机器学习导论
习题五

学号, 作者姓名, 邮箱

2021 年 6 月 23 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 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 $\{y_i\}_{i=1}^d$ 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. 由 y_i 是零均值的随机变量可得

$$\mathbb{E}[y_i] = \boldsymbol{\mu}_i^\top \mathbb{E}[\boldsymbol{x}_i] + \mathbb{E}[a_i] = \boldsymbol{u}_i^\top \boldsymbol{\mu_x} + a_i = 0, \quad i = 1, \ldots, d$$

   则 $a_i = -\boldsymbol{u}_i^\top \boldsymbol{\mu_x}, i = 1, \ldots, d$

2.

$$\mathrm{Var}\,(y_1) = \mathbb{E}\left[\left(\boldsymbol{u}_1^\top (\boldsymbol{x} - \boldsymbol{u}_x)\right)^2\right] = \mathbb{E}\left[\boldsymbol{u}_1^\top (\boldsymbol{x} - \boldsymbol{u}_x)(\boldsymbol{x} - \boldsymbol{u}_x)^\top \boldsymbol{u}_1\right] = \boldsymbol{u}_1^\top \Sigma_x \boldsymbol{u}_1$$

   则优化问题为

$$\max_{u_1 \in \mathbb{R}^D} \boldsymbol{u}_1^\top \Sigma_x \boldsymbol{u}_1 \quad \text{s.t.} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_1 = 1$$

   由拉格朗日函数

$$\mathcal{L} = \boldsymbol{u}_1^\top \Sigma_x \boldsymbol{u}_1 + \lambda_1 \left(1 - \boldsymbol{u}_1^\top \boldsymbol{u}_1\right)$$

   其中 λ_1 为拉格朗日乘子。

   令上式对 $\boldsymbol{u}_1$ 和 λ_1 求导分别等于0，可得

$$\Sigma_x \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1 \quad \text{and} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_1 = 1$$

   因此 $\boldsymbol{u}_1$ 是 λ_1 所对应的 Σ_x 的特征向量。

   要最大化 $\boldsymbol{u}_1^\top \Sigma_x \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1^\top \boldsymbol{u}_1 = \lambda_1$，则 λ_1 应为 Σ_x 最大的特征值，$\boldsymbol{u}_1$ 是 Σ_x 最大的特征向量。

3. 由 y_1, y_2 不相关，得 $\mathrm{Cov}(y_1, y_2) = \mathbb{E}(y_1 y_2) - \mathbb{E}(y_1)\mathbb{E}(y_2) = 0$。

   又 y_1, y_2 均值为0，所以 $\mathbb{E}(y_1 y_2) = \mathbb{E}(y_1)\mathbb{E}(y_2) = 0$。即

$$\mathbb{E}\left[\left(\boldsymbol{u}_1^\top (\boldsymbol{x} - \boldsymbol{u}_x)\right)\left(\boldsymbol{u}_2^\top (\boldsymbol{x} - \boldsymbol{u}_x)\right)\right] = \mathbb{E}\left[\boldsymbol{u}_1^\top (\boldsymbol{x} - \boldsymbol{u}_x)(\boldsymbol{x} - \boldsymbol{u}_x)^\top \boldsymbol{u}_2\right] = \boldsymbol{u}_1^\top \Sigma_x \boldsymbol{u}_2 = \lambda_1 \boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$$

由于 $\lambda_1 \neq 0$，因此 $\boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$。

类似于 $\boldsymbol{u}_1$ 的求解过程，求解 $\boldsymbol{u}_2$ 对应的最大化目标是 $\boldsymbol{u}_2^\top \Sigma_x \boldsymbol{u}_2$，多一个约束条件 $\boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$。即优化问题为

$$\max_{u_2 \in \mathbb{R}^D} \boldsymbol{u}_2^\top \Sigma_x \boldsymbol{u}_2 \quad \text{s.t.} \quad \boldsymbol{u}_2^\top \boldsymbol{u}_2 = 1 \quad \text{and} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$$

定义拉格朗日函数

$$\mathcal{L} = \boldsymbol{u}_2^\top \Sigma_x \boldsymbol{u}_2 + \lambda_2 \left(1 - \boldsymbol{u}_2^\top \boldsymbol{u}_2\right) + \gamma \boldsymbol{u}_1^\top \boldsymbol{u}_2$$

对 $\boldsymbol{u}_2, \lambda_2, \gamma$ 分别求导令导数等于0，得

$$\Sigma_x \boldsymbol{u}_2 + \frac{\gamma}{2} \boldsymbol{u}_1 = \lambda_2 \boldsymbol{u}_2, \quad \boldsymbol{u}_2^\top \boldsymbol{u}_2 = 1 \quad \text{and} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$$

