一、复习题

1. (1) 对原式进行变形，同时左乘与右乘 $(XX^T+\lambda I_m)$ 和 $(X^TX+\lambda I_d)$，有：

$$X(X^TX+\lambda I_d) = (XX^T+\lambda I_m)X$$

进行分配得：$XX^TX+\lambda XI_d = XX^TX+\lambda I_mX$

即：$XX^TX+\lambda X = XX^TX+\lambda X$　故式(4)均成立.

(2) 优化目标函数为：$\frac{1}{2}\|Xw+1_mb-y\|_2^2+\lambda\|w\|_2^2$

先优化 b. 且显然有 $1_m^T \cdot 1_m = m$

由于 $\lambda\|w\|_2^2$ 不包含 b. 故这里对其不必关心.

对 $\frac{1}{2}(Xw+1_mb-y)^T(Xw+1_mb-y)$ 中的 b 求导

即　$mb-1_m^Ty+1_m^TXw$. 令其等于0.

可解得　$b=\frac{1_m^T(y-Xw^*)}{m}$

再优化 w. 对 $\frac{1}{2}\|Xw+1_mb-y\|_2^2+\lambda\|w\|_2^2$ 中的 w 求导

得　$X^T(Xw+1_mb-y)+2\lambda w=0$　由于 λ 为常数. 故可解得：

$w_R^* = (X^TX+\lambda I_d)^{-1}X^T(y-1_mb)$.　　$b_R^* = \frac{1_m^T(y-Xw_R^*)}{m}$

而 $w_{LS}^* = (X^TX)^{-1}X^T(y-1_mb)$　　$b_{LS}^* = \frac{1_m^T(y-Xw_{LS}^*)}{m}$

区别在于岭回归引入了 λI_d 这一项. 以保证 W_R^* 为满秩矩阵 避免了 wX^TX 无法求逆的问题. 同时减小了模型复杂度. 降低了过拟合风险.

2.(1) ①证明 $0 \le Ent(D)$

由定义知 $Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k$ 因为 p_k 为样本所占比例. 故 $0 \le p_k \le 1$.

所以 $\log_2 p_k \le 0$, $-\log_2 p_k \ge 0$ $\therefore Ent(D) = \sum_{k=1}^{|y|} -p_k \log_2 p_k \ge 0$

在 $p_k = 1$ 即 $|y| = 1$, 仅有一类样本时取到等号. $Ent(D) = 0$

②证明 $Ent(D) \le \log_2 |y|$.

设 $f(x) = -x\log x$. $(0 \le x \le 1)$. $f'(x) = -1 - \ln x$ $f''(x) = -\frac{1}{x} < 0$ $\therefore f(x)$ 在 $[0,1]$ 上为上凸函数

由Jensen不等式得: $\frac{-\sum_{k=1}^{|y|} p_k \ln p_k}{|y|} \le f\left(\frac{\sum_{k=1}^{|y|} p_k}{|y|}\right) = -\frac{\sum_{k=1}^{|y|} p_k}{|y|} \ln\left(\frac{\sum_{k=1}^{|y|} p_k}{|y|}\right) = \frac{1}{|y|} \ln |y|$

故 $Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k \le \ln |y|$ 由 $\sum_{k=1}^{|y|} p_k = 1$ 和 $f(x) = -x\log x$ 得:

当 $p_1 = p_2 = \cdots = p_k = \frac{1}{|y|}$ 时. $Ent(D)$ 取得最大值 $Ent(D) = \log_2 |y|$.

2.(2) 由题可知. 在二分类问题中. $|y| = 2$.

①信息熵: $Ent(D) = -(p\log_2 p + [(1-p)\log_2(1-p)]) = -p\log_2 p - (1-p)\log_2(1-p)$

②基尼指数: $1 - [p^2 + (1-p)^2] = 2p - 2p^2$

③误分类错误率: $1 - \max_k p_k = 1 - \max\{p, 1-p\}$.

