3 讲座 - "FL 设计原则

第 2 讲回顾了使用数字数组存储数据点（其特征和标签）和模型参数的 ML 方法。我们还讨论了作为实用 ML 系统主要设计原则的 ERM（及其正则化）。本讲座将 ML 的基本概念从单数据集单模型设置扩展到涉及数据和模型分布式集合的 FL 应用。

第 3.2 节介绍了用于存储本地数据集集合和本地模型相应参数的经验图。第 3.3 节介绍了 FL 系统的主要设计原则。该原则是使用特定损失函数和模型的 ERM 的特例。具体而言，我们通过 "经验图 "节点来表示数据和模型。我们根据各个局部模型参数在经验图边缘上的变化，对其造成的损失进行惩罚。这种惩罚是正则化的一个实例，并将各个局部模型的训练结合起来。

3.1 学习目标

听完讲座后，您应该

- 熟悉经验图的概念、
- 了解连通性与拉普拉斯矩阵频谱的关系、

- 知道一些本地模型变化的测量方法、
- 熟悉作为配方 FL 的 GTVMin。

3.2 经验图谱及其拉普拉卡方

考虑一组本地数据集 $D^{(1)}, \ldots .^{(2)}$我们的目标是为每个本地数据集 $D^{(i)}$ 训练一个个性化模型 $H^{(i)}$ ，$i = 1, \ldots , n$。我们用一个经验图来表示这样一个本地数据集和（个人）本地模型的集合，以及它们之间的关系。图 3.1 描述了经验图的一个例子。

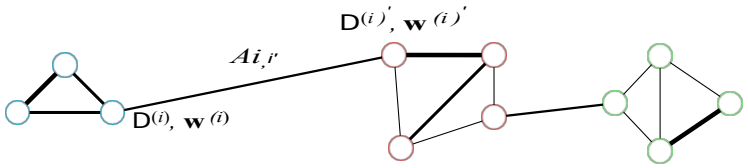


图 3.1：经验图示例，其节点 $i \in V$ 包含本地数据集 $D^{(i)}$ 和由本地模型参数 $\mathbf{w}^{(i)}$ 参数化的本地模型。

经验图 G 是一个无向加权图 $G = (V, E)$，节点 $V := \{1, ., n\}$.经验图 G 的每个节点 $i \in V$ 都带有一个本地数据集

$$D^{(i)} := \mathbf{x}^{(i,1)}, y^{(i,1)}, \ldots, \mathbf{x}^{(i,m_i)}, y^{(i,m_i)})\}. \tag{3.1}$$

这里，$\mathbf{x}^{(i,r)}$ 和 $y^{(i,r)}$ 分别表示本地数据集 $D^{(i)}$ 中第 r 个数据点的特征和标签。请注意，本地数据集的大小 m_i 可能因节点 $i \in V$ 而异。

将特征向量 $\mathbf{x}^{(i,r)}$ 和标签 $y^{(i,r)}$ 分别收集成特征矩阵 $\mathbf{X}^{(i)}$ 和标签向量 $\mathbf{y}^{(i)}$ 是很方便的、

$$\mathbf{X}^{(i)} := \mathrm{x}^{(i,1)}, \ldots, \mathbf{x}^{(i,m_i)\ T}, \text{而 } \mathbf{y} := \mathrm{y}^{(1)}, \ldots, y_{\circ}^{(m_i)\ T} \quad (3.2)$$

然后，本地数据集 $D^{(i)}$ 可以用特征矩阵 $\mathbf{X}^{(i)} \in \mathbb{R}^{m_i \times d}$ 和向量 $\mathbf{y}^{(i)} \in \mathbb{R}^{m_i}$ 来紧凑表示。

除了本地数据集 $D^{(i)}$ 之外，每个节点 $i \in G$ 还携带一个本地模型 $H^{(i)}$. Our focus is on local models that can be parametrized by local model parameters $\mathbf{w}^{(i)} \in \mathbb{R}^d$, for $i = 1, \ldots, n$. The usefulness of a specific choice for the local model parameter $\mathbf{w}^{(i)}$ is measured by a local loss function $L_i\ \mathbf{w}^{(i)}$, for $i = 1, \ldots, n$. In principle, we can use different local loss functions $L_{i'}(\cdot) \neq L_{i'}(\cdot)$ at different nodes $i, i' \in V$.

