机器学习导论 习题一

211300024, 石睿, 211300024@smail.nju.edu.cn

2023 年 3 月 22 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) 解:

$\because \frac{\partial \alpha}{\partial x} = \nabla_x \alpha$, 即对 α 求偏导

$$\therefore \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{x}} = \frac{\partial(\boldsymbol{x^T A x})}{\partial \boldsymbol{x}} = \frac{\partial((a_{11}x_1x_1 + \cdots + a_{1n}x_1x_n) + \cdots + (a_{n1}x_nx_1 + \cdots + a_{nn}x_nx_n))}{\partial \boldsymbol{x}}$$

$$= \begin{pmatrix} \frac{\partial H}{x_1} \\ \frac{\partial H}{x_2} \\ \vdots \\ \frac{\partial H}{x_n} \end{pmatrix} \qquad \text{(注: 令上面那一大长串式子为 H)}$$

$$= \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} a_{11} & \cdots & a_{n1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \boldsymbol{Ax} + \boldsymbol{A^T x}$$

(2) 解:

令 $\boldsymbol{y^T}$=f , $\boldsymbol{Ax}$=g。故此时 $\alpha = f \cdot g$

$$\therefore \frac{\partial \alpha}{\partial z} = \frac{\partial f}{\partial z} \cdot g + f \cdot \frac{\partial g}{\partial z} = \frac{\partial f}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot g + y^T \cdot \frac{\partial g}{\partial x} \cdot \frac{\partial x}{\partial z} = 1 \cdot \frac{\partial y}{\partial z} \cdot Ax + y^T \cdot A^T \cdot \frac{\partial x}{\partial z}$$
$$= \frac{\partial y}{\partial z} \cdot Ax + (Ay)^T \cdot \frac{\partial x}{\partial z}$$

(3) 解:

$$\because E = A^{-1} \cdot A$$
$$\therefore 0 = \frac{\partial E}{\partial x} = \frac{\partial (A^{-1} \cdot A)}{\partial x} = \frac{\partial A^{-1}}{\partial x} \cdot A + A^{-1} \cdot \frac{\partial A}{\partial x}$$
$$\therefore \frac{\partial A^{-1}}{\partial x} = -A^{-1} \cdot \frac{\partial A}{\partial x} \cdot A^{-1}$$

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) 解:

$\because$ 由错误率和精度的定义

$$\text{错误率：}\ E(f;D)=\frac{1}{m}\cdot\sum_{i=1}^{m}I(f(x_i\neq y_i)\qquad \text{精度：}\ acc(f;D)=\frac{1}{m}\cdot\sum_{i=1}^{m}I(f(x_i)=y_i)$$

$\because$ m=9+13+10+8=40

$\therefore N(f(x_i)=y_i)=7+9+8+6=30,\qquad \mathrm{N(f(x}_i)\neq y_i)=10$

$\therefore acc(f;D)=\frac{3}{4}\qquad E(f;D)=\frac{1}{4}$

(2) 解:

对于多分类任务，每两类组合都对应一个混淆矩阵（二分类结果）

在本题中一共有 $C_4^2=6$ 种组合，他们的查准率和查全率分别为 P_i R_i

$$\left(\begin{array}{c|c}TP=7 & FN=2\\ \hline FP=0 & TN=9\end{array}\right)\left(\begin{array}{c|c}TP=7 & FN=1\\ \hline FP=1 & TN=8\end{array}\right)\left(\begin{array}{c|c}TP=7 & FN=0\\ \hline FP=1 & TN=6\end{array}\right)$$

$$\left(\begin{array}{c|c}TP=9 & FN=0\\ \hline FP=0 & TN=8\end{array}\right)\left(\begin{array}{c|c}TP=9 & FN=1\\ \hline FP=2 & TN=6\end{array}\right)\left(\begin{array}{c|c}TP=8 & FN=1\\ \hline FP=1 & TN=6\end{array}\right)$$

$\therefore$ 可以计算出每两类的 $P_i\quad R_i$ 以及 $\bar{TP}=\frac{47}{6}\quad \bar{FN}=\frac{5}{6}\quad \bar{FP}=\frac{5}{6}\quad \bar{TN}=\frac{43}{6}$

