机器学习导论
习题五

191250026, 丁云翔, 191250026@smail.nju.edu.cn

2021 年 6 月 5 日

学术诚信

本课程非常重视学术诚信规范，助教老师和助教同学将不遗余力地维护作业中的学术诚信规范的建立。希望所有选课学生能够对此予以重视。[1]

(1) 允许同学之间的相互讨论，但是**署你名字的工作必须由你完成**，不允许直接照搬任何已有的材料，必须独立完成作业的书写过程;

(2) 在完成作业过程中，对他人工作（出版物、互联网资料）中文本的直接照搬（包括原文的直接复制粘贴及语句的简单修改等）都将视为剽窃，剽窃者成绩将被取消。**对于完成作业中有关键作用的公开资料，应予以明显引用**；

(3) 如果发现作业之间高度相似将被判定为互相抄袭行为，**抄袭和被抄袭双方的成绩都将被取消**。因此请主动防止自己的作业被他人抄袭。

作业提交注意事项

(1) 请在LaTeX模板中**第一页填写个人的姓名、学号、邮箱信息**；

(2) 本次作业需提交该pdf文件，pdf文件名格式为**学号_姓名.pdf**，例如190000001_张三.pdf，**需通过教学立方提交**。

(3) 未按照要求提交作业，或提交作业格式不正确，将会**被扣除部分作业分数**；

(4) 本次作业提交截止时间为**6月6日23:55:00。**

[1] 参考尹一通老师高级算法课程中对学术诚信的说明。

1 [30pts] PCA

$\boldsymbol{x} \in \mathbb{R}^D$是一个随机向量，其均值和协方差分别是$\boldsymbol{\mu_x} = \mathbb{E}(\boldsymbol{x}) \in \mathbb{R}^D$，$\Sigma_x = \mathbb{E}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top \in \mathbb{R}^{D\times D}$。定义随机变量$y_i = \boldsymbol{u}_i^\top \boldsymbol{x} + a_i \in \mathbb{R}, i = 1, \ldots, d \leq D$为$\boldsymbol{x}$的主成分，其中$\boldsymbol{u}_i \in \mathbb{R}^D$是单位向量($\boldsymbol{u}_i \top \boldsymbol{u}_i = 1$)，$a_i \in \mathbb{R}$，$\{y_i\}_{i=1}^d$是互不相关的零均值随机变量，它们的方差满足$\mathrm{Var}\,(y_1) \geq \mathrm{Var}\,(y_2) \geq \cdots \geq \mathrm{Var}\,(y_d)$。假设$\Sigma_x$没有重复的特征值，请证明：

1. **[5pts]** $a_i = -\boldsymbol{u}_i^\top \boldsymbol{\mu_x}, i = 1, \ldots, d$。
2. **[10pts]** $\boldsymbol{u}_1$是Σ_x最大的特征值对应的特征向量。

   提示：写出要最大化的目标函数，写出约束条件，使用拉格朗日乘子法。
3. **[15pts]** $\boldsymbol{u}_2^\top \boldsymbol{u}_1 = 0$，且$\boldsymbol{u}_2$是$\Sigma_x$第二大特征值对应的特征向量。

   提示：由$\{y_i\}_{i=1}^n$是互不相关的零均值随机变量可推出$\boldsymbol{u}_2^\top \boldsymbol{u}_1 = 0$。$\boldsymbol{u}_2^\top \boldsymbol{u}_1 = 0$ 可作为第二小问的约束条件之一。

Solution. 此处用于写解答(中英文均可)

(1)

$\{y_i\}_{i=1}^d$是互不相关的零均值随机变量，故有

$$\begin{aligned}
\mathbb{E}(y_i) = 0,\ i = 1, \ldots, d \\
\mathbb{E}(\boldsymbol{u}_i^\top \boldsymbol{x} + a_i) = 0,\ i = 1, \ldots, d \\
\boldsymbol{u}_i^\top \mathbb{E}(\boldsymbol{x}) + a_i = 0,\ i = 1, \ldots, d \\
a_i = -\boldsymbol{u}_i^\top \mathbb{E}(\boldsymbol{x}),\ i = 1, \ldots, d \\
a_i = -\boldsymbol{u}_i^\top \boldsymbol{\mu_x},\ i = 1, \ldots, d
\end{aligned}$$

