2024 秋季高级机器学习
习题二

2024.11.15

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. (1)

对于带约束问题，构造拉格朗日函数：

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

其中，λ 是拉格朗日乘子，对应于约束条件。为了找到最小值，需要对 $\mathcal{L}$ 关于 $\boldsymbol{w}$ 求导，并令导数为 0：

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = -2 \sum_{i=1}^{m} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) \boldsymbol{x}_i + \lambda q \sum_{j=1}^{d} \text{sign}(w_j) |w_j|^{q-1} \mathbf{e}_j = 0$$

这里，$\mathbf{e}_j$ 是第 j 个标准基向量。解这个方程，得到：

$$\sum_{i=1}^{m} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) \boldsymbol{x}_i = \lambda q \sum_{j=1}^{d} \text{sign}(w_j) |w_j|^{q-1} \mathbf{e}_j$$

这个方程与无约束问题中的梯度为零的条件是相同的。

因此，可以选择适当的拉格朗日乘子 λ，使得在满足约束条件情况下，二者的最优化的解相同，因此最小化两种约束问题等价。

(2)

在上述拉格朗日乘子法中，λ 表示对违反约束的惩罚程度。当 λ 增加时，违反约束的代价增加，因此解会更倾向于满足约束条件。在这种情况下，λ 与 η 之间的关系可以通过 KKT 条件来理解。

KKT 条件要求在最优解处，拉格朗日乘子 λ 必须非负，并且满足：

$$\lambda \left(\sum_{j=1}^{d} |w_j|^q - \eta \right) = 0$$

这意味着如果 $\sum_{j=1}^{d} |w_j|^q < \eta$，则 $\lambda = 0$，即没有违反约束，不需要惩罚；如果 $\sum_{j=1}^{d} |w_j|^q = \eta$，则 $\lambda > 0$，表示约束是活跃的，需要通过增加 λ 来增加违反约束的代价。

因此，η 和 λ 之间的关系是：

1）λ 是无约束问题中正则化项的权重，表示对模型复杂度的惩罚力度，λ 越大，则模型参数越稀疏

2）η 是带约束问题中的约束值，控制模型的稀疏程度

3）二者呈反比关系，λ 越大，正则项惩罚越大，模型越稀疏，η 越小。

2. **相同点**：
1. 稀疏表示：两者都假设信号可以通过少量的基向量稀疏表示。
2. 稀疏正则化：都利用稀疏约束（如 ℓ_1-范数）提高模型效率或重建质量。
3. 应用稀疏性：两者都用稀疏性处理高维数据，降低复杂性。
4. 优化目标相同：都是要求解具有稀疏表示的信号或系数。

不同点：
字典学习侧重于从数据中学习一个过完备字典，并利用该字典将信号表示为稀疏线性组合，用于特征提取和数据建模。

在字典学习中，信号通过选择字典中的一组基来进行稀疏表示，编码过程通常是通过求解稀疏系数来实现的。字典本身是可学习的，能够根据数据的特点优化表示效果。

压缩感知利用信号稀疏性在采样不足时进行信号重建，解决欠定问题，更侧重于通过合适的采样和信号恢复算法，在降低采样率的情况下实现高质量的信号重构。

压缩感知是基于稀疏性和不完全采样的理论，它通过少量的随机线性测量来恢复稀疏信号。其核心是通过非线性优化方法恢复信号。更侧重于通过合适的采样和信号恢复算法，在降低采样率的情况下实现高质量的信号重构。

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:

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

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：

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

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

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

   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. **E 步更新公式推导**

   对于未标记样本 $\boldsymbol{x}_j$ 属于各混合分布的概率 γ_{ji}，有：

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

2. **M 步更新公式推导**

   M 步的目标是基于 γ_{ji} 更新模型参数。

   由于带标记样本的标签 y_j 确定，那么可以简化 $LL(D_l \cup D_u)$ 为：

$$LL(D_l \cup D_u) = \sum_{(\boldsymbol{x}_j, y_j) \in D_l} \ln \left(\alpha_{y_j} p(\boldsymbol{x}_j | \boldsymbol{\mu}_{y_j}, \Sigma_{y_j}) \right)$$

