机器学习导论 习题一

学号, 姓名, 邮箱

2024 年 3 月 13 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. 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 $\alpha \in \mathbb{R}$, $\boldsymbol{y} \in \mathbb{R}^{m\times 1}$, $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$ 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 $\alpha = \boldsymbol{x}^\top \mathbf{A}\boldsymbol{x} + \mathbf{b}^\top \boldsymbol{x}$, 试求 $\frac{\partial \alpha}{\partial \boldsymbol{x}}$.

(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 $x_1, \cdots, x_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. 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 $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$ 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)

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)

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)

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):

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

Solution. 此处用于写解答 (中英文均可)

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)

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.