2024 春季高级机器学习
习题一

211820073 胡涂

2024.4.17

一. (50 points) **PCA**

除了最大方差和最小重构误差的解释外，还可以从矩阵的低秩近似 (low-rank approximation) 角度理解 PCA。假设 $X \in \mathbb{R}^{m\times d}$ 是已经中心化的样本矩阵. 低秩近似就是寻找一个秩为 d' 的矩阵 X'，$1 \le d' < d$ 满足:

$$\begin{aligned} \min \quad & \|X - X'\|_F^2 \\ \text{s.t.} \quad & \operatorname{rank}(X') = d' \ , \end{aligned} \tag{1}$$

其中，$\|\cdot\|_F$ 为 Frobenius 范数（F 范数）欲使 X' 的秩为 d'，一个直接的做法是寻找 $\mathbb{R}^d$ 的一个 d' 维子空间 $\mathcal{W}$，将 X 中的样本 $\boldsymbol{x}$ 投影到该子空间中。令 $W = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'})$ 是 $\mathcal{W}$ 中的一组单位正交基。

1. (12 points) 请证明 $\boldsymbol{x}$ 在子空间 $\mathcal{W}$ 上的正交投影为 $\boldsymbol{x}' = (WW^\top)\boldsymbol{x}$，并写出 X 在子空间 $\mathcal{W}$ 上的投影。
2. (12 points) 根据第一问的结果，可以把优化目标写成:

$$\begin{aligned} \min_{W} \quad & \|X - XWW^\top\|_F^2 \\ \text{s.t.} \quad & W^\top W = I \ . \end{aligned} \tag{2}$$

   关联 PCA 的优化目标，并证明其与低秩近似等价（提示：参考教材中的公式 10.15，考虑矩阵的 F 范数和矩阵的迹之间的关系）。说明 PCA 得到的低维表示和 $\boldsymbol{x}'$ 之间的关系。
3. (12 points) 请证明对矩阵 X 进行正交变换后，新矩阵的 F 范数不变。
4. (14 points) 分别对 X，X' 进行奇异值分解: $X = U_{\boldsymbol{x}}\Sigma_{\boldsymbol{x}}V_{\boldsymbol{x}}^\top$，$X' = U_{\boldsymbol{x}'}\Sigma_{\boldsymbol{x}'}V_{\boldsymbol{x}'}^\top$，令 $U = U_{\boldsymbol{x}}^\top U_{\boldsymbol{x}'}$，$V = V_{\boldsymbol{x}}^\top V_{\boldsymbol{x}'}$ 请证明 $\min \|X - X'\|_F^2$ 等价于 $\min \|\Sigma_{\boldsymbol{x}'}\|_F^2 - 2\operatorname{tr}(\Sigma_{\boldsymbol{x}}^\top U \Sigma_{\boldsymbol{x}'} V^\top)$。(提示：考虑第三问得到的结论，即正交变换不影响矩阵 F 范数)

解：

1. 设投影后的 $\boldsymbol{x}' = \sum_{i=1}^{d'} z_i w_i = W(z_1, \ldots, z_{d'}) = Wz$, $\boldsymbol{x}$ 可以被 $\boldsymbol{x}'$ 与一个正交于子空间 W 的向量 α 表示, 即 $\boldsymbol{x} = \boldsymbol{x}' + \alpha$，其中 $W^\top \alpha = 0$。
   那么 $W^\top x = W^\top \boldsymbol{x}' + W^\top \alpha = z$, $(WW^\top)x = Wz = \boldsymbol{x}'$。
   由于 X 中的 x 为行向量，因此 X 在 W 中的投影 $X' = XWW^\top$。

2. **与 PCA 的关联**:
   已知矩阵 F-范数与 trace 的关系

$$\|A\|_F^2 = tr(AA^\top)$$

于是

$$\|X - XWW^\top\|_F^2 = tr(-2XWW^\top X^\top) = -2tr(XWW^\top X^\top) = -2tr(W^\top X^\top XW)$$

因此优化问题可以改写成

$$\begin{aligned} \min_W \quad & -tr(W^\top X^\top XW) \\ \text{s.t.} \quad & W^\top W = I\ . \end{aligned} \tag{3}$$

与书中 (10.15)(10.16) 等价 (此处 X 中数据为行向量)

与低秩近似的关系:
将投影 $X' = XWW^\top$ 带入至低秩近似原形式

$$\begin{aligned} \min_W \quad & \|X - XWW^\top\|_F^2 \\ \text{s.t.} \quad & W^\top W = I \\ & rank(XWW^\top) = d' \end{aligned} \tag{4}$$

只需要证明 $rank(XWW^\top) = d'$ 恒成立

$$rank(WW^\top) = rank(W) = d'$$

假设 $rank(X) = d$

$$rank(X) + rank(WW^\top) - d \le rank(XWW^\top) \le rank(WW^\top)$$

$$d' \le rank(XWW^\top) \le d'$$

于是 $rank(XWW^\top) = d'$, (1)(3) 优化形式等价 (假设 $rank(X) = d$)。

PCA 得到的低维表示和 $\boldsymbol{x}'$ 之间的关系：PCA 得到的低维表示就是低秩近似的 x'

