机器学习导论 习题五

221300079, 王俊童, 221300079@smail.nju.edu.cn

2024 年 6 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 — `HW5.tex`;
   (b) 由 (a) 编译得到的 PDF 文件 — `HW5.pdf`;
   (c) 第四题 AdaBoost 代码 — `AdaBoost.py`;
   (d) 第四题 Random Forest 代码 — `RandomForest.py`;
   (e) 第四题绘图代码 — `main.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 **6 月 14 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] Naive Bayesian

朴素贝叶斯是一种经典的生成式模型. 请仔细学习《机器学习》第七章 7.3 节, 并完成下题.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	2	2	2	2	3	3	3	3	3	3	3	3
$X^{(2)}$	S	M	L	S	M	L	S	M	L	S	M	L	S	M	L
Y	1	1	-1	1	-1	-1	1	1	1	-1	1	1	-1	1	1

(1) **[10pts]** 使用表 1的数据训练朴素贝叶斯模型. 给定新的输入 $x = (2, M)^\top$, 试计算 $\mathbb{P}(x = 1)$ 以及 $\mathbb{P}(x = -1)$, 并判断该样本应当被分为哪一类.

(2) **[10pts]** 若使用"拉普拉斯修正"训练模型, 对于新输入 $x = (2, M)^\top$, 试计算此时的 $\mathbb{P}_\lambda(x = 1)$ 以及 $\mathbb{P}_\lambda(x = -1)$, 并判断此时该样本应当被分为哪一类.

(3) **[5pts]** 根据以上结果, 试讨论在朴素贝叶斯模型中, 使用"拉普拉斯修正"带来的好处与影响.

Solution. （1）解答如下，参考书上的 7.3 的解答过程，此处给出两个 feature：
$P(Y = 1) = \frac{10}{15}, P(Y = -1) = \frac{5}{15}$
$P(X^{(1)} = 2|Y = 1) = \frac{2}{10}, P(X^{(1)} = 2|Y = -1) = \frac{2}{5}$
$P(X^{(2)} = M|Y = 1) = \frac{4}{10}, P(X^{(2)} = M|Y = -1) = \frac{1}{5}$
所以可得：
$P(Y = 1) = \frac{10}{15} * \frac{2}{10} * \frac{4}{10} = \frac{4}{75} \approx 0.053$
$P(Y = -1) = \frac{5}{15} * \frac{2}{5} * \frac{1}{5} = \frac{2}{75} \approx 0.026$
$P(Y = 1) > P(Y = -1)$.
所以预测为正类
（2）解答如下，参考书上的 7.3 的后的拉普拉斯修正的解答过程，此处给出两个 feature：
$P(Y = 1) = \frac{10+1}{15+2} = \frac{11}{17}, P(Y = -1) = \frac{5+1}{15+2} = \frac{6}{17}$
$P(X^{(1)} = 2|Y = 1) = \frac{2+1}{10+3} = \frac{3}{13}, P(X^{(1)} = 2|Y = -1) = \frac{2+1}{5+3} = \frac{3}{8}$
$P(X^{(2)} = M|Y = 1) = \frac{4+1}{10+3} = \frac{5}{13}, P(X^{(2)} = M|Y = -1) = \frac{1+1}{5+3} = \frac{2}{8}$
所以可得：
$P(Y = 1) = \frac{11}{17} * \frac{3}{13} * \frac{5}{13} = \frac{165}{2873} \approx 0.057$
$P(Y = -1) = \frac{6}{17} * \frac{3}{8} * \frac{2}{8} = \frac{9}{272} \approx 0.033$
$P(Y = 1) > P(Y = -1)$.
所以预测为正类
（3）1. 拉普拉斯修正避免了一些样本不充分从而估值为 0 的情况（虽然在这个题目里面没有体现罢了）。先验概率不为零，可以提高准确率。
2. 拉普拉斯修正可以处理稀疏数据，在一些特征空间大的情况下，可以提高鲁棒性，避免了一些极端数据和情况。
3. 而且拉普拉斯修正改进了模型的性能，考虑了未出现过的类别，提高了性能，比如这个题，我们可以看出的是由于拉普拉斯修正这个值后面是整体的上升了。确实是产生了一定的偏差，可能会产生一种偏向。但是就这个题目来说，明显对于较小的值得改变更大。但确

