机器学习导论 习题一

学号, 姓名, 邮箱

2023 年 4 月 19 日

作业提交注意事项

1. 请在 LaTeX 模板中第一页填写个人的学号、姓名、邮箱；
2. 本次作业需提交作答后的该 pdf 文件、编程题代码 (.py 文件)；**请将二者打包为.zip 文件上传**。注意命名规则：三个文件均命名为“学号 _ 姓名”+“. 后缀”(例如“211221001_ 张三”+ “.pdf”、“.py”、“.zip”)；
3. 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**；
4. 本次作业提交地址为`here`，请大家预留时间提前上交，以防 ddl 邻近因网络等原因，出现作业未能按时提交的情况。

1 [15pts] Derivatives of Matrices

有 $\alpha \in \mathbb{R}$，$\mathbf{y} \in \mathbb{R}^{m\times 1}$，$\mathbf{x} \in \mathbb{R}^{n\times 1}$，试完成下题，并给出计算过程。

(1) **[4pts]** 此问中假设 $\mathbf{A} \in \mathbb{R}^{n\times n}$，且 $\alpha = \mathbf{x}^\top \mathbf{A}\mathbf{x}$，试求 $\frac{\partial \alpha}{\partial \mathbf{x}}$ 。

(2) **[5pts]** 此问中假设 $\mathbf{A} \in \mathbb{R}^{m\times n}$，且 $\alpha = \mathbf{y}^\top \mathbf{A}\mathbf{x}$，同时 $\mathbf{y}$、$\mathbf{x}$ 为 $\mathbf{z}$ 的函数，试求 $\frac{\partial \alpha}{\partial \mathbf{z}}$。

(3) **[6pts]** 此问中假设 $\mathbf{A} \in \mathbb{R}^{n\times n}$ 且 $\mathbf{A}$ 可逆，$\mathbf{A}$ 为 α 的函数同时 $\frac{\partial \mathbf{A}}{\partial \alpha}$ 已知。试求 $\frac{\partial \mathbf{A}^{-1}}{\partial \alpha}$。

(**Hint:** 可以参考 The Matrix Cookbook。)

Solution. 此处用于写解答 (中英文均可)

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

(2) $\frac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{x}^\top \mathbf{A}^\top \frac{\partial \mathbf{y}}{\partial \mathbf{z}} + \mathbf{y}^\top \mathbf{A} \frac{\partial \mathbf{x}}{\partial \mathbf{z}}$.

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

2 [15pts] 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 2.3.3 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（四分类）的例子，并试着给出如何对多分类的预测结果。请回答如下问题。

表 1: 类别的真实标记与预测

真实类别 \ 预测类别	第一类	第二类	第三类	第四类
第一类	7	2	1	0
第二类	0	9	0	1
第三类	1	0	8	1
第四类	1	2	1	6

(1) **[5pts]** 如表1所示，请计算该学习器的错误率、精度。

(2) **[5pts]** 请分别计算宏查准率，宏查全率，微查准率，微查全率，保留三位有效数字，并两两比较大小，这种大小关系是否恒成立?（提示：凸不等式）

(3) **[5pts]** 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，并比较大小，这种大小关系是否恒成立?（提示：调和平均）

Solution. 此处用于写解答 (中英文均可)

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.

$$\text{Acc} = \frac{7+9+8+6}{7+2+1+9+1+1+8+1+2+1+1+6} = \frac{30}{40} = 0.75.$$

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, 第一类, 第二类，第三类，第四类作为正例, 以其他类作为负例, 可以得到以下四个二分类混淆矩阵:

真实类别 \ 预测类别	正类	反类
正类	7	3
反类	2	27

真实类别 \ 预测类别	正类	反类
正类	9	1
反类	4	27

真实类别 \ 预测类别	正类	反类
正类	8	2
反类	2	28

真实类别 \ 预测类别	正类	反类
正类	6	4
反类	2	28

依照教材公式, 可以计算以下指标:

$$\begin{aligned}
\text{Macro-P} &= \frac{1}{4}\sum_{i=1}^{3} P_i = \frac{1}{4} \times \left(\frac{7}{10} + \frac{9}{12} + \frac{8}{10} + \frac{6}{8}\right) = 0.755 \\
\text{Micro-P} &= \frac{\overline{TP}}{\overline{TP} + \overline{FP}} = \frac{(7+9+8+6)/4}{(7+9+8+6+3+3+2+2)/4} = 0.750 \\
\text{Macro-R} &= \frac{1}{3}\sum_{i=1}^{3} R_i = \frac{1}{3} \times \left(\frac{7}{10} + \frac{9}{10} + \frac{8}{10} + \frac{6}{10}\right) = 0.750 \\
\text{Micro-R} &= \frac{\overline{TP}}{\overline{TP} + \overline{FN}} = \frac{(7+9+8+6)/5}{(7+9+8+6+3+1+2+4)/5} = 0.750
\end{aligned}$$

依照教材公式, 可以计算以下指标:

$$\begin{aligned}
\text{Macro-F1} &= \frac{2 \times \text{ Macro-P } \times \text{ Macro-R}}{\text{Macro-P } + \text{ Macro-R}} = \frac{2 \times 0.755 \times 0.75}{0.755 + 0.75} = 0.752 \\
\text{Micro-F1} &= \frac{2 \times \text{ Micro-P } \times \text{ Micro-R}}{\text{Micro-P } + \text{ Micro-R}} = \frac{2 \times 0.75 \times 0.75}{0.75 + 0.75} = 0.75
\end{aligned}$$

3 [15pts] 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。

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	x_8	x_9
标记	0	1	0	1	0	0	1	1	0
分类器输出值	0.4	0.9	0.7	0.4	0.2	0.8	0.8	0.6	0.5

(1) **[5pts]** 如表2所示，第二行为样例对应的真实标记，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，并写出绘图中使用到的节点 (在坐标系中的) 坐标及其对应的阈值与样例编号。

(2) **[3pts]** 根据上题中的 ROC 曲线，计算其对应的 AUC 值（请给出具体的计算步骤）。

(3) **[7pts]** 结合前两问使用的例子 (可以借助图片示意)，试证明对有限样例成立:

$$\mathrm{AUC} = \frac{1}{m^+ m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left(\mathbb{I}\left\{ f(x^+) > f(x^-) \right\} + \frac{1}{2} \mathbb{I}\left\{ f(x^+) = f(x^-) \right\} \right) \tag{3.1}$$

Solution. 此处用于写解答 (中英文均可)

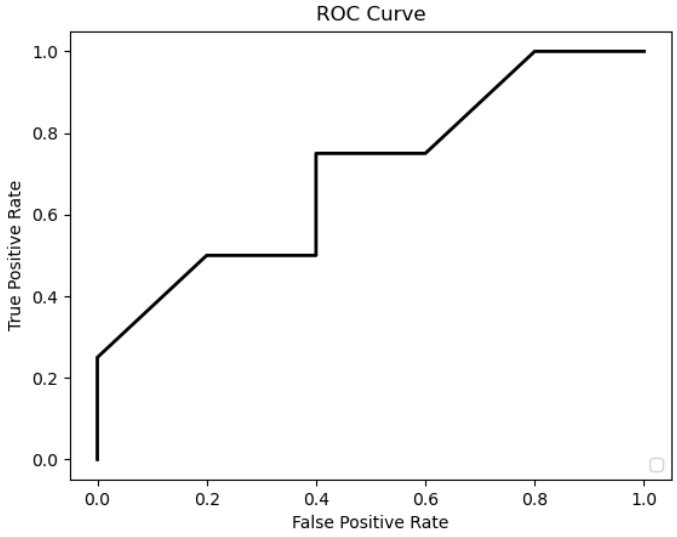

图 1: 分类器对应 ROC 曲线

(1) 本题所示的 ROC 曲线如图1所示。在绘图过程中使用到的节点坐标为:

- $(0, 0.25)$，对应阈值 0.9，对应样例 x_2;

- (0.2, 0.5)，对应阈值 0.8，对应样例 x_6 和 x_7；
- (0.4, 0.5)，对应阈值 0.7，对应样例 x_3；
- (0.4, 0.75)，对应阈值 0.6，对应样例 x_8；
- (0.6, 0.75)，对应阈值 0.5，对应样例 x_9；
- (0.8, 1.0)，对应阈值 0.4，对应样例 x_1 和 x_4；
- (1.0, 1.0)，对应阈值 0.2，对应样例 x_5.