故 $p \ge 0.5$ 时, 选择 p. 即为 $1-p$. $p < 0.5$ 时 选择 $1-p$. 即为 p.

$\therefore Error(D) = \begin{cases} 1-p, & p \ge 0.5 \\ p. & p < 0.5 \end{cases}$

(3) 由(1)可知. $Ent(D)$ 为一个上凸函数. 且 $\sum_{v=1}^{V} \frac{|D^v|}{|D|} = 1$. 数据集 D 为 D_v, $v \in [1, V]$ 之和.

故由Jensen不等式有: $Ent(D) \ge \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$

故 $Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) \ge 0$

更正: 一.2(1)中信息熵下界的证明:

设 $f(p) = -p\log_2 p$. 则有 $f'(p) = -\log_2 p - \frac{1}{\ln 2}$ $f''(p) = -\frac{1}{p\ln 2}$

$\because p \in [0,1]$ 所以 $f''(p) < 0$. 说明 $f'(p)$ 单调递减. 且 $f'(\frac{1}{e}) = 0$

当 $p = 0$ 和 1 时 $f(p) = 0$ $\therefore f(p)$ 在 $[0,1]$ 上最小值为 0. 即对任意 $-p_k\log_2 p_k$

都有 $-p_k \log_2 p_k \ge 0$ $\therefore Ent(D) \ge 0$.

3.(1) 由链式法则: $\frac{\partial L(y,\hat{y}_1)}{\partial \beta_1} = \frac{\partial L(y,\hat{y}_1)}{\partial \hat{y}_1}\cdot\frac{\partial \hat{y}_1}{\partial \beta_1}$

对于 $\frac{\partial L(y,\hat{y}_1)}{\partial \hat{y}_1} = -[y\cdot\frac{1}{\hat{y}_1} + (1-y)\cdot(-\frac{1}{1-\hat{y}_1})] = \frac{1-y}{1-\hat{y}_1} - \frac{y}{\hat{y}_1}$

由Sigmoid函数性质: $f'(x)=f(x)\cdot(1-f(x))$ $\quad \frac{\partial \hat{y}_1}{\partial\beta_1} = \hat{y}_1\cdot(1-\hat{y}_1)$

代入可得 $\frac{\partial L(y,\hat{y}_1)}{\partial \hat{y}_1} = (\frac{1-y}{1-\hat{y}_1}-\frac{y}{\hat{y}_1})(\hat{y}_1\cdot(1-\hat{y}_1)) = (1-y)\hat{y}_1 - y(1-\hat{y}_1) = \hat{y}_1 - y$

3.(2) 对于 $i=j$ 时. $\frac{\partial L(y,\hat{y})}{\partial \hat{y}_j}\cdot\frac{\partial \hat{y}_j}{\partial\beta_j} = (-\frac{y_j}{\hat{y}_j})\cdot\hat{y}_j(1-\hat{y}_j) = \hat{y}_j - y_j$.

对于 $i\neq j$ 时. $\frac{\partial L(y,\hat{y})}{\partial \hat{y}_i}\cdot\frac{\partial\hat{y}_i}{\partial\beta_j} = (-\frac{y_i}{\hat{y}_i})\cdot(-\hat{y}_i\hat{y}_j) = \hat{y}_j y_i$

且对于 $i\neq j$ 时. $y_i=0$ 故 $\frac{\partial L(y,\hat{y})}{\partial \hat{y}_i}\cdot\frac{\partial \hat{y}_i}{\partial \beta_j}=0$

所以 $\frac{\partial L(y,\hat{y})}{\partial\beta_j} = \frac{\partial L(y,\hat{y})}{\partial \hat{y}_j}\cdot\frac{\partial\hat{y}_j}{\partial\beta_j} = \hat{y}_j - y_j$

