机器学习导论 习题一

211300063, 张运吉, 211300063@smail.nju.edu.cn

2023 年 3 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 211300001_ 张三” + “.pdf”、“.py”、“.zip”);
3. 若多次提交作业, 则在命名 .zip 文件时加上版本号, 例如 211300001_ 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. 本次作业提交地址为 here, 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 [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 $\frac{\partial \mathbf{A}^{-1}}{\partial \alpha}$.

(提示: 可以参考 The Matrix Cookbook.)

Solution. 此处用于写解答 (中英文均可)

(1)

$$A = \begin{bmatrix} \alpha_1^T \\ \alpha_2^T \\ ... \\ \alpha_n^T \end{bmatrix}, Ax = \begin{bmatrix} \alpha_1^T x \\ \alpha_2^T x \\ ... \\ \alpha_n^T x \end{bmatrix} \tag{1.1}$$

则 $x^T Ax = \sum_{i=1}^{n} x_i \alpha_i^T x = a_{11}x_1^2 + a_{12}x_1x_2 + ... + a_{1n}x_1x_n + a_{21}x_1x_2 + a_{22}x_2^2 + ... + a_{2n}x_2x_n + ... + a_{n1}x_1x_n + a_{n2}x_2x_n + ... + a_{nn}x_n^2$

$\therefore \frac{\partial \alpha}{\partial \mathbf{x}_k} = \sum_{j=1}^{n} a_{kj}x_j + \sum_{i=1}^{n} a_{ik}x_i \therefore \frac{\partial \alpha}{\partial \mathbf{x}} = x^T(A^T + A)$

(2) 根据定义:$y^T Ax = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}y_i x_j$

$\therefore \frac{\partial \alpha}{\partial z_k} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}\left[y_i \frac{\partial x_j}{\partial z_k} + x_j \frac{\partial y_i}{\partial z_k}\right]$

$\therefore \frac{\partial \alpha}{\partial \mathbf{z}} = \frac{\partial \alpha}{\partial \mathbf{x}}\frac{\partial \mathbf{x}}{\partial \mathbf{z}} + \frac{\partial \alpha}{\partial \mathbf{y}}\frac{\partial \mathbf{y}}{\partial \mathbf{z}} = y^T A\frac{\partial \mathbf{x}}{\partial \mathbf{z}} + x^T A^T\frac{\partial \mathbf{y}}{\partial \mathbf{z}}$

(3) 由定义：$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$

上式两端对 α 求偏导：$\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial \alpha} + \frac{\partial \mathbf{A}^{-1}}{\partial \alpha}\mathbf{A} = \mathbf{0}$

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

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, 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

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及精度.

(2) **[5pts]** 请分别计算宏查准率, 宏查全率, 微查准率, 微查全率, 并两两比较大小.

(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.

Solution. 此处用于写解答 (中英文均可)

(1) 由图可知，样本总数为 40，分类错误的样本数为 10，所以错误率:$10/40 \times 100\% = 25\%$, 精度为:$1 - 25\% = 75\%$

(2) 把第一类当作正类，其他当作反类时：

$TP = 7, FN = 3, FP = 2, TN = 28$

$P = \frac{7}{9}, R = \frac{7}{10}$

把第二类当作正类，其他当作反类时：

$TP = 9, FN = 1, FP = 4, TN = 26$

$P = \frac{9}{13}, R = \frac{9}{10}$

把第三类当作正类，其他当作反类时：

$TP = 8, FN = 2, FP = 2, TN = 28$

$P = \frac{8}{10}, R = \frac{8}{10}$

把第四类当作正类，其他当作反类时：

$TP = 6, FN = 4, FP = 2, TN = 28$

$P = \frac{6}{8}, R = \frac{6}{10}$

所以，宏查准率和宏查全率分别为：

$macro-P=\frac{1}{4}(\frac{7}{9}+\frac{9}{13}+\frac{8}{10}+\frac{6}{8})\approx 0.7550$

$macro-R=\frac{1}{4}(\frac{7}{10}+\frac{9}{10}+\frac{8}{10}+\frac{6}{10})\approx 0.7500$