(2) 如图1所示，可以计算此时的 AUC 值如下所示：

$$\text{AUC} = 0.2 \times \left(\frac{1}{2}(0.25 + 0.5) + 0.5 + 0.75 + \frac{1}{2}(0.75 + 1) + 1\right) = 0.7$$

(3) 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

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

4 [20pts] Linear Regression

线性回归模型是一类常见的机器学习方法，其基础形式与变体常应用在回归任务中。根据《机器学习》书第 3.2 节定义，可以将收集到的 d 维数据及其标签如下表示：

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} & 1 \\ x_{21} & x_{22} & \dots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^\top & 1 \\ \mathbf{x}_2^\top & 1 \\ \vdots & \vdots \\ \mathbf{x}_m^\top & 1 \end{pmatrix}; \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

将参数项与截距项合在一起，定义为 $\hat{\boldsymbol{w}} = \left(\boldsymbol{w}^\top; b\right)^\top$。此时成立 $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{w}}$。

《机器学习》式 (3.11) 给出了最小二乘估计 (Least Square Estimator, LSE) 的闭式解：

$$\hat{\boldsymbol{w}}^*_{\mathbf{LSE}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} \tag{4.1}$$

(1) **[8pts]** (投影矩阵的性质) 容易验证，当采用最小二乘估计 $\hat{\boldsymbol{w}}^*_{\mathbf{LSE}}$ 时，成立：

$$\widehat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{w}}^*_{\mathbf{LSE}} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

记 $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$，则有 $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$。$\mathbf{H}$ 被称为“Hat Matrix”，其存在可以从空间的角度，把 $\widehat{\mathbf{y}}$ 看作是 $\mathbf{y}$ 在矩阵 $\mathbf{H}$ 空间中的投影。$\mathbf{H}$ 矩阵有着许多良好的性质。

已知此时 $\mathbf{X}$ 矩阵列满秩，$\mathbf{I}$ 为单位阵，试求 $\mathbf{I}-\mathbf{H}$ 的全部特征值并注明特征值的重数。

(**Hint:** 利用 $\mathbf{H}$ 矩阵的投影性质与对称性。)

(2) **[5pts]** (岭回归) 当数据量 m 较小或数据维度 d 较高时，矩阵 $\mathbf{X}^\top\mathbf{X}$ 可能不满秩，4.1 中的取逆操作难以实现。此时可使用岭回归代替原始回归问题，其形式如下：

$$\hat{\boldsymbol{w}}^*_{\mathbf{Ridge}} = \arg\min_{\hat{\boldsymbol{w}}} \frac{1}{2}\left(\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{w}}\|_2^2 + \lambda\|\hat{\boldsymbol{w}}\|_2^2\right) \tag{4.2}$$

试求岭回归问题的闭式解，并简述其对原问题的改进。

(3) **[7pts]** 定义 $\tilde{\mathbf{x}}_i = \left(\mathbf{x}_i^\top; 1\right)^\top$，$\widehat{y}_i = \tilde{\mathbf{x}}_i^\top\hat{\boldsymbol{w}}^*_{\mathbf{LSE}}$，$\bar{y} = \frac{1}{m}\sum_{i=1}^m y_i$。

对线性回归模型进行统计分析时，会涉及如下三个基础定义：

$$\begin{cases} \text{Total sum of squares (TSS):} & \displaystyle\sum_{i=1}^m \left(y_i - \bar{y}\right)^2 \\ \text{Regression sum of squares (RSS):} & \displaystyle\sum_{i=1}^m \left(\widehat{y}_i - y_i\right)^2 \\ \text{Explained sum of squares (ESS):} & \displaystyle\sum_{i=1}^m \left(\widehat{y}_i - \bar{y}\right)^2 \end{cases}$$

试证明 TSS = RSS + ESS。(**Hint:** 使用向量形式可以简化证明步骤。)

Solution. 此处用于写解答 (中英文均可)

(1) 本题主要用到了 Hat Matrix 的两个性质：$\mathbf{H}^2 = \mathbf{H}$ 及 $\mathbf{H}^\top = \mathbf{H}$。

