机器学习导论 习题五

211300024, 石睿, 211300024@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) 解:

$$\begin{aligned}
&\because h^*(x) = \arg\max_{c_i \in \mathcal{Y}} P(c_i|x) \\
&\therefore P(y = h^*(x)) = P(y = h^*(x)|x = x) = P(c^*|x) \\
&\therefore \text{令} c^* = h^*(x) = \arg\max_{c_i \in \mathcal{Y}} P(c_i|x) \geq P(c_i|x) \quad \text{其中} c_i \neq c^* \\
&\text{假设} P(y = h^*(x)) < \frac{1}{N} \\
&\therefore \sum_{j=1}^{N} P(c_j|x) \leq N \cdot P(y = h^*(x)|x) < N \cdot \frac{1}{N} = 1 \\
&\therefore \text{和} \sum_{j=1}^{N} P(c_j|x) = 1 \text{矛盾} \\
&\therefore P(y = h^*(x)) \geq \frac{1}{N}
\end{aligned}$$

(2) 解:

[1]

$$\because \sum_{j=1}^{N} \lambda_{ij} P(c_j|x) = \lambda_s P(c_1|x) + \cdots + 0 \cdot P(c_i|x) + \cdots + \lambda_s P(c_N|x)$$

$$\because \sum_{j=1}^{N} P(c_j|x) = 1$$

$$\therefore \lambda_s \left(P(c_1|x) + \cdots + P(c_{i-1}|x) + P(c_{i+1}|x) + \cdots + P(c_N|x) \right)$$

$$= \lambda_s \left(1 - P(c_i|x) \right)$$

$$\therefore R(c_i|x) = \begin{cases} \lambda_s - \lambda_s \cdot P(c_i|x) & c \in \mathcal{Y} \\ \lambda_r & \text{拒识}(c \notin \mathcal{Y}) \end{cases}$$

[2]

2.1分类规则

由拒识的定义，分类规则应规定成如下形式

$$h^*(x) = \begin{cases} \arg\max_{c_i \in \mathcal{Y}} P(c_i|x) & , \max_{c_i \in \mathcal{Y}} P(c_i|x) > 1 - \frac{\lambda_r}{\lambda_s} \\ \text{拒识} & , \max_{c_i \in \mathcal{Y}} P(c_i|x) \leq 1 - \frac{\lambda_r}{\lambda_s} \end{cases}$$

2.2拒识规则

若拒识的风险比把 x 分成已知类别$\mathcal{Y} = c_1, \cdots, c_N$的任何类别的风险还要小

也就是$\max\limits_{c_i \in \mathcal{Y}} \lambda_s \cdot (1 - P(c_i|x)) \geq \lambda_r$

那么就应该拒识这个样本，拒绝给出在已知类别内的分类结果

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 $H(\Theta \mid \Theta_t) = \sum_{\boldsymbol{Z}} P(\boldsymbol{Z} \mid \boldsymbol{X}, \Theta^t) \ln P(\boldsymbol{Z} \mid \boldsymbol{X}, \Theta)$, $Q(\Theta \mid \Theta^t)$ 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	0.44915	0.55085	2.24575	2.75425
HHHHTHHHHH	0.80499	0.19501	7.24491	1.75509
HTHHHHHTHH	0.73347	0.26653	5.86776	2.13224

(1) 解:

以下对参数形式做出规定:

Y_i代表第 i 轮（一共三轮）每轮结果正面朝上的硬币次数

Z_i代表第 i 轮（一共三轮）每轮是正在投 A 硬币还是 B 硬币

$$\therefore \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \quad \boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix}$$

本小问要求$P(Z_j|Y_j,\theta)$

$$\because P(Z_j|Y_j,\theta) = \frac{P(Y_j|\theta,Z_j)\cdot P(Z_j|\theta)}{P(Y_j|\theta)} = \frac{P(Y_j|\theta,Z_j)\cdot P(Z_j|\theta)}{\sum_z P(Y_j|Z_j,\theta)\cdot P(Z_j|\theta)}$$

$\because$ 此时由题$P(Z=A|\theta) = P(Z=B|\theta) = \dfrac{1}{2}$

$$\therefore P(Z_j|Y_j,\theta) = \frac{P(Y_j|\theta,Z_j)}{\sum_z P(Y_j|Z_j,\theta)}$$

