2024 秋季高级机器学习
习题二参考答案

2025.1.17

一. (30 points) **特征选择与稀疏学习**

1. (20 points) 教材中提到，为了缓解过拟合问题，可对损失函数引入正则化项。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，以平方误差为损失函数，使用 $\sum_j |w_j|^q$ 作为正则项，可以得到带正则化的误差项

$$\sum_{i=1}^{m} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda \sum_{j=1}^{d} |w_j|^q, \tag{1}$$

其中 $\boldsymbol{w}$ 是待学习参数，$\lambda > 0$ 是正则化系数。

（1）（10 points）试说明最小化以上不带约束的问题与最小化下面带约束的问题等价。（提示：可以利用拉格朗日乘子）

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{w}} & \sum_{i=1}^{m} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 \\ \text{subject to} & \sum_{j=1}^{d} |w_j|^q \leq \eta, \end{array} \tag{2}$$

（2）（10 points）在（1）的基础上，请讨论 η 和 λ 之间的联系。（提示：可以考虑 KKT 条件）

2. (10 points) 字典学习与压缩感知都有对稀疏性的利用，请你分析两者对稀疏性利用的异同点。

解：

1. 注：本题题干中 “$\lambda > 0$” 存在争议。$\lambda > 0$ 方便了第（2）小问的讨论，但是当 η 约束对原始最小二乘问题最优解没有影响时，优化问题 (1) 中 λ 取 0 才可等价，因此严格来说 $\lambda \geq 0$ 时两个优化问题等价。（例如：考虑由 $y = x$ 产生的数据集 $D = \{(1,1), (2,2)\}$，取 $q = 1$，显然无约束的最小二乘问题最优解 $w = 1$，若此时 $\eta \geq 1$，则优化问题 (2) 中的约束项不影响最优解，若要使优化问题 (1) 中最优解也为 $w = 1$，则必须要使得正则化项对优化没有影响，即有 $\lambda = 0$。）

（1）记优化问题 (1) 的最优解为 $\boldsymbol{w}_1^*$，优化问题 (2) 的最优解为 $\boldsymbol{w}_2^*$，要证等价，即证

$$\forall \lambda \geq 0, \exists \eta \geq 0, \boldsymbol{w}_1^* = \boldsymbol{w}_2^* \tag{†}$$
$$\forall \eta \geq 0, \exists \lambda \geq 0, \boldsymbol{w}_1^* = \boldsymbol{w}_2^* \tag{††}$$

记

$$L_1 = \sum_{i=1}^{m} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda \sum_{j=1}^{d} |w_j|^q,$$

最优解 $\boldsymbol{w}_1^*$ 应满足

$$\frac{\partial L_1}{\partial \boldsymbol{w}} = -2\sum_{i=1}^{m} \boldsymbol{x}_i \left(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i\right) + \lambda q \boldsymbol{w}^{q-1} \cdot \operatorname{sign}(\boldsymbol{w}) = 0 \tag{‡}$$

其中

$$\boldsymbol{w}^{q-1} = \begin{bmatrix} |w_1|^{q-1} \\ |w_2|^{q-1} \\ \vdots \\ |w_d|^{q-1} \end{bmatrix},$$

$\operatorname{sign}(\cdot)$ 表示符号函数。为优化问题 (2) 引入拉格朗日函数

$$L_2 = L(\boldsymbol{w}, \nu) = \sum_{i=1}^{m} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \nu \left(\sum_{j=1}^{d} |w_j|^q - \eta \right)$$

最优解 $\boldsymbol{w}_2^*$ 应满足的 KKT 条件为

$$\begin{cases} \nu \geq 0 \\ \nu \left(\sum_{j=1}^{d} |w_j|^q - \eta \right) = 0 \\ \sum_{j=1}^{d} |w_j|^q - \eta \leq 0 \\ \dfrac{\partial L_2}{\partial \boldsymbol{w}} = -2\sum_{i=1}^{m} \boldsymbol{x}_i \left(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i\right) + \nu q \boldsymbol{w}^{q-1} \cdot \operatorname{sign}(\boldsymbol{w}) = 0 \end{cases} \tag{‡‡}$$

