机器学习导论 习题五

学号, 姓名, 邮箱

2023 年 6 月 27 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, **请将其打包为 .zip 文件上传**. 注意命名规则, 两个文件均命名为“学号_姓名”+“. 后缀”(例如 211300001_张三”+“.pdf”、“.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 **6 月 6 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] Minimum Error Rate Determination

贝叶斯判定准则与贝叶斯最优分类器是机器学习中十分重要的概念. 请仔细阅读《机器学习》第 7 章 7.1 节, 完成如下问题.

(1) **[5pts]** 请证明课本 (7.6) 式中的贝叶斯最优分类器 $h^*(\boldsymbol{x})$ 满足

$$P(y = h^*(\boldsymbol{x})) \geq \frac{1}{N}.$$

其中 N 为类别数目, y 为样本 $\boldsymbol{x}$ 的真实标记.

(2) **[10pts]** 在实际应用场景中, 随着环境发生变化, 可能会出现模型从未见过的新类别. 由于新环境中的一些样本不属于任何已知类，已有分类器必然会给出错误的预测结果，从而可能误导人们做出错误决策. 一种方法是引入“拒识”(reject) 的概念, 允许分类器在必要情况下, 拒绝为某些样本给出分类结果, 也作为环境中可能出现新类的预警. 例如考虑 N 分类问题, 可能的类别标记为 $\mathcal{Y} = \{c_1, \cdots, c_N\}$, 将真实标记为 c_j 的样本误分类为 c_i 产生的损失为 λ_{ij}. 引入拒识的情况下, 损失的定义将扩展为:

$$\lambda_{ij} = \begin{cases} 0 & \text{若 } i = j; \\ \lambda_s & \text{若 } i \neq j; \\ \lambda_r \ (\lambda_r < \lambda_s) & \text{拒识.} \end{cases}$$

请由此给出样本 $\boldsymbol{x}$ 上条件风险 $R(c_i \mid \boldsymbol{x})$ 的表达式. 结合贝叶斯判定准则, 请给出此时的贝叶斯最优分类器 $h^\star(\boldsymbol{x})$ (包含分类规则和拒识规则), 并描述其意义.

Solution. 此处用于写解答 (中英文均可)

(1) 根据课本 (7.6) 式, 我们有贝叶斯最优分类器 $h^\star(\boldsymbol{x})$ 为

$$h^\star(\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} P(y = c \mid x),$$

结合类别数目 $|\mathcal{Y}| = N$, 可得

$$1 = \sum_{c \in \mathcal{Y}} P(y = c \mid \boldsymbol{x}) \leq N \cdot h^\star(\boldsymbol{x}),$$

即题给结论得证.

(2) 根据拒识的定义, 此时条件风险的表达式为

$$R(c_i \mid \boldsymbol{x}) = \sum_{j=1}^{N} \lambda_{ij} P(c_j \mid \boldsymbol{x}) = \begin{cases} \lambda_s(1 - P(c_i \mid \boldsymbol{x})) & i = 1, 2, \cdots, N; \\ \lambda_r & \text{拒识.} \end{cases}$$

由贝叶斯判定准则, 贝叶斯最优分类器 $h^\star(\boldsymbol{x})$ 在每个样本上都最小化条件风险:

$$h^\star(\boldsymbol{x}) = \arg\min_{i \in [N]} R(c_i \mid \boldsymbol{x}) = \begin{cases} \arg\max_{i \in [N]} P(c_i \mid \boldsymbol{x}) & \text{若 } \max_{i \in [N]} P(c_i \mid x) > 1 - \frac{\lambda_r}{\lambda_s}; \\ \text{拒识} & \text{相反.} \end{cases}$$

由此可见, 当样本属于任何已知类别的后验概率均小于阈值 λ_r/λ_s 时, 我们认为该样本可能来自于未见类别, 并拒绝给出分类结果. 相反地, 我们依旧遵循贝叶斯判定准则.

2 [35pts] Expectation Maximization

通常情况下, 模型会假设训练样本所有属性变量的值都可以观测到. 但在现实应用中, 往往会遇到属性变量不可观测的情况, 例如西瓜的根蒂脱落, 便无法观测到“根蒂”属性的取值. 在这种存在“未观测”变量的情况下, EM(Expectation-Maximization) 算法是估计参数隐变量的利器. 请仔细阅读《机器学习》第七章 7.6 节, 回答以下问题.