$\because$ 同时，由于每一轮的二项分布

$$\therefore P(Y_j = m|Z_j = A,\theta) = C_{10}^m \theta_A^m (1-\theta_A)^{10-m}$$

$$\therefore P(Y_j = m|Z_j = B,\theta) = C_{10}^m \theta_B^m (1-\theta_B)^{10-m}$$

$\therefore$ 以下为第一轮的数据，Z1=A 为例计算数值，其他组同理啦

$$P(Z_1 = A|Y_1 = 5,\theta) = \frac{C_{10}^5(\theta_A^{(0)})^5(1-\theta_A^{(0)})^{10-5}}{C_{10}^5(\theta_A^{(0)})^5(1-\theta_A^{(0)})^{10-5} + C_{10}^5(\theta_B^{(0)})^5(1-\theta_B^{(0)})^{10-5}}$$

$$= \frac{0.6^5 0.4^5}{0.6^5 0.4^5 + 0.5^{10}} = 0.44915$$

(2) 解：

【法一：带入了数据】

此时考虑三轮实验的所有数据作为 EM 算法迭代的显变量，每轮所选硬币为隐变量

$$\therefore P(\boldsymbol{Y},\boldsymbol{Z}|\theta)=\prod_{j=1}^{3}P(Y_j,Z_j|\theta)=\prod_{j=1}^{3}P(Z_j|\theta)P(Y_j|Z_j,\theta)=\frac{1}{8}\prod_{j=1}^{3}P(Y_j|Z_j,\theta)$$

$$(*)=\frac{1}{8}P(Y_1=5|Z_1,\theta)P(Y_2=9|Z_2,\theta)P(Y_3=8|Z_3,\theta)$$

$$\therefore P(\boldsymbol{Z}|\boldsymbol{Y},\theta^i)=\prod_{j=1}^{3}P(Z_j|Y_j,\theta)=\prod_{j=1}^{3}\frac{P(Y_j|Z_j,\theta)}{\sum_z P(Y_j|Z_j,\theta^i)}$$

$$(**)=P(Z_1|Y_1,\theta)P(Z_2|Y_2,\theta)P(Z_3|Y_3,\theta)$$

以下令$C_j=P(\boldsymbol{Z}=D_j|Y,\theta^i)$

对每一个C_j，因为 Y 的取值在实验中已经固定，其实只有 Z 在变化啦，一共有 8 中可能的D_j

$$D_1=\begin{pmatrix}A\\A\\A\end{pmatrix}D_2=\begin{pmatrix}A\\A\\B\end{pmatrix}D_3=\begin{pmatrix}A\\B\\A\end{pmatrix}D_4=\begin{pmatrix}A\\B\\B\end{pmatrix}$$

$$D_5=\begin{pmatrix}B\\A\\A\end{pmatrix}D_6=\begin{pmatrix}B\\A\\B\end{pmatrix}D_7=\begin{pmatrix}B\\B\\A\end{pmatrix}D_8=\begin{pmatrix}B\\B\\B\end{pmatrix}$$

带入 (1) 中计算得到的结果组合一下，得

$C_1=0.26519,C_2=0.09637,C_3=0.06424,C_4=0.02335$

$C_5=0.32524,C_6=0.11819,C_7=0.07879,C_8=0.02863$

[1]E 步

以下设$Z_i=\begin{pmatrix}m_i\\n_i\\p_i\end{pmatrix}$

$$Q(\theta,\theta^i)=\sum_{\boldsymbol{Z}}\log P(\boldsymbol{Y},\boldsymbol{Z}|\theta)P(\boldsymbol{Z}|\boldsymbol{Y},\theta^i)$$

$$=\sum_{j=1}^{8}\log\left(\frac{1}{8}P(Y_1=5|Z_1=m_j,\theta)P(Y_2=9|Z_2=n_j,\theta)P(Y_3=8|Z_3=p_j,\theta)\right)\cdot C_j$$

$$=\sum_{j=1}^{8}\left(\log\frac{1}{8}+\log P(Y_1=5|Z_1=m_j,\theta)+\log P(Y_2=9|Z_2=n_j,\theta)+\log P(Y_3=8|Z_3=p_j,\theta)\right)C_j$$