证明完毕。

(2)

由$y_i = \boldsymbol{u}_i^\top \boldsymbol{x} + a_i = \boldsymbol{u}_i^\top (\boldsymbol{x} - \boldsymbol{\mu_x})$可得

$$y_i y_i^T = \boldsymbol{u}_i^\top (\boldsymbol{x} - \boldsymbol{\mu_x})(\boldsymbol{x} - \boldsymbol{\mu_x})^\top \boldsymbol{u}_i \tag{1}$$

又$\{y_i\}_{i=1}^d$是互不相关的零均值随机变量，$\boldsymbol{x} - \boldsymbol{\mu_x}$为常量，故有$\left\{\boldsymbol{u}_i^\top\right\}_{i=1}^d$互不相关，即$\{\boldsymbol{u}_i\}_{i=1}^d$互不相关。又$\boldsymbol{u}_i \in \mathbb{R}^D$是单位向量，故$\{\boldsymbol{u}_i\}_{i=1}^d$是一组标准正交基。故(1)式为投影后样本点的协方差矩阵。又$\mathrm{Var}\,(y_i) = y_i y_i^T$，且$\mathrm{Var}\,(y_1) \geq \mathrm{Var}\,(y_2) \geq \cdots \geq \mathrm{Var}\,(y_d)$，故最大化目标函数为$\mathrm{Var}\,(y_i)$，联立(1)式，给出以下最大化目标

$$\begin{aligned}
&\max_{\boldsymbol{u}_i \in \{\boldsymbol{u}_i\}_{i=1}^d} \quad \boldsymbol{u}_i^\top (\boldsymbol{x} - \boldsymbol{\mu_x})(\boldsymbol{x} - \boldsymbol{\mu_x})^\top \boldsymbol{u}_i \\
&s.t. \quad \boldsymbol{u}_i^\top \boldsymbol{u}_i = 1, \quad \boldsymbol{u}_i \in \{\boldsymbol{u}_i\}_{i=1}^d
\end{aligned} \tag{2}$$

则有拉格朗日函数为

$$\mathbf{L}(\boldsymbol{u}_i, \lambda_i) = \boldsymbol{u}_i^\top (\boldsymbol{x} - \boldsymbol{\mu_x})(\boldsymbol{x} - \boldsymbol{\mu_x})^\top \boldsymbol{u}_i - \lambda_i (\boldsymbol{u}_i^\top \boldsymbol{u}_i - 1) \tag{3}$$

分别对$\boldsymbol{u}_i$和λ_i求导，并令它们等于0，可得

$$\frac{\partial \mathbf{L}}{\partial \boldsymbol{u}_i} = 2(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})^\top \boldsymbol{u}_i - 2\lambda_i \boldsymbol{u}_i = 0$$
$$\frac{\partial \mathbf{L}}{\partial \lambda_i} = \boldsymbol{u}_i^\top \boldsymbol{u}_i - 1 = 0$$

即

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})^\top \boldsymbol{u}_i &= \lambda_i \boldsymbol{u}_i \\ \Sigma_x \boldsymbol{u}_i &= \lambda_i \boldsymbol{u}_i \end{aligned} \tag{4}$$

可得$\boldsymbol{u}_i$为Σ_x的特征向量，λ_i为Σ_x的特征值，在(4)式左右两边同时左乘$\boldsymbol{u}_i^\top$可得

$$\begin{aligned}\boldsymbol{u}_i^\top \Sigma_x \boldsymbol{u}_i &= \boldsymbol{u}_i^\top \lambda_i \boldsymbol{u}_i \\ &= \lambda_i \boldsymbol{u}_i^\top \boldsymbol{u}_i \\ &= \lambda_i \end{aligned} \tag{5}$$

$\boldsymbol{u}_i^\top \Sigma_x \boldsymbol{u}_i$即为先前最大化的目标函数，联立(1)(2)(5)，最大化$\mathrm{Var}\,(y_i)$即最大化λ_i，由于方差$\mathrm{Var}\,(y_1)$为$\mathrm{Var}\,(y_i)$中的最大值，故特征值λ_1也为λ_i中的最大值，故有$\boldsymbol{u}_1$是Σ_x最大的特征值对应的特征向量，证明完毕。

(3)