$\therefore \frac{\partial(y,\hat{y})}{\partial\beta} = \hat{y} - y$. 其中 $\hat{y}$为softmax输出的概率向量

y为真实类别的标签向量

一. 3.(3) 对于二分类问题而言. 两者无本质区别. 下面以输入β_1为例:

Sigmoid函数: $\hat{y}_1 = \frac{1}{1+e^{-\beta_1}}$

Softmax函数: $\hat{y}_1 = \frac{e^{\beta_1}}{e^{\beta_1}+e^{\beta_2}} = \frac{1}{1+e^{-(\beta_1-\beta_2)}}$

由于(X_1-X_2)可以用Z_1代替. 即可写为: $\hat{y}_1(z_1) = \frac{1}{1+e^{-z_1}}$

这与Sigmoid形式完全相同. 所以理论上两者没有区别. 但是对于实际的输入输出来说, Sigmoid输出一个值$\hat{y}_1$. 表示正例的概率. 反例的概率则为 $1-\hat{y}_1$.

Softmax在二分类中输出两个值$\hat{y}_1$和$\hat{y}_2$. 分别对应正例和反例的概率

且满足$\hat{y}_1+\hat{y}_2=1$.

二.PCA降维

1. ①对任意非零向量 $v\in R^d$. $v^T\hat{X}^T\hat{X}v=(\hat{X}v)^T(\hat{X}v)=\|\hat{X}v\|^2\geq 0$

故 $\hat{X}^T\hat{X}$ 是半正定矩阵. 同理任意非零向量 $u\in R^m$

$u^T\hat{X}\cdot\hat{X}^Tu=(\hat{X}^Tu)^T\cdot(\hat{X}^Tu)=\|\hat{X}^Tu\|^2\geq 0$ 故 $\hat{X}\hat{X}^T$ 也是半正定矩阵.

②对 $\hat{X}$ 进行奇异值分解. 有 $\hat{X}=U\Sigma V^T$. 故 $\hat{X}^T\hat{X}=V\Sigma^TU^TU\Sigma V^T$

由于 U 是正交矩阵. 故 $U^TU=I$ 所以 $\hat{X}^T\hat{X}=V\Sigma^T\Sigma V^T$

故其特征值为 $\Sigma^T\Sigma$ 的对角元素. 即奇异值的平方.

同理 $\hat{X}\hat{X}^T=U\Sigma\Sigma^TU^T$. 特征值为 $\Sigma\Sigma^T$ 的对角元素. 也是奇异值的平方.

故 $\hat{X}^T\hat{X}$ 和 $\hat{X}\hat{X}$ 的非零特征值相同.

③为了高效. 由于 $\hat{X}\hat{X}^T$ 的维度较小. 可以先用其推导出 $\hat{X}^T\hat{X}$ 的特征值.

再对其进行特征值分解.

二.2.

①关系: 奇异值是 $\hat{X}^T\hat{X}$ 或 $\hat{X}\hat{X}^T$ 的特征值的平方根

即 $\sigma_i=\sqrt{x_i}$. 其中 σ_i 为 $\hat{X}$ 的奇异值, x_i 为 $\hat{X}^T\hat{X}$ 或 $\hat{X}\hat{X}^T$ 的特征值.

②若获得了 $\hat{X}$ 的奇异值分解. 即有 $\hat{X}=U\Sigma V^T$

样本的协方差矩阵 $\hat{X}\cdot\hat{X}^T=U\Sigma V^T\cdot V\Sigma^TU^T=U\Sigma\Sigma^TU^T$ (忽略常数因子).

由于矩阵 U 正交. $UU^T=I$. $UU^{-1}=I$ 故 $U^T=U^{-1}$

故 $\hat{X}\hat{X}^T=U\Sigma\Sigma^TU^{-1}$. 为特征值分解.

U 中对应的列向量. u_i 即为对应的特征向量.

