2024 秋季高级机器学习
习题一

221300079 王俊童

2024.10.9

一. (30 points) **机器学习导论复习题 (前九章)**

1. (10 points) 给定包含 m 个样例的数据集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}$, 其中 $y_i\in\mathbb{R}$ 为 $\boldsymbol{x}_i$ 的实数标记, $\boldsymbol{x}_i=(x_{i1};x_{i2};\cdots;x_{id})\in\mathbb{R}^d$。针对数据集 D 中的 m 个示例, 教材 3.2 节所介绍的 "线性回归" 模型要求该线性模型的预测结果和其对应的标记之间的误差之和最小:

$$\begin{aligned}(\boldsymbol{w}^*,b^*)&=\frac{1}{2}\arg\min_{(\boldsymbol{w},b)}\sum_{i=1}^{m}\left(f\left(\boldsymbol{x}_i\right)-y_i\right)^2\\&=\frac{1}{2}\arg\min_{(\boldsymbol{w},b)}\sum_{i=1}^{m}\left(y_i-\left(\boldsymbol{w}^\top\boldsymbol{x}_i+b\right)\right)^2,\end{aligned}\tag{1}$$

即寻找一组权重 $(\boldsymbol{w},b)$, 使其对 D 中示例预测的整体误差最小。定义 $\boldsymbol{y}=[y_1;\ldots;y_m]\in\mathbb{R}^m$, 且 $\boldsymbol{X}=\left[\boldsymbol{x}_1^\top;\boldsymbol{x}_2^\top;\cdots;\boldsymbol{x}_m^\top\right]\in\mathbb{R}^{m\times d}$, 线性回归的优化过程可以使用矩阵进行表示:

$$\begin{aligned}(\boldsymbol{w}^*,b^*)&=\arg\min_{\boldsymbol{w},b}\frac{1}{2}\left(\boldsymbol{X}\boldsymbol{w}+\mathbf{1}_m b-\boldsymbol{y}\right)^\top\left(\boldsymbol{X}\boldsymbol{w}+\mathbf{1}_m b-\boldsymbol{y}\right)\\&=\arg\min_{\boldsymbol{w},b}\frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{w}+\mathbf{1}_m b-\boldsymbol{y}\right\|_2^2,\end{aligned}\tag{2}$$

其中, $\mathbf{1}_m\in\mathbb{R}^m$ 为元素全为 1 、长度为 m 的向量。在实际问题中, 我们常常会遇到示例相对较少, 而特征很多的场景。在这类情况中如果直接求解线性回归模型, 较少的示例无法获得唯一的模型参数, 会具有多个模型能够" 完美" 拟合训练集中的所有样例。此外, 模型很容易过拟合。为缓解这些问题, 常引入正则化项 $\Omega(\boldsymbol{w})$, 通常形式如下:

$$\boldsymbol{w}_{\text{Ridge}}^*,b_{\text{Ridge}}^*=\arg\min_{\boldsymbol{w},b}\frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{w}+\mathbf{1}_m b-\boldsymbol{y}\right\|_2^2+\lambda\Omega(\boldsymbol{w}),\tag{3}$$

其中, $\lambda>0$ 为正则化参数。正则化表示了对模型的一种偏好, 例如 $\Omega(\boldsymbol{w})$ 一般对模型的复杂度进行约束, 因此相当于从多个在训练集上表现同等预测结果的模型中选出模型复杂度最低的一个。考虑岭回归问题, 即设置正则项 $\Omega(\boldsymbol{w})=\|\boldsymbol{w}\|_2^2$。

(1) (3 points) 请证明对于任何矩阵 $\boldsymbol{X}\in\mathbb{R}^{m\times d}$, 下式均成立

$$\left(\boldsymbol{X}\boldsymbol{X}^\top+\lambda\boldsymbol{I}_m\right)^{-1}\boldsymbol{X}=\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_d\right)^{-1}.\tag{4}$$

(2) (7 points) 请给出岭回归的最优解 $\boldsymbol{w}_{\text{Ridge}}^*$ 和 b_{Ridge}^* 的闭式解表达式, 并使用矩阵形式表示, 分析其最优解和原始线性回归最优解 $\boldsymbol{w}_{\text{LS}}^*$ 和 b_{LS}^* 的区别。

2. (10 points) 教材 4.2 节中给出度量样本集合纯度的常用指标, 从而衍生出决策树属性选择的常用准则。假设决策树分类问题中标记空间 $\mathcal{Y}$ 的大小为 $|\mathcal{Y}|$，训练集 D 中第 k 类样本所占比例为 $p_k(k=1,2,\ldots,|\mathcal{Y}|)$ 。请回答以下问题:

(1) (3 points) 信息熵 $\mathrm{Ent}(D)$ 定义如下

$$\mathrm{Ent}(D)=-\sum_{k=1}^{|\mathcal{Y}|}p_k\log_2 p_k, \tag{5}$$

请证明信息熵的上下界为

$$0\leq \mathrm{Ent}(D)\leq \log_2|\mathcal{Y}|, \tag{6}$$

并给出等号成立的条件。

(2) (3 points) 除信息熵外，教材中也介绍了基尼指数衡量纯度，定义如下

$$\sum_{k=1}^{|\mathcal{Y}|}\sum_{k'\neq k}p_k p_{k'}=1-\sum_{k=1}^{|\mathcal{Y}|}p_k^2, \tag{7}$$

由于在决策树叶结点中使用包含样例最多的类别作为其预测结果，因此也可使用误分类错误率

$$1-\max_k p_k, \tag{8}$$

作为衡量指标。请给出二分类问题（$|\mathcal{Y}|=2$，正类所占比例为 p, 负类为 $1-p$）下三种衡量标准的表达式。

(3) (4 points) 在 ID3 决策树的生成过程中, 需要计算信息增益以生成新的结点。设离散属性 a 有 V 个可能取值 $\left\{a^1,a^2,\cdots,a^V\right\}$ ，请参考教材 4.2.1 节相关符号的定义证明:

$$\mathrm{Gain}(D,a)=\mathrm{Ent}(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Ent}\left(D^v\right)\geq 0, \tag{9}$$

即信息增益非负。

3. (10 points) 给定训练集 $D=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2),\ldots,(\boldsymbol{x}_m,\boldsymbol{y}_m)\}$. 其中 $\boldsymbol{x}_i\in\mathbb{R}^d,\boldsymbol{y}_i\in\mathbb{R}^l$ 表示输入示例由 d 个属性描述，输出 l 维实值向量. 教材图 5.7 给出了一个有 d 个输入神经元、l 个输出神经元、q 个隐层神经元的多层神经网络，其中输出层第 j 个神经元的阈值用 θ_j 表示，隐层第 h 个神经元的阈值用 γ_h 表示。输入层第 i 个神经元与隐层第 h 个神经元之间的连接权为 v_{ih} ，隐层第 h 个神经元与输出层第 j 个神经元之间的连接权为 w_{hj} 。记隐层第 h 个神经元接收到的输入为 $\alpha_h=\sum_{i=1}^{d}v_{ih}x_i$, 输出层第 j 个神经元接收到的输入为 $\beta_j=\sum_{h=1}^{q}w_{hj}b_h$ ，其中 b_h 为隐层第 h 个神经元的输出。

不同任务中神经网络的输出层往往使用不同的激活函数和损失函数，本题介绍几种常见的激活和损失函数，并对其梯度进行推导。

(1) (3 points) 在二分类问题中 ($l=1$), 标记 $y\in\{0,1\}$ ，一般使用 Sigmoid 函数作为激活函数，使输出值在 $[0,1]$ 范围内，使模型预测结果可直接作为概率输出. Sigmoid 函数的输出一般配合二元交叉熵损失函数使用，对于一个训练样本 $(\boldsymbol{x},y)$ 有

$$\ell\left(y,\hat{y}_1\right)=-\left[y\log\left(\hat{y}_1\right)+(1-y)\log\left(1-\hat{y}_1\right)\right], \tag{10}$$

记 $\hat{y}_1$ 为模型将样本判断为正例的预测概率，请计算 $\frac{\partial \ell(y,\hat{y}_1)}{\partial \beta_1}$。

（2）（5 points）当 $l>1$，网络的预测结果为 $\hat{\boldsymbol{y}} \in \mathbb{R}^l$，其中 $\hat{y}_i$ 表示输入被预测为第 i 类的概率。对于第 i 类的样本，其标记 $\boldsymbol{y} \in \{0,1\}^l$，有 $y_i=1, y_j=0, j \neq i$。对于一个训练样本 $(\boldsymbol{x}, \boldsymbol{y})$，交叉熵损失函数 $\ell(\boldsymbol{y}, \hat{\boldsymbol{y}})$ 的定义如下

$$\ell(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_{j=1}^{l} y_j \log \hat{y}_j, \tag{11}$$

在多分类问题中，一般使用 Softmax 函数作为输出层激活函数，其计算公式如下

$$\hat{y}_j = \frac{e^{\beta_j}}{\sum_{k=1}^{l} e^{\beta_k}}, \tag{12}$$

易见 Softmax 函数输出的 $\hat{\boldsymbol{y}}$ 符合 $\sum_{j=1}^{l} \hat{y}_j = 1$，所以可以直接作为每个类别的概率. Softmax 函数输出一般配合交叉熵损失函数使用，请计算 $\frac{\partial \ell(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial \boldsymbol{\beta}}$。

