机器学习导论 习题一

211300044, 吴羽珩, 2559280859@qq.com

2023 年 3 月 20 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)
我们有 111122222

$$\frac{\partial(a^T b)}{\partial x} = \frac{\partial a^T}{\partial x} b + \frac{\partial b^T}{\partial x} a$$

(1)

$$\alpha = \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} x_y x_j$$

所以对于向量 $\mathbf{x}$ 的第 k 个元素 x_k, 有偏导数

$$\frac{\partial \alpha}{\partial x_k} = \sum_{j=1}^{n} a_{kj} x_j + \sum_{i=1}^{n} a_{ik} x_i$$

由此可得偏导数为:

$$\frac{\partial \alpha}{\partial x} = \frac{\partial (x^T \mathbf{A} x)}{\partial x} = \frac{\partial x^T}{\partial x} \mathbf{A} x + \frac{\partial x^T \mathbf{A}^T}{\partial x} x = (\mathbf{A} + \mathbf{A}^T) x$$

(2)

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

(3) 由定义: $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$
上式两端对 α 求偏导:

$$\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \alpha} + \frac{\partial \mathbf{A}^{-1}}{\partial \alpha} \mathbf{A} = 0$$

所以

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

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.

此处用于写解答 (中英文均可)

Solution.

(1) 学习器的精度为:$\frac{3}{4} = 75\%$, 学习器的错误率为 $1 -$ 精度 $= \frac{1}{4} = 25\%$

(2) 四个混淆矩阵:1 类, 非 1 类;2 类, 非 2 类;3 类, 非 3 类;4 类, 非 4 类

宏查准率为:

$$\text{macro-P} = \frac{1}{4}(\frac{7}{9} + \frac{9}{13} + \frac{8}{10} + \frac{6}{8}) = 0.755$$
$$\text{macro-R} = \frac{1}{4}(\frac{7}{10} + \frac{9}{10} + \frac{8}{10} + \frac{6}{10}) = 0.750$$

经过计算四个混淆矩阵对应位置元素的平均值后:

$\overline{TP} = 7.5$,$\overline{FP} = 2.5$,$\overline{FN} = 2.5$

所以有:

$$\text{micro-P} = 0.750 \text{micro-R} = 0.750$$

由此发现宏查准率大于宏查全率 微查准率和微查全率相等

(3) 由书上的公式:

$$\text{macro-F1} = 0.7525$$

$$\text{micro-F1} = 0.7500$$

宏 F1 度量大于微 F1 度量.

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:

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

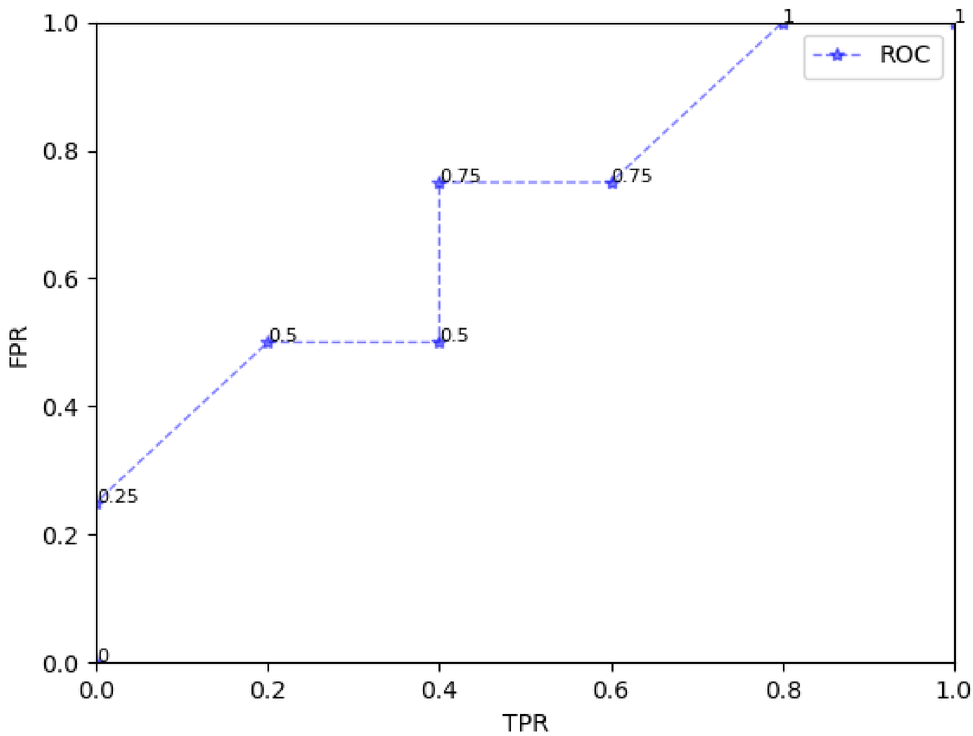

图 1: ROC 曲线图

Solution. 此处用于写解答 (中英文均可)

