机器学习导论 习题五

211300063, 张运吉, 211300063@smail.nju.edu.cn

2023 年 6 月 6 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) 证明如下:

$$\begin{aligned} P\left(h^\star(\boldsymbol{X}) \mid \boldsymbol{x}\right) &= P\left(Y = h^\star(\boldsymbol{x}) \mid \boldsymbol{X} = \boldsymbol{x}\right) \\ &= \max_{c \in \mathcal{Y}} P(Y = c \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \frac{1}{N} N \max_{c \in \mathcal{Y}} P(Y = c \mid \boldsymbol{X} = \boldsymbol{x}) \\ &\geq \frac{1}{N} \sum_{c \in \mathcal{Y}} P(Y = c \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \frac{1}{N} \end{aligned}$$

其中第二个等式是根据贝叶斯最优分类器的定义，也就是公式 (7.6).

$$\begin{aligned} P\left(y = h^\star(\boldsymbol{x})\right) &= \sum_{\boldsymbol{x}} P\left(Y = h^\star(\boldsymbol{X}) \mid \boldsymbol{X} = \boldsymbol{x}\right) P(\boldsymbol{X} = \boldsymbol{x}) \\ &\geq \frac{1}{N} \sum_{\boldsymbol{x}} P(\boldsymbol{X} = \boldsymbol{x}) \\ &= \frac{1}{N} \end{aligned}$$

其中第一个等式是根据全概率公式得来.

(2) 添加 reject 的情况下:

$$R(c \mid \boldsymbol{x}) = \sum_{j=1}^{N} \lambda_{ij} P\left(c_j \mid \boldsymbol{x}\right) = \begin{cases} (1 - P(c \mid \boldsymbol{x}))\lambda_s & c \in \mathcal{Y} \\ \lambda_r & \text{reject!} \end{cases}$$

根据贝叶斯最优判别准则，我们要最小化 $R(c \mid \boldsymbol{x})$.
因此, 当 $\lambda_r \leqslant \min\limits_{c \in \mathcal{Y}} (1 - P(c \mid \boldsymbol{x}))\lambda_s$ 时，也即 $\max\limits_{c \in \mathcal{Y}} P(c \mid \boldsymbol{x}) \leq 1 - \frac{\lambda_r}{\lambda_s}$ 时，reject!
综合起来，有：

$$h^{\star}(\boldsymbol{x}) = \begin{cases} \text{reject!} & \text{if } \max\limits_{c \in \mathcal{Y}} P(c \mid \boldsymbol{x}) \leq 1 - \frac{\lambda_r}{\lambda_s} \\ \arg\max\limits_{c \in \mathcal{Y}} P(c \mid \boldsymbol{x}) & \text{otherwise} \end{cases}$$

引入“拒识”相当于引入了“像”和“不像”的边界，只有在 $\max\limits_{c \in \mathcal{Y}} P(c \mid \boldsymbol{x}) > 1 - \frac{\lambda_r}{\lambda_s}$ 时，我们才会根据分类规则 $\arg\max\limits_{c \in \mathcal{Y}} P(c \mid \boldsymbol{x})$ 把样本进行分类，否则拒绝分类. 这样做的好处是符合直觉的，只有当一个样本足够像某一类时我们才有较大的把握对其分类，如果某个样本不像标记空间中的任何一类，那么我们是没有把握对其正确分类的. 这样做可以提高分类器的性能。当拒识的损失较小时，分类器更倾向于将未知类别的样本拒绝，从而提供了一种预警机制，避免误导人们做出错误决策.

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 $\boldsymbol{z} = (z_1, z_2)$ 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 $H(\Theta \mid \Theta_t) = \sum_{\boldsymbol{Z}} P(\boldsymbol{Z} \mid \boldsymbol{X}, \Theta^t) \ln(\boldsymbol{Z} \mid \boldsymbol{X}, \Theta)$, $Q(\Theta \mid \Theta^t)$ 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	44.91%	55.09%	2.2457	2.7543
HHHHTHHHHH	80.50%	19.50%	7.2449	1.7551
HTHHHHHTHH	73.35%	26.65%	5.8677	2.1323

(2) 由极大似然估计:

$$\theta_A^{(1)} = \frac{2.2457 + 7.2449 + 5.8677}{2.2457 + 7.2449 + 5.8677 + 4.5177} = 0.7727$$
$$\theta_B^{(1)} = \frac{2.7543 + 1.7551 + 2.1323}{2.7543 + 1.7551 + 2.1323 + 3.4823} = 0.6560$$