比对式 (‡) 与式 (‡‡)，对于 (†)，取 $\nu = \lambda, \eta = \sum_{j=1}^{d} \left|w_{1j}^*\right|^q$ 即可使得 $\boldsymbol{w}_1^*$ 满足优化问题 (2)；对于 (††)，取 $\lambda = \nu$ 即可使得 $\boldsymbol{w}_2^*$ 满足优化问题 (1)，因此等价。

（2）当 $\lambda > 0$ 时，正则化约束存在，因此优化问题 (1) 中最优解可表示成为关于 λ 的函数 $\boldsymbol{w}_1^*(\lambda)$，由第（1）小问中等价讨论可知，取 $\nu = \lambda > 0$，因此有 $\eta = \sum_{j=1}^{d} \left|w_{1j}^*(\lambda)\right|^q$ 。

当 $\lambda = 0$ 时，正则化约束消失，优化问题 (1) 退化成一个无约束的最小二乘问题，由第（1）小问中讨论并结合 (‡‡)，取 $\nu = \lambda = 0$，$\eta \geq \sum_{j=1}^{d} \left|w_{1j}^*(\lambda)\right|^q$ 仍可使两问题等价。

2. 相同点：都基于信号或数据在某个基底或字典下具有稀疏表示的假设，并利用了稀疏表示在提取信号本质特征方面的优势。

   不同点：字典学习旨在从给定的训练数据中学习一个字典，使得原始数据可以在该字典下获得最稀疏的表示，而压缩感知关注的是如何利用信号本身所具有的稀疏性，从部分观测样本中恢复原信号。字典学习需要优化字典，而压缩感知中通常测量矩阵和稀疏基（类似于字典）已知。

二. (40 points) **半监督学习**

生成式方法 (generative methods) 是直接基于生成式模型的方法。此类方法假设所有数据 (无论是否有标记) 都是由同一个潜在的模型“生成”的。这个假设使得我们能通过潜在模型的参数将未标记数据与学习目标联系起来，而未标记数据的标记则可看作模型的缺失参数，通常可基于 EM 算法进行极大似然估计求解。我们接下来探究高斯混合模型的参数估计过程。

给定有标记样本集 $D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_l, y_l)\}$ 和未标记样本集 $D_u = \{\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \ldots, \boldsymbol{x}_{l+u}\}$, $l \ll u, \quad l+u=m$。假设所有样本独立同分布，且都是由同一个高斯混合模型生成的。用极大似然法来估计高斯混合模型的参数 $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leq i \leq N\}$，$D_l \cup D_u$ 的对数似然是:

$$\begin{aligned} LL(D_l \cup D_u) = & \sum_{(\boldsymbol{x}_j, y_j) \in D_l} \ln \left(\sum_{i=1}^{N} \alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \cdot p(y_j \mid \Theta = i, \boldsymbol{x}_j) \right) \\ & + \sum_{\boldsymbol{x}_j \in D_u} \ln \left(\sum_{i=1}^{N} \alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right) \end{aligned} \tag{3}$$

上式由两项组成：基于有标记数据 D_l 的有监督项和基于未标记数据 D_u 的无监督项。我们将用 EM 算法求解高斯混合模型参数。

1. (10 points) **E 步更新公式**：根据当前模型参数计算未标记样本 $\boldsymbol{x}_j$ 属于各高斯混合成分的概率为：

$$\gamma_{ji} = \frac{\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{i=1}^{N} \alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \tag{4}$$

   请尝试推导上式。

2. (30 points) **M 步更新公式**：基于 γ_{ji} 更新模型参数，其中 l_i 表示第 i 类的有标记样本数目为：

$$\boldsymbol{\mu}_i = \frac{1}{\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i} \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \boldsymbol{x}_j + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{x}_j \right), \tag{5}$$

$$\begin{aligned} \boldsymbol{\Sigma}_i = & \frac{1}{\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i} \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \right. \\ & \left. + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \right), \end{aligned} \tag{6}$$

$$\alpha_i = \frac{1}{m} \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i \right). \tag{7}$$