进一步计算可得:$\overline{TP}=7.5,\overline{FN}=2.5,\overline{FP}=2.5$

所以，微查准率和微查全率分别为：

$micro-P=\frac{\overline{TP}}{\overline{TP}+\overline{FP}}=0.7500$

$micro-R=\frac{\overline{TP}}{\overline{TP}+\overline{FN}}=0.7500$

macro-P $>$ micro-P, macro-R=micro-R

(3) $macro-F1=\frac{2\times macro-P\times macro-R}{macro-P+macro-R}\approx 0.7525$

$micro-F1=\frac{2\times micro-P\times micro-R}{micro-P+micro-R}=0.7500$

$macro-F1>micro-F1$

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:

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

Solution. 此处用于写解答 (中英文均可)

(1) 列出表格：

样例编号	none	2	6	7	3	8	9	1	4	5
FPR	0.0	0.0	0.2	0.2	0.4	0.4	0.6	0.8	0.8	1.0
TPR	0.0	0.25	0.5	0.5	0.5	0.75	0.75	1.0	1.0	1.0
阈值	1.0	0.9	0.8	0.8	0.7	0.6	0.5	0.4	0.4	0.2

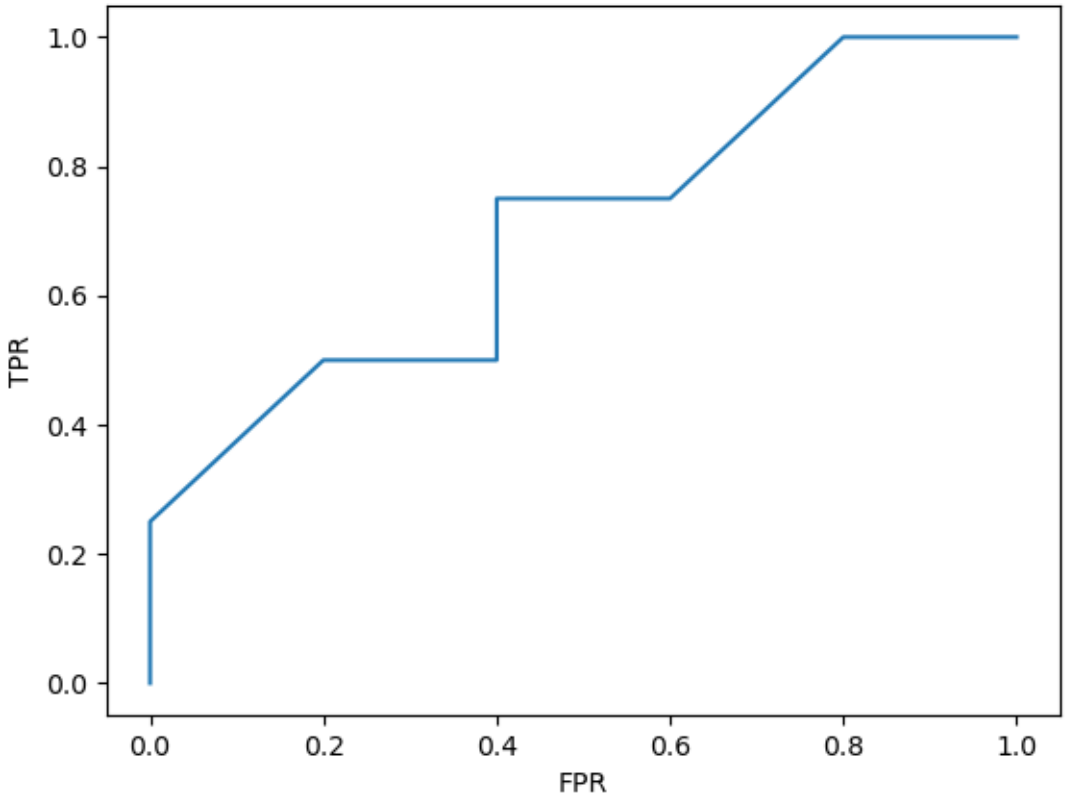

图 1: ROC

(2)

$$\begin{aligned} AUC &= \frac{1}{2}\sum_{i=1}^{7}(x_{i+1}-x_i)(y_{i+1}+y_i) \\ &= 0.075+0.1+0.15+0.175+0.2 \\ &= 0.7 \end{aligned} \tag{3.2}$$

(3) 根据 AUC 的计算公式可以得知，AUC 是累加 ROC 曲线上相邻两个点与 x 轴围成的梯形的面积。

考虑点 (x_i, y_i) 和 (x_{i+1}, y_{i+1})，若 $x_i = x_{i+1}$ 则对应梯形的面积为 0.