实这个方法使得样本估计更加趋于实际值了，更加平滑。
4. 拉普拉斯修正可以影响极端的概率值，一些频繁出现的，概率影响比较小，但有一些没怎么出现的，概率影响大，这个题就可以看出来。
5. 当数据非常稀疏或者样本量很小的时候，拉普拉斯修正还能够影响最终的类别判断先验概率。
6. 拉普拉斯修正增强了假设特征和类别独立，但这个不总是成立，可能影响性能。

2 [25pts] Nearest Neighbor

假设数据集 $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ 是从一个以 $\mathbf{0}$ 为中心的 p 维单位球中独立均匀采样而得到的 n 个样本点. p 维单位球可以表示为:

$$B = \{\mathbf{x} : \|\mathbf{x}\|^2 \leq 1\} \subset \mathbb{R}^p. \tag{2.1}$$

其中, $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$, $\langle \mathbf{x}, \mathbf{x} \rangle$ 是 $\mathbb{R}^p$ 空间中向量的内积. 在本题中, 我们将探究原点 O 与其最近邻 (1-NN) 的距离 d^*, 以及 d^* 与 p 之间的关系. O 与其 1-NN 之间的距离定义为:

$$d^* := \min_{1 \leq i \leq n} \|\mathbf{x}_i\|, \tag{2.2}$$

不难发现 d^* 是一个随机变量, 因为 $\mathbf{x}_i$ 是随机产生的.

(1) **[5pts]** 当 $p = 1$ 且 $t \in [0, 1]$ 时, 请计算 $\mathbb{P}(d^* \leq t)$, 即随机变量 d^* 的累积分布函数 (Cumulative Distribution Function, **CDF**).

(2) **[7pts]** 请写出 d^* 的 **CDF** 的一般公式, 即当 $p \in \{1, 2, 3, ...\}$ 时 d^* 对应的取值.

(**Hint:** 半径为 r 的 p 维球体积是: $V_p(r) = \frac{(r\sqrt{\pi})^p}{\Gamma(p/2+1)}$, 其中, $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(1) = 1$, 且有 $\Gamma(x+1) = x\Gamma(x)$ 对所有的 $x > 0$ 成立; 并且对于 $n \in \mathbb{N}^*$, 有 $\Gamma(n+1) = n!$.)

(3) **[8pts]** 请求解随机变量 d^* 的中位数, 请写成关于 n 和 p 的函数.

(**Hint:** 即使得 $\mathbb{P}(d^* \leq t) = 1/2$ 成立时的 t 值)

(4) **[5pts]** 结合以上问题, 谈谈你关于 n 和 p 以及它们对 1-NN 的性能影响的理解.

Solution. (1) 考虑 p=1 的时候其实一个在一条线上面的一个区间。所以由于 X_i 这个东西在 [-1,1] 上面是一个独立且均匀的采样, 我们可以得到:

$$P(d^* \leq t) = 1 - P(|X_i| > t)$$

又由于乘法原则, 原式子可以化简为:

$$P(d^* \leq t) = 1 - (1-t)^n$$

(2) 考虑一个 p 维的球, 根据 hint 可以得到

$$V_p(r) = \frac{(r\sqrt{\pi})^p}{\Gamma(\frac{p}{2}+1)}$$

这个问题的 p 维度可以考虑成一个大球包含了一个小球, 数据分布可能在这个 p 维度的两个球之间的距离, 那么这个分布可以表示为:

$$\frac{V_p(r)}{V_p(1)} = \frac{(r\sqrt{\pi})^p}{\Gamma(\frac{p}{2}+1)} * \frac{(\Gamma(\frac{p}{2}+1)}{(\sqrt{\pi})^p} = t^p$$

根据第一问的乘法原则, 带入可以得到:

$$P(d^* \leq t) = 1 - (1-t^p)^n$$

(3) 根据 hint 可以得到:

$$P(d* \le t) = \frac{1}{2}$$

进行反解即可得到:

$$1 - (1 - t^p)^n = \frac{1}{2}$$

$$t = \sqrt[p]{1 - \sqrt[n]{\frac{1}{2}}}$$

(4)n 代表的数据的样本数量，而 p 代表的维度。

1. 其实可以看出，如果 p 越大，维度越高，采样更密集, 在高维空间中的 1-NN 效果不会特别的好。样本更稀疏，距离更远。有可能需要更多的样本来支撑。计算效率也没有那么高。