标记数据部分不需要显式引入后验概率 γ_{ji} 由于其类别已知，因此又有：

$$\sum_{(\boldsymbol{x}_j,y_j)\in D_l} \ln\left(\alpha_{y_j} p(\boldsymbol{x}_j|\boldsymbol{\mu}_{y_j},\Sigma_{y_j})\right) = \sum_{(\boldsymbol{x}_j,y_j)\in D_l} \sum_{i=1}^{N} \delta(y_j=i)\ln(\alpha_i p(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\Sigma_i))$$

因此对标记数据和未标记数据进行对齐，可以得到 M 步的优化目标函数：

$$Q(\Theta|\Theta^{(t)}) = \sum_{\boldsymbol{x}_j\in D_l\cup D_u} \sum_{i=1}^{N} \gamma_{ji}\ln(\alpha_i p(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\Sigma_i)),$$

其中对带标记数据，$\gamma_{ji}=\delta(y_j=i)$, 对无标记数据，后验概率 $\gamma_{ji}=p(z_j=i|\boldsymbol{x}_j,\Theta^{(t)})$。现在可以对参数 α_i, $\boldsymbol{\mu}_i$, Σ_i 分别求偏导来得到参数更新公式：

(1) $\boldsymbol{\mu}_i$：

$$\frac{\partial Q(\Theta|\Theta^{(t)})}{\partial \boldsymbol{\mu}_i} = \sum_{\boldsymbol{x}_j\in D_l\cup D_u} \gamma_{ji}\Sigma_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i) = 0$$

因此：

$$\boldsymbol{\mu}_i = \frac{\sum_{\boldsymbol{x}_j\in D_l\cup D_u}\gamma_{ji}\boldsymbol{x}_j}{\sum_{\boldsymbol{x}_j\in D_l\cup D_u}\gamma_{ji}}$$

对于标记数据 D_l, 权重 $\gamma_{ji}=1$, 对于无标记数据 D_u，权重由 γ_{ji} 给出，因此有：

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

(2) $\boldsymbol{\Sigma}_i$：

$$\frac{\partial Q(\Theta|\Theta^{(t)})}{\partial \boldsymbol{\Sigma}_i} = -\frac{1}{2}\sum_{\boldsymbol{x}_j\in D_l\cup D_u}\left(\gamma_{ji}\left(\boldsymbol{\Sigma}_i^{-1} - \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}_i^{-1}\right)\right)$$

因此：

$$\boldsymbol{\Sigma}_i = \frac{\sum_{\boldsymbol{x}_j\in D_l\cup D_u}\gamma_{ji}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top}{\sum_{\boldsymbol{x}_j\in D_l\cup D_u}\gamma_{ji}}$$

对标记数据与未标记数据进行对齐可得：

$$\boldsymbol{\Sigma}_i = \frac{\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top + \sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top}{\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji} + \sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}1}$$

(3) α_i：

$$\frac{\partial Q(\Theta|\Theta^{(t)})}{\partial \alpha_i} = \sum_{\boldsymbol{x}_j\in D_l\cup D_u}\frac{\gamma_{ji}}{\alpha_i} = 0$$

因此：

$$\alpha_i = \frac{\sum_{\boldsymbol{x}_j\in D_l\cup D_u}\gamma_{ji}}{\sum_{j=1}^{m}1} = \frac{\sum_{\boldsymbol{x}_j\in D_l\cup D_u}\gamma_{ji}}{m}$$

对标记数据与未标记数据进行对齐可得：

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

综上所述，我们得到了高斯混合模型的 M 步更新公式。

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 方法描述**参考 Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021)LoRA: Low-Rank Adaptation of Large Language Models 其中的 Abstract 和下面的图示可知：

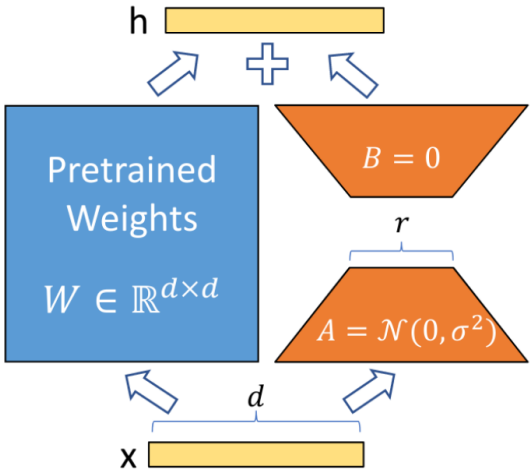

Figure 1: Our reparametrization. We only train A and B.