由第(2)问证明过程中已得到的结论：$\{\boldsymbol{u}_i\}_{i=1}^d$互不相关，可知$\boldsymbol{u}_i^\top \boldsymbol{u}_1 = 0,\ \boldsymbol{u}_i \in \{\boldsymbol{u}_i\}_{i=2}^d$，特别地，当$i=2$时，$\boldsymbol{u}_2^\top \boldsymbol{u}_1 = 0$。将$\boldsymbol{u}_i^\top \boldsymbol{u}_1 = 0,\ \boldsymbol{u}_i \in \{\boldsymbol{u}_i\}_{i=2}^d$加入约束条件，则此时$\boldsymbol{u}_1$不在最优解范围中，由此可得新的优化目标和拉格朗日函数

$$\begin{aligned}\max_{\boldsymbol{u}_i \in \{\boldsymbol{u}_i\}_{i=2}^d} \quad & \boldsymbol{u}_i^\top (\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})^\top \boldsymbol{u}_i \\ s.t. \quad & \boldsymbol{u}_i^\top \boldsymbol{u}_i = 1, \quad \boldsymbol{u}_i \in \{\boldsymbol{u}_i\}_{i=1}^d \\ & \boldsymbol{u}_i^\top \boldsymbol{u}_1 = 0, \quad \boldsymbol{u}_i \in \{\boldsymbol{u}_i\}_{i=2}^d \end{aligned} \tag{6}$$

$$\mathbf{L}(\boldsymbol{u}_i, \lambda_i, \beta_i) = \boldsymbol{u}_i^\top (\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})^\top \boldsymbol{u}_i - \lambda_i(\boldsymbol{u}_i^\top \boldsymbol{u}_i - 1) - \beta_i \boldsymbol{u}_i^\top \boldsymbol{u}_1 \tag{7}$$

分别对$\boldsymbol{u}_i$, λ_i和β_i求导，并令它们等于0，可得

$$\frac{\partial \mathbf{L}}{\partial \boldsymbol{u}_i} = 2(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})^\top \boldsymbol{u}_i - 2\lambda_i \boldsymbol{u}_i - \beta_i \boldsymbol{u}_1 = 0$$
$$\frac{\partial \mathbf{L}}{\partial \lambda_i} = \boldsymbol{u}_i^\top \boldsymbol{u}_i - 1 = 0$$
$$\frac{\partial \mathbf{L}}{\partial \beta_i} = \boldsymbol{u}_i^\top \boldsymbol{u}_1 = 0$$

即

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})^\top \boldsymbol{u}_i - \frac{1}{2}\beta_i \boldsymbol{u}_1 &= \lambda_i \boldsymbol{u}_i \\ \Sigma_x \boldsymbol{u}_i - \frac{1}{2}\beta_i \boldsymbol{u}_1 &= \lambda_i \boldsymbol{u}_i \end{aligned} \tag{8}$$

在(8)式左右两边同时左乘$\boldsymbol{u}_1^\top$可得

$$\begin{aligned}\boldsymbol{u}_1^\top\Sigma_x\boldsymbol{u}_i-\boldsymbol{u}_1^\top\frac{1}{2}\beta_i\boldsymbol{u}_1&=\boldsymbol{u}_1^\top\lambda_i\boldsymbol{u}_i\\ \boldsymbol{u}_1^\top\Sigma_x\boldsymbol{u}_i-\frac{1}{2}\beta_i\boldsymbol{u}_1^\top\boldsymbol{u}_1&=\lambda_i\boldsymbol{u}_1^\top\boldsymbol{u}_i\\ \boldsymbol{u}_1^\top\Sigma_x\boldsymbol{u}_i&=\frac{1}{2}\beta_i\end{aligned}\tag{9}$$

发现$\boldsymbol{u}_1^\top\Sigma_x\boldsymbol{u}_i$为标量，又$\Sigma_x$为对称矩阵，故将$\boldsymbol{u}_1^\top\Sigma_x\boldsymbol{u}_i$转置得

$$\begin{aligned}(\boldsymbol{u}_1^\top\Sigma_x\boldsymbol{u}_i)^\top&=\frac{1}{2}\beta_i\\ \boldsymbol{u}_i^\top\Sigma_x\boldsymbol{u}_1&=\frac{1}{2}\beta_i\end{aligned}\tag{10}$$