经验图还包含一对（不同）节点 $i, i' \in V$ 之间的无向边 $\{i, i'\} \in E$。我们使用无向边 $\{i, i'\} \in E$ 来耦合相应局部模型 $H^{(i)}$ ， $H^{(i')}$ 的训练。这种耦合的强度由正边权重 $A_{i,i'} > 0$ 决定。耦合是通过惩罚本地模型参数 $\mathbf{w}^{(i)}$ 和 $\mathbf{w}^{(i')}$ 之间的差异来实现的（见第 3.3 节）。

There are different ways to measure the discrepancy between the local model parameters $\mathbf{w}^{(i)}, \mathbf{w}^{(i')}$ at connected nodes $\{i, i'\} \in E$. For example, we can use some cost or penalty function $\phi\ \mathbf{w}^{(i)} - \mathbf{w}^{(i')}$ that satisfies basic requirements such as being a (semi-)norm [32].

根据第 3.3 节中介绍的设计原则，惩罚值 $\phi(\cdot)$ 的选择对 FL 方法的计算和统计特性有着至关重要的影响。除非另有说明，我们使用 $\phi(\cdot) := \|\cdot\|_2^2$

也就是说，我们用欧几里得距离的平方 $\mathbf{w}^{(i)} - \mathbf{w}^{(i')}{}^2$ 来衡量一条边 $\{i,$

$i'\}\in E$ 上局部模型参数之间的差异。将经验值中所有边的差异（按边的权重加权）相加

图中得出了本地模型参数的总变化量

$$\sum_{\{i,i'\}\in E} A_{i,i'} \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(i')} \right\|_2^2. \tag{3.3}$$

经验图 G 的连通性可以用节点 $i \in V$ 周围的加权节点度来表征

$$d^{(i)} := \sum_{i' \in N^{(i)}} A_{i,i'}。 \tag{3.4}$$

这里，我们使用了节点 $i \in V$ 的邻域 $N^{(i)} := \{i' \in V : \{i,i'\} \in E\}$。$G$ 的连接性的全局特征是最大加权节点度

$$d_{\text{最大}} := \max_{i \in V} d^{(i)} \overset{(3.4)}{=} \text{最大值}_{i \in V} \sum_{i' \in N^{(i)}} A_{i,i'}。 \tag{3.5}$$

除了节点度（最大值），我们还可以分析连接性

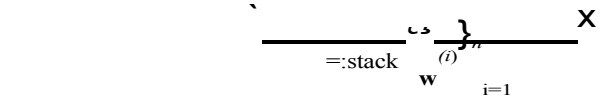
通过 G 的拉普拉斯矩阵 $\mathbf{L}$ 的特征值和特征向量$(G) \in R^{n \times n}$。拉普拉斯矩阵的元素定义如下（见图 3.2）

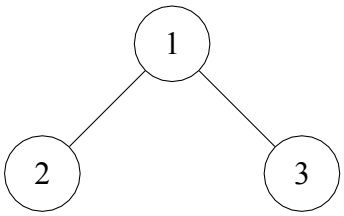
$$L_{i,i'}^{(G)} := \begin{cases} -A_{i,i'} & \text{for } i \neq i', \{i,i'\} \in E \\ \sum_{i'' \neq i} A_{i,i''} & \text{对于 } i = i' \\ 0 & \text{否则} \end{cases} \tag{3.6}$$

拉普拉斯矩阵是对称的psd矩阵，这源于同一性

$$\mathbf{w}^T (\mathbf{L}^{(G)} \otimes \mathbf{I}) \mathbf{w} = \sum_{\{i,i'\} \in E} A_{i,i'} \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(i')} \right\|_2^2$$

$$\text{对于任意 } \mathbf{w} := \left(\mathbf{w}^{(1)^T}, \ldots, \mathbf{w}^{(n)^T} \right)^T \tag{3.8}$$





$$\mathbf{L}^{(G)} = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \tag{3.7}$$

图 3.2： 左图：有三个节点的经验图 G 示例 $i = 1, 2, 3$.右图G 的拉普拉斯矩阵 $\mathbf{L}^{(G)} \in \mathbb{R}^{3\times 3}$ 。

作为一个 psd 矩阵，$\mathbf{L}^{(G)}$ 具有一个 EVD

$$\mathbf{L}^{(G)} = \sum_{j=1}^{n} \lambda_j \mathbf{u}^{(j)} \mathbf{u}^{(j)T}, \tag{3.9}$$

特征值越来越有序

$$0 = \lambda_1 \mathbf{L}^{(G)} \leq \lambda_2 \mathbf{L}^{(G)} \leq \ldots \leq \lambda_n \mathbf{L}^{(G)} . \tag{3.10}$$

According to (3.8), we can measure the total variation of local model parameters by stacking them into a single vector $\mathbf{w} \in \mathbb{R}^{nd}$ and computing the quadratic form $\mathbf{w}^T \mathbf{L}^{(G)} \otimes \mathbf{I}_{d\times d} \mathbf{w}$.