3. 假设 $Q \in \mathbb{R}^{m\times m}$,$Q^TQ = QQ^T = I$, 那么对 X 的列空间进行正交变换 QX,F 范数为

$$\|QX\|_F = \sqrt{tr(QXX^TQ^T)} = \sqrt{tr(XQQ^TX^T)} = \sqrt{tr(XX^T)} = \|X\|_F$$

对行空间的正交变换与上述证明类似。

4.
$$\begin{aligned}
&\|X - X'\|_F^2 = \|U_{\boldsymbol{x}}\Sigma_{\boldsymbol{x}}V_{\boldsymbol{x}}^\top - U_{\boldsymbol{x}'}\Sigma_{\boldsymbol{x}'}V_{\boldsymbol{x}'}^\top\|_F^2 = \\
&tr(U_{\boldsymbol{x}}\Sigma_{\boldsymbol{x}}\Sigma_{\boldsymbol{x}}^\top U_{\boldsymbol{x}}^\top) + tr(U_{\boldsymbol{x}'}\Sigma_{\boldsymbol{x}'}\Sigma_{\boldsymbol{x}'}^\top U_{\boldsymbol{x}'}^\top) - 2tr(U_{\boldsymbol{x}}\Sigma_{\boldsymbol{x}}V_{\boldsymbol{x}}^\top V_{\boldsymbol{x}'}\Sigma_{\boldsymbol{x}'}^\top U_{\boldsymbol{x}'}^\top) = \\
&\|U_{\boldsymbol{x}}\Sigma_{\boldsymbol{x}}\|_F^2 + \|U_{\boldsymbol{x}'}\Sigma_{\boldsymbol{x}'}\|_F^2 - 2tr(\Sigma_{\boldsymbol{x}}^\top U\Sigma_{\boldsymbol{x}'}V^\top) = \\
&\|\Sigma_{\boldsymbol{x}}\|_F^2 + \|\Sigma_{\boldsymbol{x}'}\|_F^2 - 2tr(\Sigma_{\boldsymbol{x}}^\top U\Sigma_{\boldsymbol{x}'}V^\top)
\end{aligned} \tag{5}$$

由于 $\|\Sigma_x\|_F^2$ 是一个常数，因此 $\min\|X - X'\|_F^2$ 等价于 $\min\|\Sigma_{\boldsymbol{x}'}\|_F^2 - 2\,\mathrm{tr}(\Sigma_{\boldsymbol{x}}^\top U\Sigma_{\boldsymbol{x}'}V^\top)$

二. (10 points) **降维与度量学习的应用**

在机器学习中，往往可以把模型分成表示学习（学习高维数据的低维表示）和分类器学习（在低维数据上学习一个分类器）两个部分。数据降维可以看成是一种表示学习的方法，因此一种评估降维方法效果的方式为，先对数据进行降维，然后训练一个分类器，最后比较分类效果的优劣。常用的指标有如下几种：

1. k 近邻分类器的精度。
2. 线性分类精度。
3. 最近类中心分类精度。

请描述并比较这三种评价指标，并简要说明它们各自的优劣。

解：

- k-NN Accuracy: 在特征空间中找到测试样本最近的 k 个训练样本，然后根据这 k 个样本的标签通过多数投票的方式来预测测试样本的标签。
  - 优势
    * 实现简单
    * 没有对样本分布与模式做过多假设
    * 可以容易地扩展到多分类问题
  - 劣势
    * 预测时间复杂度 $O(nd)$，随着样本数增加与维数的增加，模型推断的计算开销大
    * k 设置较小的时候，容易被异常/噪声点影响

- 线性分类 Accuracy: 假设样本线性可分，通过训练线性模型对降维后样本进行分类，得出准确率
  - 优势
    - 计算效率高：一旦模型被训练，分类新数据的速度很快。
    - 易于解释：模型的权重可以提供对特征重要性的直观理解。
  - 劣势
    - 强假设：对样本做了线性可分的假设，表达能力有限
- 最近类中心分类 Accuracy：使用每一类的样本，通过某种指标计算类中心向量（一般是均值），通过计算样本与类中心向量的距离分类。
  - 优势
    - 简单高效：一旦类中心被计算，新样本的分类非常快速。
    - 鲁棒性较高：通过平均化（如 kmedios 的中心选取方法），减少了噪声和异常点的影响。
  - 劣势
    - 如果类内方差很大或类形态不规则，性能会受影响。

三. (40 points) **稀疏学习**

习题一中涉及到了 PCA 的矩阵低秩近似角度理解。Robust PCA 在此基础上增加了一个变量和正则项：

$$\begin{aligned} \min_{X',E} \quad & \operatorname{rank}(X') + \lambda \|E\|_0 \\ s.t. \quad & X = X' + E \end{aligned} \tag{6}$$

