机器学习导论
习题五

191300020, 黄彦骁, AdrianHuang@smail.nju.edu.cn

2021 年 6 月 6 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. 由于 y_i 为零均值随机变量，所以有 $\mathbb{E}(y_i) = \mathbb{E}(\boldsymbol{u}_i x + a_i) = \boldsymbol{u}_i \mathbb{E}(x) + a_i = 0$, 便可推出 $a_i = -\boldsymbol{u}_i \boldsymbol{\mu_x}$。

2. 首先推导任意 $\boldsymbol{u}_i, \boldsymbol{u}_j, i \neq j$ 之间两两正交，由于 $\{y_i\}_{i=1}^d$ 为互不相关的零均值随机变量，故有：

$$Cov(y_i, y_j) = Cov(\boldsymbol{u}_i(\boldsymbol{x} - \boldsymbol{\mu_x}), \boldsymbol{u}_j(\boldsymbol{x} - \boldsymbol{\mu_x})) = \boldsymbol{u}_i^\top \boldsymbol{u}_j Cov(\boldsymbol{x} - \boldsymbol{\mu_x}, \boldsymbol{x} - \boldsymbol{\mu_x}) = 0, i \neq j$$

   所以可以得到 $\boldsymbol{u}_i^\top \boldsymbol{u}_j = 0, i \neq j$，即 $\boldsymbol{u}_i, \boldsymbol{u}_j, i \neq j$ 之间两两正交。
   而我们在 PCA 的过程需要最大化投影后的样本点方差，即 $\mathbf{tr}(\boldsymbol{y}\boldsymbol{y}^\top)$，其中 $\boldsymbol{y}$ 为降维之后坐标组成的向量，设 $\boldsymbol{U} = \{\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_d^\top\}$. 优化问题为：

$$\begin{aligned} &\mathbf{min}_{\boldsymbol{U}} \ -\mathbf{tr}(\boldsymbol{U}^\top \hat{\boldsymbol{x}} \hat{\boldsymbol{x}}^\top \boldsymbol{U}) \\ &\mathbf{s.t.}\ \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I} \end{aligned}$$

   使用拉格朗日乘子法对问题进行求解有：

$$\hat{\boldsymbol{x}}^\top \hat{\boldsymbol{x}} \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \Rightarrow \Sigma_x \boldsymbol{u}_i = \lambda \boldsymbol{u}_i$$

   即我们需要找到的最优解为 Σ_x 的特征值对应的特征向量。
   而后我们有：

$$Var(y_i) = \boldsymbol{u}_i^\top Var(\hat{\boldsymbol{x}}) \boldsymbol{u}_i = \boldsymbol{u}_i^\top Var(\boldsymbol{x}) \boldsymbol{u}_i = \boldsymbol{u}_i^\top \Sigma_x \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i^\top \boldsymbol{u}_i = \lambda_i$$

   故最大特征值就对应最大的方差，也即 $\boldsymbol{u}_1$ 对应最大特征值的特征向量。

3. 由第二问可知 $\boldsymbol{u}_2$ 对应第二大特征值的特征向量。

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. 对数似然：

$$LL(D) = ln(\prod_{j=1}^{N} g(x_j)) = \sum_{j=1}^{N} ln(\sum_{i=1}^{K} \pi_i g_i(x_j))$$

2. 我们令:

$$\gamma_{ij} = \frac{\pi_i g_i(x_j)}{\sum_{l=1}^{K} \pi_l g_l(x_j)}$$

Require: : 样本集 $\{x_1, x_2, \ldots, x_N\}$, 高斯模型参数 K。

过程:

初始化模型参数 $\pi_i, \mu_i, 1 \le i \le K, \sigma^2$

repeat

  for $j = 1, 2, \ldots, N$ **do**

    计算出每一个样本对应出现的后验分布 γ_{ij}

  end for

  for $i = 1, 2, \ldots, K$ **do**

    计算新的 $\mu_i\prime = \frac{\sum_{j=1}^{N} \gamma_{ij} x_j}{\sum_{j=1}^{N} \gamma_{ij}}$