   请根据先前给出的对数似然函数，计算推导出以上 3 个参数的更新公式。

解：

1. 未标记样本 $\boldsymbol{x}_j$ 属于第 i 个高斯混合成分的概率为

$$\begin{aligned} \gamma_{ji} &= p(\Theta = i \mid \boldsymbol{x}_j) \\ &= \frac{p(\Theta = i, \boldsymbol{x}_j)}{p(\boldsymbol{x}_j)} \\ &= \frac{p(\Theta = i) p(\boldsymbol{x}_j \mid \Theta = i)}{\sum_{k=1}^{N} p(\Theta = k) p(\boldsymbol{x}_j \mid \Theta = k)} \\ &= \frac{\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{k=1}^{N} \alpha_k \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \end{aligned}$$

2. 注：严格来说求解第三个更新公式时还应考虑混合系数的约束 $\alpha_i \geq 0$，部分同学遗漏

$$
\begin{aligned}
\frac{\partial LL(D_l \cup D_u)}{\partial \boldsymbol{\mu}_i} =& \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{\partial \ln\left(\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)\right)}{\partial p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \cdot \frac{\partial p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\partial \boldsymbol{\mu}_i} \\
&+ \frac{\partial}{\partial p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \left(\sum_{\boldsymbol{x}_j \in D_u} \ln\left(\sum_{k=1}^{N} \alpha_k \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right) \right) \cdot \frac{\partial p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\partial \boldsymbol{\mu}_i} \\
=& \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right) \\
&+ \sum_{\boldsymbol{x}_j \in D_u} \frac{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{k=1}^{N} \alpha_k \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right) \\
=& \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \left[\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\right] + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \left[\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\right] \\
=& \boldsymbol{\Sigma}_i^{-1} \left[-\boldsymbol{\mu}_i \left(l_i + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \right) + \left(\sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{x}_j + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \boldsymbol{x}_j \right) \right]
\end{aligned}
$$

令 $\frac{\partial LL(D_l \cup D_u)}{\partial \boldsymbol{\mu}_i} = \mathbf{0}$ 即得第一个更新公式。

$$
\begin{aligned}
\frac{\partial LL(D_l \cup D_u)}{\partial \boldsymbol{\Sigma}_i} =& \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{\partial \ln\left(\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)\right)}{\partial p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \cdot \frac{\partial p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\partial \boldsymbol{\Sigma}_i} \\
&+ \frac{\partial}{\partial p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \left(\sum_{\boldsymbol{x}_j \in D_u} \ln\left(\sum_{k=1}^{N} \alpha_k \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right) \right) \cdot \frac{\partial p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\partial \boldsymbol{\Sigma}_i} \\
=& \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \cdot \frac{1}{2} \left[-\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top} \boldsymbol{\Sigma}_i^{-1} \right] \\
&+ \sum_{\boldsymbol{x}_j \in D_u} \frac{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{k=1}^{N} \alpha_k \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)} \cdot \frac{1}{2} \left[-\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top} \boldsymbol{\Sigma}_i^{-1} \right] \\
=& \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{1}{2} \left[-\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top} \boldsymbol{\Sigma}_i^{-1} \right] \\
&+ \sum_{\boldsymbol{x}_j \in D_u} \frac{1}{2} \gamma_{ji} \left[-\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top} \boldsymbol{\Sigma}_i^{-1} \right] \\
=& \frac{1}{2} \boldsymbol{\Sigma}_i^{-1} \left[-\boldsymbol{\Sigma}_i \left(l_i + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \right) + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top} \right. \\
&\left. + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top} \right] \boldsymbol{\Sigma}_i^{-1}
\end{aligned}
$$

令 $\frac{\partial LL(D_l \cup D_u)}{\partial \boldsymbol{\Sigma}_i} = \mathbf{0}$ 即得第二个更新公式。

考虑到混合系数的约束 $\alpha_i \geq 0$ ，$\sum_{i=1}^{N} \alpha_i = 1$，记拉格朗日函数

$$L = L(\boldsymbol{\alpha}, \nu, \boldsymbol{\lambda}) = LL(D_l \cup D_u) + \sum_{i=1}^{N} \lambda_i(-\alpha_i) + \nu \left(\sum_{i=1}^{N} \alpha_i - 1 \right)$$