2.1 [5pts] EM with Coin Flips

考虑简单的抛硬币问题. 现有两枚硬币 A 和 B，正面朝上的概率分别为 θ_A, θ_B, 结果朝上记为 H (head), 朝下记为 T (tail). 独立地进行 N 轮实验, 在第 k 轮实验中, 以均等概率选择一枚硬币 $Z_k \in \{A, B\}$ 并重复抛掷 M 次, 其中硬币朝上的次数 X_k 为可观测变量, 而选择的硬币类型 Z_k 为隐变量不可观测. 我们将使用 EM 算法, 迭代一次, 对参数 $\theta = (\theta_A, \theta_B)$ 进行估计, 使用的实验数据如表1所示. 具体而言共 3 轮实验, 每轮选取的硬币记为 z_i $(i = 1, 2, 3)$, 抛掷 10 次并记录结果, 硬币朝上的次数记为 x_i $(i = 1, 2, 3)$.

(1) **[2pts]** **E** 步 (Expectation): 假设参数的初始值 $\theta^0 = (0.6, 0.5)$. 请结合实验数据, 推断出隐变量取值 $\boldsymbol{z} = (z_1, z_2, z_3)$ 的分布, 即推断出第 i 轮实验 $(i = 1, 2, 3)$ 中抛掷硬币 A、硬币 B 各自的概率, 完善表1的第 2-3 列.

(2) **[3pts]** **M** 步 (Maximization): 根据隐变量取值 $\boldsymbol{z}$ 的分布, 对参数 θ 进行极大似然估计. 请完善表1的第 4-5 列, 给出 EM 算法迭代一次后的参数估计值 $\theta^1 = (\theta_A^1, \theta_B^1)$.

2.2 [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 $\Sigma_i = \epsilon^2 \mathbf{I}, i \in [k]$. 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 $\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2 \neq \|\boldsymbol{x}_i - \boldsymbol{\mu}_{k'}\|^2, \forall i \in [n], k \neq k'$ 成立.

(3) **[10pts]** 请证明: 随着 $\epsilon^2 \to 0$, 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.

2.3 [20pts] Convergence Analysis

EM 算法广泛应用于机器学习等其他领域, 其中一个原因是它拥有着良好的理论保障: 随着 **E** 步与 **M** 步的迭代执行直至收敛, 已观测数据的对数“边际似然” $LL(\Theta \mid \boldsymbol{X})$ 将单调非减. 沿用《机器学习》7.6 节中的符号定义, 我们将试图证明该结论.

(4) **[5pts]** 请证明在 **E** 步中, $LL(\Theta|\boldsymbol{X})$ 可以被分拆为两项:

$$LL(\Theta \mid \boldsymbol{X}) = Q(\Theta \mid \Theta^t) - H(\Theta \mid \Theta^t),$$

其中 $H(\Theta \mid \Theta_t) = \sum_{\boldsymbol{Z}} P(\boldsymbol{Z} \mid \boldsymbol{X}, \Theta^t) \ln P(\boldsymbol{Z} \mid \boldsymbol{X}, \Theta)$, $Q(\Theta \mid \Theta^t)$ 的定义见课本 (7.36) 式.

(5) **[10pts]** 请证明 $H(\Theta \mid \Theta^t)$ 满足以下性质:

$$\Theta^t = \arg\max_{\Theta} H(\Theta \mid \Theta^t).$$

(提示: 使用 Jensen 不等式)

(6) **[5pts]** 请证明在 EM 算法的迭代过程中, 已观测数据关于当前参数 Θ^t 的对数“边际似然”单调非减, 即

$$LL(\Theta^{t+1} \mid \boldsymbol{X}) \geq LL(\Theta^t \mid \boldsymbol{X}).$$

Solution. 此处用于写解答 (中英文均可)

表 1: 实验数据

抛掷结果	选择 A 的概率	选择 B 的概率	A 朝上次数的期望值	B 朝上次数的期望值
HTTTHHTHTH	0.45	0.55	2.2	2.8
HHHHTHHHHH	0.80	0.20	7.2	1.8
HTHHHHHTHH	0.73	0.27	5.9	2.1