    计算新的 $(\sigma')^2 = \frac{1}{K} \sum_{i=1}^{K} \frac{\sum_{j=1}^{N} \gamma_{ij} (x_j - \mu_i')^\top \mathbf{I} (x_j - \mu_i')}{\sum_{j=1}^{N} \gamma_{ij}}$

    计算新的 $\pi_i' = \frac{\sum_{j=1}^{N} \gamma_{ij}}{N}$

  end for

  更新模型参数

until 满足停止条件

$C_i = \emptyset (1 \le i \le K)$

for j=1,2...,N **do**

  根据 $\lambda_j = \arg max_{i \in \{1,2,\ldots,K\}} \gamma_{ij}$ 确定簇标记 λ_j，同时将相应的 x_j 划入相应的簇。

end for

3. 当 σ^2 为一个常数且接近于 0 时，有：

$$\begin{aligned}\gamma_{ij} &= \frac{\pi_i g_i(x_j)}{\sum_{l=1}^{K} \pi_l g_l(x_j)} \\ &= \frac{\pi_i exp(-\frac{1}{2\sigma^2}(x_j-\mu_i)^\top(x_j-\mu_i))}{\sum_{l=1}^{K} \pi_l exp(-\frac{1}{2\sigma^2}(x_j-\mu_l)^\top(x_j-\mu_l))}\end{aligned}$$

如果样本 x_j 属于地 k 类的概率最大，那么该样本离第 k 类的中心点距离非常近，$(x_j-\mu_k)^\top(x_j-\mu_k)$ 会无限趋近于 0，则有：

$$exp(-\frac{1}{2\sigma^2}(x_j-\mu_k)^\top(x_j-\mu_k)) \to 1$$

$$exp(-\frac{1}{2\sigma^2}(x_j-\mu_i)^\top(x_j-\mu_i)) \to 0$$

$$\gamma_{kj} \to 1, \gamma_{ij} \to 0, i = 1, 2, \ldots, k, i \neq k$$

则该 EM 算法变成了一个硬聚类，也就时 K-均值聚类。

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. 1. GradientBoosting 和常见的 Boosting 算法类似，通过将多个性能一般的模型组合起来来达到一个较好的性能。模型的训练通过反复选择一个负梯度的方向来对目标函数进行优化。和 AdaBoost 相同的在于 Gradient Boosting 也是重复选择一个性能一般的模型并且每次基于先前模型的表现进行调整。不同的是，AdaBoost 是通过提升错分数据点的权重来修补模型的不足，但 Gradient Boosting 是通过计算负梯度来修补模型的不足。因此相比 AdaBoost, Gradient Boosting 对目标函数的种类有更多的包容性。

2. 随机森林在决策树 Bagging 训练过程中引入了随机属性选择，大大减少了最优属性选择过程的计算量。而正常情况下决策树耗时最长的部分即为最优属性选择，因而随机森林比普通决策树 Bagging 训练速度要快。

3. 设 B 个变量分别为 $X_1, X_2 \ldots X_B$，有相关系数得 $Cov(X_i, X_j) = \rho\sigma^2$。推导其均值的方差有：

$$\begin{aligned}
Var(\frac{X_1 + \ldots + X_B}{B}) &= \frac{1}{B^2}(Var(X_1) + Var(X_2 + \ldots + X_B) + 2Cov(X_1, (X_2 + \ldots + X_B))) \\
&= \frac{1}{B^2}(Var(X_1) + Var(X_2 + \ldots + X_B) + 2(Cov(X_1, X_2) + \ldots + Cov(X_1, X_B))) \\
&= \frac{1}{B^2}(Var(X_1) + 2\sum_{i=2}^{B} Cov(X_1, X_i) + Var(X_2 + \ldots + X_B)) \\
&= \frac{1}{B^2}(\sum_{i=1}^{B} Var(X_B) + 2\sum_{j \neq k} Cov(X_j, X_k)) \\
&= \frac{1}{B^2}(B\sigma^2 + \rho B(B-1)\sigma^2) \\
&= \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2
\end{aligned}$$

故求得方差为 $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$.

参考文献

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