(1) 阈值与对应的样例编号分别为：

样例	NULL	x_2	x_7, x_7	x_3	x_8	x_9	x_1, x_4	x_5
阈值	1	0.9	0.8	0.7	0.6	0.5	0.4	0.2

(2) AUC 的值即为上述 ROC 曲线与 x 轴所围成的面积:

$$\begin{aligned} S &= \frac{1}{2}\sum_{i=1}^{8}(x_{i+1}-x_i)(y_i+y_{i+1}) \\ &= 0.7 \end{aligned}$$

(3) 注意到 AUC 的物理意义为正样本预测结果大于负样本预测结果的概率, 他反映了分类器对于样本的排序能力, 换言之, 就是随机拿出一个正样本和一个负样本, 正样本的预测结果比负样本的预测结果大的概率：如果我们把所有的正样本和负样本都比较一遍, 那么按照题干, 正样本有 m_+ 个, 负样本有 m_- 个, 那么一共有 $m_+ * m_-$ 对, 而题干中的双重求和符号 (可以理解为双重循环) 即遍历每一对, 统计正样本预测值大于负样本预测值的一共有多少对, 值得注意的是, 对于正样本预测值等于负样本预测值的情况记为 0.5, 所以我们得到了题干中的公式

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

计算证明: 在 ROC 曲线当中，每条横线都对应着至少一个标签为反例的样本，同样，每条垂线都对应着至少一个标签为正例的样本，斜线则对应着多个预测值相同的正或反样本. 每增加一个正例, 在 ROC 图中 y 轴的投影长度为 $\frac{1}{m^+}$, 每增加一个负例, 在 x 轴的投影长度为 $\frac{1}{m^-}$ 所以对于某一个梯形来说

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

所以所有梯形的面积和为 AUC 的值, 即为:

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

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:

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

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) 由于

$$\mathbf{H}^T = \mathbf{H}, \mathbf{H}^2 = \mathbf{H}$$

我们得知 H 矩阵是一个对称幂等矩阵, 同理

$$(\mathbf{I}-\mathbf{H})^T=\mathbf{I}-\mathbf{H},(\mathbf{I}-\mathbf{H})^2=\mathbf{I}-\mathbf{H}$$

可知 $\mathbf{I}-\mathbf{H}$ 也是一个对称幂等矩阵. 我们先看 $\mathbf{H}$ 矩阵, 利用对称幂等矩阵的性质可知 (特征值的定义易证), $\mathbf{H}$ 矩阵的特征值为 1,0.

$\mathbf{H}$ 是一个 $m\cdot m$ 的矩阵, 设 $\mathbf{H}$ 矩阵的秩为 $r(\mathbf{H})$, 则特征值 1 的重数为 $r(\mathbf{H})$, 特征值 0 的重数为 $m-r(\mathbf{H})$, 下求 $r(\mathbf{H})$:

由于矩阵 $\mathbf{X}$ 是一个列满秩的矩阵, 所以 $r(\mathbf{X})=d+1$, 由矩阵秩的性质 $r(\mathbf{X}^T\mathbf{X})=r(\mathbf{X})=d+1$ 且 $\mathbf{X}^T\mathbf{X}$ 是一个 $(d+1)\cdot(d+1)$ 的矩阵, 所以 $\mathbf{X}^T\mathbf{X}$ 是一个方阵且可逆, 又由于 $\mathbf{X}^T$ 行满秩,$\mathbf{X}$ 列满秩, 所以

$$\mathbf{H}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

的秩为 $d+1$

所以 $\mathbf{I}-\mathbf{H}$ 的特征值为 0(重数为 d+1), 1(重数为 m-d-1)

(2) 设

$$L(w)=(\mathbf{X}w-y)^T(\mathbf{X}w-y)+\lambda w^Tw$$

我们要求 $\hat{\boldsymbol{w}}$ minimize $L(w)$, 求

$$\frac{\partial L(w)}{\partial w}=2\mathbf{X}^T\mathbf{X}w-2\mathbf{X}^TY+2\lambda Ew$$

另其为 0, 解得

$$w=(\mathbf{X}^T\mathbf{X}+\lambda E)^{-1}\mathbf{X}^Ty$$

岭回归也是一种线性回归, 只不过在算法建立回归方程的时候加上正则化的限制, 从而达到解决过拟合的效果, 通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法。