由(4)式，可知$\Sigma_x\boldsymbol{u}_1=\lambda_1\boldsymbol{u}_1$，故(10)式等价于

$$\begin{aligned}\boldsymbol{u}_i^\top\lambda_1\boldsymbol{u}_1&=\frac{1}{2}\beta_i\\ \lambda_1\boldsymbol{u}_i^\top\boldsymbol{u}_1&=\frac{1}{2}\beta_i\\ 0&=\frac{1}{2}\beta_i\\ \beta_i&=0\end{aligned}$$

故(8)式可化简为

$$\Sigma_x\boldsymbol{u}_i=\lambda_i\boldsymbol{u}_i\tag{11}$$

可得$\boldsymbol{u}_i$为Σ_x的特征向量，λ_i为Σ_x的特征值，在(11)式左右两边同时左乘$\boldsymbol{u}_i^\top$可得

$$\begin{aligned}\boldsymbol{u}_i^\top\Sigma_x\boldsymbol{u}_i&=\boldsymbol{u}_i^\top\lambda_i\boldsymbol{u}_i\\ &=\lambda_i\boldsymbol{u}_i^\top\boldsymbol{u}_i\\ &=\lambda_i\end{aligned}\tag{12}$$

$\boldsymbol{u}_i^\top\Sigma_x\boldsymbol{u}_i$即为先前最大化的目标函数，联立(1)(6)(12)，最大化$\mathrm{Var}\left(y_i\right)$即最大化$\lambda_i$，由于约束条件中有$\boldsymbol{u}_i^\top\boldsymbol{u}_1=0$，所以参数取值范围为$\boldsymbol{u}_i\in\{\boldsymbol{u}_i\}_{i=2}^d$，故$\boldsymbol{u}_i\neq\boldsymbol{u}_1$，方差$\mathrm{Var}\left(y_2\right)$为$\mathrm{Var}\left(y_i\right)$中除$\mathrm{Var}\left(y_1\right)$以外的最大值，故特征值$\lambda_2$也为$\lambda_i$中除$\lambda_1$之外的的最大值，即第二大特征值，故$\boldsymbol{u}_2$是$\Sigma_x$第二大特征值对应的特征向量，证明完毕。

2 [30pts] Clustering

考虑p维特征空间里的混合模型

$$g(x)=\sum_{k=1}^{K}\pi_k g_k(x)$$

其中$g_k=N\left(\mu_k,\mathbf{I}\cdot\sigma^2\right)$，$\mathbf{I}$是单位矩阵，$\pi_k>0$，$\sum_k\pi_k=1$。$\left\{\mu_k,\pi_k\right\},k=1,\ldots,K$和$\sigma^2$是未知参数。

设有数据$x_1,x_2,\ldots,x_N\sim g(x)$,

1. **[10pts]** 请写出数据的对数似然。

2. **[15pts]** 请写出求解极大似然估计的EM算法。

3. **[5pts]** 请简要说明如果σ的值已知，并且$\sigma \to 0$，那么该EM算法就相当于K-means聚类。

Solution. 此处用于写解答(中英文均可)

(1)

设数据$x_1, x_2, \ldots, x_N \sim g(x)$均属于样本集$D$，则数据的对数似然为

$$\begin{aligned} LL(D) &= \ln\left(\prod_{j=1}^{N} g(x_j)\right) \\ &= \sum_{j=1}^{N} \ln g(x_j) \\ &= \sum_{j=1}^{N} \ln\left(\sum_{i=1}^{K} \pi_i \cdot g_i(x_j)\right) \end{aligned} \tag{13}$$

(2)

令随机变量$z_j \in \{1, 2, ..., K\}$表示生成样本x_j的高斯混合成分，其取值未知。z_j的先验概率$P(z_j = i)$对应于π_i $(i = 1, 2, ..., K)$.根据贝叶斯定理，z_j的后验分布对应于

$$\begin{aligned} p_M(z_j = i \mid x_j) &= \frac{P(z_j = i) \cdot p_M(x_j \mid z_j = i)}{p_M(x_j)} \\ &= \frac{\pi_i \cdot g_i(x_j)}{\sum_{l=1}^{K} \pi_l \cdot g_l(x_j)} \end{aligned} \tag{14}$$