∴ 由查准率 P 以及查全率 R 的定义 $P = \frac{TP}{TP+FP}$　　$R = \frac{TP}{TP+FN}$ 可得

宏查准率

$$macro - P = \frac{1}{n} \cdot \sum_{i=1}^{n} P_i = \frac{2161}{2376}$$

宏查全率

$$macro - R = \frac{1}{n} \cdot \sum_{i=1}^{n} R_i = \frac{1959}{2160}$$

微查准率

$$micro - P = \frac{\bar{TP}}{\bar{TP} + \bar{FP}} = \frac{47}{52}$$

微查全率

$$micro - R = \frac{\bar{TP}}{\bar{TP} + \bar{FN}} = \frac{47}{52}$$

(3) 解:

宏 F_1 度量

$$macro - F_1 = \frac{2 \cdot macro - P \cdot macro - R}{macro - P + macro - R} \approx \frac{1.6497}{1.8164} \approx 0.9082 \quad \text{(保留了 4 位小数)}$$

微 F_1 度量

$$micro - F_1 = \frac{2 \cdot micro - P \cdot micro - R}{micro - P + micro - R} = \frac{47}{52} \approx 0.9039 \quad \text{(保留了 4 位小数)}$$

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) 解:

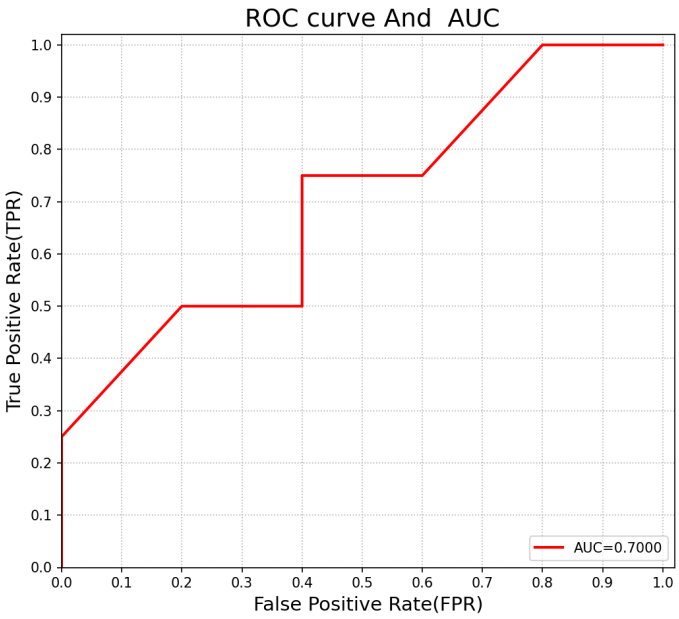
对数据重新整理如右表格所示

x_2	x_6 and x_7	x_3	x_8	x_9	x_1 and x_4	x_5
0.9	0.8 and 0.8	0.7	0.6	0.5	0.4 and 0.4	0.2
1	0 and 1	0	1	0	0 and 1	0

对阈值的所有情况进行遍历，结果如下表格所示

> 0.9	$0.8 - 0.9$	$0.7 - 0.8$	$0.6 - 0.7$	$0.5 - 0.6$	$0.4 - 0.5$	$0.2 - 0.4$	< 0.2
$TPR = 0$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{3}{4}$	1	1
$FPR = 0$	0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1

故对以上得到的 8 个节点 (FPR_i, TPR_i)，绘制到 ROC 中，如下图所示（利用 matplotlib）



(2) 解:

$$AUC = \frac{1}{2} \cdot \sum_{i=1}^{m-1} (x_{i+1} - xi) \cdot (y_i + y_{i+1}) \qquad \therefore AUC = \frac{3}{40} + \frac{1}{10} + \frac{3}{20} + \frac{7}{40} + \frac{1}{5} = \frac{7}{10}$$

(3) 解:

令原式为 A

$$\begin{aligned} A &= \frac{1}{m^+ \cdot m^-} \cdot \sum_{x^- \in D^-} \sum_{x^+ \in D^+} \left(I(f(x^+) > f(x^-)) + \frac{1}{2} \cdot I(f(x^+) = f(x^-)) \right) \\ &= \sum_{x^- \in D^-} \left(\frac{1}{m^+} \cdot \frac{1}{m^-} \cdot \sum_{x^+ \in D^+} I(f(x^+) > f(x^-)) + \frac{1}{2} \cdot \frac{1}{m^+} \cdot \frac{1}{m^-} \cdot \sum_{x^+ \in D^+} I(f(x^+) = f(x^-)) \right) \\ &= \sum_{x^- \in D^-} \frac{1}{2} \cdot \frac{1}{m^-} \cdot \left(\frac{2}{m^+} \cdot \sum_{x^+ \in D^+} I(f(x^+) > f(x^-)) + \frac{1}{m^+} \cdot \sum_{x^+ \in D^+} I(f(x^+) = f(x^-)) \right) \\ &= \sum_{x^- \in D^-} \frac{1}{2} \cdot \frac{1}{m^-} \cdot B \quad \text{（注：B 的定义如下式）} \end{aligned} \tag{3.2}$$