(1) 以第一轮实验为例, 抛掷结果 $HTTTHHTHTH$ 中, 硬币出现 5 次朝上、5 次朝下. 若选择的是硬币 A, 出现该结果的概率为 $P(x_1 = 5, z_1 = A, \theta^0) = (\theta_A^0)^5(1-\theta_A^0)^5 = 0.000796$; 若选择的是硬币 B, 出现该结果的概率为 $P(x_1 = 5, z_1 = B, \theta^0) = (\theta_B^0)^5(1-\theta_B^0)^5 = 0.000977$. 因此

$$P(z_1 = A \mid x_1 = 5, \theta^0) = \frac{0.000796}{0.000796 + 0.000977} = 0.45,$$
$$P(z_1 = B \mid x_1 = 5, \theta^0) = \frac{0.000977}{0.000796 + 0.000977} = 0.55.$$

同理可以求出第二、第三轮实验的结果, 如表1所示.

(2) 仍旧以第一轮实验为例. 抛掷硬币十次, 因选择硬币 A、硬币 B 的概率分别为 $0.45, 0.55$, 故它们正反面出现的总次数的期望值分别为 $4.5, 5.5$ 次. 结合第一轮实验结果中硬币 5 次朝上、5 次朝下, 可得硬币 A、硬币 B 朝上次数的期望值分别为 $2.25, 2.75$ 次, 朝下次数的期望值分别为 $2.25, 2.75$ 次. 第二、第三轮实验的结果同理可求.

最终所有待求值的结果如表1. 利用硬币在三轮实验中朝上次数的期望值的总和, 以及朝下次数的期望值的总和, 我们便得到了两枚硬币正面朝上概率新的估计:

$$\theta_A^1 = \frac{2.2 + 7.2 + 5.9}{(2.2 + 7.2 + 5.9) + (2.2 + 0.8 + 1.5)} = 0.77,$$
$$\theta_B^1 = \frac{2.8 + 1.8 + 2.1}{(2.8 + 1.8 + 2.1) + (2.8 + 0.2 + 0.5)} = 0.66.$$

(3) GMM 的 **E** 步计算每个样本属于每个高斯成分的后验概率 γ_{ji}, 代入 $\Sigma_i = \epsilon^2\mathbf{I}$:

$$\begin{aligned}\gamma_{ji} &= \frac{\alpha_i \cdot \exp\left(-\frac{\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2}{2\epsilon^2}\right)}{\sum_{l=1}^{k} \alpha_l \cdot \exp\left(-\frac{\|\boldsymbol{x}_j - \boldsymbol{\mu}_l\|^2}{2\epsilon^2}\right)} \\ &= \frac{1}{1 + \sum_{l \neq i} \frac{\alpha_l}{\alpha_i} \exp\left(\frac{\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 - \|\boldsymbol{x}_j - \boldsymbol{\mu}_l\|^2}{2\epsilon^2}\right)}.\end{aligned}$$

下面进行分类讨论.

(a) $\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|^2 \le \|\boldsymbol{x}_j-\boldsymbol{\mu}_l\|^2, \forall l \ne i$. 随着 $\epsilon^2 \to 0$, 我们有

$$\frac{\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|^2-\|\boldsymbol{x}_j-\boldsymbol{\mu}_l\|^2}{2\epsilon^2} \to -\infty, \quad \forall l \ne i,$$

于是可得

$$\gamma_{ji}=\frac{1}{1+\sum_{l\ne i}\frac{\alpha_l}{\alpha_i}\exp\left(\frac{\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|^2-\|\boldsymbol{x}_j-\boldsymbol{\mu}_l\|^2}{2\epsilon^2}\right)} \to \frac{1}{1+0}=1.$$

(b) 存在混合成分 k', 使得 $\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|^2 > \|\boldsymbol{x}_j-\boldsymbol{\mu}_{k'}\|^2$. 那么对于该混合成分, 我们有

$$\exp\left(\frac{\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|^2-\|\boldsymbol{x}_j-\boldsymbol{\mu}_{k'}\|^2}{2\epsilon^2}\right) \to \infty \ \text{ 随着 } \ \epsilon^2 \to 0,$$