(3) 令 γ_{ji} 表示样本 x_j 由第 i 个高斯分布生成的后验概率. 则:

$$\begin{aligned}
\gamma_{ji} &= \frac{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{\ell=1}^{k} \alpha_\ell \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell\right)} \\
&= \frac{\frac{\alpha_i}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)}}{\sum_{\ell=1}^{k} \frac{\alpha_\ell}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_\ell|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_\ell)^\top \boldsymbol{\Sigma}_\ell^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_\ell)}} \\
&= \frac{\alpha_i}{\sum_{\ell=1}^{k} \alpha_\ell \exp\left(\frac{\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 - \|\boldsymbol{x}_j - \boldsymbol{\mu}_\ell\|^2}{2|\epsilon|^{2n}}\right)}
\end{aligned}$$

记 $i^\star = \arg\min_l \|\boldsymbol{x}_j - \boldsymbol{\mu}_l\|^2$.

- $i \neq i^\star$ 时:
  $\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 - \|\boldsymbol{x}_j - \boldsymbol{\mu}_{i^\star}\|^2 > 0$, 又 $\epsilon^2 \to 0$, 即 $|\epsilon| \to 0$.
  $\therefore \exp\left(\frac{\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 - \|\boldsymbol{x}_j - \boldsymbol{\mu}_{i^\star}\|^2}{2|\epsilon|^{2n}}\right) \to +\infty$
  根据题意 $\exists \delta > 0$ 使得对于选择各个混合成分的概率有 $\alpha_i > \delta, \forall i \in [k]$, 故:

$$\begin{aligned}
\sum_{\ell=1}^{k} \alpha_\ell \exp\left(\frac{\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 - \|\boldsymbol{x}_j - \boldsymbol{\mu}_\ell\|^2}{2|\epsilon|^{2n}}\right) &\geq \delta \sum_{\ell=1}^{k} \exp\left(\frac{\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 - \|\boldsymbol{x}_j - \boldsymbol{\mu}_\ell\|^2}{2|\epsilon|^{2n}}\right) \\
&> \delta \exp\left(\frac{\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 - \|\boldsymbol{x}_j - \boldsymbol{\mu}_{i^\star}\|^2}{2|\epsilon|^{2n}}\right) \\
&\to +\infty
\end{aligned}$$

$\therefore \gamma_{ji} \to 0$

- $i = i^\star$ 时:
  有 $\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 - \|\boldsymbol{x}_j - \boldsymbol{\mu}_\ell\|^2 < 0$, 进而有 $\frac{\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 - \|\boldsymbol{x}_j - \boldsymbol{\mu}_\ell\|^2}{2|\epsilon|^{2n}} \to -\infty$
  $\exp\left(\frac{\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 - \|\boldsymbol{x}_j - \boldsymbol{\mu}_{i^\star}\|^2}{2|\epsilon|^{2n}}\right) \to 0$

$$\sum_{\ell=1}^{k} \alpha_\ell \exp\left(\frac{\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 - \|\boldsymbol{x}_j - \boldsymbol{\mu}_\ell\|^2}{2|\epsilon|^{2n}}\right) \to \alpha_{i^\star} \exp\left(\frac{\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 - \|\boldsymbol{x}_j - \boldsymbol{\mu}_{i^\star}\|^2}{2|\epsilon|^{2n}}\right)$$
$$= \alpha_{i^\star}$$

$\therefore \gamma_{ji} \to \frac{\alpha_i}{\alpha_{i^*}} = 1$

综上,

$$\lim_{\epsilon \to 0} \gamma_{ji} = \begin{cases} 0 & \text{if } \|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 \geq \|\boldsymbol{x}_j - \boldsymbol{\mu}_l\|^2, \forall l \in [k] \\ 1 & \text{else} \end{cases}$$

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。因此，k 均值聚类算法可以看作是高斯混合聚类的一种特例。