则有

$$\begin{aligned}
\frac{\partial L}{\partial \alpha_i} &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{\partial \ln \left(\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \right) \right)}{\partial \alpha_i} + \frac{\partial}{\partial \alpha_i} \left(\sum_{\boldsymbol{x}_j \in D_u} \ln \left(\sum_{k=1}^{N} \alpha_k \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \right) \right) \right) - \lambda_i + \nu \\
&= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{1}{\alpha_i} + \sum_{\boldsymbol{x}_j \in D_u} \frac{p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \right)}{\sum_{k=1}^{N} \alpha_k \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \right)} - \lambda_i + \nu \\
&= \frac{l_i}{\alpha_i} + \sum_{\boldsymbol{x}_j \in D_u} \frac{\gamma_{ji}}{\alpha_i} - \lambda_i + \nu
\end{aligned}$$

考虑到 KKT 条件，可得以下式子

$$\begin{cases}
\dfrac{\partial L}{\partial \alpha_i} = \dfrac{l_i}{\alpha_i} + \displaystyle\sum_{\boldsymbol{x}_j \in D_u} \frac{\gamma_{ji}}{\alpha_i} - \lambda_i + \nu = 0 \\
\displaystyle\sum_{i=1}^{N} \alpha_i - 1 = 0 \\
\lambda_k \alpha_k = 0 \ (1 \leq k \leq N, k \in \mathbb{Z})
\end{cases} \tag{*}$$

结合 (*) 中第一和第三个式子，有

$$l_i + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} = \alpha_i(\lambda_i - \nu) = -\alpha_i \nu \tag{**}$$

考虑所有的 N 个成分，并结合 (*) 中第二个式子有

$$-\sum_{i=1}^{N} \alpha_i \nu = -\nu = \sum_{i=1}^{N} \left(l_i + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \right) = l + \sum_{\boldsymbol{x}_j \in D_u} \sum_{i=1}^{N} \frac{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \right)}{\sum_{k=1}^{N} \alpha_k \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \right)} = l + u = m$$

将 $\nu = -m$ 代入式 (**) 即得第三个更新公式。

三. (30 points) **方法讨论**

1. (10 points) LoRA（Low-Rank Adaptation）是当前常见的模型微调技术之一，它通过在预训练模型的基础上引入低秩矩阵来调整模型参数，从而实现对模型的微调。请先对 LoRA 方法进行描述，并讨论 LoRA 有作用的原因（可以结合教材第 11 章内容进行讨论）。
2. (20 points) 给定 $D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_l, y_l)\}$ 和 $D_u = \{\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \ldots, \boldsymbol{x}_{l+u}\}$，$l \ll u$，且 $l + u = m$。我们可将其映射为一个图，数据集中每个样本对应于图中一个结点，若两个样本之间的相似度很高 (或相关性很强)，则对应的结点之间存在一条边，边的 “强度” (strength) 正比于样本之间的相似度 (或相关性)。我们先基于 $D_l \cup D_u$ 构建一个图 $G = (V, E)$，其中结点集 $V = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l, \boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_{l+u}\}$，边集 E 可表示为一个亲和矩阵 (affinity matrix)，常基于高斯函

数定义为

$$(W)_{ij} = \begin{cases} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\right), & \text{if } i \neq j; \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

其中 $i, j \in \{1, 2, \ldots, m\}$，$\sigma > 0$ 是用户指定的高斯函数带宽参数。

在上述情景中，我们可将有标记样本所对应的结点想象为染过色，而未标记样本所对应的结点尚未染色，于是，半监督学习就对应于“颜色”在图上扩散或传播的过程。该算法亦被称为标记传播方法 (label propagation)。我们接下来仅考虑二分类场景，希望从图 $G = (V, E)$ 学得一个实值函数 $f : V \to \mathbb{R}$，其对应的分类规则为：$y_i = \text{sign}(f(\boldsymbol{x}_i))$，$y_i \in \{-1, +1\}$，并定义关于 f 的“能量函数”(energy function)：

$$E(f) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (W)_{ij} \left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\right)^2 \tag{9}$$