（3）（2 points）分析在二分类中使用 Softmax 激活函数和 Sigmoid 激活函数的联系与区别。

解：

1. (1) 要证明 $\left(\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_m\right)^{-1}\boldsymbol{X} = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X} + \lambda \boldsymbol{I}_d\right)^{-1}$. 根据题目可知 $\left(\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_m\right)$ 和 $\left(\boldsymbol{X}^\top\boldsymbol{X} + \lambda \boldsymbol{I}_d\right)$ 两者是满秩的，所以可逆性可以得到证明。
   然后首先左乘 $\left(\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_m\right)$，然后右乘 $\left(\boldsymbol{X}^\top\boldsymbol{X} + \lambda \boldsymbol{I}_d\right)$。即可得到：$\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X} + \lambda \boldsymbol{I}_d\right) = \boldsymbol{X}\left(\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_m\right)$, 所以对于任意的 X，均可证明其对于这个式子均成立。
   (2) 可得原问题如下：$\boldsymbol{w}^*_{\text{Ridge}}, b^*_{\text{Ridge}} = \arg\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} + \mathbf{1}_m b - \boldsymbol{y}\|_2^2 + \lambda||\boldsymbol{w}||_2^2$
   首先化简原问题：$\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + 2b\boldsymbol{w}^T\boldsymbol{X}^T\mathbf{1}_m - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + mb^2 - 2mb\boldsymbol{y}^T\mathbf{1}_m + \boldsymbol{y}^T\boldsymbol{y}) + \lambda\boldsymbol{w}^T\boldsymbol{w}$
   对 w 求偏导，用拉格朗日乘子法得到:$(\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I}_d)\boldsymbol{w} = \boldsymbol{X}^T(\boldsymbol{y} - \mathbf{1}_m b)$
   $\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I}_d)^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \mathbf{1}_m b)$
   对 b 求偏导：$b = \frac{1}{m}\mathbf{1}_m^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$
   把 W 得到的偏导带入 b 可以得到：$(\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I}_d)\boldsymbol{w} = \boldsymbol{X}^T(y - \mathbf{1}_m\frac{1}{m}\mathbf{1}_m^T(\boldsymbol{y} - \boldsymbol{w}\boldsymbol{X}))$
   可解:$\boldsymbol{w}^*_{\text{Ridge}} = (\boldsymbol{X}^T(E - \frac{1}{m}\mathbf{1}\mathbf{1}_m)\boldsymbol{X} + 2\lambda\boldsymbol{I}_d)^{-1}\boldsymbol{X}^T(E - \frac{1}{m}\mathbf{1}\mathbf{1}_m)\boldsymbol{y}$
   带入 b 可得：$\boldsymbol{b}^*_{\text{Ridge}} = \frac{1}{m}\mathbf{1}_m^T(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^T(E - \frac{1}{m}\mathbf{1}\mathbf{1}_m)\boldsymbol{X} + 2\lambda\boldsymbol{I}_d)^{-1}\boldsymbol{X}^T(E - \frac{1}{m}\mathbf{1}\mathbf{1}_m)\boldsymbol{y})$

   对于原始闭式解来说，跟上式基本差不多，少了正则化项：
   可解:$\boldsymbol{w}^*_{\text{LS}} = (\boldsymbol{X}^T(E - \frac{1}{m}\mathbf{1}\mathbf{1}_m)\boldsymbol{X})^{-1}\boldsymbol{X}^T(E - \frac{1}{m}\mathbf{1}\mathbf{1}_m)\boldsymbol{y}$
   带入 b 可得：$\boldsymbol{b}^*_{\text{LS}} = \frac{1}{m}\mathbf{1}_m^T(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^T(E - \frac{1}{m}\mathbf{1}\mathbf{1}_m)\boldsymbol{X})^{-1}\boldsymbol{X}^T(E - \frac{1}{m}\mathbf{1}\mathbf{1}_m)\boldsymbol{y})$

   对于岭回归和原线性回归来说，岭回归多了一个 $2\lambda\mathbf{1}_m$, 而就是这个正则化后的项使得原始式子在 w 的最优解时那个逆变得可逆了，因为如果不加这个正则化项，原解并不一定是可逆的，就不一定有稳定解。当 $\lambda > 0$ 的时候，w 的范数可能就较小，可以防止过拟合，特别是在特征数量较多而样本数量较少的情形下。同理对于 b，b 的闭式解收到 w 的影响岭回归里面引入的正则化项也可以影响到 b 从而使得整个解答更加稳定。