这里取要求的前 k 个向量, 即 $u_1, u_2\cdots u_k$ 组成的投影矩阵 W^*

③优势: 1) 适用于高维数据, 计算开销少; 当 $d\gg m$ 时. 特征值的分解

成本过高. 而 SVD 可以直接分解.

2) SVD 数值稳定. 而特征值分解不行.

3) 对于某些稀疏数据. thin SVD 的计算量和内存开销更小.

二.3. 中心化将数据平移至均值为零的状态, 这意味着去除了数据的偏移, 使得所有主成分的计算基于数据的内部结构和相对分布, 而不是受到原始数据均值位置的影响. 如果不进行中心化, PCA 可能会错误地将均值位置作为主成分的一个方向, 从而偏离对数据方差的真实描述。

二.4. 首先对数据进行中心化处理:

$$\bar{x}_1 = \frac{3+4+4+6+3}{5} = 4 \qquad \bar{x}_2 = \frac{2+3+2+3+0}{5} = 2$$

计算中心化后的数据矩阵 $\hat{X}^T = \begin{pmatrix} -1 & 0 & 0 & 2 & -1 \\ 0 & 1 & 0 & 1 & -2 \end{pmatrix}$

$$\hat{X}^T\hat{X} = \begin{pmatrix} -1 & 0 & 0 & 2 & -1 \\ 0 & 1 & 0 & 1 & 2 \end{pmatrix}\begin{pmatrix} -1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 2 & 1 \\ -1 & -2 \end{pmatrix} = \begin{pmatrix} 6 & 4 \\ 4 & 6 \end{pmatrix}$$

协方差矩阵 C 为: $C = \frac{1}{n-1}\hat{X}^T\hat{X} = \begin{pmatrix} \frac{3}{2} & 1 \\ 1 & \frac{3}{2} \end{pmatrix}$

求其特征值与特征向量: $|C-\lambda I| = 0$ 即 $\begin{vmatrix} \frac{3}{2}-\lambda & 1 \\ 1 & \frac{3}{2}-\lambda \end{vmatrix} = 0$

即 $(\frac{3}{2}-\lambda)^2 - 1 = 0$ $\quad \therefore \lambda = \frac{5}{2}, \frac{1}{2}$

$(C - \frac{5}{2}I)v_1 = 0 \Rightarrow \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$

解得: 特征向量 $v_1 = (1,1)^T$. 对应特征值 $\lambda_1 = \frac{5}{2}$

$(C - \frac{1}{2}I)v_2 = 0 \Rightarrow \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$

解得: 特征向量 $v_2 = (1,-1)^T$. 对应特征值 $\lambda_2 = \frac{1}{2}$

因此投影矩阵为: $W = (v_1 v_2) = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$

$\therefore$ 降维后的数据为 $Z = W^T\hat{X}^T = \begin{pmatrix} -1 & 1 & 0 & 3 & -3 \\ -1 & -1 & 0 & 1 & 1 \end{pmatrix}$

三. 度量学习应用:

1. (1) 对标准马氏距离的M进行讨论, 对协方差矩阵Σ进行特征值分解.

$\Sigma = V\Lambda V^{-1}$ 由于V是一个正交矩阵. 有 $V^{-1} = V^T$.

故可写成 $\Sigma = V\Lambda V^T$ $\Sigma^{-1} = V\Lambda^{-1}V^T$. Λ的对角元素为特征值, 对应主成分的方差.

重写马氏距离: $dist^2_{mah}(x_i, x_j) = (x_i - x_j)^T V\Lambda^{-1}V^T(x_i - x_j)$

在这种表示中, 引入 Λ^{-1}. 使得每个特征的贡献大都会被其方差标准化.

消去了变量之间的相关性.

由于Σ对角元素表示各个特征值的方差. 故使用 Λ^{-1} (Σ^{-1}) 对每个特征标准化

将每个特征转换为相同的度量单位, 消去了数据的量纲影响.

