机器学习导论
习题五

191300023, 李安琦, 486488604@qq.com

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**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 $(\boldsymbol{u}_i \top \boldsymbol{u}_i = 1)$，$a_i \in \mathbb{R}$，$\{y_i\}_{i=1}^n$ 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 Σ_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 Σ_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.

(1) 证明：根据已知条件，$\{y_i\}_{i=1}^n$ 是互不相关的零均值随机变量。

那么 $\mathbb{E}(y_i) = \boldsymbol{u}_i^\top \mathbb{E}(\boldsymbol{x}) + a_i = 0, i = 1, \ldots, d \leq D$。

由题目知，$\mathbb{E}(\boldsymbol{x}) = \boldsymbol{\mu_x}$, 那么我们便可以得到：$a_i = -\boldsymbol{u}_i^\top \boldsymbol{\mu_x}, i = 1, \ldots, d$。

(2) 证明：对于任意的一个 y_i, 如果我们要最大化它的方差 $\mathrm{Var}(y_i) = \boldsymbol{u}_i^\top \Sigma_x \boldsymbol{u}_i$，那么可以得到以下结论：

$$\begin{aligned} \min \quad & -\boldsymbol{u}_i^\top \Sigma_x \boldsymbol{u}_i \\ s.t. \quad & \boldsymbol{u}_i^\top \boldsymbol{u}_i = 1 \quad i = 1, 2, 3 \cdot \cdot \cdot, d \end{aligned}$$

根据拉格朗日乘子法，可得：$L(\boldsymbol{u}_i, \lambda) = -\boldsymbol{u}_i^\top \Sigma_x \boldsymbol{u}_i + \lambda(\boldsymbol{u}_i^\top \boldsymbol{u}_i - 1)$。在对 $\boldsymbol{u}_i$ 求导后，可得：$\Sigma_x \boldsymbol{u}_i = \lambda \boldsymbol{u}_i$。这里可以看出 λ 是 Σ_x 的特征值，而 $\boldsymbol{u}_i$ 是特征向量。

而 $\mathrm{Var}(y_i) = \boldsymbol{u}_i^\top \Sigma_x \boldsymbol{u}_i = \lambda \boldsymbol{u}_i^\top \boldsymbol{u}_i = \lambda$。

也就是说 Σ_x 最大的特征值对应于 $\mathrm{Var}(y_i)$ 最大值。

根据题目得到：$\mathrm{Var}(y_1) \geq \mathrm{Var}(y_2) \geq \cdots \geq \mathrm{Var}(y_d)$，那么 $\boldsymbol{u}_1$ 是 Σ_x 最大的特征值对应的特征向量。

(3) 证明：令 $\boldsymbol{z} = \boldsymbol{x} - \mathbb{E}(\boldsymbol{x})$。对于 $\forall 1 \leq i, j \leq d$, 我们都有 $\mathsf{Cov}(y_i, y_j) = \mathbb{E}(y_i y_j) - \mathbb{E}(y_i) * \mathbb{E}(y_j) = \mathbb{E}(y_i y_j) = (\boldsymbol{u}_i^\top \boldsymbol{z}) * (\boldsymbol{u}_j^\top \boldsymbol{z}) = 0$。

因此可得，对于 $\forall 1 \leq i \leq d$, 都有 $(\boldsymbol{u}_i^\top \boldsymbol{z}) = 0$，又因为 $\boldsymbol{z}$ 均值为 0，$\boldsymbol{u}_i$ 都是单位向量，因此可知 $\boldsymbol{u}_1 \cdot \cdot \cdot \boldsymbol{u}_d$ 为一组正交单位基，于是可得 $\boldsymbol{u}_2^\top \boldsymbol{u}_1 = 0$。

仿照第二问的思路，对条件进行约束，寻找 $\mathrm{Var}(y_i)$ 的最大值，其中 $i = 2, \cdot \cdot \cdot\ d$。由此可以得到，$\boldsymbol{u}_2$ 是 Σ_x 第二大特征值对应的特征向量。

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 $g_k = N(\mu_k, \mathbf{I}\cdot\sigma^2)$，$\mathbf{I}$ 是单位矩阵，$\pi_k > 0$，$\sum_k \pi_k = 1$。$\{\mu_k, \pi_k\}, k = 1, \ldots, K$ 和 σ^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.

(1) 由已知，我们令 $D = \{x_1, x_2, \ldots, x_N\}$。那么 $LL(D) = \ln(\prod_{j=1}^{N} g(x_j)) = \sum_{j=1}^{N} \ln(\sum_{i=1}^{K} \pi_i g_i(x_j))$。并且可知 $g_k = N(\mu_k, \mathbf{I}\cdot\sigma^2)$，则 $g_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu_k)^T\mathbf{I}(x-\mu_k)}$，代入即可。