其中 $\|\cdot\|_0$ 为零范数。λ 为正则化参数。为了解该优化问题,我们考虑它的凸松弛(Convex Relaxation):

$$\begin{aligned} \min_{X',E} \quad & \|X'\|_* + \lambda \|E\|_1 \\ s.t. \quad & X = X' + E \end{aligned} \tag{7}$$

其中 $\|\cdot\|_*$ 为核范数（Nuclear Norm）。使用增广拉格朗日方法（Augmented Lagrangian Method）处理约束条件，可以得到：

$$\min_{X',E} \quad \|X'\|_* + \lambda \|E\|_1 + \langle Y, X - X' - E \rangle + \frac{\mu}{2} \|X - X' - E\|_F^2 \tag{8}$$

其中 Y 为拉格朗日乘子。此处省略后续的推导过程和收敛性分析，求解该优化问题的交替求解算法中的 python 代码片段如下：

```
1  ...
2  Xk = np.zeros(self.X.shape)
3  Ek = np.zeros(self.X.shape)
4  Yk = np.zeros(self.X.shape)
5  while (err > _tol) and iter_ < max_iter:
6      Xk = self.nuclear_prox(self.X - Ek + self.mu_inv * Yk, self.mu_inv)
7      Ek = self.L1_prox(self.X - Xk + self.mu_inv * Yk, self.mu_inv * self.lmbda)
8      Yk = Yk + self.mu * (self.X - Xk - Ek)
9      err = self.frobenius_norm(self.X - Xk - Ek)
10     iter_ += 1
11     if (iter_ % iter_print) == 0 or iter_ == 1 or iter_ > max_iter or err <= _tol:
12         print('iteration: {0}, error: {1}'.format(iter_, err))
13 ...
```

1. (10 points) Robust PCA 增加的变量 E 和正则项对模型有什么作用?
2. (15 points) 代码片段中第 6 行和第 7 行调用的方法实现了什么优化方法解了哪两个优化问题?(写出优化方法并分别写出优化问题)
3. (15 points) 你认为 Robust PCA 具有哪些实际应用场景? 在这些应用场景中有什么优势?(举出三个具体例子,并简要说明在这些场景中的优势,多于三个批改时以前三个为准)

解:

1. 增加的变量 E 用于表示原始数据 X 与降维后数据 X' 的噪声部分。由于低维数据收到噪声扰动最终呈现的维数可能会增加,添加正则项将优化目标转为同时最小化降维后数据的维数与噪声(结构风险),在优化函数中经验风险表现为 $rank(X')$,结构风险表现为 $\lambda\|E\|_0$。

2. 对于增广拉格朗日方法,根据优化问题的可分解性对整个优化问题实现 ADMM 求解。对于两个分解的优化问题,实现了近端梯度下降法对 X' 和 E 分别进行优化。[1] 对 X' 的优化采用核近端梯度优化

$$X'_{k+1} = \arg\min_{X'} \ \|X'\|_* + \lambda\|E_k\|_1 + \langle Y_k, X - X' - E_k\rangle + \frac{\mu}{2}\|X - X' - E_k\|_F^2$$

对 E 的优化采用 L1 近端梯度优化。

$$E_{k+1} = \arg\min_{E} \ \|X'_k\|_* + \lambda\|E\|_1 + \langle Y_k, X - X'_k - E\rangle + \frac{\mu}{2}\|X - X'_k - E\|_F^2$$

3. 1. 异常检测:假设异常点是不可被压缩的或不能从低维映射空间有效地被重构的。正常的数据是大量的且高度相似、相关的(低秩),而异常数据是稀疏,但会破坏数据的低秩性,符合 robust PCA 的优化假设(噪声是稀疏的,但强弱不影响),且相较于其他传统异常点检测算法(如基于聚类的)假设较少,便于直接进行优化 [3]。

2. 视频前景/背景分离：静态少变动的背景是低秩的，而动态变化的前景/异常噪声是稀疏的，使用 robust PCA 对图像矩阵数据进行分解可以得到低秩的背景矩阵 [2]。
3. 图像/各类去噪：相较于传统的 PCA，robust PCA 没有对噪声的分布进行假设，能够适用于更多场景的噪声去除应用。例如在在医学影像如 MRI 或 CT 扫描中，Robust PCA 能够帮助去除噪声和伪影，同时保留关键的结构信息。这是因为影像中的主要结构通常是连续且具有高度相关性（低秩），而噪声和伪影则是局部的、稀疏的。使用 Robust PCA，医生和诊断师可以获得更清晰、准确的影像，从而提高诊断的准确性和效率。[2]。

参考文献

[1] 最优化之 robust pca. `https://www.cnblogs.com/quarryman/p/robust_pca.html`, 2015. [Accessed 22-04-2024].

[2] Thierry Bouwmans, Sajid Javed, Hongyang Zhang, Zhouchen Lin, and Ricardo Otazo. On the applications of robust pca in image and video processing. *Proceedings of the IEEE*, 106(8):1427–1457, 2018.

[3] Huan Xu, Constantine Caramanis, and Shie Mannor. Outlier-robust pca: The high-dimensional case. *IEEE Transactions on Information Theory*, 59(1):546–572, 2013.