(3.8) 的一个直接结果是，任何一组相同的局部模型参数

$$\mathbf{w} = stack\{c\} = \mathbf{c}^T, \ldots, \mathbf{c}^T{}^T, \text{有一些 } \mathbf{c} \in \mathbb{R}^d \tag{3.11}$$

是 $\mathbf{L}^{(G)} \otimes \mathbf{I}$ 的特征向量，相应的特征值为 $\lambda_1 = 0$（见 (3.10)）。因此，

任何经验图的拉普拉斯矩阵都是奇异的（不可逆转的）。

拉普拉斯矩阵的第二个特征值 λ_2 提供了有关 G 连接结构的大量信

息。[4]

[4]谱图理论的很多内容都致力于分析不同图构造的 λ_2 [33, 34]。

- 考虑 $\lambda_2 = 0$ 的情况：在这里，我们可以找到（除 (3.11) 以外）另一个特征向量

$$\mathbf{w}\tilde{} = \text{stack } \mathrm{w}^{(i)}\}_{i=1}^{n}, \text{ 对于某个 } i, \mathbf{w}^{(i)} \neq \mathbf{w}^{(i')}, i' \in V、 \tag{3.12}$$

的 $\mathbf{L}^{(G)} \otimes \mathbf{I}$ 的特征值等于 0。在这种情况下，图 G 是不连通的，也就是说，我们可以找到节点的两个子集（分量），它们之间没有任何边。然后，我们可以将相同（非零）的向量 $\mathbf{c} \in \mathrm{R}^d \setminus \{0\}$ 分配给属于同一连通组件 C 的所有节点 i，而将零向量 $\mathbf{0}$ 分配给其他节点 $i \in V \setminus C$，从而得到特征向量。

- 另一方面，如果 $\lambda_2 > 0$，那么 G 是连通的。此外，λ_2 的值越大，G 中节点之间的连通性就越强。接下来，我们将展示如何通过特性 (3.8) 使这一含糊的说法更加精确。

RHS 上的总变化 (3.8) 是连接性的度量：如果局部模型参数 $\mathbf{w}^{(i)}$ 不同，则增加一条边会增加总变化（3.8）。此外，我们还可以将总变化下限定为 [35]

$$\sum_{\{i,\, i'\} \in E} A_{i,i'} \cdot \cdot \mathbf{w}^{(i)} - \mathbf{w}^{(i')} \cdot \cdot_2^2 \geq \lambda_2 \sum_{i=1}^{n} \cdot \cdot \mathbf{w}^{(i)} - \mathbf{avg}\{\mathbf{w}^{(i)}\} \cdot \cdot_2^2. \tag{3.13}$$

这里，$\mathbf{avg}\{w^{(i)}\} := (1/n) \sum_{i=1}^{n} w^{(i)}$ 是所有局部模型参数的平均值。

界值 (3.13) 源自 (3.8) 和库朗特-费舍尔-韦尔矩阵 $\mathbf{L}^{(G)} \otimes \mathbf{I}$ 的特征值的最小-最大特性[36, Thm.

(3.13) RHS 上的量 $\sum_{i=1}^{n} \| \mathbf{w}^{(i)} - \mathbf{avg}\{\mathbf{w}\}^{(i)n} \|_2^2$ 有一个"-"。

有趣的几何解释：它是原点的欧几里得平方准则。

叠加局部模型参数 $\mathbf{w} := \left(\mathbf{w}^{(1)^T}, \ldots, \mathbf{w}^{(n)^T}\right)^T$

子空间的正交补集上

$$S := \mathbf{1} \otimes \mathbf{a} : \mathbf{a} \in \mathbb{R}^d = \left(\mathbf{a}^T, \ldots, \mathbf{a}^T\right)^T, \text{对于某个 } \mathbf{a} \in \mathbb{R}^d \subseteq \mathbb{R}^{dn}。 \tag{3.14}$$