假设 $x_i \neq x_{i+1}$, 即表示阈值变化后假正例率变大了，不妨考虑 $x_{i+1} - x_i = \frac{1}{m^-}$ 时的情况：

此时梯形较短的底边长度为:

$$\sum_{x^+\in D^+}(\frac{1}{m^+}\mathbb{I}\left\{f(x^+) > f(x^-)\right\}) \tag{3.3}$$

较长的底边长度为：

$$\sum_{x^+\in D^+}(\frac{1}{m^+}\mathbb{I}\left\{f(x^+) > f(x^-)\right\} + \frac{1}{m^+}\mathbb{I}\left\{f(x^+) = f(x^-)\right\}) \tag{3.4}$$

梯形的面积:

$$\begin{aligned} S &= \frac{1}{2}(x_{i+1}-x_i)\left[\sum_{x^+\in D^+}(\frac{2}{m^+}\mathbb{I}\left\{f(x^+) > f(x^-)\right\} + \frac{1}{m^+}\mathbb{I}\left\{f(x^+) = f(x^-)\right\})\right] \\ &= \frac{1}{m^+m^-}\sum_{x^+\in D^+}(\mathbb{I}\left\{f(x^+) > f(x^-)\right\} + \frac{1}{2}\mathbb{I}\left\{f(x^+) = f(x^-)\right\}) \end{aligned} \tag{3.5}$$

因此，累加梯形的面积：

$$\begin{aligned} AUC &= \sum_{x^-\in D_-} S \\ &= \sum_{x^-\in D_-}\frac{1}{m^+m^-}\sum_{x^+\in D^+}(\mathbb{I}\left\{f(x^+) > f(x^-)\right\} + \frac{1}{2}\mathbb{I}\left\{f(x^+) = f(x^-)\right\}) \\ &= \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}(\mathbb{I}\left\{f(x^+) > f(x^-)\right\} + \frac{1}{2}\mathbb{I}\left\{f(x^+) = f(x^-)\right\}) \end{aligned} \tag{3.6}$$

证毕。

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》第三章 3.2 节中的定义, 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.

(提示: 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:

$$\begin{cases} \text{Total sum of squares (SST):} & \displaystyle\sum_{i=1}^m (y_i - \bar{y})^2 \\ \text{Regression sum of squares (SSR):} & \displaystyle\sum_{i=1}^m (\widehat{y}_i - y_i)^2 \\ \text{Residual sum of squares (SSE):} & \displaystyle\sum_{i=1}^m (\widehat{y}_i - \bar{y})^2 \end{cases}$$

试证明 SST = SSR + SSE. (提示: 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) $\because H^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T$

$\therefore H$ 是幂等矩阵

同理 $(I-H)^2 = I^2 - 2IH + H^2 = I - 2H + H = I - H$

$\therefore I - H$ 也是幂等矩阵

根据幂等矩阵的性质:$I - H$ 的特征值为 0 和 1.

$\because X$ 列满秩,$\therefore r(X) = d + 1$

$\therefore r(X^TX) = r(X) = d + 1$

$\therefore X^TX$ 是一个方阵且可逆

又 $\because X$ 列满秩，行满秩 $\therefore r(H) = d + 1$

$\therefore I - H$ 的特征值为 0(重数为 d+1) 和 1(重数为 m-d-1)

(2)

$$\begin{aligned} L(\hat{\boldsymbol{w}}) &= \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{w}}\|_2^2 + \lambda\|\hat{\boldsymbol{w}}\|_2^2 \\ &= (\mathbf{X}\hat{\boldsymbol{w}} - \mathbf{y})^T(\mathbf{X}\hat{\boldsymbol{w}} - \mathbf{y}) + \lambda\hat{\boldsymbol{w}}^T\hat{\boldsymbol{w}} \\ &= \hat{\boldsymbol{w}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{w}} - \mathbf{y}^T\mathbf{X}\hat{\boldsymbol{w}} - \hat{\boldsymbol{w}}^TX\mathbf{y} + y^T \end{aligned} \tag{4.3}$$

所以:

$$\frac{\partial L(\hat{\boldsymbol{w}})}{\partial \hat{\boldsymbol{w}}} = 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{w}} - 2X^Ty + 2\lambda\hat{\boldsymbol{w}} \tag{4.4}$$