因为只需要对θ求偏导即可，所以和θ无关的项可以去掉

以下只计算$Q(\theta,\theta^i)$的等价形式

$$
\begin{aligned}
Q(\theta,\theta^{i}) &= (C_1+C_2+C_3+C_4)\log P(Y_1=5|Z_1=A,\theta)\\
&+(C_5+C_6+C_7+C_8)\log P(Y_1=5|Z_1=B,\theta)\\
&+(C_1+C_2+C_5+C_6)\log P(Y_2=9|Z_2=A,\theta)\\
&+(C_3+C_4+C_7+C_8)\log P(Y_2=9|Z_2=B,\theta)\\
&+(C_1+C_3+C_5+C_7)\log P(Y_3=8|Z_3=A,\theta)\\
&+(C_2+C_4+C_6+C_8)\log P(Y_3=8|Z_3=B,\theta)
\end{aligned}
$$

[2]M 步

带入以上算出的所有的常数 C，求偏导计算即可，此处就不展开啦

$step1: \dfrac{\partial Q(\theta,\theta^0)}{\partial \theta_A}=0$

解得$15.35834-15.858834\theta_A-4.51766\theta_A=0$

$\therefore \theta_A^1=0.77271$

$step2: \dfrac{\partial Q(\theta,\theta^0)}{\partial \theta_B}=0$

解得$5.69614-5.696214\theta_B-3.245962\theta_B=0$

$\therefore \theta_B^1=0.65603$

【法二：对上面的计算结果进行化简】

$E(A_u)=E(A_u^1)+E(A_u^2)+E(A_u^3)=15.35842$

$E(A_d)=E(A_d^1)+E(A_d^2)+E(A_d^3)=4.51768$

$E(B_u)=E(B_u^1)+E(B_u^2)+E(B_u^3)=6.64158$

$E(B_d)=E(B_d^1)+E(B_d^2)+E(B_d^3)=3.48232$

$\theta_A^1=\dfrac{E(A_u)}{E(A_u)+E(A_d)}=0.77271$

$\theta_B^1=\dfrac{E(B_u)}{E(B_u)+E(B_d)}=0.65603$

(3) 解:

[1] 【从高斯混合分布出发考虑】

$\because Q(\theta,\theta^i)=\sum_z \log P(\boldsymbol{Y},\boldsymbol{Z}|\theta)P(\boldsymbol{Z}|\boldsymbol{Y},\theta^i)$

由题，$P_{\mathcal{M}}(x)=\sum_{i=1}^{k}\alpha_i P(x|\mu_i,\Sigma_i)$已知

它把样本集 D 划分成 k 个簇$C=C_1,C_2,\cdots,C_k$

【显】此时依第 i 个高斯模型所产生的的概率分布是已知的，是显数据。

【隐】反映数据 x j 来自第 i 个高斯模型的数据是未知的，是隐数据。

更多定义如下

1 $\lambda_j=\arg\max_{i\in\{1,2,\cdots,k\}}\gamma_{ji}$ 其中λ_j指示数据 x j 是由哪个高斯分布产生的

2 $\gamma_{ji}=P(Z_j=i|x=x_j,\theta^{(i)})$表示数据 x j 由第 i 个高斯分布产生的概率

3 Z_j表示数据 x j 被哪个高斯分布产生，是隐变量

4 x表示数据 x 的产生，是显变量

$$\begin{aligned}
&\gamma_{ji}=P(Z_j=i|x=x_j,\theta^{(m)})\\
&=\frac{P(Z_j=i|\theta^m)P(x=x_j|Z_j=i,\theta^m)}{P_{\mathcal{M}}(x_j|\theta^m)}\\
&=\frac{\alpha_i\frac{1}{(2\pi)^n|\epsilon|}exp(\frac{-1}{2\epsilon^2}(x_j-\mu_i)^T(x_j-\mu_i))}{\sum_{p=1}^{k}\alpha_p P(x=x_j|\theta^m)}\\
&=\frac{\alpha_i\cdot exp(\frac{-1}{2\epsilon^2}\left\|x_j-\mu_i\right\|^2)}{\sum_{p=1}^{k}\alpha_k\cdot exp(\frac{-1}{2\epsilon^2}\left\|x_j-\mu_p\right\|^2)}\\
&=\frac{\alpha_i}{\sum_{p=1}^{k}\alpha_k\cdot exp(\frac{-1}{2\epsilon^2}\left\|x_j-\mu_p\right\|^2-\left\|x_j-\mu_i\right\|^2)}
\end{aligned}$$