第一个式子左乘以 $\boldsymbol{u}_1^\top$ 可得，$\boldsymbol{u}_1^\top \Sigma_x \boldsymbol{u}_2 + \frac{\gamma}{2} \boldsymbol{u}_1^\top \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1^\top \boldsymbol{u}_2 + \frac{\gamma}{2} = \lambda_2 \boldsymbol{u}_1^\top \boldsymbol{u}_2$，因此 $\gamma = 2\left(\lambda_2 - \lambda_1\right) \boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$，故 $\Sigma_x \boldsymbol{u}_2 = \lambda_2 \boldsymbol{u}_2$。则最大化的目标为 $\boldsymbol{u}_2^\top \Sigma_x \boldsymbol{u}_2 = \lambda_2 = \mathrm{Var}\left(y_2\right)$。

因为 Σ_x 没有重复的特征值，所以当 λ_2 为 Σ_x 第二大的特征值，$\boldsymbol{u}_2$ 为 Σ_x 第二大的特征值对应的特征向量时，$\boldsymbol{u}_2^\top \Sigma_x \boldsymbol{u}_2$ 取得最大值。

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 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. 此处用于写解答(中英文均可)

1.

$$\begin{aligned} LL(D) &= \log\left(\prod_{i=1}^{N} g\left(x_i\right)\right) \\ &= \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k g_k\left(x_i\right)\right) \end{aligned}$$

2. 算法仿照书210页图9.6，其中关键的公式：

   E步，样本 x_i 由第 k 个成分生成的概率

$$\gamma_{ik} = \frac{\hat{\pi}_k \hat{g}_k\left(x_i\right)}{\sum_{l=1}^{K} \hat{\pi}_l \hat{g}_l\left(x_i\right)}$$

   M步，计算新均值向量

$$\hat{\mu}_k = \frac{\sum_{i=1}^{N} \gamma_{ik} x_i}{\sum_{i=1}^{N} \gamma_{ik}}.$$

   一般计算新协方差矩阵，

$$\Sigma_k = \frac{\sum_{i=1}^{N} \gamma_{ik}\left(x_i - \hat{\mu}_k\right)\left(x_i - \hat{\mu}_k\right)^T}{\sum_{i=1}^{N} \gamma_{ik}}$$

   不过这里方差是标量，因此应为

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik}\left(x_i - \hat{\mu}_k\right)^T\left(x_i - \hat{\mu}_k\right)}{pN}$$

   计算新混合系数

$$\hat{\pi}_k = \frac{\sum_{i=1}^{N} \gamma_{ik}}{N}$$

3.

$$\gamma_{ik} = \frac{\exp\left(-\frac{1}{2\sigma^2}\left\|x_i - \mu_k\right\|^2\right)}{\sum_{l=1}^{K} \exp\left(-\frac{1}{2\sigma^2}\left\|x_i - \mu_l\right\|^2\right)}$$

   当 $\sigma \to 0$, 对于离 x_i 最近的成分 k，$\gamma_{ik} \to 1$，对于其他成分 $\gamma_{ik} \to 0$。

   则E步等价于k-means中划入均值向量距离最近的簇。M步等价于重新计算新的簇的均值。

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) 共同点：都是用弱学习器逐步拟合没学好的部分，组合在一起成为一个较好的集成模型。

不同点：Adaboost调整没学好的样本的权重，GradientBoosting调整target拟合残差。

言之有理亦可。

(2) 随机森林相对于Bagging决策树的关键区别在于，在选择划分属性时，首先随机选择一个属性集的子集，再在这个子集中寻找最优属性。

由于一般随机选择的属性子集规模比所有属性集小(如推荐值$k = \log_2 d$)，训练时只需考察这个较小的子集，从而训练速度更快。

(3)

$$\begin{aligned}\operatorname{Var}\left(\frac{\sum_{i=1}^{B} X_i}{B}\right) &= \frac{1}{B^2}\operatorname{Var}\left(\sum x_i\right) \\ &= \frac{1}{B^2}\sum_{i=1}^{B}\operatorname{Var}\left(X_i\right) + \frac{1}{B^2}\sum_{i\neq j}^{B}\operatorname{Cov}\left(X_i, X_j\right) \\ &= \frac{\sigma^2}{B} + \frac{B-1}{B}\sigma^2\rho \\ &= \sigma^2\rho + \frac{1-\rho}{B}\sigma^2.\end{aligned}$$

(备注：随着B的增加，第二项逐渐消失，但第一项消不掉，所以树与树之间的相关性限制了平均带来的好处。)

参考文献

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