事实上，$\mathbf{w} \in \mathbb{R}^{nd}$ 在 S 上的投影 $\mathbf{P}_S \mathbf{w}$ 明确给定为

$$\mathbf{P}_S \mathbf{w} = \left(\mathbf{a}^T, \ldots, \mathbf{a}^T\right)^T, \mathbf{a} = \mathbf{avg}\{\mathbf{w}\}_{i=1}^{(i)n}。 \tag{3.15}$$

而在正交补集 $S^{\perp}$ 上的投影则明确为

$$\mathbf{P}_{S^{\perp}} = \mathbf{w} - \mathbf{P}_S \mathbf{w} = \text{堆栈}\left(\mathbf{w}^{(i)} - \mathbf{avg}\{\mathbf{w}\}_{i=1}^{n}\right)_{i=1}^{(i)n}。 \tag{3.16}$$

将给定的经验图 G 替换为等价的全连接经验图 G'（见图 3.3）可能会很方便。图 G' 的每一对不同节点 i 之间都有一条边，i、i'

$$E' = \{i, i'\}, \text{有一些 } i, i' \in V, i \neq i'。$$

对于任意边 $\{i, i'\} \in E$，边权重选为 $A'_{i,i'} = A_{i,i'}$；如果原始经验图 G 不包含节点 i, i' 之间的边，则边权重选为 $A'_{i,i'} = 0$。请注意，经验图的无向边 E 编码了本地数据集之间相似性的对称概念：如果节点 i 处的本地数据集 $D^{(i)}$ 与节点 i 处的本地数据集 $D^{(i')}$ 相似'，即 $\{i, i'\} \in E$，那么也是本地数据集 $D^{(i')}$ 与本地数据集 $D^{(i)}$ 相似。

3.3 广义总变异最小化

考虑某个经验图 G，其节点 $i \in V$ 包含本地数据集

$D^{(i)}$ 和由向量 $\mathbf{w}^{(i)} \in \mathrm{R}^{d}$ 参数化的局部模型。我们学习这些

图 3.3：左图：由 $n = 4$ 个节点组成的经验图 G。右图等价全连接经验图 G'，节点相同，边权重不为零，对于 $\{i, i'\} \in E$，$A'_{i,i'} = A_{i,i'}$；对于 $\{i, i'\} \notin E$，$A'_{i,i'} = 0$。

通过最大限度地减少局部模型参数的局部损失，同时确保其总体变化较小。要求学习到的局部模型参数有较小的总体变化，可以使它们在连接良好的节点（"集群"）上（近似）保持恒定。

为了在局部损耗和总体变化之间取得最佳平衡，我们解决了广义总体变化（GTV）最小化问题、

$$\{\mathbf{w}^{(i)}\}_{i=1}^{n} \in \operatorname*{argmin}_{\{\mathbf{w}^{(i)}\}_{i=1}^{n}} \sum_{i \in \mathcal{V}} L_i\left(\mathbf{w}^{(i)}\right) + \alpha \sum_{\{i, i'\} \in \mathcal{E}} A_{i,i'} \left\|\mathbf{w}^{(i)} - \mathbf{w}^{(i')}\right\|_2^2 \quad \text{(GTVMin)}. \tag{3.17}$$

请注意，GTVMin 是 RERM 的一个实例：正则因子是局部模型参数在经验图的加权边 $A_{i,i'}$ 上的总变化。显然，经验图是基于 GTVMin

方法的重要设计选择。这一选择可由基于 GTVMin 的 FL 系统的计算方面和统计方面来指导。

有些应用领域允许利用领域专业知识来猜测经验图的有用选择。如果本地数据集是在不同的地理位置生成的，我们可以使用基于近邻图的经验图。

数据生成器（如 FMI 气象站）之间的大地距离。第 7 讲还将讨论经验图学习方法，这种方法以数据驱动的方式确定边权重 $A_{i,i'}$，即直接从本地数据集 $D^{(i)}$ ， $D^{(i')}$ 中确定边权重。

Let us now consider the special case of GTVMin with local models being a linear model. For each node $i \in V$ of the empirical graph, we want to learn