(2) 是的, 在数据维度大于样本数量或特征存在共线性时, 或者数据中的某些特征

的方差为0时. 就会出现协方差矩阵不可逆. 应该这样应对:

① 正则化协方差矩阵, 添加一个小的正则项, 即 $\Sigma_{reg} = \Sigma + \lambda I$

② 使用PCA降维, 挑选其主要成分, 去除冗余或低方差的特征再计算.

一. (3)

① 非负性: 由于M是半正定矩阵. $M \geq 0$. 故 $(x_i - x_j)^T M (x_i - x_j) \geq 0$

故 $dist_{mah}(x_i, x_j) \geq 0$ 成立.

② 同一性: 当 $x_i = x_j$ 时. 显然 $(x_i - x_j)^T M (x_i - x_j) = 0$ $dist_{mah}(x_i, x_j) = 0$

当 $dist_{mah}(x_i, x_j) = 0$ 时. 即 $(x_i - x_j)^T M (x_i - x_j) = 0$

由于M为半正定矩阵. 所以 $x_i - x_j = 0$ 即 $x_i = x_j$ 成立.

③ 对称性: 由于 $\|x_i - x_j\|_2 = \|x_j - x_i\|_2$

故显然有: $(x_i - x_j)^T M (x_i - x_j) = (x_j - x_i)^T M (x_j - x_i)$

即 $dist_{mah}(x_i - x_j) = dist_{mah}(x_j - x_i)$ 故满足对称性

④ 直递性: 将M分解成 $V\Lambda^{-1}V^T$. 由于 Λ^{-1} 为对角阵. 故将 V^{-1} 分解成 PP^T

$$
\begin{aligned}
\text{故 } dist^2_{mah}(x_i, x_j) &= (x_i - x_j)^T V P P^T V^T (x_i - x_j) \\
&= (V^T x_i - V^T x_j) \cdot P \cdot P^T (V^T x_i - V^T x_j) \\
&= (P^T V^T x_i - P^T V^T x_j)^T (P^T V^T x_i - P^T V^T x_j) \\
&= dist^2(P^T V^T x_i, P^T V^T x_j)
\end{aligned}
$$

同理有: $dist_{mah}(x_i, x_k) = dist(P^T V^T x_i, P^T V^T x_k)$

$dist_{mah}(x_k, x_j) = dist(P^T V^T x_k, P^T V^T x_j)$

由欧氏距离性质有: $dist(P^T V^T x_i, P^T V^T x_j) \leq dist(P^T V^T x_i, P^T V^T x_k) + dist(P^T V^T x_k, P^T V^T x_j)$

故有: $dist_{mah}(x_i, x_j) \leq dist_{mah}(x_i, x_k) + dist_{mah}(x_k, x_j)$.

三.1.4

对于 LDA 优化目标为:

$$J(W)=\frac{\mathrm{tr}(W^\top S_b W)}{\mathrm{tr}(W^\top S_w W)}$$

其中, W 是降维投影矩阵, S_b 为类间散度矩阵, S_w 为类内散度矩阵.
在马氏距离的度量中, LDA 可以看作是通过选择一个 M 矩阵来衡量样本之间的距离, 对于 LDA, 降维后的马氏距离中的矩阵 M 可以写为,

$$M_{\mathrm{LDA}}=S_w^{-1}$$

PCA 的目标函数可以表示为:

$$J(W)=\mathrm{tr}(W^\top S_t W)$$

其中, S_t 是数据的总散度矩阵 (也就是协方差矩阵), W 是投影矩阵.
在马氏距离的度量中, PCA 可以看作是通过选择一个矩阵 M 来衡量样本之间的距离, 对于 PCA, 降维后的马氏距离中的矩阵 M 可以写为:

$$M_{\mathrm{PCA}}=S_t^{-1}$$

LDA 和 PCA 两者都是线性降维的方法, 并且在度量学习中也都可以用马氏距离来解释, 都是通过选择不同的矩阵 M 来实现数据的投影与度量, 最终在降维后的空间中计算样本之间的欧氏距离, 不同点在于 LDA 是有监督的降维方法, 而 PCA 是无监督的降维方法. LDA 对应的矩阵 M 是类内散度矩阵的逆, 即 $M_{\mathrm{LDA}}=S_w^{-1}$, 表示样本所属类别的关系, 特别是类内的相关性, PCA 对应的矩阵 M 是协方差矩阵的逆, 即 $M_{\mathrm{PCA}}=S_t^{-1}$, 只关注数据在各个方向上的总方差, 最大化数据的方差, 以保留尽可能多的信息量.

2.

题1：LMNN的损失函数可以表示为：$\varepsilon(L)=(1-\mu)\varepsilon_{pull}(L)+\mu\varepsilon_{push}(L)$

其中：$\varepsilon_{pull}(L)=\sum_{j\sim i}\|L(\vec{x}_i-\vec{x}_j)\|^2$ $\varepsilon_{push}(L)=\sum_{i,j\sim i}\sum_{l}(1-y_{il})[1+\|L(\vec{x}_i-\vec{x}_j)\|^2-\|L(\vec{x}_i-\vec{x}_l)\|^2]_+$

这其中 $j\sim i$ 表示样本 j 是样本 i 的邻居，y_{il} 是一个指示函数，用于区分是否为不同类别，$[z]_+=\max(0,z)$ 表示 hinge 损失。

①类内紧密性部分的梯度：

令马氏距离矩阵 $M=L^TL$，则这部分可以表示为：

$$\varepsilon_{pull}(M)=\sum_{j\sim i}(\vec{x}_i-\vec{x}_j)^TM(\vec{x}_i-\vec{x}_j)$$

$$\frac{\partial\varepsilon_{pull}(M)}{\partial M}=\sum_{j\sim i}(\vec{x}_i-\vec{x}_j)(\vec{x}_i-\vec{x}_j)^T$$

②类间分离性部分的梯度：这部分包含 hinge 损失，只有在违背约束的情况下才会对梯度有贡献。损失形式为：$\varepsilon_{push}(M)=\sum_{i,j\sim i}\sum_{l}(1-y_{il})[1+(\vec{x}_i-\vec{x}_j)^TM(\vec{x}_i-\vec{x}_j)-(\vec{x}_i-\vec{x}_l)^TM(\vec{x}_i-\vec{x}_l)]_+$

$$\frac{\partial\varepsilon_{push}(M)}{\partial M}=\sum_{i,j\sim i}\sum_{l}(1-y_{il})\mathbb{1}_{[z]_+>0}[(\vec{x}_i-\vec{x}_j)(\vec{x}_i-\vec{x}_j)^T-(\vec{x}_i-\vec{x}_l)(\vec{x}_i-\vec{x}_l)^T]$$

其中 $\mathbb{1}_{[z]_+>0}$ 是指示函数，表示当 hinge 损失大于0时，该项对梯度有贡献。

因此 LMNN 损失函数对于 M 的梯度为：

$$\frac{\partial\varepsilon(M)}{\partial M}=(1-\mu)\sum_{j\sim i}(\vec{x}_i-\vec{x}_j)(\vec{x}_i-\vec{x}_j)^T+\mu\sum_{i,j\sim i}\sum_{l}(1-y_{il})\mathbb{1}_{[z]_+>0}[(\vec{x}_i-\vec{x}_j)(\vec{x}_i-\vec{x}_j)^T-(\vec{x}_i-\vec{x}_l)(\vec{x}_i-\vec{x}_l)^T]$$

$$=\sum_{i,j\sim i}\left[(1-\mu)(\vec{x}_i-\vec{x}_j)(\vec{x}_i-\vec{x}_j)^T+\mu\sum_{l}(1-y_{il})\mathbb{1}_{[z]_+>0}[(\vec{x}_i-\vec{x}_j)(\vec{x}_i-\vec{x}_j)^T-(\vec{x}_i-\vec{x}_l)(\vec{x}_i-\vec{x}_l)^T]\right]$$