(4) 证明如下:

$$\begin{aligned}
Q\left(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}_t\right) - H\left(\boldsymbol{\Theta} \mid \Theta^t\right) &= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\Theta}^t} LL(\boldsymbol{\Theta} \mid \boldsymbol{X}, \boldsymbol{Z}) - \sum_{\boldsymbol{Z}} P\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t\right) \ln P(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}) \\
&= \sum_{\boldsymbol{Z}} P\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t\right) \left(LL(\boldsymbol{\Theta} \mid \boldsymbol{X}, \boldsymbol{Z}) - \ln P(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta})\right) \\
&= \sum_{\boldsymbol{Z}} P\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t\right) \left(\ln P(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\Theta}) - \ln P(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta})\right) \\
&= \sum_{\boldsymbol{Z}} P\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t\right) \left(\ln \frac{P(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\Theta})}{P(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta})}\right) \\
&= \sum_{\boldsymbol{Z}} P\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t\right) \left(\ln P(\boldsymbol{X} \mid \boldsymbol{\Theta})\right) \\
&= \left(\ln P(\boldsymbol{X} \mid \boldsymbol{\Theta})\right) \sum_{\boldsymbol{Z}} P\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t\right) \\
&= \ln P(\boldsymbol{X} \mid \boldsymbol{\Theta}) \\
&= LL(\boldsymbol{\Theta} \mid \boldsymbol{X})
\end{aligned}$$

(5) 证明如下:

$$\begin{aligned}
H\left(\boldsymbol{\Theta} \mid \Theta^t\right) - H\left(\boldsymbol{\Theta}^t \mid \boldsymbol{\Theta}^t\right) &= \sum_{\boldsymbol{Z}} P\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t\right) \ln P(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}) - \sum_{\boldsymbol{Z}} P\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t\right) \ln P\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t\right) \\
&= \sum_{\boldsymbol{Z}} P\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t\right) \ln \frac{P(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta})}{P\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t\right)} \\
&\leq \ln \sum_{\boldsymbol{Z}} P\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t\right) \frac{P(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta})}{P\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t\right)} \\
&= \ln \sum_{\boldsymbol{Z}} P(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}) \\
&= 0
\end{aligned}$$

其中不等式是使用了 Jensen 不等式，当且仅当 $P(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}) = P\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t\right)$ 即 $\boldsymbol{\Theta} = \boldsymbol{\Theta}^t$ 时等号成立.

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

(6) 由 (5) 可知 $H\left(\boldsymbol{\Theta}^{t+1} \mid \boldsymbol{\Theta}^t\right) \leq H\left(\boldsymbol{\Theta}^t \mid \boldsymbol{\Theta}^t\right)$

由 M 步更新规则 $\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} Q\left(\boldsymbol{\Theta} \mid {}^t\right)$ 可知 $Q\left(\boldsymbol{\Theta}^{t+1} \mid \boldsymbol{\Theta}^t\right) \geq Q\left(\boldsymbol{\Theta}^t \mid \boldsymbol{\Theta}^t\right)$

因此:

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

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:

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

(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 $H^\star(\boldsymbol{x})$, $\operatorname{sign}(H^\star(\boldsymbol{x}))$ 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) 证明如下:

$$\begin{aligned}\mathcal{D}_{t+1}(\boldsymbol{x}) &= \frac{\mathcal{D}_t \exp\left(-\alpha_t f(\boldsymbol{x}) h_t(\boldsymbol{x})\right)}{Z_t} \\ &= \frac{\exp\left(-\alpha_t f(\boldsymbol{x}) h_t(\boldsymbol{x})\right)}{Z_t} \times \frac{\exp\left(-\alpha_{t-1} f(\boldsymbol{x}) h_{t-1}(\boldsymbol{x})\right)}{Z_{t-1}} \times \mathcal{D}_{t-1} \\ &= \cdots \\ &= \mathcal{D}_1(\boldsymbol{x}) \prod_{s=1}^{t} \frac{\exp\left(-\alpha_s f(\boldsymbol{x}) h_s(\boldsymbol{x})\right)}{Z_s} \\ &= \frac{1}{m} \prod_{s=1}^{t} \frac{\exp\left(-\alpha_s f(\boldsymbol{x}) h_s(\boldsymbol{x})\right)}{Z_s} \\ &= \frac{1}{m} \frac{\exp\left(-\sum_{s=1}^{t} \alpha_s f(\boldsymbol{x}) h_s(\boldsymbol{x})\right)}{\prod_{s=1}^{t} Z_s} \\ &= \frac{\exp\left(-y \sum_{s=1}^{t} \alpha_s h_s(\boldsymbol{x})\right)}{m \prod_{s=1}^{t} Z_s}\end{aligned}$$