由题，在$i\neq p$的时候，上式中两范数的平方不相等

因$\lambda_j=\arg\max_{i\in\{1,2,\cdots,k\}}\gamma_{ji}$ 其中λ_j

$\therefore \lambda_j=\arg\max_{i\in[k]}\frac{\alpha_i}{\sum_{p=1}^{k}\alpha_k\cdot exp(\frac{-1}{2\epsilon^2}\left\|x_j-\mu_p\right\|^2-\left\|x_j-\mu_i\right\|^2)}$

$\because$ 随着$\epsilon^2->0$

所以上式中的以 e 为底的指数函数可能会趋向于正无穷或者 0，取决于两范数平方相减后正负号

[2] 【从 K-means 更新的角度考虑】

$$\gamma_{ij}=\begin{cases}1 & \left\|x_i-\mu_j\right\|^2\leq\left\|x_i-\mu_{j'}\right\|,\forall j'\\ 0 & otherwise\end{cases}$$

$\therefore$ 由上式，一定存在j^*,使$j=\arg\min_j\left\|x_i-\mu_j\right\|^2$

$\therefore$ 令$j^*=i^*$

$\therefore$ 对此$i^*,\left\|x_i-\mu_{i^*}\right\|-\left\|x_i-\mu_p\right\|^2\leq 0$

[3] 【以下证明，只有在$i^* = i$的时候，才能让γ_{ji}取得最大值】

假设$i \neq i^*$

$\therefore$ 一定存在 p，使得在分母累加的时候

$\alpha_p \cdot exp(\frac{1}{2\epsilon^2}(\|x_i - \mu_i\| - \|x_j - \mu_{i^*}\|))$

$= \alpha_p \cdot e^{\infty}$

$\therefore$ 此时 i 一定不能是可以让整体值最大的 i

$\because$ 在$i = i^*$的时候带入没有分母是正无穷的情况

$\therefore \lambda_j = i^*$

也即在$\epsilon^2 -> 0$的时候，高斯混合聚类中的 E 步会收敛到 k-means 算法中簇划分的更新规则上！

(4) 解：

$$Q(\theta|\theta^t) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^t} LL(\theta|\boldsymbol{X},\boldsymbol{Z})$$

$$= \sum_{\boldsymbol{Z}} P(\boldsymbol{Z}|\theta^t,\boldsymbol{X}) LL(\theta|\boldsymbol{X},\boldsymbol{Z})$$

$$= \sum_{\boldsymbol{Z}} P(\boldsymbol{Z}|\theta^t,\boldsymbol{X}) \log P(\boldsymbol{X},\boldsymbol{Z}|\theta)$$

$$\because P(\boldsymbol{X}|\theta) = \frac{P(\boldsymbol{X},\boldsymbol{Z}|\theta)}{P(\boldsymbol{Z}|\boldsymbol{X},\theta)}$$

$$\therefore logP(\boldsymbol{X}|\theta) = logP(\boldsymbol{X},\boldsymbol{Z}|\theta) - \log P(\boldsymbol{Z}|\boldsymbol{X},\theta)$$

$$\therefore Q(\theta|\theta^t) = \sum_{\boldsymbol{Z}} P(\boldsymbol{Z}|\theta^t,\boldsymbol{X}) \left(logP(\boldsymbol{X}|\theta) + \log P(\boldsymbol{Z}|\boldsymbol{X},\theta)\right)$$

$$= \sum_{\boldsymbol{Z}} P(\boldsymbol{Z}|\theta^t,\boldsymbol{X}) \log P(\boldsymbol{X}|\theta) + H(\theta|\theta^t)$$

$$= \log P(\boldsymbol{X}|\theta) \cdot \left(\sum_{\boldsymbol{Z}} P(\boldsymbol{Z}|\theta^t,\boldsymbol{X})\right) + H(\theta|\theta^t)$$