结合题给假设 $\alpha_m \ge \delta > 0, \forall m \in [k]$, 可得

$$\gamma_{ji} \le \frac{1}{1+\frac{\alpha_{k'}}{\alpha_i}\exp\left(\frac{\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|^2-\|\boldsymbol{x}_j-\boldsymbol{\mu}_{k'}\|^2}{2\epsilon^2}\right)} \to 0 \ \text{ 随着 } \ \epsilon^2 \to 0.$$

综上所述, 当 $\epsilon \to 0$ 时, 高斯混合聚类的 **E** 步规则会收敛至 k 均值算法中簇划分的更新规则, 即根据距离最近的均值向量确定样本的簇标记.

(4) 首先对已观测数据的对数“边际似然” $LL(\Theta \mid \boldsymbol{X})$ 进行分拆:

$$LL(\Theta \mid \boldsymbol{X}) = \ln P(\boldsymbol{X} \mid \Theta) = \ln\left(\frac{P(\boldsymbol{X},\boldsymbol{Z} \mid \Theta)}{P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta)}\right) = \ln P(\boldsymbol{X},\boldsymbol{Z} \mid \Theta) - \ln P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta).$$

此时计算边际似然 $LL(\Theta \mid \boldsymbol{X})$ 关于隐变量分布 $P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta^t)$ 的期望, 便能得证:

$$\begin{aligned}\ln P(\boldsymbol{X} \mid \Theta) &= \sum_{\boldsymbol{Z}} P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta^t)\ln P(\boldsymbol{X},\boldsymbol{Z} \mid \Theta) - \sum_{\boldsymbol{Z}} P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta^t)\ln P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta)\\ &= Q(\Theta \mid \Theta^t) - H(\Theta \mid \Theta^t).\end{aligned}$$

(5) 我们只需证明 $\forall \Theta, H(\Theta^t \mid \Theta^t) - H(\Theta \mid \Theta^t) \ge 0$:

$$\begin{aligned}H(\Theta^t \mid \Theta^t) - H(\Theta \mid \Theta^t) &= \mathbb{E}_{\boldsymbol{Z}\mid\boldsymbol{X},\Theta^t}\left[\ln P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta^t)\right] - \mathbb{E}_{\boldsymbol{Z}\mid\boldsymbol{X},\Theta^t}\left[\ln P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta)\right]\\ &= -\mathbb{E}_{\boldsymbol{Z}\mid\boldsymbol{X},\Theta^t}\left[\ln\left(\frac{P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta)}{P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta^t)}\right)\right]\\ &\ge -\ln\left\{\mathbb{E}_{\boldsymbol{Z}\mid\boldsymbol{X},\Theta^t}\left[\ln\left(\frac{P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta)}{P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta^t)}\right)\right]\right\}\\ &= -\ln\left[\sum_{\boldsymbol{Z}} P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta^t)\cdot\frac{P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta)}{P(\boldsymbol{Z} \mid \boldsymbol{X},\Theta^t)}\right] = 0,\end{aligned}$$

其中不等式利用了函数 $\varphi(x) = -\ln x$ 的凸性以及 Jensen 不等式.

(6) 由第五小问中的结论, 可知新得到的参数值 Θ^{t+1} 满足 $H(\Theta^{t+1} \mid \Theta^t) \le H(\Theta^t \mid \Theta^t)$. 进一步地, 根据 EM 算法中 **M** 步的定义, 可得

$$\Theta^{t+1} = \arg\max_{\Theta} Q(\Theta;\Theta^t) \quad\Longrightarrow\quad Q(\Theta^{t+1} \mid \Theta^t) \ge Q(\Theta^t \mid \Theta^t).$$

结合第四小问中的结论, 便能最终证得对数“边际似然”的单调非减性:

$$\begin{aligned}LL(\Theta^{t+1} \mid \boldsymbol{X}) &= Q(\Theta^{t+1} \mid \Theta^t) - H(\Theta^{t+1} \mid \Theta^t)\\ &\ge Q(\Theta^t \mid \Theta^t) - H(\Theta^t \mid \Theta^t) = LL(\Theta^t \mid \boldsymbol{X}).\end{aligned}$$

3 [30pts] Boosting

Boosting 算法有序地训练一批弱学习器进行集成得到一个强学习器, 其中最著名的代表便是 AdaBoost. 该算法通过迭代地调整训练样本分布, 可以使得经验误差会随着学习轮数 T 指数级下降. 不仅如此, AdaBoost 还拥有很好的泛化性能保障, 其泛化误差在经验误差达到最小后仍然能持续地降低. 本题将针对 AdaBoost 算法展开更加深入的讨论.