在优化过程中，每次更新 M 后，可以通过以下操作保持其对称性。

$$M=\frac{1}{2}(M+M^T)$$

这样在梯度下降的过程中任何对 M 的非对称扰动都被对称化。

题2：可以将目标函数重写为 $\|\hat{M}-M\|_F^2 = tr((\hat{M}-M)^2) = tr(\hat{M}^2) - 2tr(\hat{M}M) + tr(M^2)$

由于 $tr(M^2)$ 是常数，只需最小化 $f(\hat{M}) = tr(\hat{M}^2) - 2tr(\hat{M}M)$

∵ $\hat{M}$ 是半正定对称矩阵，对 $\hat{M}$ 进行特征值分解：$\hat{M} = P\hat{\Lambda}P^T$

其中：P 是由 $\hat{M}$ 的特征向量组成的正交矩阵，

$\hat{\Lambda} = diag(\hat{\lambda}_1, \cdots \hat{\lambda}_d)$ 是 $\hat{M}$ 的特征值组成的对角矩阵，且 $\hat{\lambda}_i \geq 0$

由于 $tr(AB) = tr(BA)$，且 $P^TP = I$，因此有：$tr(\hat{M}^2) = tr((P\hat{\Lambda}P^T)(P\hat{\Lambda}P^T)^T) = tr(P\hat{\Lambda}^2P^T) = tr(\hat{\Lambda}^2)$

$tr(\hat{M}M) = tr(P\hat{\Lambda}P^TQ\Lambda Q^T) = tr(\hat{\Lambda}P^TQ\Lambda Q^TP)$

令 $W = P^TQ$，则 W 是正交矩阵，且 $tr(\hat{M}M) = tr(\hat{\Lambda}W\Lambda W^T)$

因此目标函数变为：$f(\hat{M}) = tr(\hat{\Lambda}^2) - 2tr(\hat{\Lambda}W\Lambda W^T)$

展开 $tr(\hat{\Lambda}W\Lambda W^T)$，其中 W_{ij} 是矩阵 W 的元素：$tr(\hat{\Lambda}W\Lambda W^T) = \sum_{i=1}^{d}\sum_{j=1}^{d}\hat{\lambda}_i\lambda_j(W_{ij})^2$

由于 $\hat{\lambda}_i \geq 0$，且 W 为正交矩阵，因此 $\sum_{j=1}^{d}(W_{ij})^2 = 1$，而对于 W_{ij}，为了最小化 $f(\hat{M})$，

需要最大化 $\sum_{i,j}\hat{\lambda}_i\lambda_j(W_{ij})^2$，即寻找 W_{ij}，使得 $\sum_{i,j}\hat{\lambda}_i\lambda_j(W_{ij})^2$ 最大化。

注意到当 $W = I$ 时 $(W_{ij})^2 = \delta_{ij}$（克罗内克函数），即 $W_{ij} = \delta_{ij}$，此时有：

$\sum_{i,j}\hat{\lambda}_i\lambda_j(W_{ij})^2 = \sum_{i=1}^{d}\hat{\lambda}_i\lambda_i$ 这是能取到的最大值，对于任意正交矩阵 W，

根据 Schur 凸性的性质：$\sum_{i=1}^{d}\hat{\lambda}_i\lambda_i \geq \sum_{i,j}\hat{\lambda}_i\lambda_j(W_{ij})^2$

因此，为了最小化 $f(\hat{M})$ 应选择 $P = Q$，即：$W = I$。

当 $P = Q$ 时，目标函数变为：$f(\hat{M}) = \sum_{i=1}^{d}\hat{\lambda}_i^2 - 2\sum_{i=1}^{d}\hat{\lambda}_i\lambda_i = \sum_{i=1}^{d}(\hat{\lambda}_i^2 - 2\hat{\lambda}_i\lambda_i) = \sum_{i=1}^{d}(\hat{\lambda}_i - \lambda_i)^2 - \lambda_i^2$