(2) 令随机变量 $z_j \in \{1, 2, \cdots, K\}$ 表示生成样本 x_j 的高斯混合成分。

则 z_j 的先验概率 $P(z_j = i) = \pi_i$。

根据贝叶斯定理，z_j 的后验分布对应于 $g(z_j = i|x_j) = \frac{P(z_j=i)\cdot g(x_j|z_j=i)}{g(x_j)} = \frac{\pi_i \cdot g_i(x_j)}{\sum_{l=1}^{K} \pi_l \cdot g_l(x_j)}$。令 $g(z_j = i|x_j) = \gamma_{ji}$。

若要使 $LL(D)$ 最大化，那么由 $\frac{\partial LL(D)}{\partial \mu_i} = 0$，有 $\sum_{j=1}^{N} \frac{\pi_i g_i(x_j)}{\sum_{l=1}^{K} \pi_l \cdot g_l(x_j)}(x_j - \mu_i) = 0$。因此 $\mu_i = \frac{\sum_{j=1}^{N} \gamma_{ji} x_j}{\sum_{j=1}^{N} \gamma_{ji}}$。

同理，由 $\frac{\partial LL(D)}{\partial \sigma} = 0$ 可得 $\sigma = \frac{1}{K}\sum_{i=1}^{K} \frac{\sum_{j=1}^{N} \gamma_{ji}(x_j-\mu_i)(x_j-\mu_i)^T \cdot \mathbf{I}}{2\sum_{j=1}^{N} \gamma_{ji}}$。

对于混合系数 π_i，除了最大化 $LL(D)$，还需满足 $\pi_i \geq 0$，$\sum_{i=1}^{K} \pi_i = 1$。考虑 $LL(D)$ 的拉格朗日形式：$LL(D) + \lambda(\sum_{i=1}^{K} \pi_i - 1)$.

对 π_i 的导数为 0，有 $\sum_{j=1}^{N} \frac{\pi_i g_i(x_j)}{\sum_{l=1}^{K} \pi_l \cdot g_l(x_j)} + \lambda = 0$。

两边同时乘以 π_i，对所有混合成分求和可知 $\lambda = -N$，有 $pi_i = \frac{1}{N}\sum_{j=1} N\gamma_{ji}$。

因此，EM 算法的伪代码如下所示：

初始化高斯混合分布的模型参数 $\{(\pi_i, \mu_i, \sigma)|1 \leq i \leq K\}$

repeat

for j = 1, 2 $\cdots$,N do

$\gamma_{ji} = g_i(x_j)(1 \leq i \leq K)$

end for

for i = 1, 2 $\cdots$,K do

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

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

$\pi_i^1 = \frac{1}{N}\sum_{j=1} N\gamma_{ji}$
end for
将 $\{(\pi_i, \mu_i, \sigma)|1 \le i \le K\}$ 进行更新
until 满足停止条件
$C_i = \emptyset(1 \le i \le K)$
for j = 1, 2 $\cdots$,N do
确定 x_j 的簇标记 λ_j
将 x_j 划入相应的簇：$C_{\lambda_j} = C_{\lambda_j} \cup \{x_j\}$
end for

(3) 答：由于 σ 为定值，且其趋近于 0，这就导致对于样本的分类比较确定，不至于说一个样本有可能属于 A 类，也有可能属于 B 类。又因为 K-means 和该 EM 算法分类受初始值影响，都可能限于局部最优解并且类别的个数都要靠猜测，因此该 EM 算法就相当于 K-means 聚类

3 [40pts] Ensemble Methods

(1) **[10pts]** GradientBoosting[**?**] 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.

(1) 相同点：两者都是重复选择一个表现一般的模型并基于先前的表现而调整。

不同点：AdaBoost 将前一个模型预测成功的点划分低权重，失败的点划分高权重，并根据此重新建模，每一次生成的子模型都在想办法弥补上一次生成的子模型没有成功预测到的样本点，或者说是弥补上一子模型所犯的错误。Gradient Boosting 通过负梯度来识别问题，通过计算负梯度来改进模型，每一个模型都是对前一个模型所犯错误的补偿并且不能对基本算法进行选择。

(2) 与 Bagging 中基学习器的 “多样性” 仅通过样本扰动而来不同，随机森林中基学习器的多样性不仅来自于样本扰动，还来自属性扰动，这就使得最终集成的泛化性能可通过个体学习器之间差异度的增加而进一步提升。

(3) 证明：不妨设随机变量为 $X_1, X_2, \cdots, X_B$，那么由题目可知 $\mathsf{Cov}(X_i, X_j), 1 \le i, j \le B$ 相等。那么均值可以表示为 $\frac{\sum_{i=1}^{B} X_i}{B}$，其方差为 $\frac{1}{B^2}\mathrm{Var}(X_1 + X_2 + \cdots + X_B) = \frac{1}{B^2}(\sum_{i=1}^{B}\mathrm{Var}(X_i) +$

$2\sum_{1\le i<j\le B}\mathrm{Cov}(X_i,X_j)) = \frac{1}{B^2}(B\mathrm{Var}(X_1)+B(B-1)\mathrm{Cov}(X_1,X_2)) = \frac{1}{B^2}(B\sigma^2+B(B-1)\rho\sigma^2) = \rho\sigma^2+\frac{1-\rho}{B}\sigma^2$ 。综上得证。