令 $\frac{\partial L(\hat{\boldsymbol{w}})}{\partial \hat{\boldsymbol{w}}} = 0$，得：

$$\hat{\boldsymbol{w}} = (\mathbf{X}^T\mathbf{X} + \lambda E)^{-1}\mathbf{X}^T\mathbf{y} \tag{4.5}$$

岭回归在原问题得基础上增加了 $L2$ 正则项 $\lambda\hat{\boldsymbol{w}}^T\hat{\boldsymbol{w}}$，使得每个变量得权重不会太大。当某些特征权重比较大的时候，自变化变化一点，就会导致因变量变化很大，使得方差变大，有过拟合得风险。因此，岭回归在对局部特征较为明显的数据进行回归分析的时候有利于避免过拟合的现象，使得结果更为可靠。

(3)

$$\begin{aligned} SST &= \sum_{i=1}^{m}(y_i - \bar{y}_i)^2 \\ &= \sum_{i=1}^{m}(y_i - \widehat{y}_i + \widehat{y}_i - \bar{y}_i)^2 \\ &= \sum_{i=1}^{m}(\widehat{y}_i - y_i)^2 + \sum_{i=1}^{m}(\widehat{y}_i - \bar{y}_i)^2 + 2\sum_{i=1}^{m}(y_i - \widehat{y}_i)(\widehat{y}_i - \bar{y}_i) \end{aligned} \tag{4.6}$$

因此，欲证 SST=SSR+SSE, 只需证 $\sum_{i=1}^{m}(y_i - \widehat{y}_i)(\widehat{y}_i - \bar{y}_i) = 0$

已知 $\widehat{y}_i = \tilde{\mathbf{x}}_i^{\top}\hat{\boldsymbol{w}}^*_{\mathbf{LSE}}$ 因为采用最小二乘估计，所以 $\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}} = 2\mathbf{X}^T(\mathbf{X}\hat{\boldsymbol{w}} - \mathbf{y}) = 0$

$\mathbf{X}\hat{\boldsymbol{w}} - \mathbf{y} = 0$

$\widehat{y} - \mathbf{y} = 0$

$\therefore \sum_{i=1}^{m}(y_i - \widehat{y}_i)(\widehat{y}_i - \bar{y}_i) = 0$

SST=SSR+SSE, 证毕。

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) 如下图：

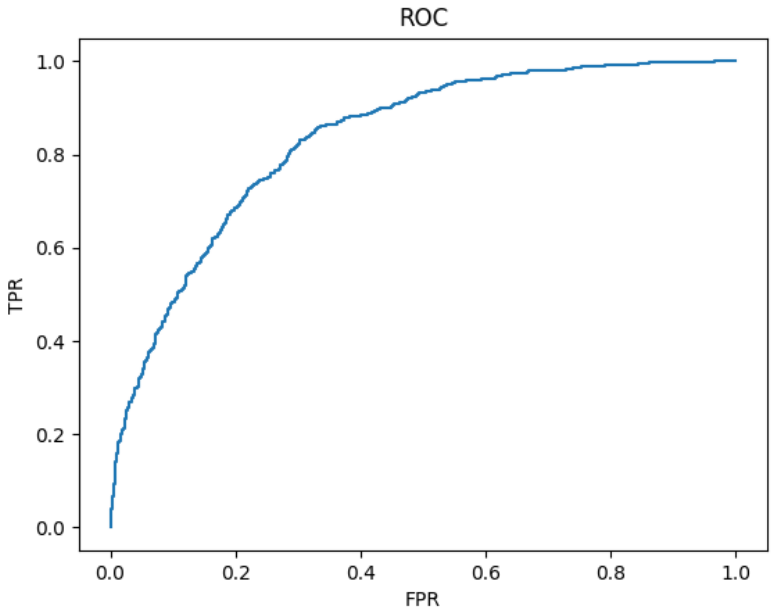

图 2: ROC of test set

模型精度：0.7621,AUC:0.8323

(2) 优点：牛顿迭代法只需进行二阶泰勒展开，收敛速度较快。

缺点：1. 涉及计算矩阵的逆，计算比较困难

2. 要求 Hessian 矩阵必须可逆

3. 局部收敛，若初始值 $\hat{\boldsymbol{w}}_0$ 选择不当可能会导致无法收敛