(2) 规范化因子 Z_t 的作用是确保 $\mathcal{D}_{t+1}$ 是一个分布, 即保证 $\sum_{\boldsymbol{x}} \mathcal{D}_{t+1}(\boldsymbol{x}) = 1$.

$$\begin{aligned}\sum_{\boldsymbol{x}} \mathcal{D}_{t+1}(\boldsymbol{x}) &= \sum_{\boldsymbol{x}} \frac{\mathcal{D}_t(\boldsymbol{x})}{Z_t} \times \begin{cases} \exp\left(-\alpha_t\right), & \text{if } h_t(\boldsymbol{x}) = f(\boldsymbol{x}) \\ \exp\left(\alpha_t\right), & \text{if } h_t(\boldsymbol{x}) \neq f(\boldsymbol{x}) \end{cases} \\ &= \frac{1}{Z_t} \left(\sum_{h_t(\boldsymbol{x}) = f(\boldsymbol{x})} \mathcal{D}_t(\boldsymbol{x}) e^{-\alpha_t} + \sum_{h_t(\boldsymbol{x}) \neq f(\boldsymbol{x})} \mathcal{D}_t(\boldsymbol{x}) e^{\alpha_t} \right) \\ &= \frac{1}{Z_t} \left((1 - \epsilon_t) e^{-\alpha_t} + \epsilon_t e^{\alpha_t} \right) \\ &= \frac{2\sqrt{\epsilon_t (1 - \epsilon_t)}}{Z_t}\end{aligned} \tag{3.1}$$

最后一个等式是把 $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$ 代入得到.

因此 $\frac{2\sqrt{\epsilon_t(1-\epsilon_t)}}{Z_t} = 1$, 解得: $Z_t = 2\sqrt{\epsilon_t (1 - \epsilon_t)}$.

(3) 证明如下:

$$\begin{aligned}
\hat{R}_D(f) &= \frac{1}{m}\sum_{i=1}^{m} 1_{y_i f(\boldsymbol{x}_i)\leq 0} \\
&\leq \frac{1}{m}\sum_{i=1}^{m} \exp\left(-y_i f\left(\boldsymbol{x}_i\right)\right) \\
&= \sum_{i=1}^{m} \mathcal{D}\left(\boldsymbol{x}_i\right) \prod_{t=1}^{T} Z_t \\
&= \prod_{t=1}^{T} 2\sqrt{\epsilon_t\left(1-\epsilon_t\right)} \\
&= \prod_{t=1}^{T} \sqrt{(1-2\gamma_t)(1+2\gamma_t)} \quad \left(\gamma_t = \frac{1}{2}-\epsilon_t\right) \\
&= \prod_{t=1}^{T} \sqrt{1-4\gamma_t^2} \\
&\leq \prod_{t=1}^{T} \sqrt{\exp(-4\gamma_t^2)} \quad (1+x \leqslant \exp(x)) \\
&= \prod_{t=1}^{T} \exp\left(-2\gamma_t^2\right) \\
&= \exp\left(-2\sum_{t=1}^{T}\left(\frac{1}{2}-\epsilon_t\right)^2\right)
\end{aligned}$$

进一步地,若对于任意 $t \in [T], \gamma \leq \left(\frac{1}{2}-\epsilon_t\right)$, 则 $\gamma^2 \leq \left(\frac{1}{2}-\epsilon_t\right)^2, \sum_{t=1}^{T}\gamma^2 \leq \sum_{t=1}^{T}\left(\frac{1}{2}-\epsilon_t\right)^2$.

$$\begin{aligned}
\hat{R}_D(f) &\leq \exp\left(-2\sum_{t=1}^{T}\left(\frac{1}{2}-\epsilon_t\right)^2\right) \\
&\leq \exp\left(-2\sum_{t=1}^{T}\gamma^2\right) \\
&= \exp\left(-2\gamma^2 T\right)
\end{aligned}$$