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

$\therefore$ 在公式 3.2 中,$\frac{1}{2}$ 是梯形公式中的【系数】，$\frac{1}{m^-}$ 是梯形公式中的【高】（横坐标单位长度）

$\therefore$ 在公式 3.3 中，左侧大括号中的内容为梯形的【上底】。表示在当前阈值下，真正例的个数。再乘以梯形的高（$\frac{1}{m^+}$），即表示当前阈值下，真正例率的大小。也即在 ROC 曲线下，

固定 FPR，来看 TPR 的大小。
∴ 在公式 3.3 中，右侧大括号中的内容为梯形的【下底】。表示在缩小了阈值后，原来的真正例的个数再加上阈值缩小后同时对真正例以及假正例的增加个数。再乘以梯形的高（$\frac{1}{m^+}$），即表示当前阈值下，真正例率的大小。也即在 ROC 曲线下，固定 TPR，来看 FPR 的大小。

$$\text{由 AUC 的定义} \qquad AUC = \frac{1}{2} \cdot \sum_{i=1}^{m-1} (x_{i+1} - x_i) \cdot \ y_i + y_{i+1} \tag{3.4}$$

∴，公式 3.2 和 3.4 都表达了 AUC 是 ROC 曲线和 x 轴（FPR 轴）围成的面积，所以原式得证

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) 解:

$\because$ 已知 H 是幂等矩阵，即$H^2 = H$

$\therefore$ 由矩阵特征值定义，设$H \cdot \xi = \lambda \cdot \xi$

$$\therefore \begin{cases} H^2 = H \cdot (H\xi) = \lambda \cdot H\xi = \lambda \cdot \lambda \cdot \xi = \lambda^2 \cdot \xi \\ H^2\xi = H\xi = \lambda \cdot \xi \end{cases}$$

$\therefore \lambda \cdot (\lambda - 1) \cdot \xi = 0 \; \xi \neq 0$

$\therefore \lambda = 0, 1$

$\because X$列满秩

$\therefore X^T$行满秩，即 X 的行向量是张成的空间的一组基，且$X^T X$满秩

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

$\therefore r(H) = d + 1$

$\therefore$ 由高等代数中的定理，H 是 m 阶方阵，相似于 r(H) 个 1 和 m-r(H) 个 0 组成主对角线的对角矩阵

$\therefore$ H 有 d+1 重特征值 1，m-d-1 重特征值 0

(2) 解:

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

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

$$\because \text{设} f(\hat{\boldsymbol{w}}) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\hat{\boldsymbol{w}} - \hat{\boldsymbol{w}}^T\mathbf{X}^T\mathbf{y} + \hat{\boldsymbol{w}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{w}} + \lambda\hat{\boldsymbol{w}}^T\hat{\boldsymbol{w}}$$

$$\therefore \text{令} \frac{\partial f}{\partial \hat{\boldsymbol{w}}} = -X^T\mathbf{y} - \mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{w}} + 2\lambda\hat{\boldsymbol{w}} = 0$$

$$\therefore \mathbf{X}^T\mathbf{y} = (\lambda I + \mathbf{X}^T\mathbf{X})\hat{\boldsymbol{w}}$$

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

把最小二乘法（对回归问题的无偏估计）转化为岭回归问题（有偏估计方法）

【1】以损失部分信息为代价, 通过引入正则化项，使原问题有闭式解。

【2】可变参数的引入，可以通过对可变参数的变化，看“岭迹”在哪个值的时候比较稳定，以获得更符合实际，更可靠的回归方法。解决了因 $\mathbf{x}^T\mathbf{X}$ 行列式的值接近于 0，而计算其误差很大的情况发生。

(2) 解:

$$
\begin{aligned}
SST &= \sum_{i=1}^{m}(y_i-\bar{y})^2 = \sum_{i=1}^{m}((y_i-\hat{y}_i)+(\hat{y}_i-\bar{y}))^2 \\
&= \sum_{i=1}^{m}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{m}(\hat{y}_i-\bar{y})^2 + 2\sum_{i=1}^{m}((y_i-\hat{y}_i)\cdot(\hat{y}_i-\bar{y})) \\
&= SSR + SSE + 2\sum_{i=1}^{m}((y_i-\hat{y}_i)\cdot(\hat{y}_i-\bar{y}))
\end{aligned}
$$