2. 如果 n 比较小，如果有一些干扰或者噪声数据，会使得 1-NN 的效果变差，因为只有少部分样本可以拿来进行运算。n 小了对于一定维度下来说计算负担也是不小的，n 小了有时候也不是好事，比如如果距离远，这种计算开销还是大。

反正一句话，这个 n 和这个 p 呢，还是要稍微的合理才行，既不能太小了，有可能效果不好，也不能太大了，这样的话可能会增加计算负担，应该合理选择。

3 [25pts] K-means and EM Algorithm

EM (Expectation-Maximization) 算法是存在“未观测”变量的情况下估计参数隐变量的利器. 请仔细阅读《机器学习》第九章以及第七章 7.6 节, 回答以下问题.

3.1 [10pts] K-means and GMM

在《机器学习》9.4.3 节中, 我们在聚类问题下推导了高斯混合模型 (GMM) 的 EM 算法, 即高斯混合聚类. 沿用该小节中的记号, 我们考虑一种简化后的高斯混合模型, 其中高斯混合分布共由 k 个混合成分组成, 且每个混合成分拥有相同的协方差矩阵 $\Sigma_i = \varepsilon^2\mathbf{I}, i \in [k]$. 假设 $\exists\delta > 0$ 使得对于选择各个混合成分的概率有 $\alpha_i \geq \delta, \forall i \in [k]$, 并且在高斯混合聚类的迭代过程中始终有 $\|\boldsymbol{x}_i - \mu_k\|^2 \neq \|\boldsymbol{x}_i - \mu_{k'}\|^2$ for $\forall i \in [n], k \neq k'$ 成立.

(1) **[10pts]** 请证明: 随着 $\varepsilon^2 \to 0$, 高斯混合聚类中的 **E** 步会收敛至 k 均值聚类算法中簇划分的更新规则, 即每个样本点仅指派给一个高斯成分. 由此可见, k 均值聚类算法是高斯混合聚类的一种特例.

3.2 [15pts] EM for Survival Analysis

生存分析 (Survival Analysis) 是一类重要的研究问题. 考虑如下图.(不给图 latex 编译不了，我把图删了..) 所示场景, 医院收集了病人接受治疗后的生存时间数据, 并在时刻 $T = a$ 停止了收集. 假设病人接受治疗后的生存时间服从正态分布分布 $\mathcal{N}(\theta, 1)$. 若一共有 m 个病人参与实验, 其中在 $T = a$ 之前死亡的人数为 n, 收集其生存时间数据为 $\mathbf{X} = \{x_1, \cdots, x_n, \underbrace{a, \cdots, a}_{m-n \text{ 个} a}\}$, 希望使用 EM 算法估计 θ.

(2) **[10pts] E** 步: (**Hint:** observed dataset **X** implies that $z_i \geq a, i = 1, \cdots, m-n$.)

(a) **[2pts]** 记 $\mathcal{N}(0,1)$ 的 CDF 为 $\Phi(\cdot)$, 直接写出似然函数 $L(\mathbf{X};\theta)$.

(b) **[3pts]** 记未观测生存时间数据为 $\mathbf{Z} = \{z_1, \cdots, z_{m-n}\}$. 试求对数似然函数 $\log L(\mathbf{X}, \mathbf{Z}; \theta)$.

(c) **[5pts]** 试求 $f(z_i \mid \mathbf{X}, \theta_t)$, 并依此写出 $Q(\theta \mid \theta_t)$.

(3) **[5pts] M** 步: 记 $\mathcal{N}(0,1)$ 的 PDF 为 $\phi(\cdot)$, 试求 θ 的更新公式 (使用 $\phi(\cdot)$, $\Phi(\cdot)$ 表示).