由 $\mathbf{H}^2 = \mathbf{H}$，可知对于矩阵 $\mathbf{H}$ 的所有特征值 λ 及其对应的特征向量 $\mathbf{x}$，成立：

$$\mathbf{H}^2\mathbf{x} = \mathbf{H}\mathbf{x} = \lambda\mathbf{x} \tag{4.3}$$

同时易推导：

$$\mathbf{H}^2\mathbf{x} = \mathbf{H}(\lambda\mathbf{x}) = \lambda\mathbf{H}\mathbf{x} = \lambda^2\mathbf{x} \tag{4.4}$$

综合4.3和4.4，可知成立：

$$(\lambda - \lambda^2)\mathbf{x} = 0$$

因此可解得 $\lambda = 0$ 或 1。现考虑其重数如下：

已知 $\mathbf{H}$ 为实对称矩阵，故可将其进行特征值分解如下：

$$\mathbf{H} = \mathbf{Q}\Lambda\mathbf{Q}^\top \tag{4.5}$$

因为 $\text{rank}(\mathbf{H}) = \text{rank}(\Lambda)$，且已知 Λ 的主对角元只能为 0 或 1。故可知 $\mathbf{H}$ 矩阵的秩即为特征值 1 的重数，为 $d+1$ 重（$\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{H}\mathbf{X}) \leq \text{rank}(\mathbf{H}) \leq \text{rank}(\mathbf{X})$，故 $\text{rank}(\mathbf{H}) = \text{rank}(\mathbf{X}) = d+1$）；因此 0 为 $m-d-1$ 重。

接着，利用4.5，可知矩阵 $\mathbf{I}-\mathbf{H}$ 可以特征值分解如下：

$$\mathbf{I} - \mathbf{H} = \mathbf{Q}\mathbf{I}\mathbf{Q}^\top - \mathbf{Q}\Lambda\mathbf{Q}^\top = \mathbf{Q}(\mathbf{I} - \Lambda)\mathbf{Q}^\top$$

因此，矩阵 $\mathbf{I}-\mathbf{H}$ 的特征值也是 0 和 1，且重数正好与矩阵 $\mathbf{H}$ 对应的重数反过来。

综上所述，矩阵 $\mathbf{I}-\mathbf{H}$ 的特征值为 0 ($d+1$ 重) 和 1 ($m-d-1$ 重)。

(2) 假设：

$$f(\hat{\boldsymbol{w}}) = \frac{1}{2}\left(\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{w}}\|_2^2 + \lambda\|\hat{\boldsymbol{w}}\|_2^2\right)$$

则4.2可写为：

$$\hat{\boldsymbol{w}}^*_{\mathbf{Ridge}} = \arg\min_{\hat{\boldsymbol{w}}} f(\hat{\boldsymbol{w}})$$

对 $f(\hat{\boldsymbol{w}})$ 求梯度，可得：

$$\begin{aligned}\frac{\partial f}{\partial \hat{\boldsymbol{w}}} &= -\mathbf{X}^\top\mathbf{y} + \left(\mathbf{X}^\top\mathbf{X}\right)\hat{\boldsymbol{w}} + \lambda\hat{\boldsymbol{w}} \\ &= -\mathbf{X}^\top\mathbf{y} + \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)\hat{\boldsymbol{w}}\end{aligned}$$

由 $\frac{\partial f}{\partial \hat{\boldsymbol{w}}} = 0$，可知：

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

岭回归针对原问题的主要改进在于，使得取逆操作一定可以成功，提升了模型鲁棒性；使用结构风险最小化，一定程度上降低了过拟合风险等（提到其中任何一点，或其他正确结论就可得分）。