即$p_M(z_j = i \mid x_j)$给出了样本x_j由第i个高斯混合成分生成的后验概率，将其简记为γ_{ji} $(i = 1, 2, ..., K)$。若参数$\{(\pi_i, \mu_i) \mid 1 \le i \le k,\ \mathbf{I} \cdot \sigma^2\}$能使(13)式最大化，则由$\frac{\partial LL(D)}{\partial \mu_i} = 0$有

$$\sum_{j=1}^{N} \frac{\pi_i \cdot g_i(x_j)}{\sum_{l=1}^{K} \pi_l \cdot g_l(x_j)} (x_j - \mu_i) = 0 \tag{15}$$

由(14)式和$\gamma_{ji} = p_M(z_j = i \mid x_j)$，有

$$\mu_i = \frac{\sum_{j=1}^{N} \gamma_{ji} x_j}{\sum_{j=1}^{N} \gamma_{ji}} \tag{16}$$

类似的，由$\frac{\partial LL(D)}{\partial \mathbf{I} \cdot \sigma^2} = 0$可得

$$\mathbf{I} \cdot \sigma^2 = \frac{1}{K} \sum_{i=1}^{K} \frac{\sum_{j=1}^{N} \gamma_{ji} (x_j - \mu_i)(x_j - \mu_i)^\top}{\sum_{j=1}^{N} \gamma_{ji}} \tag{17}$$

对于混合系数π_i，除了要最大化$LL(D)$，还需满足$\pi_k > 0$，$\sum_{k=1}^{K} \pi_k = 1$，考虑$LL(D)$的拉格朗日形式

$$LL(D) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1\right) \tag{18}$$

其中λ为拉格朗日乘子。由(18)式对π_k的导数为0，有

$$\sum_{j=1}^{N} \frac{g_i(x_j)}{\sum_{l=1}^{K} \pi_l \cdot g_l(x_j)} + \lambda = 0 \tag{19}$$

两边同时乘以π_i，对所有混合成分求和可知$\lambda = -N$，有

$$\pi_i = \frac{1}{N} \sum_{j=1}^{N} \gamma_{ji} \tag{20}$$

由上述推导即可获得求解高斯混合模型极大似然估计的EM算法：

1. 初始化高斯混合分布的模型参数$\{(\pi_i, \mu_i) \mid 1 \leq i \leq k,\ \mathbf{I} \cdot \sigma^2\}$
2. **repeat**
3. **for** $j = 1, 2, ..., N$ **do**
4. 根据(14)式计算x_j由各混合成分生成的后验概率，即$\gamma_{ji} = p_M(z_j = i \mid x_j)\ (1 \leq i \leq K)$。
5. **end for**
6. $(\mathbf{I} \cdot \sigma^2)' = 0$
7. **for** $i = 1, 2, ..., K$ **do**
8. 计算新均值向量：$\mu_i' = \frac{\sum_{j=1}^{N} \gamma_{ji} x_j}{\sum_{j=1}^{N} \gamma_{ji}}$
9. 累加新协方差矩阵：$(\mathbf{I} \cdot \sigma^2)' = (\mathbf{I} \cdot \sigma^2)' + \frac{1}{K} \frac{\sum_{j=1}^{N} \gamma_{ji} (x_j - \mu_i)(x_j - \mu_i)^\top}{\sum_{j=1}^{N} \gamma_{ji}}$
10. 计算新混合系数：$\pi_i' = \frac{1}{N} \sum_{j=1}^{N} \gamma_{ji}$
11. **end for**
12. 将模型参数$\{(\pi_i, \mu_i) \mid 1 \leq i \leq k,\ \mathbf{I} \cdot \sigma^2\}$更新为$\{(\pi_i', \mu_i') \mid 1 \leq i \leq k,\ (\mathbf{I} \cdot \sigma^2)'\}$
13. **until** 满足停止条件

算法停止后得到的模型参数即为极大似然估计的解。

(3)