3.1 [15pts] AdaBoost Empirical Error Bound

考虑训练集 $D = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_m, y_m)\}$, $y_m \in \{-1, +1\}$, 参照《机器学习》第八章图 8.3 的变量定义, 我们将证明如下定理: AdaBoost 迭代 T 轮后返回的分类器 f, 经验误差满足

$$\hat{R}_D(f) = \frac{1}{m}\sum_{i=1}^{m} 1_{y_i f(\boldsymbol{x}_i) \le 0} \le \exp\left[-2\sum_{t=1}^{T}\left(\frac{1}{2} - \epsilon_t\right)^2\right].$$

进一步地, 若对于任意的 $t \in [T]$, $\gamma \le (\frac{1}{2} - \epsilon_t)$, 那么有

$$\hat{R}_D(f) \le \exp(-2\gamma^2 T).$$

(1) **[5pts]** 请证明数据分布 D_t 的调整过程满足:

$$\mathcal{D}_{t+1}(\boldsymbol{x}) = \frac{e^{-y\sum_{s=1}^{t}\alpha_s h_s(\boldsymbol{x})}}{m\prod_{s=1}^{t} Z_s}, \quad \forall t \in [T].$$

(2) **[5pts]** 请证明规范化因子 Z_t 与基学习器误差 ϵ_t 的关系:

$$Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}, \quad \forall t \in [T].$$

(3) **[5pts]** 利用前两问的结论, 完成题给定理的证明.

(提示: 使用不等式 $\mathbb{I}(u \le 0) \le \exp(-u)$, $\forall u \in \mathbb{R}$)

3.2 [15pts] Multi-Class AdaBoost

AdaBoost 的应用场景可以从二分类拓展到多分类, 一种经典的扩展方法为 SAMME (Stage-wise Additive Modeling using a Multi-class Exponential loss function). 该算法首先将样本的标记 $c \in [K]$ 编码为 K 维向量 $\boldsymbol{y}$, 其中目标类别对应位置的值为 1, 其余类别对应位置的值为 $-\frac{1}{K-1}$, 即

$$y_k = \begin{cases} 1, & \text{若 } c = k, \\ -\frac{1}{K-1}, & \text{若 } c \ne k. \end{cases}$$

同时, 基学习器的输出 $h_t(\boldsymbol{x})$ 为 K 维向量, 不失一般性可以约束 $h_t(\boldsymbol{x})$ 的各个维度和为零. 记基学习器的线性组合为 $H(\boldsymbol{x})$, SAMME 使用的多分类指数损失函数为:

$$\ell_{\text{multi-exp}}(H|\mathcal{D}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[e^{-\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{y}_k[H(\boldsymbol{x})]_k}\right] = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[e^{-\frac{1}{K}\boldsymbol{y}^\top H(\boldsymbol{x})}\right].$$

(4) 考虑优化问题如下

$$\begin{aligned}\min_{H(\boldsymbol{x})} \quad & \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{x}} \exp\left(-\frac{1}{K}(Y_1 H(\boldsymbol{x})_1 + \cdots + Y_K H(\boldsymbol{x})_K)\right) \\ \text{s.t.} \quad & H(\boldsymbol{x})_1 + \cdots H(\boldsymbol{x})_K = 0.\end{aligned}$$

请证明对于最优解 $H^\star(\boldsymbol{x})$, $\arg\max_{k\in[K]} H^\star(\boldsymbol{x})_k$ 达到了贝叶斯最优错误率, 即 SAMME 使用的多分类指数损失函数是 0/1 损失函数的一致的替代损失函数.