综上

$$\therefore \log P(\boldsymbol{X}|\theta) = LL(\theta|\boldsymbol{X}) = Q(\theta|\theta^t) - H(\theta|\theta^t)$$

(5) 解:

本题即要说明$\theta^t = \arg\max\limits_{\theta} H(\theta|\theta^t)$

$H(\theta|\theta^t) - H(\theta^t|\theta^t) = \sum\limits_{\boldsymbol{Z}} P(\boldsymbol{Z}|\boldsymbol{X},\theta^t)\log P(\boldsymbol{Z}|\boldsymbol{X},\theta) - \sum\limits_{\boldsymbol{Z}} P(\boldsymbol{Z}|\boldsymbol{X},\theta^t)\log P(\boldsymbol{Z}|\boldsymbol{X},\theta^t)$

$= \sum\limits_{\boldsymbol{Z}} P(\boldsymbol{Z}|\boldsymbol{X},\theta^t)\log\left(\dfrac{P(\boldsymbol{Z}|\boldsymbol{X},\theta)}{P(\boldsymbol{Z}|\boldsymbol{X},\theta^t)}\right)$

接下来应用琴生不等式

$\log\sum\limits_j \lambda_j y_j \geq \sum\limits_j \lambda_j \log y_j$

$\therefore H(\theta|\theta^t) - H(\theta^t|\theta^t)$

$\leq \sum\limits_{\boldsymbol{Z}} \log\left(\dfrac{P(\boldsymbol{Z}|\boldsymbol{X},\theta)\cdot P(\boldsymbol{Z}|\boldsymbol{X},\theta^t)}{P(\boldsymbol{Z}|\boldsymbol{X},\theta^t)}\right)$

$= \log\left(\sum\limits_{\boldsymbol{Z}} P(\boldsymbol{Z}|\boldsymbol{X},\theta)\right)$

$= \log 1 = 0$

等号当且仅当$\theta = \theta^t$时成立

(6) 解:

$$\begin{cases} \text{由 (4) 中可知} & LL(\theta|\boldsymbol{X}) = Q(\theta|\theta^t) - H(\theta|\theta^t) \\ \text{由 EM 更新规则可知} & \theta^{t+1} = \arg\max_\theta Q(\theta|\theta^t) \\ \text{由（5）中可知} & \theta^t = \arg\max_\theta H(\theta|\theta^t) \end{cases}$$

$\therefore LL(\theta^{t+1}|\boldsymbol{X}) - LL(\theta^t|\boldsymbol{X})$

$= \left[Q(\theta^{t+1}|\theta^t) - H(\theta^{t+1}|\theta^t)\right] - \left[Q(\theta^t|\theta^t) - H(\theta^t|\theta^t)\right]$

$= \left[Q(\theta^{t+1}|\theta^t) - Q(\theta^t|\theta^t)\right] + \left[H(\theta^t|\theta^t) - H(\theta^{t+1}|\theta^t)\right]$

≥ 0

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

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

(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 $H^\star(\boldsymbol{x})$, $\arg\max_{k\in[K]} H^\star(\boldsymbol{x})_k$ 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) 解:

$$\begin{aligned}
&\text{由书中定义}\mathcal{D}_t(x) = \mathcal{D}(x) \cdot \frac{exp(-f(x) \cdot H_{t-1}(x))}{\mathbb{E}_{x\sim\mathcal{D}}\left(exp(-f(x) \cdot H_{t-1}(x))\right)} \\
&\therefore \mathcal{D}_{t+1}(x) = \mathcal{D}(x) \cdot \frac{exp(-f(x) \cdot H_t(x))}{\mathbb{E}_{x\sim\mathcal{D}}\left(exp(-f(x) \cdot H_t(x))\right)} \\
&= \mathcal{D}(x) \cdot \frac{exp(-f(x) \cdot H_{t-1}(x)) \cdot exp(-f(x)\alpha_t h_t(x))}{\mathbb{E}_{x\sim\mathcal{D}}\left(exp(-f(x) \cdot H_t(x))\right)} \\
&= \mathcal{D}_t(x) \cdot \frac{\mathbb{E}_{x\sim\mathcal{D}}\left(exp(-f(x) \cdot H_{t-1}(x))\right)}{exp(-f(x) \cdot H_{t-1}(x))} \frac{exp(-f(x) \cdot H_{t-1}(x)) \cdot exp(-f(x)\alpha_t h_t(x))}{\mathbb{E}_{x\sim\mathcal{D}}\left(exp(-f(x) \cdot H_t(x))\right)} \\
&= \mathcal{D}_t(x) \exp(-f(x)\alpha_t h_t(x)) \frac{1}{Z_t} \\
&\because \mathcal{D}_1(x) = \frac{\mathcal{D}(x) exp(-f(x))}{\mathbb{E}_{x\sim\mathcal{D}} exp(-f(x))} = \frac{\mathcal{D}(x) exp(-f(x))}{m \cdot \mathcal{D}(x) \cdot exp(-f(x))} = \frac{1}{m}
\end{aligned}$$