线性假设 $h^{(i)}(\mathbf{x}) := \mathbf{w}^{(i)T}\mathbf{x}$ 的参数 $\mathbf{w}^{(i)}$ 。

通过平均平方误差损失计算参数质量

$$L_i\left(\mathbf{w}^{(i)}\right) := (1/m_i)\sum_{r=1}^{m_i}\left(y^{(i,r)} - \mathbf{w}^{(i)T}\mathbf{x}^{(i,r)}\right)^2 \overset{(3.2)}{=} (1/m_i)\left\|\mathbf{y}^{(i)} - \mathbf{X}^{(i)}\mathbf{w}^{(i)}\right\|_2^2 . \quad (3.18)$$

将 (3.18) 插入 (3.17)，就得到了以下用于训练局部线性模型的

GTVMin 实例、

$$\{\hat{\mathbf{w}}^{(i)}\}_{i\in V} \in \text{最小值}_{\{\mathbf{w}^{(i)}\}_{i\in V}} \sum_{i\in V}(1/m_i)\left\|\mathbf{y}^{(i)} - \mathbf{X}^{(i)}\mathbf{w}^{(i)}\right\|_2^2 + \alpha \sum_{\{i,i'\}\in E} A_{i,i'}\left\|\mathbf{w}^{(i)} - \mathbf{w}^{(i')}\right\|_2^2 . \quad (3.19)$$

根据恒等式 (3.8) 可以使用拉普拉斯矩阵 $\mathbf{L}^{(G)}$ 将 (3.19) 重写为

$$\hat{\mathbf{w}}_{(i)} \in \underset{\mathbf{w}=\text{stack}\{\mathbf{w}^{(i)}\}_{i\in V}}{\operatorname{argmin}} \sum_{i}(1/m_i)\left\|\mathbf{y}^{(i)} - \mathbf{X}^{(i)(i)}\mathbf{w}\right\|_2^2 + \alpha \mathbf{w}^T \mathbf{L}^{(G)} \otimes \mathbf{I}_d \mathbf{w}. \quad (3.20)$$

让我们将（3.20）中的目标函数重写为

$$\mathbf{w}^{T}\begin{bmatrix}\mathbf{Q}^{(1)} & 0 & \cdots & 0\\ 0 & \mathbf{Q}^{(2)} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \mathbf{Q}^{(n)}\end{bmatrix}+\alpha\mathbf{L}^{(G)}\otimes\mathbf{I}\mathbf{w}+\left[\alpha^{(1)^{T}},\ldots,\alpha^{(n)^{T}}\right]\mathbf{w} \tag{3.21}$$

与 $\mathbf{Q}^{(i)}=(1/m)_i\mathrm{X}\,\mathrm{X}^{(i)^{T}(i)}$，和 $\mathbf{q}^{(i)}:=(-2/m)_i\mathrm{X}\,\mathrm{y}$。$^{(i)^{T}(i)}$

因此，与线性回归（2.7）和脊回归（2.27）一样，GTVMin（3.20）（对于局部线性模型 $H^{(i)}$）也是最小化一个凸二次函数、

$$\mathrm{w}^{\wedge(i)}\}_{i=1}^{n}\in\underset{\mathrm{w}=\mathrm{stack}\ \{\mathbf{w}^{(i)}\}_{i=1}^{n}}{\operatorname{argmin}}\ \mathbf{w}^{T}\mathbf{Q}\mathbf{w}+\mathbf{q}^{T}\mathbf{w}。 \tag{3.22}$$

在这里，我们使用了 psd 矩阵

$$\mathbf{Q}:=\begin{bmatrix}\mathbf{Q}^{(1)} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}^{(2)} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}^{(n)}\end{bmatrix}+\alpha\mathbf{L}^{(G)}\otimes I，\ \mathbf{Q}^{(i)}:=(1/m)_i\ \mathbf{X}^{(i)^{T}}\mathbf{X}^{(i)} \tag{3.23}$$

和向量

$$\mathbf{q}:=\mathrm{q}^{(1)^{T}},..,q^{(n)^{T}T},\ \mathbf{q}^{(i)}:=(-2/m)X_i{}^{(i)^{T}(i)}\ \mathrm{y}。 \tag{3.24}$$