(提示: 使用拉格朗日乘子法)

Solution. 此处用于写解答 (中英文均可)

(1) AdaBoost 算法中，样本分布每一轮的更新公式为

$$\mathcal{D}_{t+1}(\boldsymbol{x}_i) = \frac{\mathcal{D}_t(\boldsymbol{x}_i) e^{-\alpha_t y_i h_t(\boldsymbol{x}_i)}}{Z_t}.$$

重复地代入样本分布的更新公式，便可以证明题给结论:

$$\begin{aligned}\mathcal{D}_{t+1}(\boldsymbol{x}_i) = \frac{\mathcal{D}_t(\boldsymbol{x}_i) e^{-\alpha_t y_i h_t(\boldsymbol{x}_i)}}{Z_t} &= \frac{\mathcal{D}_{t-1}(\boldsymbol{x}_i) e^{-\alpha_{t-1} y_i h_{t-1}(\boldsymbol{x}_i)} e^{-\alpha_t y_i h_t(\boldsymbol{x}_i)}}{Z_{t-1} Z_t} \\ &= \frac{e^{-y_i \sum_{s=1}^t \alpha_s h_s(\boldsymbol{x}_i)}}{m \prod_{s=1}^t Z_s}, \quad \forall t \in [T].\end{aligned}$$

(2) 根据归一化因子 Z_t 以及基学习器误差 ϵ_t 的定义，我们有

$$\begin{aligned}Z_t = \sum_{i=1}^m \mathcal{D}_t(i) e^{-\alpha_t y_i h_t(\boldsymbol{x}_i)} &= \sum_{i: y_i h_t(\boldsymbol{x}_i) = +1} \mathcal{D}_t(\boldsymbol{x}_i) e^{-\alpha_t} + \sum_{i: y_i h_t(\boldsymbol{x}_i) = -1} \mathcal{D}_t(\boldsymbol{x}_i) e^{\alpha_t} \\ &= (1-\epsilon_t) e^{-\alpha_t} + \epsilon_t e^{\alpha_t}.\end{aligned}$$

进一步带入基学习器权重 α_t 的定义, 便可知

$$Z_t = (1-\epsilon_t)\sqrt{\frac{\epsilon_t}{1-\epsilon_t}} + \epsilon_t \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} = 2\sqrt{\epsilon_t(1-\epsilon_t)}.$$

(3) 首先利用不等式 $\mathbb{I}(u \le 0) \le \exp(-u)$, $\forall u \in \mathbb{R}$, 可以获得 AdaBoost 经验误差的上界:

$$\hat{R}_D(f) = \frac{1}{m}\sum_{i=1}^m 1_{y_i f(\boldsymbol{x}_i) \le 0} \le \frac{1}{m}\sum_{i=1}^m e^{-y_i f(\boldsymbol{x}_i)}.$$

AdaBoost 迭代 T 轮后返回的分类器 f 为 T 个基学习器 $\{h_t\}_{t=1}^T$ 的加权集成, 于是

$$\begin{aligned}\hat{R}_D(f) \le \frac{1}{m}\sum_{i=1}^m e^{-y_i f(\boldsymbol{x}_i)} &= \frac{1}{m}\sum_{i=1}^m e^{-y_i \sum_{t=1}^T \alpha_t h_t(\boldsymbol{x}_i)} \\ &= \frac{1}{m}\sum_{i=1}^m \left[m \prod_{t=1}^T Z_t\right] \mathcal{D}_{T+1}(\boldsymbol{x}_i) = \prod_{t=1}^T Z_t,\end{aligned}$$

其中第二个等式利用了第一问的结论. 再通过第二问的结论, 我们便能证明 AdaBoost 算法经验误差的指数收敛率:

$$\begin{aligned}\prod_{t=1}^T Z_t = \prod_{t=1}^T 2\sqrt{\epsilon_t(1-\epsilon_t)} = \prod_{t=1}^T \sqrt{1 - 4(\frac{1}{2}-\epsilon_t)^2} &\le \prod_{t=1}^T \exp\left[-2(\frac{1}{2}-\epsilon_t)^2\right] \\ &= \exp\left[-2\sum_{t=1}^T \left(\frac{1}{2}-\epsilon_t\right)^2\right].\end{aligned}$$