请尝试利用上述的条件，推导出未标记节点的函数值 f_u 的预测公式。你的答案可以写为矩阵乘法的形式。

解：

1. LoRA 中，预训练的网络权重 $W_0 \in \mathbb{R}^{d \times k}$ 会固定不动，选择引入可更新参数的矩阵 $B \in \mathbb{R}^{d \times r}$，$A \in \mathbb{R}^{r \times k}$，其中 $r \ll \min(d, k)$。微调时网络更新方式为

   $$W_0 + \Delta W = W_0 + BA$$

   有效的原因：(1) LoRA 的低秩分解大幅降低待优化参数量，加速训练。(2) 预训练模型已经拥有了一定的基础能力，为了更好适应下游任务实际需要更新的参数空间可能是低维的，LoRA 的低秩分解能够选择捕捉重要的特征。与全参数训练相比，在下游任务上微调性能也不会显著降低。

2. 注：此处求解过程与教材 13.4 中内容相同，部分同学采用课外文献求解得到结果，也可以

$$\begin{aligned} E(f) &= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\mathbf{W})_{ij} \left(f\left(\boldsymbol{x}_i\right) - f\left(\boldsymbol{x}_j\right)\right)^2 \\ &= \frac{1}{2} \left(\sum_{i=1}^{m} d_i f^2\left(\boldsymbol{x}_i\right) + \sum_{j=1}^{m} d_j f^2\left(\boldsymbol{x}_j\right) - 2 \sum_{i=1}^{m} \sum_{j=1}^{m} (\mathbf{W})_{ij} f\left(\boldsymbol{x}_i\right) f\left(\boldsymbol{x}_j\right) \right) \\ &= \sum_{i=1}^{m} d_i f^2\left(\boldsymbol{x}_i\right) - \sum_{i=1}^{m} \sum_{j=1}^{m} (\mathbf{W})_{ij} f\left(\boldsymbol{x}_i\right) f\left(\boldsymbol{x}_j\right) \\ &= \boldsymbol{f}^{\mathrm{T}} (\mathbf{D} - \mathbf{W}) \boldsymbol{f} \\ &= \left(\boldsymbol{f}_l^{\mathrm{T}} \boldsymbol{f}_u^{\mathrm{T}}\right) \left(\begin{bmatrix} \mathbf{D}_{ll} & \mathbf{0}_{lu} \\ \mathbf{0}_{ul} & \mathbf{D}_{uu} \end{bmatrix} - \begin{bmatrix} \mathbf{W}_{ll} & \mathbf{W}_{lu} \\ \mathbf{W}_{ul} & \mathbf{W}_{uu} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{f}_l \\ \boldsymbol{f}_u \end{bmatrix} \\ &= \boldsymbol{f}_l^{\mathrm{T}} \left(\mathbf{D}_{ll} - \mathbf{W}_{ll}\right) \boldsymbol{f}_l - 2 \boldsymbol{f}_u^{\mathrm{T}} \mathbf{W}_{ul} \boldsymbol{f}_l + \boldsymbol{f}_u^{\mathrm{T}} \left(\mathbf{D}_{uu} - \mathbf{W}_{uu}\right) \boldsymbol{f}_u \end{aligned}$$

其中

$$\begin{cases} \boldsymbol{f} = \left(\boldsymbol{f}_l^\top \boldsymbol{f}_u^\top\right)^\top \\ \boldsymbol{f}_l = \left(f\left(\boldsymbol{x}_1\right); f\left(\boldsymbol{x}_2\right); \cdots; f\left(\boldsymbol{x}_l\right)\right) \\ \boldsymbol{f}_u = \left(f\left(\boldsymbol{x}_{l+1}\right); f\left(\boldsymbol{x}_{l+2}\right); \cdots; f\left(\boldsymbol{x}_{l+u}\right)\right) \\ \mathbf{D} = \operatorname{diag}\left(d_1, d_2, \cdots, d_{l+u}\right) \\ d_i = \sum_{j=1}^{l+u} (\mathbf{W})_{ij} \end{cases}$$

令 $\frac{\partial E(f)}{\partial \boldsymbol{f}_u} = \mathbf{0}$，可得

$$\boldsymbol{f}_u = \left(\mathbf{D}_{uu} - \mathbf{W}_{uu}\right)^{-1} \mathbf{W}_{ul} \boldsymbol{f}_l$$