(3) 本题使用向量表达可以简化证明。

可以将 TSS, RSS, ESS 用向量形式表示如下：

$$\begin{cases} \text{TSS} &= \sum_{i=1}^{m} (y_i - \bar{y})^2 = \left(\mathbf{y} - \mathbf{1} \cdot \frac{\mathbf{1}^\top \mathbf{y}}{m}\right)^\top \left(\mathbf{y} - \mathbf{1} \cdot \frac{\mathbf{1}^\top \mathbf{y}}{m}\right) = \mathbf{y}^\top \left(\mathbf{I} - \frac{\mathbf{1} \cdot \mathbf{1}^\top}{m}\right) \mathbf{y} \\ \text{RSS} &= \sum_{i=1}^{m} (\widehat{y}_i - y_i)^2 = (\mathbf{H}\mathbf{y} - \mathbf{y})^\top (\mathbf{H}\mathbf{y} - \mathbf{y}) = \mathbf{y}^\top (\mathbf{I} - \mathbf{H}) \mathbf{y} \\ \text{ESS} &= \sum_{i=1}^{m} (\widehat{y}_i - \bar{y})^2 = \left(\mathbf{H}\mathbf{y} - \mathbf{1} \cdot \frac{\mathbf{1}^\top \mathbf{y}}{m}\right)^\top \left(\mathbf{H}\mathbf{y} - \mathbf{1} \cdot \frac{\mathbf{1}^\top \mathbf{y}}{m}\right) = \mathbf{y}^\top \left(\mathbf{H} - \frac{\mathbf{1} \cdot \mathbf{1}^\top}{m}\right) \mathbf{y} \end{cases}$$

从三者的向量形式容易看出，成立 TSS = RSS + ESS。
关键在于 $\mathbf{H1} = \mathbf{1}$，因为 $\mathbf{1}$ 是 $\mathbf{X}$ 的最后一列。

5 [35pts] Logistic Regression in Practice

对数几率回归 (Logistic Regression, 简称 LR) 是实际应用中非常常用的分类学习算法。

(1) **[30pts]** 请编程实现二分类的 LR, 要求采用牛顿法进行优化求解。详细编程题指南请参见链接：`here`。

请将绘制好的 ROC 曲线放在解答处，并记录模型的精度与 AUC (保留 4 位小数)。

(2) **[5pts]** 试简述在对数几率回归中，相比梯度下降方法，使用牛顿法的优点和缺点。

Solution. 此处用于写解答 (中英文均可)

(1) AUC 值为 0.8323，精度为 0.7621。ROC 曲线绘制如下：

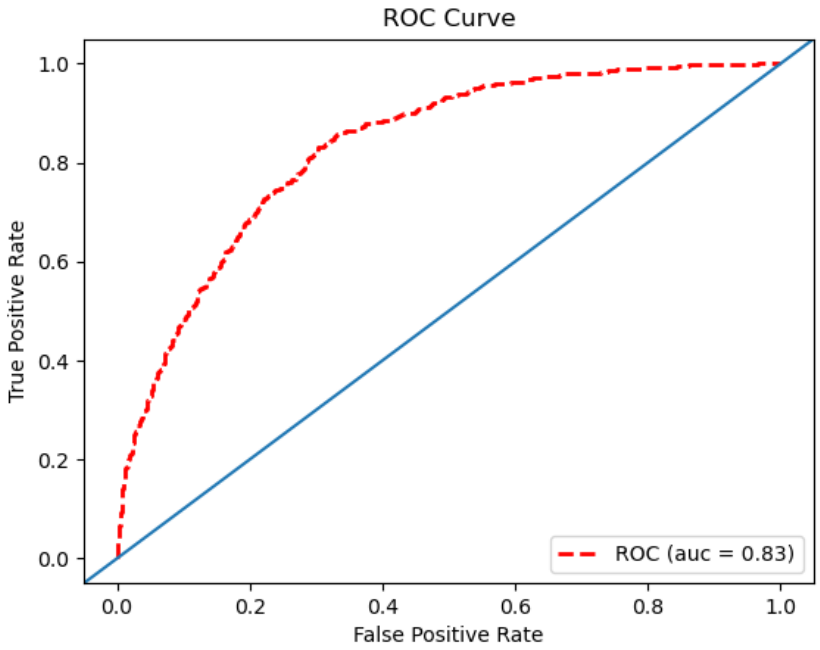

图 2: 对数几率回归模型在测试集上的 ROC 曲线

(2) 参考答案：

优点是利用了二阶信息，收敛轮数更少（说收敛速度快不准确）等。

缺点是每一步的计算量变大，取逆操作可能存在问题等。