(4) 记优化问题的目标函数为 F, 可以被化简为

$$\begin{aligned}
F(H(\boldsymbol{x})) &= \sum_{k=1}^{K} \exp\left(-\frac{1}{K}(Y_1 H(\boldsymbol{x})_1 + \cdots + Y_K H(\boldsymbol{x})_K)\right) P(c=k \mid \boldsymbol{x}) \\
&= \sum_{k=1}^{K} \exp\left(-\frac{1}{K}\left(H(\boldsymbol{x})_k - \frac{1}{K-1}\sum_{j\neq k} H(\boldsymbol{x})_j\right)\right) P(c=k \mid \boldsymbol{x}) \\
&= \sum_{k=1}^{K} \exp\left(-\frac{H(\boldsymbol{x})_k}{K-1}\right) P(c=k \mid \boldsymbol{x}),
\end{aligned}$$

其中第一个等式基于条件期望的定义, 第二个等式利用了 SAMME 算法对样本标记的编码方式, 而第三个等式基于优化问题的约束条件. 为等式约束引入拉格朗日乘子 λ, 可得拉格朗日函数

$$L(H(\boldsymbol{x}), \lambda) = \sum_{k=1}^{K} \exp\left(-\frac{H(\boldsymbol{x})_k}{K-1}\right) P(c=k \mid \boldsymbol{x}) - \lambda(H(\boldsymbol{x})_1 + \cdots + H(\boldsymbol{x})_K)$$

令 $\nabla_{H(\boldsymbol{x})_1} L = \cdots = \nabla_{H(\boldsymbol{x})_K} L = \nabla_\lambda L = 0$, 可得

$$\begin{aligned}
-\frac{1}{K-1}\exp\left(-\frac{H^\star(\boldsymbol{x})_1}{K-1}\right) P(c=1 \mid \boldsymbol{x}) - \lambda = 0, \\
\vdots \\
-\frac{1}{K-1}\exp\left(-\frac{H^\star(\boldsymbol{x})_K}{K-1}\right) P(c=K \mid \boldsymbol{x}) - \lambda = 0, \\
H(\boldsymbol{x})_1^\star + \cdots + H(\boldsymbol{x})_K^\star = 0.
\end{aligned}$$

求解后便能得到

$$H^\star(\boldsymbol{x})_k = (K-1)\left(\log P(c=k \mid \boldsymbol{x}) - \frac{1}{K}\sum_{k'=1}^{K} \log P\left(c=k' \mid \boldsymbol{x}\right)\right), \quad k=1,\ldots,K.$$

这便意味着

$$\arg\max_{k\in[K]} H^\star(\boldsymbol{x})_k = \arg\max_{k\in[K]} P(c=k \mid \boldsymbol{x}),$$

即 $\arg\max_{k\in[K]} H^\star(\boldsymbol{x})_k$ 达到了贝叶斯最优错误率.

4 [20pts] Bagging

考虑回归学习任务 $f:\mathbb{R}^d \to \mathbb{R}$. 假设已经训练得到 M 个基学习器 $\hat{f}_1(\boldsymbol{x}), \hat{f}_2(\boldsymbol{x}), \cdots, \hat{f}_M(\boldsymbol{x})$. 我们可以将基学习器的预测值看作真实值加上偏差项

$$\hat{f}_m(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon_m(\boldsymbol{x}), \quad \forall m \in [M],$$

每个基学习器的期望平方误差即为 $\mathbb{E}_{\boldsymbol{x}}\left[\epsilon_m(\boldsymbol{x})^2\right]$. 所有基学习器的期望平方误差的均值为

$$E_{avg} = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\boldsymbol{x}}\left[\epsilon_m(\boldsymbol{x})^2\right].$$

与此同时, M 个基学习器通过集成得到的 Bagging 模型为

$$\hat{f}_{bag}(\boldsymbol{x}) = \frac{1}{M}\sum_{m=1}^{M}\hat{f}_m(\boldsymbol{x}),$$

于是该 Bagging 模型在单个样本上的误差为

$$\epsilon_{bag}(\boldsymbol{x}) = \hat{f}_{bag}(\boldsymbol{x}) - f(\boldsymbol{x}) = \frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\boldsymbol{x}),$$

其期望平方误差即为

$$E_{bag} = \mathbb{E}_{\boldsymbol{x}}\left[\epsilon_{bag}(\boldsymbol{x})^2\right].$$

