2024 秋季高级机器学习
习题二

221300079 王俊童

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）对于这个问题，我们用 lagrange 乘子法对带约束优化进行求导：
原式子可以写作：

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

$$\frac{\partial L}{\partial \boldsymbol{w}} = -2\sum_{i=1}^{m}(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)\boldsymbol{x}_i + \mu q \sum_{j=1}^{d} |w_j|^{q-1} sign(w_j)\boldsymbol{e_j} = 0$$

其中 e_j 是第 j 个 w 的标准基向量。
同理我们也对于无约束优化做求导:

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

根据 kkt 条件可以得到，若 $\mu = \lambda$ 的时候，两个算式对于 w 的驻点是一样的，而且若解 λ 和 μ 可以得到他们是一样的。所以可以知道经过 lagrange 乘子法的检验，他们确实是等价问题。

（2）对于这个问题，我们可以得到 η 和 λ 之间的联系基本有以下：
若 λ 变大，那么 w 就会稀疏，那第二项就会更小，因此，λ 决定了正则化项对损失函数的影响，η 则控制了约束条件的大小。通过 kkt 条件来看：有互补松弛条件

$$\mu(\sum_{min}^{max}|w_j|^q-\eta)=0$$

如果 $\sum_{j=1}^{d}|w_j|^q=\eta$，满足驻点条件，此时 $\lambda=\mu$
如果 $\sum_{j=1}^{d}|w_j|^q<\eta$，则对应的 $\mu=0$。正则化项在此刻没啥用。

2. 相同点：
(1) 对于这两个问题，他们都追求稀疏性：字典学习是:$X=D\alpha$, 压缩感知是:$X=\Psi\theta$, 这两个式子里面 θ,α 都是稀疏的。
(2) 它们都通过优化问题来利用稀疏性，字典学习通过稀疏表示优化字典，而压缩感知通过稀疏恢复优化信号。都是通过减少不必要的数据来保留重要数据，从而减少数据量。
不同点:
(1) 侧重点不一样：
字典学习更侧重于学习一个合适的字典，这个字典能够很好地对数据进行稀疏表示。而压缩感知主要侧重于从少量的测量数据中恢复原始的稀疏信号。
(2) 方法不一样:
字典学习通常通过迭代算法来更新字典 D 和稀疏系数矩阵 α 如 K - SVD 算法。而压缩感知则重点在于求解欠定方程组，用 L0 范数最小化等方法来解决。

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 $\{(\alpha_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\mid 1\le i\le N\}$，$D_l\cup D_u$ 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：

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

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

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

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. 证明如下，首先明确问题为: 设 z_j 是隐变量，可以属于 $1, 2, .., n$, 意为分为 n 个类别。

$$\gamma_{ji} = p(z_j = i | \boldsymbol{x_j}) = \frac{p(z_j = i, \boldsymbol{x_j})}{p(\boldsymbol{x_j})}$$

对于这个 $p(\boldsymbol{x_j})$, 可以经过以下推导:

$$p(\boldsymbol{x_j}) = \sum_{i=1}^{N} p(z_j = i, \boldsymbol{x_j}) = \sum_{i=1}^{N} p(z_j = i) p(\boldsymbol{x_j} | z_j = i)$$

此处 $\alpha_i = p(z_j = i)$ 表示样本来自第 i 个高斯混合分布的先验概率。
$p(\boldsymbol{x_j} | z_j = i)$ 表示第 $\boldsymbol{x_j}$ 个样本在第 i 类隐变量上的分布，对于这个式子，可以写作:$p(\boldsymbol{x_j} | z_j = i) = p(\boldsymbol{x_j} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ 所以

$$p(\boldsymbol{x_j}) = \sum_{i=1}^{N} \alpha_i p(\boldsymbol{x_j} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

那么同理可得：分子的组成其实是分母的其中一部分，也就是 $\alpha_i p(\boldsymbol{x_j} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ 所以综上所述:

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

2. 对于整个式子我们可以稍微化简一下，首先是对于 D_l, 说明在已经知道标签和特征的情况下，我们有下面式子成立，说明对于特征 yj，只有 xj 与之匹配:

$$\sum_{(\boldsymbol{x}_j, y_j) \in D_l} \ln \left(\sum_{i=1}^{N} \alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \cdot p(y_j \mid \Theta = i, \boldsymbol{x}_j) \right) = \sum_{(\boldsymbol{x}_j, y_j) \in D_l} \ln(\alpha_{yj} \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_{yj}, \boldsymbol{\Sigma}_{yj}))$$

同理对于 D_u

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

(2.1) 首先证明 $\boldsymbol{\mu}_i$ 这个公式：可由解下面这个公式得到：

$$\frac{\partial LL(D_l \cup D_u)}{\partial \boldsymbol{\mu}_i} = 0$$

对于这个式子我们分开求导:

$$\frac{\partial LL(D_l)}{\partial \mu_i} = \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{1}{p(x_j|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})} \cdot \frac{p(x_j|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})}{\mu_i} = \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \Sigma_i^{-1}(x_j - \mu_i)$$

$$\frac{\partial LL(D_u)}{\partial \mu_i} = \sum_{\boldsymbol{x} \in D_u} \gamma_{ij} \cdot \Sigma_i^{-1}(x_j - \mu_i)$$

所以综上，将二者相加等于 0，即可得到:

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

移项可得:

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

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.

2.