(4) 原问题的拉格朗日函数:

$$\begin{aligned}
L(\boldsymbol{x},\lambda) &= \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{x}} \exp\left(-\frac{1}{K}\left(Y_1 H(\boldsymbol{x})_1 + \cdots + Y_K H(\boldsymbol{x})_K\right)\right) - \lambda\left(H(\boldsymbol{x})_1 + \cdots H(\boldsymbol{x})_K\right) \\
&= \sum_{i=1}^{K}\left(P(y=k \mid \boldsymbol{x}) \exp\left(-\frac{1}{K}\left(Y_1 H(\boldsymbol{x})_1 + \cdots + Y_K H(\boldsymbol{x})_K\right)\right) - \lambda H(\boldsymbol{x})_k\right) \\
&= \sum_{i=1}^{K}\left(P(y=k \mid \boldsymbol{x}) \exp\left(-\frac{1}{K}\left(H(\boldsymbol{x})_k - \frac{1}{K-1}\sum_{j\neq k} H(\boldsymbol{x})_j\right)\right) - \lambda H(\boldsymbol{x})_k\right) \\
&= \sum_{i=1}^{K}\left(P(y=k \mid \boldsymbol{x}) \exp\left(-\frac{1}{K}\left(H(\boldsymbol{x})_k - \frac{1}{K-1}(0 - H(\boldsymbol{x})_k)\right)\right) - \lambda H(\boldsymbol{x})_k\right) \\
&= \sum_{i=1}^{K}\left(P(y=k \mid \boldsymbol{x}) \exp\left(-\frac{H(\boldsymbol{x})_k}{K-1}\right) - \lambda H(\boldsymbol{x})_k\right)
\end{aligned}$$

求导可得:

$$\frac{\partial}{\partial H(\boldsymbol{x})_k}\mathcal{L}(\boldsymbol{x},\lambda) = -\frac{1}{K-1}P(y=k\mid\boldsymbol{x})\exp\left(-\frac{H(\boldsymbol{x})_k}{K-1}\right)-\lambda$$

令导数为 0, 可得:

$$H^\star(\boldsymbol{x})_k = (K-1)(\ln P(y=k\mid\boldsymbol{x}) - \ln(-(K-1)\lambda))$$

因为 $\sum_{k=1}^{K} H^\star(\boldsymbol{x})_k = 0$
所以 $\ln(-(K-1)\lambda) = \frac{1}{K}\ln\prod_{k=1}^{K}P(y=k\mid\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K}\ln P(y=k\mid\boldsymbol{x})$
所以 $H^\star(\boldsymbol{x})_k = (K-1)\left(\ln P(y=k\mid\boldsymbol{x}) - \frac{1}{K}\sum_{i=1}^{K}\ln P(y=i\mid\boldsymbol{x})\right)$
学习器的输出是一个 K 维向量，根据题目的定义，应该选择输出向量中最大的分量对应的位置作为预测类别，即预测类别 $\arg\max_{k\in[K]} H^\star(\boldsymbol{x})_k = \arg\max_{k\in[K]}\ln P(y=k\mid\boldsymbol{x}) = \arg\max_{k\in[K]} P(y=k\mid\boldsymbol{x})$.
说明 $\arg\max_{k\in[K]} H^\star(\boldsymbol{x})_k$ 达到了贝叶斯最优分类率，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.

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) 记 $p(x)$ 为样本分布的概率密度函数，个体学习器相互独立:

$$\forall m \neq l, \mathbb{E}_{\boldsymbol{x}}\left[\epsilon_m(\boldsymbol{x})\epsilon_l(\boldsymbol{x})\right] = 0$$

也即

$$\forall m \neq l, \int_{\boldsymbol{x}} p(\boldsymbol{x})\epsilon_m(\boldsymbol{x})\epsilon_\ell(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$$

$$
\begin{aligned}
\mathbb{E}_{\text{bag}} &= \int_{\boldsymbol{x}} p(\boldsymbol{x})\epsilon_{\text{bag}}^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \int_{\boldsymbol{x}} p(\boldsymbol{x})\frac{1}{M^2}\left(\sum_{m=1}^{M}\epsilon_m(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x} \\
&= \frac{1}{M^2}\int_{\boldsymbol{x}} p(\boldsymbol{x})\sum_{m=1}^{M}\sum_{l=1}^{M}\epsilon_m(\boldsymbol{x})\epsilon_\ell(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{l=1}^{M}\int_{\boldsymbol{x}} p(\boldsymbol{x})\epsilon_m(\boldsymbol{x})\epsilon_\ell(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\int_{\boldsymbol{x}} p(\boldsymbol{x})\epsilon_m^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \frac{2}{M^2}\sum_{1\le m<\ell\le M}\int_{\boldsymbol{x}} p(\boldsymbol{x})\epsilon_m(\boldsymbol{x})\epsilon_\ell(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\int_{\boldsymbol{x}} p(\boldsymbol{x})\epsilon_m^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}_x\left[\epsilon_m^2(\boldsymbol{x})\right] \\
&= \frac{1}{M}\boldsymbol{E}_{\text{avg}}
\end{aligned}
$$