Solution. （1）根据高斯混合聚类的定义，其 E 步为:

$$p_M(z_j = i|x_j) = \frac{P(z_j = i)p_M(x_j|z_j = i)}{p_M(x_j)}$$

$$\gamma_{ji} = p_M(z_j = i|x_j) = \frac{\alpha_i\mathcal{N}(x_i|\mu_i, \epsilon^2\mathbf{I})}{\sum_{j=1}^k \alpha_j\mathcal{N}(x_i|\mu_j, \epsilon^2\mathbf{I})}$$

当 $\epsilon^2 \to 0$, 这个高斯分布会变得特别的尖，所有的值基本都靠近在 μ 附近，所以对于任意的 x_i, 当 μ_i 是所有 μ_j 中离这个 x_i 最近的哪一个是，这个 $\mathcal{N}(x_i|\mu_i, \epsilon^2\mathbf{I})$ 会特别的大，反正比起来其他的 $\mathcal{N}(x_i|\mu_j, \epsilon^2\mathbf{I})$ 会特别明显。
又由于这个题目要求 $\alpha_i \geq \delta > 0$，且保证 $\|\boldsymbol{x}_i - \mu_k\|^2 \neq \|\boldsymbol{x}_i - \mu_{k'}\|^2$ for $\forall i \in [n], k \neq k'$, 这个 γ_{ji} 将趋向于 1 当且仅当最近的那一个的时候，其余情况都是趋近于 0. 根据 9.4.1 中的

聚类簇的划分方法，这个确实是相等的，找的是相距最近的那一个。随着这个 $\epsilon^2 \to 0$ 可以看出高斯混合聚类的 E 步会收敛到 k 均值聚类算法中簇划分的规则。所以可以得到 k 均值聚类算法是高斯混合聚类的一个特例。

(2)

(a) 似然函数如下:

$$L(\mathbf{X};\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}} \cdot \left[\prod_{j=n+1}^{m} (1-\Phi(a-\theta))\right]$$

(b) 首先根据这个写出似然函数:

$$L(\mathbf{X},\mathbf{Z};\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}} \cdot \prod_{j=1}^{m-n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z_j-\theta)^2}{2}}$$

对上面的式子取对数:

$$\log L(\mathbf{X},\mathbf{Z};\theta) = \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi) - \frac{(x_i-\theta)^2}{2}\right] + \sum_{j=1}^{m-n} \left[-\frac{1}{2}\log(2\pi) - \frac{(z_j-\theta)^2}{2}\right]$$

(c) 首先是这个 f 的理解，如下所示:

$$f(z_i \mid \mathbf{X},\theta_t) = \frac{f(z_j)}{P(z_j \geq a)}$$

$$f(z_i \mid \mathbf{X},\theta_t) = \frac{\phi(z_i-\theta_t)}{1-\Phi(a-\theta_t)}, \quad z_i \geq a$$

其中这个 ϕ 代表 PDF。

所以根据 EM 算法可以得到相应的 Q。

$$Q(\theta \mid \theta_t) = E_{\mathbf{Z}|\mathbf{X},\theta_t}[\log L(\mathbf{X},\mathbf{Z};\theta)] = \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi) - \frac{(x_i-\theta)^2}{2}\right] + \sum_{j=1}^{m-n} E_{z_j|\mathbf{X},\theta_t} \left[-\frac{1}{2}\log(2\pi) - \frac{(z_j-\theta)^2}{2}\right]$$

(3) 根据 M 步的更新公式，首先对于 θ 求导可以得到我们想要的最大化。

$$0 = \frac{dQ(\theta|\theta_t)}{d\theta} = \sum_{i=1}^{n} -(x_i-\theta) + \sum_{j=1}^{m-n} E_{z_j|\mathbf{X},\theta_t}[-(z_j-\theta)]$$

解上述方程得到 θ 的更新公式。然后计算在条件密度 $f(z_i \mid \mathbf{X},\theta_t)$ 下 z_i 的期望。根据截断正态分布的性质，该条件期望可以表示为:

$$E_{z_i|\mathbf{X},\theta_t}[z_i] = \frac{\int_a^\infty z_i\phi(z_i)dz_i}{1-\Phi(a-\theta_t)} = \theta_t + \frac{\phi(a-\theta_t)}{1-\Phi(a-\theta_t)}$$

带入 $Q(\theta|\theta_t)$ 的导数中并计算得到 θ 的更新公式为:

$$\theta_{t+1} = \frac{\sum_{i=1}^{n} x_i + (m-n)(\theta_t + \frac{\phi(a-\theta_t)}{1-\Phi(a-\theta_t)})}{m}$$

通过迭代这个公式，我们就可以更新参数 θ 直至收敛。

4 [25pts] Ensemble Methods

在本题中, 我们尝试使用 AdaBoost 与 Random Forest 这两种经典的集成学习的方法进行分类任务. 本次实验使用的数据集为 UCI 二分类数据集 Adult (Census Income).
关于编程题的详细说明, 请参考: 编程题指南.pdf.