综上，不断累乘带入即可得到以下答案

$$\begin{aligned}
&\therefore \mathcal{D}_{t+1}(x) = \frac{exp(-f(x)\sum_{s=1}^{t} \alpha_s h_s(x))}{m \cdot \prod_{s=1}^{t} Z_s} \\
&= \frac{exp(-y\sum_{s=1}^{t} \alpha_s h_s(x))}{m \cdot \prod_{s=1}^{t} Z_s}
\end{aligned}$$

(2) 解:

$$
\begin{aligned}
Z_t &= \frac{\mathbb{E}_{x\sim\mathcal{D}}\left(exp(-f(x)H_t(x))\right)}{\mathbb{E}_{x\sim\mathcal{D}}\left(exp(-f(x)H_{t-1}(x))\right)}\\
&= \frac{\sum_{i=1}^{|D|}\mathcal{D}(x_i)\cdot exp(-f(x_i)H_{t-1}(x_i))\cdot exp(-f(x_i)\alpha_t h_t(x))}{\mathbb{E}_{x\sim\mathcal{D}}\left(exp(-f(x)H_{t-1}(x))\right)}\\
&= \sum_{i=1}^{|D|}\mathcal{D}_t(x_i)\cdot exp(-f(x_i)\alpha_t h_t(x))\\
&= \mathbb{E}_{x\sim\mathcal{D}_t}\left(exp(-f(x)\alpha_t h_t(x))\right)
\end{aligned}
$$

$\because$ 在更新$\mathcal{D}_t(x)$之前，$\alpha_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$已经算出来了，可以直接带入呢

$$
\begin{aligned}
\therefore Z_t &= \mathbb{E}_{x\sim\mathcal{D}_t}\left(exp(-f(x)\alpha_t h_t(x))\right)\\
&= \mathbb{E}_{x\sim\mathcal{D}_t}\left((e^{\alpha_t})^{-f(x)h_t(x)}\right)\\
&= \mathbb{E}_{x\sim\mathcal{D}_t}\left((\frac{1-\epsilon_t}{\epsilon_t})^{-\frac{1}{2}f(x)h_t(x)}\right)\\
&= \sum_{i=1}^{|\mathcal{D}|}\mathcal{D}_t(x_i)\cdot\left(\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}\cdot\mathbf{I}(f(x)\neq h_t(x))+\sqrt{\frac{\epsilon_t}{1-\epsilon_t}}\cdot\mathbf{I}(f(x)=h_t(x))\right)
\end{aligned}
$$

$$
\therefore \boldsymbol{Z}_t = \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}\cdot P_{x\sim\mathcal{D}_t}(f(x)\neq h_t(x)|x)+\sqrt{\frac{\epsilon_t}{1-\epsilon_t}}\cdot P_{x\sim\mathcal{D}_t}(f(x)=h_t(x)|x)
$$

$$
\because \epsilon_t = P_{x\sim\mathcal{D}_t}(f(x)\neq h_t(x)|x)
$$

$$
\begin{aligned}
\therefore Z_t &= \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}\cdot(1-\epsilon_t)+\sqrt{\frac{\epsilon_t}{1-\epsilon_t}}\cdot\epsilon_t\\
&= 2\sqrt{\epsilon_t\cdot(1-\epsilon_t)}
\end{aligned}
$$

