取决于新样本的预测值与现有样本的预测值分布的关系。

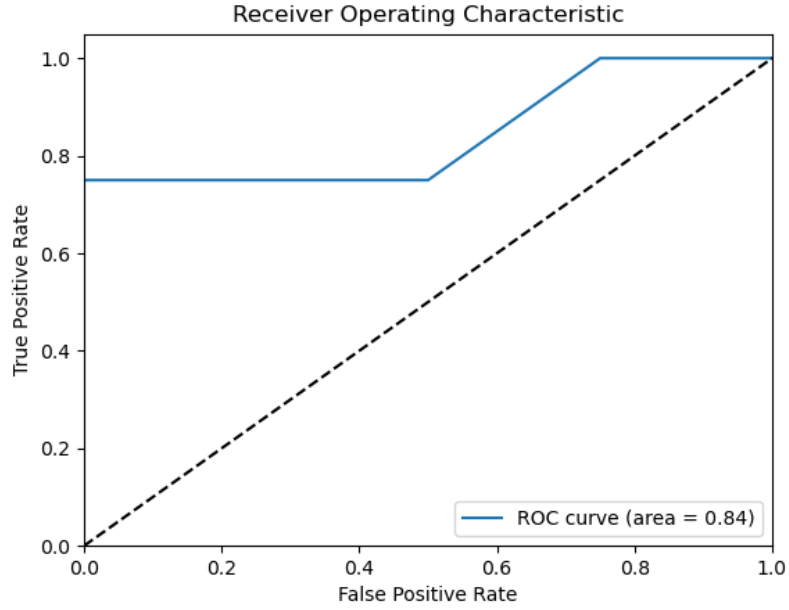


图 5: ROC Curve

(4) 证明如下:

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, 则对应应在 ROC 曲线上的坐标为 $(x, y + \frac{1}{m^+})$ 。

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

$$\text{AUC} = \frac{1}{m^+ m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left(\mathbb{I}\{f(x^+) > f(x^-)\} + \frac{1}{2} \mathbb{I}\{f(x^+) = f(x^-)\} \right). \quad (4.2)$$

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)

(1) 代码实现如下:

```
p = 1 / (1 + math.e ** (-X @ beta))
p=p.reshape(1,-1)[0]
result = X.T @ np.diag(p * (1 - p)) @ X
```

(2) 代码实现如下:

```
# Update parameter `self.beta` using Newton's method
new_beta = self.beta - np.linalg.inv(self._partial_2_l(self.beta, X, y)) @
self._partial_l(self.beta, X, y)
```

(3) 代码实现如下:

```
p = 1 / (1 + math.e ** (-X @ self.beta))
y_prob = np.copy(p)
```

(4) 代码实现如下:

```
fpr, tpr, thresholds = roc_curve(y_test, y_prob)
auc_value = auc(fpr, tpr)

plt.plot(fpr, tpr, label='ROC curve' % auc_value)
plt.xlabel('FPR')
plt.ylabel('TPR')
plt.title('ROC Curve')
plt.legend(loc="lower right")
```

AUC in test set: 0.8292

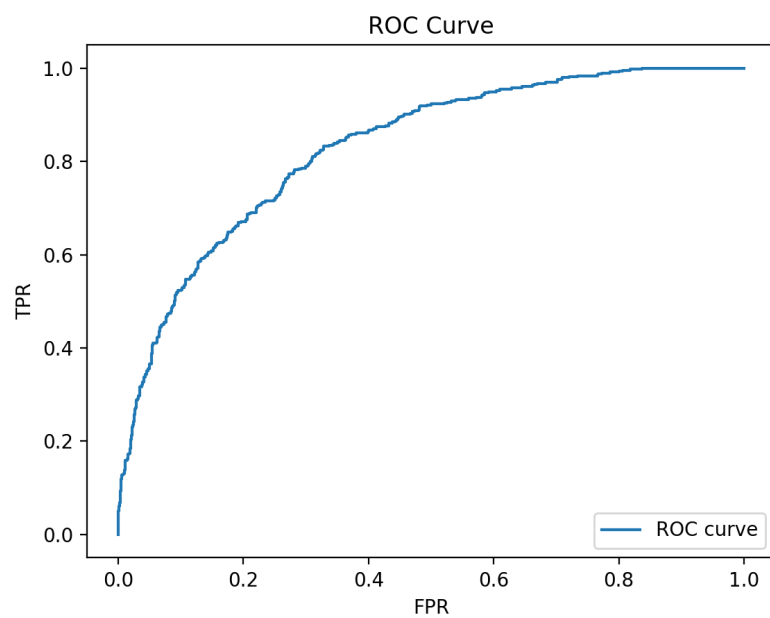


图 6: ROC Curve

Acknowledgments

允许与其他同样未完成作业的同学讨论作业的内容, 但需在此注明并加以致谢; 如在作业过程中, 参考了互联网上的资料或大语言模型的生成结果, 且对完成作业有帮助的, 亦需注明并致谢.