2. (1) 首先证明下界：由于 log 函数定义在 0 到正无穷上而概率约束全部都大于 0，所以信息熵肯定是一个大于零的数据。取到 0 的情况可以是假如只有一个标记空间里面的 p 刚好为 1，其

余的全是 0，那么根据定义可以证明得到整个式子相加等于 0。或者 k=1 的时候，pk 就是 1，log1 等于 0，总体也为 0.
其次证明下界：首先说明，当所有的概率取值相等都是 $\frac{1}{|y|}$ 的时候，取到最大值 $log_2|y|$.
证明如下：可以证明若存在两个概率 x 和 1-x，可以根据信息熵的定义证明：$z = -(xlogx + (1-x)log(1-x))$,$z' = -log\frac{x}{1-x}$。当 $x = \frac{1}{2}$ 时，最大，$0 \le x < \frac{1}{2}$, 单调递增，$\frac{1}{2} \le x < 1$, 单调递减。那么根据这个道理，每次我们取最大和最小的概率分别作为 x_1, x_2, 可得 $-(plogp + (x_1 + x_2 - p)log(x_1 + x_2 - p)) > -(x_1logx_1 + x_2logx_2)$. 那么每次都可以做这种合并，做了 n-1 次之后，由于概率约束相加总和为 1，可以得到结果为 $-plogp$, 当 p 为 $\frac{1}{|y|}$ 的时候，结果为上界答案。

(2) 假设一个概率是 p，另一个是 1-p：
信息熵：$Ent(D) = -plogp - (1-p)log(1-p)$
Gini index:$Gini(p) = 1 - (p^2 + (1-p)^2) = 2p(1-p)$
误分类错误率:$Err = 1 - max(p, 1-p)$. 当 p 大于 0.5 的时候, 取 1-p。当 p 小于 0.5 的时候, 取 p

(3) 可得这个-log2 是一个凸函数，根据 jensen 不等式:$f(\sum_{i=1}^{n} \lambda_i x_i) \le \sum_{i=1}^{n} \lambda_i f(x_i)$. 对于这个 $Ent(D^v) = -\sum_{k=1}^{|y|} p_k^v log_2 p_k^v$, $\sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \sum_{k=1}^{|y|} p_k^v log_2 p_k^v$
由 jensen 可得：$\sum_{v=1}^{V} \frac{|D^v|}{|D|} \sum_{k=1}^{|y|} p_k^v log_2 p_k^v \le \sum_{k=1}^{|y|} (\sum_{v=1}^{V} \frac{|D^v|}{|D|} p_k^v) log_2 (\sum_{v=1}^{V} \frac{|D^v|}{|D|} p_k^v)$
而 $(\sum_{v=1}^{V} \frac{|D^v|}{|D|} p_k^v) = p_k$, 所以 $\sum_{k=1}^{|y|} (p_k) log_2 (p_k) = Ent(D)$
所以可得 $\text{Gain}(D, a) = \text{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \text{Ent}\,(D^v) \ge 0$

3. (1) 由于激活函数是 sigmoid：$f(x)' = f(x)(1 - f(x))$ 可以得到。由链式法则可以得到:
$\frac{\partial l}{\partial \beta_1} = \frac{\partial l}{\partial \hat{y}_1} \frac{\partial \hat{y}_1}{\partial \beta_1} = \frac{\hat{y}_1 - y_1}{\hat{y}_1(1-\hat{y}_1)} * \hat{y}_1(1 - \hat{y}_1) = \hat{y}_1 - y_1$
(2) 可以得到 l 的偏导数如下:$\frac{\partial l}{\partial \hat{y}_j} = \frac{-y_j}{\hat{y}_j}$
$\hat{y}_j = \frac{e^{\beta_j}}{\sum_{k=1}^{l} e^{\beta_k}}$ 对于这个函数求导可以分情况，若 $\hat{y}_j$ 对于 β_i 求导：
当 i=j：$\frac{\partial \hat{y}_j}{\partial \beta_j} = \frac{e^{\beta_j} \sum_{k=1}^{l} e^{\beta_k} - e^{\beta_j} e^{\beta_j}}{(\sum_{k=1}^{l} e^{\beta_k})^2} = \hat{y}_j(1 - \hat{y}_j)$
当 i 不等于 j:$\frac{\partial \hat{y}_j}{\partial \beta_j} = -\frac{e^{\beta_j} e^{\beta_i}}{(\sum_{k=1}^{l} e^{\beta_k})^2} = -\hat{y}_j \hat{y}_i$
所以同上，链式法则可以得到:
$\frac{\partial l}{\partial \beta_i} = \sum_{l}^{j=1} \frac{\partial l}{\partial \hat{y}_j} \frac{\partial \hat{y}_j}{\partial \hat{\beta}_i}$
当 i=j:$\frac{\partial l}{\partial \beta_i} = \hat{y}_j - y_j$
当 i 不等于 j:$-y_j \hat{y}_i$
所以如果要求：$\frac{\partial \ell(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial \boldsymbol{\beta}} = [\frac{\partial l}{\partial \beta_1}, ..., \frac{\partial l}{\partial \beta_l}] = -[y_1(1 - \hat{y}_1, ..., y_n(1 - \hat{y}_n))]$
(3) 联系：从数学形式上看，如果将 Softmax 函数应用于二分类问题，设 $\beta_1 = \beta, \beta_2 = 0$ 差不多是这个形式。且对于 softmax:$\hat{y}_1 = \frac{e^{\beta}}{e^{\beta} + e^{0}} = \frac{1}{1 + e^{-\beta}}$. 恰好就是 Sigmoid 函数的形式。这表明在二分类的特定情况下，Softmax 函数可以退化为与 Sigmoid 函数形式相同的表达式. 都可以映射到 0 和 1 区间。
区别：sigmoid 输出是一个值，表示某一类的概率，但是 softmax 明确的表示了两类的概率。且 sigmoid 计算简单，softmax 复杂度高。且搭配不一样的损失函数效果不一样。