若σ的值已知，并且$\sigma \to 0$，所以此时EM算法只需确定$\mu_i, \pi_i\ (1 \leq i \leq k)$，由于方差$\sigma \to 0$，所以每类样本的分布非常集中，无限趋近于该类样本的均值，所以在E步中可以直接指派离样本点最近的均值向量所在的成分给该样本而无需对每个混合成分计算概率（即直接选取均值向量离样本点最近的混合成分使其$\gamma_{ji} = 1$，其他成分的$\gamma_{ji} = 0$），故参数π_i也不需要估计了，在M步中只需更新$\mu_i' = \frac{\sum_{j=1}^{N} \gamma_{ji} x_j}{\sum_{j=1}^{N} \gamma_{ji}} = \frac{\sum_{x \in C_i} x}{|C_i|}$，故此时EM算法就相当于K-means聚类。

3 [40pts] Ensemble Methods

(1) **[10pts]** GradientBoosting[Friedman, 2001] 是一种常用的 Boosting 算法，请简要分析其与 AdaBoost 的异同。

(2) **[10pts]** 请简要说明随机森林为何比决策树 Bagging 集成的训练速度更快。

(3) **[20pts]** Bagging 产生的每棵树是同分布的，那么 B 棵树均值的期望和其中任一棵树的期望是相同的。因此，Bagging 产生的偏差和其中任一棵树的偏差相同，Bagging 带来的性能提升来自于方差的降低。

我们知道，方差为 σ^2 的 B 个独立同分布的随机变量，其均值的方差为 $\frac{1}{B}\sigma^2$。如果这些随机变量是同分布的，但不是独立的，设两两之间的相关系数 $\rho > 0$，请推导均值的方差为 $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$。

Solution. 此处用于写解答(中英文均可)

(1)

相同之处在于二者都属于Boosting算法；不同的个体学习器之间存在相关关系，即当前个体学习器的训练是基于之前轮的学习结果的；都是基于加性模型，给每个个体分类器分配了权重，最后的输出结果是所有个体学习器结果的的加权和。

不同之处在于GradientBoosting是通过梯度下降法来优化损失函数，每一轮都在先前轮训练的模型的损失函数的负梯度方向上训练一个新的个体学习器来减少损失；而AdaBoost是以类似牛顿迭代法来优化损失函数，通过在每一轮给分类错误的样本增加权重，从而在下一轮基于调整权重后的分布训练下一个个体学习器，提高对分类错误的样本的预测表现，从而减小损失。

(2)

因为在个体决策树的构建过程中，决策树Bagging集成使用的是“确定型”决策树，在选择划分属性时要对当前节点的所有属性进行考察，从而选出一个最优属性进行划分，故训练速度较慢；而随机森林使用的是“随机型”决策树，只需考察当前节点属性集合的一个子集，并从中选出一个最优属性进行划分，考察范围比决策树Bagging集成更小，故训练速度更快。

(3)

这些随机变量是同分布的，且该分布的方差为σ^2，故由相关系数的定义$\rho = \frac{COV(x_i,x_j)}{\sqrt{D(x_i)}\sqrt{D(x_j)}}$可知

$$\rho = \frac{COV(x_i, x_j)}{\sigma^2},\ 1 \le i \le B,\ 1 \le j \le B,\ i \ne j \tag{21}$$

由方差和协方差的运算性质，可得均值的方差为

$$\begin{aligned} D(\mu) &= D(\frac{\sum_{i=1}^{B} x_i}{B}) \\ &= \frac{1}{B^2} D(\sum_{i=1}^{B} x_i) \\ &= \frac{1}{B^2}\left(\sum_{i=1}^{B} D(x_i) + \sum_{i=1}^{B}\sum_{j=1,j\ne i}^{B} COV(x_i, x_j)\right) \end{aligned} \tag{22}$$

由(21)式可知$COV(x_i, x_j) = \rho\sigma^2$，又$D(x_i) = \sigma^2$，故(22)式等价于

$$\begin{aligned} D(\mu) &= \frac{1}{B^2}\left(\sum_{i=1}^{B}\sigma^2 + \sum_{i=1}^{B}\sum_{j=1,j\neq i}^{B}\rho\sigma^2\right) \\ &= \frac{1}{B^2}\left(B\sigma^2 + B(B-1)\rho\sigma^2\right) \\ &= \frac{1}{B}\left(\sigma^2 + (B-1)\rho\sigma^2\right) \\ &= \frac{1}{B}\left(B\rho\sigma^2 + (1-\rho)\sigma^2\right) \\ &= \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \end{aligned} \tag{23}$$

参考文献

[Friedman, 2001] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.