(1) **[10pts]** 请参考《机器学习》中对 AdaBoost 与 Random Forest 的介绍, 使用决策树作为基分类器, 实现 AdaBoost 分类器与 Random Forest 分类器.

(2) **[10pts]** 请基于上述实现, 通过 5-折交叉验证, 探究基学习器数量对集成效果的影响. (请在报告中附上绘制的折线图, 并简要论述分类器数量对分类效果的影响.)

(3) **[5pts]** 请分别汇报最优超参数 (即: 基学习器数量) 下, 两种模型在测试集上的 AUC 指标 (结果保留三位小数).

Solution. （1）我已在代码里面实现所有功能。
Random Forest AUC = 0.898
AdaBoost AUC = 0.745
（2）绘制的 20 轮的基学习器图如下:
效果分析，发现在基学习器一开始上升的时候，最低的 AUC 是一轮训练的结果，但是随

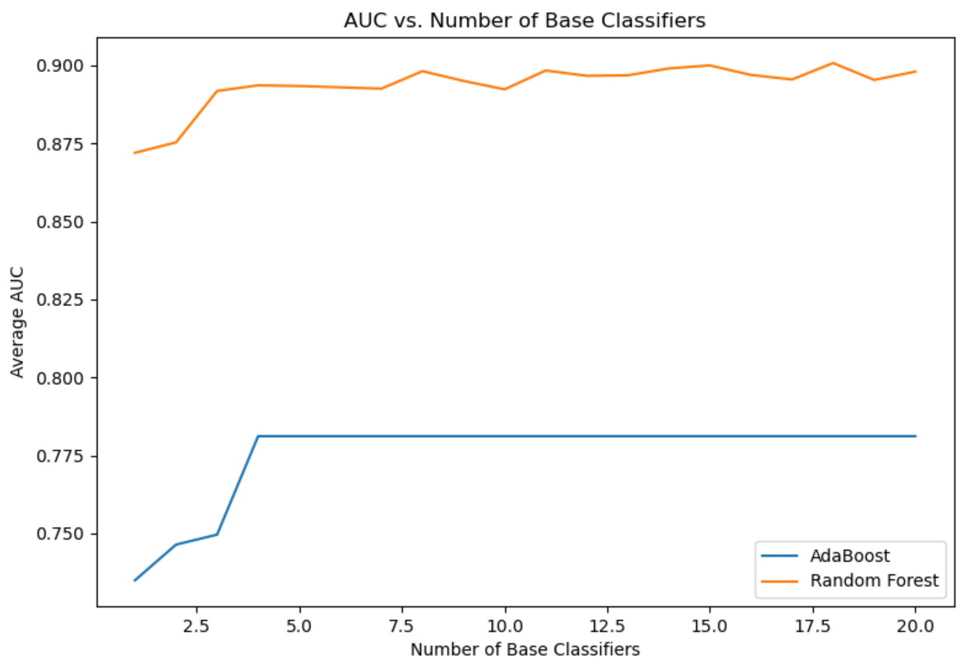

图 1: evaluation

着基学习器的增加，他的训练效果会不断地变好，最后都趋于平稳。特别是这个 adaboost，在第五轮之后基本就不动了。非常的离谱。事实情况说明分类器似乎并不是一直会越多越

好，一开始确实是有一个上升的趋势，但是到了后面，分类器自己的分类情况是有所限制的，所以会导致在一个值附近波动，多了的分类器算是一种“过拟合”了吧，所以还是要选择一个最好的情况，目前看来既不是开头也不是结尾。

（3）根据 20 轮的训练结果反馈，最优的超参数如下:

(AUC, 超参数)

(0.7811619280989419, 5-20)

(0.9007296864925012, 18)

但其实可能从第五轮开始这个 adaboost 就不变了。

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, 且对完成作业有帮助的, 亦需注明并致谢.

感谢人工智能学院 221300004 王晨阳对我的帮助和极具启发式的意见。