由于 λ_i^2 是常数，所以只需最小化 $g(\hat{\lambda}_i) = \sum_{i=1}^{d}(\hat{\lambda}_i - \lambda_i)^2$

需要指明的是，此时并未假设 $\hat{\Lambda}$ 是特定的对角矩阵，而是通过优化过程得出了应选择 $P = Q$，从而使得 $\hat{M}$ 与 M 在同一基下进行比较。现在需要在约束 $\hat{\lambda}_i \geq 0$ 下，最小化 $g(\lambda_i)$，对于每一个 i 这是一个简单的优化问题：$\min_{\hat{\lambda}_i \geq 0}(\hat{\lambda}_i - \lambda_i)^2$

解为：若 $\lambda_i \geq 0$，则最优解为 $\hat{\lambda}_i = \lambda_i$ 因此最优 $\hat{\lambda}_i$ 为：$\hat{\lambda}_i = max(\lambda_i, 0)$

若 $\lambda_i < 0$，则最优解为 $\lambda_i = 0$ 代入最优特征值

$\hat{\Lambda} = \Lambda^+ = diag(\hat{\lambda}_1, \cdots \hat{\lambda}_d) = diag(max(\lambda_1, 0), max(\lambda_2, 0) \cdots max(\lambda_d, 0))$.

∵ $P = Q$ ∴ 最优半正定矩阵为：$\hat{M} = P\hat{\Lambda}P^T = Q\Lambda^+Q^T$.

题3：因为 $M=P^TP$ ∴距离度量可表示为：

$$dist^2_{mah}(x_i,x_j)=(x_i-x_j)^T P^T P(x_i-x_j)=\|P(x_i-x_j)\|^2$$

LMNN可以转化为关于P的无约束优化问题：

$$L(P)=\sum_{(i,j)}\|P^T(x_i-x_j)\|^2+c\sum_{(i,j,k)}[1+\|P^T(x_i-x_j)\|^2-\|P^T(x_i-x_k)\|^2]_+$$

其中 $[\cdot]_+$ 表示取非负部分，$c=\frac{\mu}{1-\mu}$ 代表近邻保持项和排斥项之间的权重比例.

首先考虑第一项，$L_1(P)=\sum_{(i,j)}\|P^T(x_i-x_j)\|^2$. 对P求导：

$$\nabla L_1(P)=\frac{\partial}{\partial P}\|P^T(x_i-x_j)\|^2=2P(x_i-x_j)(x_i-x_j)^T.$$

再考虑排斥项：$L_2(P)=c\sum_{(i,j,k)}[1+\|P^T(x_i-x_j)\|^2-\|P^T(x_i-x_k)\|^2]_+$

该项对 $\|P^T(x_i-x_k)\|^2$ 和 $\|P^T(x_i-x_j)\|^2$ 求导. 梯度推导同上.

$$\therefore \nabla L_2(P)=2c\sum_{(i,j,k)}I[1+\|P^T(x_i-x_j)\|^2-\|P^T(x_i-x_k)\|^2>0]$$
$$\cdot P((x_i-x_j)(x_i-x_j)^T-(x_i-x_k)(x_i-x_k)^T)$$

两部分梯度相加最终梯度为：

$$\nabla L(P)=2\sum_{(i,j)}P(x_i-x_j)(x_i-x_j)^T+2c\sum_{(i,j,k)}I[1+\|P^T(x_i-x_j)\|^2-\|P^T(x_i-x_k)\|^2>0]$$
$$\cdot P((x_i-x_j)(x_i-x_j)^T-(x_i-x_k)(x_i-x_k)^T)$$

由于 $\nabla L(P)$ 中包含指示符函数I的项使得梯度不再是线性的

使得该问题关于P是非凸的.