(3) 解:

$$\text{由题目给出的提示}\hat{R}_D(f) \le \frac{1}{m} \cdot \sum_{i=1}^{m} exp(-f(x)H_T(x))$$

由（1）中的结论

$$\hat{R}_D(f) \le \frac{1}{m} \sum_{i=1}^{m} \mathcal{D}_{t+1}(x_i) \cdot m \cdot \prod_{j=1}^{T} Z_j$$

$$= \prod_{j=1}^{T} Z_j \sum_{i=1}^{m} \mathcal{D}_{t+1}(x_i)$$

$$= \prod_{j=1}^{T} Z_j$$

$$\therefore \hat{R}_D(f) \le \prod_{j=1}^{T} Z_j = \prod_{j=1}^{T} 2\sqrt{\epsilon_j \cdot (1-\epsilon_j)} = \prod_{t=1}^{T} 2 \cdot \sqrt{-\epsilon_t^2 - \epsilon_t}$$

$$\because x \le e^{x-1}$$

$$\therefore \sqrt{x} \le (e^{x-1})^{\frac{1}{2}} = e^{\frac{x-1}{2}}$$

$$\therefore \text{令}\sqrt{x} = \sqrt{4 \cdot (\epsilon_t^2 + \epsilon_t)}$$

$$\therefore \hat{R}_D(f) \le \prod_{t=1}^{T} e^{-2 \cdot (\epsilon_t - \frac{1}{2})^2} = e^{-2 \cdot \sum_{t=1}^{T} (\frac{1}{2} - \epsilon_t)^2}$$

(4) 解:

$$\text{由题}l_{multi-exp}(H|\mathcal{D}) = \sum_{i=1}^{m} \mathcal{D}(x_i) \cdot e^{-\frac{1}{K} y_i^T H(x_i)}$$

$$\text{其中}y_i = \begin{pmatrix} -\frac{1}{K-1} \\ \vdots \\ 1 \\ \vdots \\ -\frac{1}{K-1} \end{pmatrix} \quad H(x_i) = \begin{pmatrix} H(x_i)_1 \\ \vdots \\ H(x_i)_K \end{pmatrix}$$

[1] 【对优化问题引入拉格朗日乘子】$\lambda \in \mathcal{R}$

$$L(H(x), \lambda)$$

$$= l_{multi-exp}(H|D) + \lambda \cdot (H(x)_1 + \cdots + H(x)_K)$$

$$= \sum_{i=1}^{m} \mathcal{D}(x_i) \cdot e^{-\frac{1}{K} y_i^T H(x_i)} + \lambda \left(H(x)_1 + \cdots + H(x)_K\right)$$

$\because L(H(x),\lambda)$是关于 H(x) 的凸函数

而且，优化问题不含有不等式约束。

所以由凸优化问题 +Slater 条件，强凸性满足。KKT 条件转化为充要条件

因为此时 H(x) 是一个 K 维向量，所以书中直接对 H(x) 求偏导的方法不再可以使用

同时因为此时 H(x) 每个分量的值会对结果产生影响，故以下对 H(x) 的第 q 个分量进行求偏导

[2] 【先把 L 整理成和 H(x) 的各个分量相关的表达式】

$$\sum_{i=1}^{m}\mathcal{D}(x_i)e^{\frac{-1}{K}y_i^T H(x_i)}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{K}P(c_i=j|x_i)\cdot exp\left(-\frac{1}{K}(\frac{-1}{K-1}\sum_{p=1,p\neq j}^{K}H(x_i)_p + H(x)_j)\right)$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{K}P(c_i=j|x_i)\cdot exp\left(\frac{1}{K\cdot(K-1)}\sum_{p=1,p\neq j}^{K}H(x_i)_p - \frac{1}{K}H(x)_j\right)$$

$\therefore L(H(x),\lambda)$

$$= \sum_{i=1}^{m}\sum_{j=1}^{K}P(c_i=j|x_i)\cdot exp\left(\frac{1}{K\cdot(K-1)}\sum_{p=1,p\neq j}^{K}H(x_i)_p - \frac{1}{K}H(x)_j\right) + \lambda\cdot\sum_{p=1}^{K}H(x)_p$$