(3) 将左侧 SST 变为

$$\sum_{i=1}^{m}(y_i-\widehat{y}_i+\widehat{y}_i-\bar{y})^2$$

经过化简整理得: 要证明 $SST=SSE+SSR$, 只须证

$$\sum(\widehat{y}_i-\bar{y}_i)(y_i-\widehat{y}_i)=0$$

最小二乘法的原理为将误差的平方和最小化, 令

$$\widehat{y}_i=\beta_0+\beta_1x_i$$

则

$$S=\sum(y_i-\widehat{y}_i)^2=\sum(y_i-\beta_0-\beta_1x_i)^2$$

我们就是要找到 β_0,β_1 使 S 最小. 令

$$\frac{\partial S}{\partial\beta_0}=0$$

化简后得到

$$\sum y_i - \widehat{y}_i = 0 \quad ,(1)$$

我们再求

$$\frac{\partial S}{\partial \beta_1} = 0$$

化简后可以得到

$$\sum x_i(y_i - \widehat{y}_i) = 0$$

又因为

$$\widehat{y}_i = \beta_0 + \beta_1 x_i$$

所以将

$$x_i = \frac{1}{\beta_1}(\widehat{y}_i - \beta_0)$$

带入上面的方程式整理得

$$\frac{1}{\beta_1}\sum \widehat{y}_i(y_i - \widehat{y}_i) - \frac{\beta_0}{\beta_1}\sum \widehat{y}_i(y_i - \widehat{y}_i) = 0$$

由 (1) 可知第二项为 0, 所以我们得到

$$\sum \widehat{y}_i(y_i - \widehat{y}_i) = 0 \quad ,(2)$$

最终根据 $(2) - \bar{y} * (1)$ 可得

$$\sum(\widehat{y}_i - \bar{y}_i)(y_i - \widehat{y}_i) = 0$$

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)

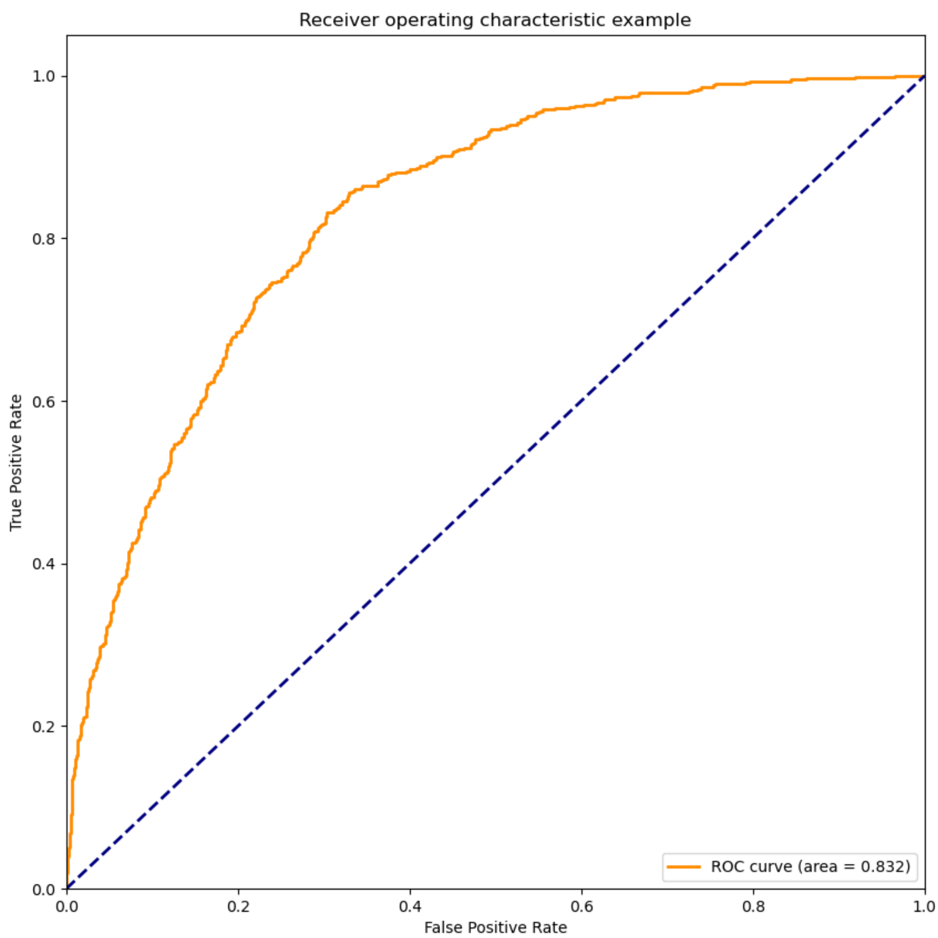

图 2: ROC of test set

Accuracy in test set: 0.7621

AUC in test set: 0.8323

(2) 牛顿法是通过求解目标函数的一阶导数为 0 时的参数, 进而求出目标函数最小值时的参数。

优点: 收敛速度很快,hessian 矩阵的逆在迭代过程中不断减小, 可以起到逐步减小步长的效果。

缺点:hessian 矩阵的逆计算复杂, 代价比较大。牛顿法是局部收敛的.

梯度下降法: 对于凸函数问题可以全局最优, 但是收敛速度比较慢