以下说明上式右侧分项为 0 即可。

即说明 $\sum_{i=1}^{m}((\hat{y}_i-y_i)\cdot(\hat{y}_i-\bar{y})) = \sum_{i=1}^{m}(\hat{y}_i-y_i)\cdot\hat{y}_i - \sum_{i=1}^{m}(\hat{y}_i-y_i)\cdot\bar{y} = 0$

$$
\sum_{i=1}^{m}(\hat{y}_i-y_i)\cdot\hat{y}_i = \hat{\mathbf{y}}^T(\hat{\mathbf{y}}-\mathbf{y}) = (\mathbf{X}^T H\mathbf{y})^T\cdot(\mathbf{X}^T H\mathbf{y}-y) = 0 \tag{4.3}
$$

$$
\begin{aligned}
&\bar{y} = \frac{1}{m}\cdot\sum_{i=1}^{m}y_i = \bar{\mathbf{X}^T}H\mathbf{y} = \frac{1}{m}\sum_{i=1}^{m}\tilde{x_i}^T H\mathbf{y} \\
&\sum_{i=1}^{m}(\hat{y}_i-y_i)\cdot\bar{y} = \bar{y}\cdot(\sum_{i=1}^{m}\hat{y}_i - \sum_{i=1}^{m}y_i) = \bar{y}\cdot(\sum_{i=1}^{m}\hat{y}_i - m\bar{\mathbf{X}^T}H\mathbf{y}) = \bar{y}\cdot(\sum_{i=1}^{m}\hat{y}_i - m\bar{\mathbf{X}^T}H\mathbf{y}) \\
&= \bar{y}\cdot(\sum_{i=1}^{m}\hat{y}_i - \sum_{i=1}^{m}\tilde{x}_i H\mathbf{y}) = 0
\end{aligned} \tag{4.4}
$$

由 4.3 和 4.4 式两个 =0 的式子，可以得到 SST=SSR+SSE

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).

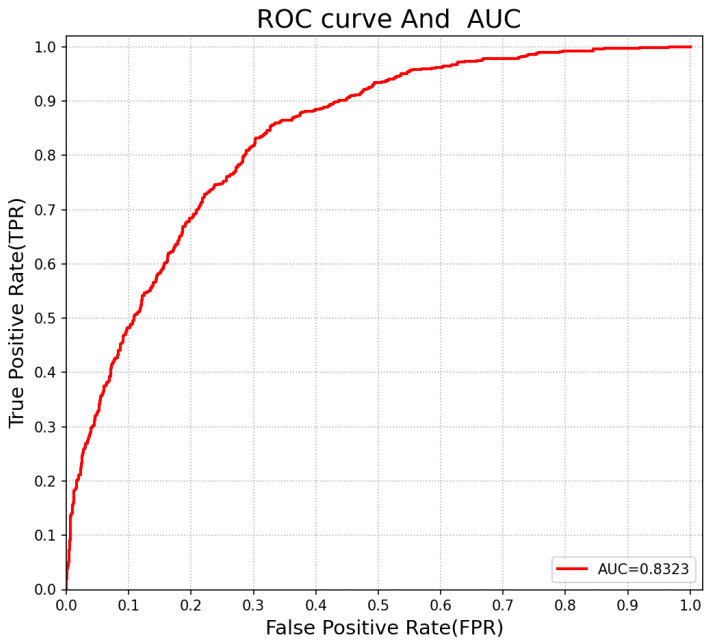
(2) **[5pts]** 试简述在对数几率回归中, 相比梯度下降方法, 使用牛顿法的优点和缺点.

Solution. 此处用于写解答 (中英文均可)

(1) 解:

模型精度：0.7626，此时阈值设定为 0.44，因为在训练集的 label 中$\frac{m^+}{m^-} = 0.44$

AUC:0.8323



(2) 解:

牛顿法 vs 梯度下降

优点：牛顿法中的逼近方式是二阶收敛的，即用二次曲面拟合所处位置。而梯度下降将是一阶收敛。

【牛顿法在对模拟精度有要求的时候可以更快做到】

缺点：牛顿法中的逼近公式中每一步需要计算 Hessian 矩阵的逆矩阵，并且和梯度做乘法，计算复杂而梯度下降只需要计算梯度

【牛顿法计算复杂度较高，不适合特征值较多的情况】