Figure 1: 论文中的图示

LoRA（Low-Rank Adaptation）是一种模型微调技术，其核心思想是在保持预训练模型权重不变的情况下，通过在某些特定的层（如线性层）中引入可训练的低秩矩阵来调整模型的行为。这种方法特别适用于深度学习模型，尤其是那些在大规模数据集上预训练的模型，如 Transformer 架构。

在 LoRA 中，对于一个给定的预训练权重矩阵 $W \in \mathbb{R}^{d\times d}$，模型的权重更新不是直接在 W 上进行，而是通过添加一个低秩矩阵 ΔW 来实现，其中 $\Delta W = A \cdot B$。这里，$A \in \mathbb{R}^{d\times r}$ 和 $B \in \mathbb{R}^{r\times d}$ 是两个较小的矩阵，且 r 远小于 d。这样的设计显著减少了需要学习的参数数量，从 d^2 减少到 $2rd$，从而降低了训练的计算和时间成本。

LoRA 有作用的原因

1. 参数效率：由于 $A \in \mathbb{R}^{d\times r}$, $B \in \mathbb{R}^{r\times d}$，且 $r \ll d$，因此需要微调学习的参数数量从原本的 d^2 减少到了 $2rd$，大大降低了训练开销与时间成本，去除了冗余的信息密度，通过低维矩阵来更有效地捕捉任务相关的特定方向。这使得在资源有限的情况下，可以对大型模型进行有效的微调。

2. 防止遗忘：LoRA 还能够在一定程度上防止模型在微调过程中的“遗忘”现象，保持模型在目标域之外任务上的性能。这是因为 LoRA 仅对低秩矩阵进行训练，而保持原始权重矩阵不变，从而减少了对预训练知识的破坏。

3. 稀疏学习：教材第 11 章提到，稀疏学习中提到样本具有稀疏表达形式时，对学习任务有不少好处，例如线性支持向量机之所以能在文本数据上有很好的性能，恰是由于文本数据在使用上述的字频表示后具有高度的稀疏性，使大多数问题变得线性可分。LoRA 通过低秩分解实现参数的稀疏化，使得模型在保持性能的同时，减少了参数的冗余，提高了存储和计算效率。

4. 推理时无额外开销：在推理时，对于使用 LoRA 的模型来说，可直接将原预训练模型权重与训练好的 LoRA 权重合并，因此在推理时不存在额外开销。

综上所述，LoRA 通过低秩适应技术，有效地减少了模型微调时的参数量，同时保持了模型性能，并减少了推理时的额外开销，这使得 LoRA 成为一种有效的模型微调技术。

2. 对于关于 f 的能量函数，其定义上是对于标记不一致样本间的惩罚项，因此最小化能量函数就可以使得 f 的标记更精确。

   由拉普拉斯矩阵变换：

   $$L = D - W, \quad D_{ii} = \sum_j W_{ij},$$

   因此有：

   $$E(f) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} W_{ij}(f(x_i) - f(x_j))^2 = \frac{1}{2} f^\top L f, \quad f = [f_1, f_2, \ldots, f_m]^\top$$

   令

   $$f = \begin{bmatrix} f_l \\ f_u \end{bmatrix}, \quad L = \begin{bmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{bmatrix}$$

   因此

   $$E(f) = \frac{1}{2} f^\top L f = \frac{1}{2}\left(f_l^\top L_{ll} f_l + 2 f_l^\top L_{lu} f_u + f_u^\top L_{uu} f_u\right)$$

   又因为对于标记已知的部分（即 $f_l^\top L_{ll} f_l$）在优化过程中不会变化，为常量，因此只需要考虑与 u（无标记）相关部分即可。

   因此当最小化 $E(f)$ 时，可对 f_u 求偏导为 0，有：

   $$\frac{\partial E(f)}{\partial f_u} = \frac{1}{2}\left(2 L_{uu} f_u + 2 f_l^\top L_{lu}\right) = 0$$

   可得：

   $$f_u = -L_{uu}^{-1} f_l^\top L_{lu}$$