二. (30 points) **PCA 降维**

教材 10.3 节介绍了主成分分析 (Principal Component Analysis, PCA) 方法对数据进行降维。本题考察 PCA 相关的线性代数基础知识以及基本操作。给定 d 维空间中 m 个样本构成的矩阵为

$$X = [x_1^\top; \ldots; x_m^\top] \in \mathbb{R}^{m \times d}, \tag{13}$$

$\hat{X} \in \mathbb{R}^{m \times d}$ 为 X 中心化后得到的矩阵。根据教材 10.3 节讨论，严格的协方差矩阵具有 $\frac{1}{m-1}$ 因子，由于常数对本题分析结果无影响，所以在本题的讨论中忽略该常数因子。

1. (6 points) $\hat{X}^\top\hat{X}$ 和 $\hat{X}\hat{X}^\top$ 为什么是半正定矩阵？二者的特征值有什么联系？受此启发，请思考当特征维度远大于样本个数时 ($d \gg m$)，使用特征值分解求解 PCA 应如何执行将更加高效？
2. (6 points) 奇异值分解定义如下：

   令 $\hat{X} \in \mathbb{R}^{m\times d}$，则存在正交矩阵 $U \in \mathbb{R}^{m\times m}$ 和 $V \in \mathbb{R}^{d\times d}$ 使得：

   $$\hat{X} = U\Sigma V^\top, \tag{14}$$

   其中

   $$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}, \tag{15}$$

   并且 $\Sigma_1 = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$，其对角线元素按数值大小排序：

   $$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0, \quad r = \mathrm{rank}(\hat{X}), \tag{16}$$

   当矩阵 $\hat{X}$ 的秩 $r = \mathrm{rank}(\hat{X}) < h$ 时，奇异值分解可以进行截尾，从而可简化为：

   $$\hat{X} = U_r\Sigma_r V_r^\top, \tag{17}$$

   式中

   $$U_r = (u_1, u_2, \ldots, u_r), V_r = (v_1, v_2, \ldots, v_r), \Sigma_r = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r), \tag{18}$$

   这种奇异值分解方式，被称为薄奇异值分解 (Thin SVD)。

   在实现 PCA 时，往往使用奇异值分解 (SVD) 而非特征值分解求解。请说明奇异值与特征值的关系，如果可以获得 $\hat{X}$ 的奇异值分解，应如何使用 PCA 对 $\hat{X}$ 进行降维？请分析使用 SVD 求解 PCA 相比于使用特征值分解求解 PCA 的优势。
3. (8 points) PCA 的一个重要步骤是将误差路径最小化重构误差，请说明为什么在最小化重构误差之前需要对数据进行中心化。
4. (10 points) 针对以下样本矩阵 (包含 5 个示例, 每个示例 2 维)，请对其进行主成分分析，将样本降至二维，并写出详细计算过程。

   $$X^\top = \begin{pmatrix} 3 & 4 & 4 & 6 & 3 \\ 2 & 3 & 2 & 3 & 0 \end{pmatrix} \tag{19}$$