(1) **[5pts]** 假设个体学习器相互独立: $\forall m \neq l, \mathbb{E}_{\boldsymbol{x}}[\epsilon_m(\boldsymbol{x})\epsilon_l(\boldsymbol{x})] = 0$. 在这种理想情形下, 请证明 E_{avg} 与 E_{bag} 满足

$$E_{bag} = \frac{1}{M}E_{avg}.$$

(2) **[10pts]** 现实任务中, 基学习器相互独立通常无法满足. 假设 $\epsilon_1(\boldsymbol{x}), \cdots, \epsilon_M(\boldsymbol{x})$ 满足 $\mathbb{E}[\epsilon_m(\boldsymbol{x})] = \mu, \mathrm{var}[\epsilon_m(\boldsymbol{x})] = \sigma^2, \forall m \in [M]$, 且彼此之间的线性相关系数均为 ρ. 请证明

$$\mathrm{var}[\epsilon_{bag}(\boldsymbol{x})] = \rho\sigma^2 + \frac{1-\rho}{M}\sigma^2.$$

可见随着基学习器数量 M 增多, Bagging 模型误差的方差将主要受制于基学习器之间的相关性. 请简要叙述随机森林算法是如何降低基决策树之间的相关性的.

(3) **[5pts]** 请证明无需对 $\epsilon_1(\boldsymbol{x}), \cdots, \epsilon_M(\boldsymbol{x})$ 做任何假设, $E_{bag} \leq E_{avg}$ 始终成立.

(提示: 使用 Jensen 不等式)

Solution. 此处用于写解答 (中英文均可)

(1) 代入 E_{bag} 的定义, 我们有

$$E_{bag} = \mathbb{E}_{\boldsymbol{x}}\left[\epsilon_{bag}(\boldsymbol{x})^2\right] = \frac{1}{M^2}\mathbb{E}_{\boldsymbol{x}}\left[\sum_{i=1}^{M}\sum_{j=1}^{M}\epsilon_i(\boldsymbol{x})\epsilon_j(\boldsymbol{x})\right]$$

$$= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}_{\boldsymbol{x}}\left[\epsilon_m(\boldsymbol{x})^2\right] = \frac{1}{M}E_{avg},$$

其中第三个等式利用了个题学习器相互独立的条件.

(2) 根据线性相关系数的定义, 我们有

$$\rho = \frac{\mathbb{E}\left[(\epsilon_i(\boldsymbol{x}) - \mu)(\epsilon_j(\boldsymbol{x}) - \mu)\right]}{\sigma^2} \quad \Longrightarrow \quad \mathbb{E}[\epsilon_i(\boldsymbol{x})\epsilon_j(\boldsymbol{x})] = \rho\sigma^2 + \mu^2, \forall i \neq j.$$

于是 Bagging 模型误差的方差为

$$\begin{aligned}
\mathrm{var}[\epsilon_{bag}(\boldsymbol{x})] &= \mathrm{var}\left[\frac{1}{M}\sum_{i=1}^{M}\epsilon_i(\boldsymbol{x})\right] = \frac{1}{M^2}\left\{\mathbb{E}\left[\left(\sum_{i=1}^{M}\epsilon_i(\boldsymbol{x})\right)^2\right] - \mathbb{E}\left[\sum_{i=1}^{M}\epsilon_i(\boldsymbol{x})\right]^2\right\} \\
&= \frac{1}{M^2}\left\{\sum_{i=1}^{M}\mathbb{E}\left[\epsilon_i(\boldsymbol{x})^2\right] + \sum_{i\neq j}\mathbb{E}\left[\epsilon_i(\boldsymbol{x})\epsilon_j(\boldsymbol{x})\right] - M^2\mu^2\right\} \\
&= \rho\sigma^2 + \frac{1-\rho}{M}\sigma^2.
\end{aligned}$$

(3) 由于函数 $\varphi(x) = x^2$ 是凸函数, 由 Jensen 不等式可得

$$\begin{aligned}
E_{bag} &= \mathbb{E}_{\boldsymbol{x}}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\boldsymbol{x})\right)^2\right] \leq \mathbb{E}_{\boldsymbol{x}}\left[\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\boldsymbol{x})^2\right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\boldsymbol{x}}\left[\epsilon_m(\boldsymbol{x})^2\right] = E_{avg}.
\end{aligned}$$