[3] 【使用 KKT 条件的充要条件，五条中的最后一条】

以下引入中间变量$Q(x,j)$

$$Q(x,j) = exp\left(\frac{1}{K(K-1)}\sum_{p=1,p\neq j}^{K}H(x)_p - \frac{1}{K}H(x)_j\right)$$

$$\frac{\partial L(H(x),\lambda)}{\partial H(x)_q} = \sum_{j=1,j\neq q}^{K}P_{x\sim\mathcal{D}}(c=j|x)\cdot\frac{1}{K(K-1)}\cdot Q(x,j) - P_{x\sim\mathcal{D}}(c=q|x)\frac{1}{K}Q(x,q)+\lambda$$

令$\dfrac{\partial L(H(x),\lambda)}{\partial H(x)_q} = 0$

化简可以得到

$$H^*(x)_q = (K-1)\log P_{x\sim\mathcal{D}}(c=q|x)$$

$$-(K-1)\log\left(\sum_{j=1,j\neq q}^{K}P_{x\sim\mathcal{D}}(c=j|x)\cdot Q(x,j)\cdot exp(\frac{-1}{K(K-1)}\sum_{p=1,p\neq q}^{K}H(x)_p)\right)$$

$\therefore$ 发现后面那一项和 H(x) 的第 q 个分量无关

$\therefore \arg\max_{q\in[K]} H^*(x)_q = \arg\max_{q\in[K]} P_{x\sim\mathcal{D}}(c=q|x)$

$= \arg\max_{q\sim[K]} P_{x\in\mathcal{D}}(f(x)=q|x)$

结果达到了贝叶斯最优分类错误率

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) 解:

$$
\begin{aligned}
E_{bag} &= \mathbb{E}_x\left((\epsilon_{bag}(x))^2\right) \\
&= \mathbb{E}_x\left((\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(x))^2\right) \\
&= \mathbb{E}_x\left(\frac{1}{M^2}(\sum_{m=1}^{M}\epsilon_M(x)^2 + 2\cdot\sum_{1\le i\le j\le M, i\ne j}\epsilon_i(x)\cdot\epsilon_j(x))\right) \\
&= \frac{1}{M^2}\mathbb{E}_x\left(\sum_{m=1}^{M}\epsilon_M(x)^2\right) + \frac{2}{M^2}\left(\sum_{1\le i\le j\le M, i\ne j}\mathbb{E}_x(\epsilon_i(x)\cdot\epsilon_j(x))\right)
\end{aligned}
$$

由独立性假设

$$
\therefore E_{bag} = \frac{1}{M^2}\mathbb{E}_x\left(\sum_{m=1}^{M}\epsilon_M(x)^2\right) = \frac{1}{M}E_{avg}
$$

(2) 解:

$$
[1]Var(\epsilon_{bag}(x)) = Var(\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(x))
$$

$$
\begin{aligned}
&= \frac{1}{M^2}\cdot Var(\sum_{m=1}^{M}\epsilon_m(x)) \\
&= \frac{1}{M^2}\left(\sum_{m=1}^{M}Var(\epsilon_m(x)) + 2\sum_{1\le i\le j\le M, i\ne j}Cov(\epsilon_i(x),\epsilon_j(x))\right) \\
&= \frac{1}{M^2}\left(M\sigma^2 + 2\frac{M(M-1)}{2}\rho\sigma^2\right) \\
&= \rho\sigma^2 + \frac{1}{M}\sigma^2(1-\rho)
\end{aligned}
$$

[2]随机森林在决策树的训练过程中引入了随机属性选择,

对每个结点，不直接从其所拥有的属性集合中选，而在此集合中随机选包含 k 个属性的子集,

再从此 k 个中选最优的属性作为划分.

故同时有样本扰动 + 属性扰动，使个体学习器的差异度增加

(3) 解:

$$
\begin{aligned}
&\because E_{avg} = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_x\left((\epsilon_m(x))^2\right) \\
&\because E_{bag} = \mathbb{E}_x\left((\epsilon_{bag}(x))^2\right) \\
&= \mathbb{E}_x\left((\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(x))^2\right) \\
&\leq \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_x\left((\epsilon_m(x))^2\right) \\
&= E_{avg}
\end{aligned}
$$