解：

1. 对于任意的非零 y，我们可以得到:$y^T(\hat{X}^T\hat{X})y = (y\hat{X})^T Xy$, 考虑到向量内积的形式，这个式子可以写作:$||\hat{X}y||^2 \geq 0$, 所以这个是半正定的。对于 $\hat{X}\hat{X}^T$ 的证明同上。
   对于两者的特征值，我们可以得到，假设:$(\hat{X}^T\hat{X})v = \lambda v$, 同时左乘 $\hat{X}$,$\hat{X}\hat{X}^T\hat{X}v = \lambda\hat{X}v$ 化简得到:$(\hat{X}\hat{X}^T)(\hat{X}v) = \lambda(\hat{X}v)$，说明 $\hat{X}^T\hat{X}$ 的特征值也是 $\hat{X}\hat{X}^T$ 的特征值，反之亦然。这样就说明了这两个矩阵具有相同的非零特征值，假设有 r 个，对于第一个矩阵是一个 d*d 维度的，那就是 d-r 个零特征值，第二个矩阵是 m*m 维度的那就是 m-r 个特征值。
   由于这个规律，我们面对 d 维度远远大于 m 维度的时候，我们可以先解这个 m*m 维度的矩阵，因为剩下的全是 0 特征值，那么这样计算复杂度就会大大的降低。

2. 首先说明奇异值和特征值的关系，奇异值的平方等于特征值, 证明如下:
$A = U\Sigma V^T, (A^TA)v = \lambda v$，由于 U，V 都是正交矩阵，所以将奇异值分解结果带入特征值分解可以得到: $V^T\Sigma^T UU\Sigma V^T = \Sigma^T\Sigma v = \lambda v$, 所以 $\sigma_i^2 = \lambda_i$.
如果可以获得矩阵的奇异值分解结果，我们可以知道对于一个 m*d 的矩阵来说，d*d 的分解矩阵 V 保留了其主成分的信息。所以假设我们要取前 k 个维度的特征（降维到 k）可以选取前 k 个最大奇异值对应的 V 中的列向量，然后 $\hat{X}_{new} = \hat{X}V_k$, 其中 $V_k = (v_1, ..., v_k)$. 这样就可以完成 PCA 对于矩阵的降维。
优势：SVD 的应用比特征值更广泛。SVD 的稳定性比特征值更好，因为如果有接近 0 的情形的时候，SVD 处理矩阵比特征值更加稳定。且若对于一个大矩阵 d*d 维度像第一问提到的那样，SVD 的分解速率更快。且 SVD 获得的信息更多，多了一个 U 矩阵。

3. 最小化重构误差可以表示为:$||X - \hat{X}||^2$, 对于为什么要进行中心化:
1. 简化计算，如果不对数据进行中心化，协方差矩阵的计算会变得复杂。协方差矩阵反映了数据之间的联系程度。本质上 $Cov(X) = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})(x_i - \bar{x})^T$，如果不进行中心化，计算协方差时需要考虑每个数据点的绝对坐标，更复杂。
2. 可以确保主成分方向的合理性：如果不中心化，主成分方向可能会偏向于数据中心所在的位置，而不是真正反映数据内部变化最大的方向。而中心化后，数据的中心移到原点，主成分方向能够更准确地反映数据在各个维度上的相对变化情况。
3. 确保降维数据的合理性：如果有一个特征特别大，那就有可能导致小特征被忽视。中心化可以保证公平.
4. 标准化之后有利于后续算法的收敛.

4. 首先可以计算均值如下：$\bar{X}_1 = \frac{3+4+4+6+3}{5} = 4, \bar{X}_2 = \frac{2+3+2+3+0}{5} = 2$, 所以中心化后的矩阵是:

$$X^\top = \begin{pmatrix} -1 & 0 & 0 & 2 & -1 \\ 0 & 1 & 0 & 1 & -2 \end{pmatrix} \tag{20}$$

所以 $Cov(X) = \frac{1}{n-1}\sum_{i=1}^{m}(x_i - \bar{x})(x_i - \bar{x})^T$，此处 n=5：

$$Cov = \begin{pmatrix} \frac{3}{2} & 1 \\ 1 & \frac{3}{2} \end{pmatrix} \tag{21}$$

然后做特征值分解:$Cov - \lambda I = 0$
$\lambda = [2.5, 0.5]$
然后解得特征向量：$(Cov - \lambda I)v = 0$, 其矩阵可表示为

$$v = \begin{pmatrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{pmatrix} \tag{22}$$

因为其本身就是 2 维度的，这里投影矩阵就是这个 v，然后将其与中心化后矩阵相乘即可:

$$X^\top = \begin{pmatrix} -0.707 & 0.707 & 0 & 2.121 & -2.121 \\ 0.707 & 0.707 & 0 & -0.707 & -0.707 \end{pmatrix} \tag{23}$$