(2) 证明如下:

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\left[\epsilon_{\text{bag}}^2(\boldsymbol{x})\right] =& \frac{1}{M^2}\sum_{m=1}^{M}\int_{\boldsymbol{x}} p(\boldsymbol{x})\epsilon_m^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \frac{2}{M^2}\sum_{1\le m<\ell\le M}\int_{\boldsymbol{x}} p(\boldsymbol{x})\epsilon_m(\boldsymbol{x})\epsilon_\ell(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
=& \frac{1}{M^2}\sum_{m=1}^{M}\left(\mathbb{E}_{\boldsymbol{x}}^2\left[\epsilon_m(\boldsymbol{x})\right] + \mathrm{Var}_{\boldsymbol{x}}\left(\epsilon_m(\boldsymbol{x})\right)\right) \\
& + \frac{2}{M^2}\sum_{1\le m<\ell\le M}\left(\mathbb{E}_{\boldsymbol{x}}^2\left[\epsilon_m(\boldsymbol{x})\right] + r_{\boldsymbol{x}}\left(\epsilon_m(\boldsymbol{x}),\epsilon_\ell(\boldsymbol{x})\right)\sqrt{\mathrm{Var}_{\boldsymbol{x}}^2\left(\epsilon_m(\boldsymbol{x})\right)\mathrm{Var}_{\boldsymbol{x}}^2\left(\epsilon_\ell(\boldsymbol{x})\right)}\right) \\
=& \frac{1}{M^2}\sum_{m=1}^{M}\left(\mu^2+\sigma^2\right) + \frac{2}{M^2}\sum_{1\le m<\ell\le M}\left(\mu^2+\rho\sigma^2\right) \\
=& \frac{1}{M}\left(\mu^2+\sigma^2\right) + \frac{M-1}{M}\left(\mu^2+\rho\sigma^2\right) \\
=& \mu^2+\rho\sigma^2+\frac{1-\rho}{M}\sigma^2 \\
\mathbb{E}_{\boldsymbol{x}}\left[\epsilon_{\text{bag}}\left(\boldsymbol{x}\right)\right] =& \int_{\boldsymbol{x}} p(\boldsymbol{x})\epsilon_{\text{bag}}\left(\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x} \\
=& \frac{1}{M}\int_{\boldsymbol{x}} p(\boldsymbol{x})\sum_{m=1}^{M}\epsilon_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
=& \frac{1}{M}\sum_{m=1}^{M}\int_{\boldsymbol{x}} p(\boldsymbol{x})\epsilon_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
=& \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\boldsymbol{x}}\left[\epsilon_m(\boldsymbol{x})\right] \\
=& \mu
\end{aligned}
$$

$\therefore \operatorname{Var}\left[\epsilon_{\text{bag}}\left(\boldsymbol{x}\right)\right] = \mathbb{E}_{\boldsymbol{x}}\left[\epsilon_{\text{bag}}^2\left(\boldsymbol{x}\right)\right] - \mathbb{E}_{\boldsymbol{x}}^2\left[\epsilon_{\text{bag}}\left(\boldsymbol{x}\right)\right] = \rho\sigma^2 + \frac{1-\rho}{M}\sigma^2$
随机森林算法主要通过以下方法降低基决策树之间的相关性:

(a) 随机采样训练样本：对于每棵基决策树，随机森林算法从原始训练集中进行有放回地随机采样，生成一个新的训练集。这样做的目的是引入样本的随机性，使得每棵决策树的训练集略有不同。

(b) 随机选择特征子集：对于每棵基决策树的每个节点，在进行分裂操作时，随机森林算法只考虑一个随机选择的特征子集来进行最优特征的选择。这样做的目的是引入特征的随机性，限制每棵决策树的特征选择，从而减少决策树之间的相关性。

(c) 构建多棵决策树：随机森林算法基于上述两个随机性来源构建多棵决策树。每棵决策树都是基于不同的训练集和随机选择的特征子集独立构建的。这样做的目的是使得每棵决策树具有一定的差异性，减少它们之间的相关性。

(3) 证明如下:

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