三. (40 points) **度量学习应用**

度量学习旨在学习一个适用于某个任务的距离度量，等价于为实现某个距离度量找到合适的特征变换。

1. (20 points) 教材 10.6 节介绍了马氏距离:

$$dist_{\text{mah}}^2(x_i, x_j) = (x_i - x_j)^\top M(x_i - x_j) = \|x_i - x_j\|_M^2, \tag{24}$$

在标准的马氏距离中，M 为样本协方差矩阵的逆 Σ^{-1}。而在度量学习中，M 是一个可学习的半正定矩阵 ($M \succeq 0$)，度量学习的过程可以看成一个优化 M 的过程。请回答以下问题：

1. (6 points) 标准的马氏距离去除了变量之间的相关性，并且与量纲无关。结合 PCA 中关于协方差矩阵的相关知识，请解释马氏距离为什么有这些优点。(提示：可将协方差矩阵进行特征值分解，重写 M 及上式)
2. (2 points) 标准马氏距离中 M 为协方差矩阵的逆，是否存在某些情况下协方差矩阵不可逆，应该如何应对这个问题?
3. (4 points) 不同于人工设定 M，度量学习在给定目标函数的条件下优化出半正定矩阵 M。结合教材 9.3 节对距离度量的介绍，请说明马氏距离是否是标准的距离度量（是否满足距离度量的四个性质)?
4. (8 points) 教材 3.4 节介绍的监督降维方法线性判别分析 LDA 以及 10.3 节介绍的无监督降维方法主成分分析 PCA 均可视为特殊的度量学习方法。简单来说，首先对样本进行降维，并在降维后空间中计算样本之间的欧氏距离作为距离度量。参考教材中的定义，类内散度矩阵 S_w 为每个类别的散度矩阵之和，类间散度矩阵 S_b 为每个类别与类中心的协方差矩阵。请写出 LDA 和 PCA 对应的马氏距离中的 M，并说明 LDA 和 PCA 的异同。(提示：将两种方法与度量学习进行关联)

2. (20 points) 度量学习方法一般需学习一个半正定的距离度量矩阵, 其目标函数是一个半正定规划 (Semi-Definite Programming, SDP) 问题, 是一类特殊的凸优化问题。

注：半正定规划有以下形式

$$\min_{X \in \mathcal{S}_+} \langle C, X \rangle$$
$$\text{s.t.} \quad \langle A_k, X \rangle \le b_k, \quad k = 1, \dots, m,$$

其中 X, C, A_k ($k = 1, \dots, m$) 均为 $d \times d$ 方阵。$\langle A, B \rangle = \mathrm{tr}(A^\top B) = \sum_{i=1}^{d} \sum_{j=1}^{d} A_{ij} B_{ij}$，$\mathcal{S}_+$ 表示半正定矩阵的集合。

本题以 LMNN 为例，探究度量学习的优化方式。

1. (5 points) 相比于线性或二次优化，半正定优化的求解较为缓慢。请推导 LMNN 损失函数对于 M 的梯度。若要保证 M 求解后为对称矩阵，M 需要如何初始化?
2. (10 points) 使用梯度下（5 points）降法求解 M 时，需保证 M 满足半正定约束。常见的做法是使用投影梯度下降（Projected Gradient Descent, PGD）方法在每次更新 M 时将其投影到半正定矩阵集合 $\mathcal{S}_+$ 中。半正定投影等价于求解如下问题：

$$\arg\min_{\hat{M}} \|\hat{M} - M\|_F^2 \quad \text{s.t.} \quad \hat{M} \in \mathcal{S}_+, \tag{25}$$

假设对称矩阵 M 的特征值分解为 $M = Q\Lambda Q^\top$，其中 $QQ^\top = I$ 为正交矩阵，Λ 为特征值构成的对角矩阵。请证上述问题的解为 $\hat{M} = Q\Lambda^+ Q^\top$，其中 Λ^+ 表示将 Λ 中的非负元素不变，负元素置零。

3. (5 points) 将任意半正定矩阵分解为投影矩阵，即 $M = PP^\top$，则 LMNN 可转化为关于 P 的无约束优化问题。请推导 LMNN 损失关于 P 的梯度。该问题是凸优化问题吗?

解：

1. (1) 协方差矩阵特征值分解可以表示为:$\Sigma = Q\Lambda Q^T, \Lambda = diag(\lambda_1, ..., \lambda_d)$.
所以可以得到：$dist^2_{\text{mah}}(x_i, x_j) = (x_i - x_j)^\top M (x_i - x_j) = (x_i - x_j)^\top \Sigma^{-1} (x_i - x_j) = (x_i - x_j)^\top Q \Lambda^{-1} Q^T (x_i - x_j)$
令 $y_i = Q^T x_i, y_j = Q^T x_j$, 则原来的式子可以写为:$dist^2_{\text{mah}}(x_i, x_j) = (y_i - y_j)^T \Lambda^{-1} (y_i - y_j) = \sum_{k=1}^{d} \frac{(x_i - x_j)^2}{\lambda_k}$.

首先其认为标准马氏距离去除了变量之间的相关性，因为其进行了 $y_i = Q^T x_i, y_j = Q^T x_j$ 的变换，从而目前是在一个新的坐标系下面的距离计算，去除了相关性。
其次其认为与量纲无关是因为 λ_k 本身作为特征值反映了数据的一定特性，但是现在输入维度都被其特征值调整，使得马氏距离和量纲没有关系了，因为比如一个数据比较大，那么其特征值肯定也很大，一除之后肯定会被调整掉。
(2) 对于协方差矩阵不可逆的情况，我认为有两种解决思路，第一个是像岭回归一样，引入一个很小的正则化项，既不会对于特征值排序和计算结果产生影响，也能够保证矩阵可逆。当然还有一个思路，就是对于不可逆的地方，将维度降下去，降到我们需要的满足可逆情况的矩阵，再进行计算。
(3) 根据标准距离度量来看，具有这四个性质：非负性，同一性，对称性，直递性。
首先证明非负性:$dist^2_{\text{mah}}(x_i, x_j) = (x_i - x_j)^\top M(x_i - x_j)$, 由于 M 矩阵是一个半正定的，令 $y = (x_i - x_j)$,$y^T My \geq 0$, 当且仅当 $x_i = x_j$ 时候，等于 0.
同一性: 当 $x_i = x_j$ 的时候, $dist^2_{\text{mah}}(x_i, x_j) = (x_i - x_j)^\top M(x_i - x_j) = 0$; 反之若 $dist^2_{\text{mah}}(x_i, x_j) = (x_i - x_j)^\top M(x_i - x_j) = 0$ 由于 M 是个半正定矩阵，只可能在 $x_i = x_j$ 时候整个式子为 0.
对称性:
$dist^2_{\text{mah}}(x_i, x_j) = (x_i - x_j)^\top M(x_i - x_j) = ((x_j - x_i)^T)^T M(x_j - x_i) = dist^2_{\text{mah}}(x_j, x_i)$
直递性 (三角不等式):$dist^2_{\text{mah}}(x_i, x_j) \leq dist^2_{\text{mah}}(x_i, x_k) + dist^2_{\text{mah}}(x_k, x_j)$
令 $a = (x_i - x_k), b = (x_k - x_j), (x_i - x_j) = a + b$，所以 $dist^2_{\text{mah}}(x_i, x_j) = (a + b)^T M(a + b)$,$dist^2_{\text{mah}}(x_i, x_k) = a^T Ma$,$dist^2_{\text{mah}}(x_k, x_j) = b^T Mb$
则可以得到,$dist^2_{\text{mah}}(x_i, x_j) = a^T Ma + 2a^T Mb + b^T Mb$,
由柯西不等式得到:$a^T Mb \leq \sqrt{(a^T Mb)(a^T Mb)}$, 所以可以得到原式子:
$dist^2_{\text{mah}}(x_i, x_j) \leq a^T Ma + \sqrt{(a^T Mb)(a^T Mb)} + b^T Mb = (\sqrt{a^T Ma}\sqrt{b^T Mb}) = dist^2_{\text{mah}}(x_i, x_k) + dist^2_{\text{mah}}(x_k, x_j)$, 所以 QED。
(4) 对于多类 LDA 来说，有 $S_w = \sum_{i=1}^n S_w i, S_b = \sum_{i=1}^n (\mu_i - \mu)^T (\mu_i - \mu)$，对于这个的解如下 $S_b W = \lambda S_w W$, 为了对应马氏距离，其 M 可以这样表示 $M = S_w^{-1} S_b$
对于 PCA 来说，$M = \Sigma^{-1}$.
相同点：他们都是将数据从高维降维到低维的技术，并且在降维之后都是用样本在低维空间的欧式距离作为距离度量，并选取不同的 M 矩阵来反映其特点的。
不同点：PCA 是无监督的，考虑的是最大可分和最小重构，而 LDA 是一种监督的降维方法，目的是最大化类间散度和最小类内散度。且他们的 M 选取不一样，PCA 考虑的是特征值分解之后的特征值更多，而 LDA 考虑的是数据之间的聚类和簇之间的关系，是一种 Global 和 local 的衡量方式。

2. (1)
(2)
(3)