3.3.1 GTVMin 的计算方面

第 5 讲将应用优化方法求解 GTVMin，从而产 生 实用的 FL 算法。 不同的 GTVMin 实例对不同的

类优化方法。例如，使用可变损失函数可以应用基于梯度的方法（见第 4 讲）来求解 GTVMin。

另一类重要的损失函数是我们可以高效计算邻近算子的函数

接近点 $\mathbf{prox}_{L,\rho}(\mathbf{w}) := \operatorname{argmin}_{w' \in \mathbb{R}^d} L(\mathbf{w}') + (\rho/2)\|w' - \mathbf{w}\|_2^2$ 对于某个 $\rho > 0$。

一些学者将可以轻松计算 $\mathbf{prox}_{L,\rho}(\mathbf{w})$ 的函数 L 称为*简单*或近似函数 [37]。使用近似损失函数的 GTVMin 可以通过近似算法非常高效地求解 [38]。

除了影响优化方法的选择，GTVMin 的基础设计选择还决定了特定优化方法所需的计算量。例如，使用边缘相对较少的经验图（"稀疏图"）通常会降低计算复杂度。事实上，第 5 讲讨论了基于 GTVMin 的算法，其计算量与经验图中的边数成正比。

现在让我们来考虑用 GTVMin (3.19) 训练局部线性模型的计算问题。如上所述，这个实例等同于求解 (3.22)。(3.22) 的任何解 $\mathbf{w}$（以及 (3.19)）都具有零梯度条件的特征

$$Q\hat{w} = -(1/2)\mathbf{q}, \tag{3.25}$$

$\mathbf{Q}$, $\mathbf{q}$ 的定义见 (3.23) 和 (3.24)。如果 (3.25) 中的矩阵 $\mathbf{Q}$ 是可逆的，则

(3.25) 以及 GTVMin 实例 (3.19) 的解是唯一的，其值为 $\mathbf{w} = (-1/2)\mathbf{Q}^{-1} q_{\circ}$

矩阵 $\mathbf{Q}$ 的大小（见 (3.23)）与经验图 G 中的节点数成正比，在互联网规模的应用中，节点数可能达到数百万（甚至数十亿）。对于这种大型系统，我们通常无法使用直接矩阵反演方法（如基于高斯消元法的方法）来计算 $\mathbf{Q}$。[15] 相反，我们通常需要采用迭代法 [39，40]。

我们将在第 4 讲中讨论基于梯度的迭代法。从局部模型参数的初始选择 $\widehat{\mathbf{w}}^{(0)} = \widehat{\mathbf{w}}^{(1)}, \ldots, \widehat{\mathbf{w}}^{(n)}$，这些方法重复(梯度步骤的变体）、

$$\widehat{\mathbf{w}}_{k+1} := \widehat{\mathbf{w}}_k - \eta \left(2\mathbf{Q}\widehat{\mathbf{w}}_k + \mathbf{q} \right) \quad \text{for } k = 0, 1, \ldots$$

梯度步骤的结果是更新本地模型参数 $\mathbf{w}^{(i)}$，其中 $\widehat{\mathbf{w}}$ 我们叠成

$$\widehat{\mathbf{w}}_{k+1} := \left(\left(\widehat{\mathbf{w}}^{(1)} \right)^T, \ldots, \left(\widehat{\mathbf{w}}^{(n)} \right)^T \right)^T.$$

我们根据某种停止标准（见第 4 讲），重复足够次数的梯度步骤。

3.3.2 GTVMin 的统计方面

对于使用 GTVMin（3.17）的解作为本地模型参数的 FL 系统，我们能说些什么呢？为了回答这个问题，我们在第 2 讲的 ERM 统计分析中使~~用了局部数据集概率模型。~~ 特别是，我们使用了一种 i.i.d. 假设的变体

：每个本地[5] 您认为需要多少次算术运算（加法、乘法）？

反转任意矩阵 $Q \in R^{?^{d \times d}}$

数据集 $D^{(i)}$，由特征和标签均为 i.i.d. RV 实现的数据点组成

$$\mathbf{y}^{(i)} = \mathbf{x}^{(i,1)}, \ldots, \mathbf{x}^{(i,m_i)} \text{ 本地特征矩阵 } \mathbf{X}^{(i)} \mathbf{w}^{(i)} + \boldsymbol{\varepsilon}^{(i)} \text{，其中 } \mathbf{x}^{(i,r)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}), \boldsymbol{\varepsilon}^{(i)} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \tag{3.26}$$

与概率模型（2.13）（我们用于分析 ERM）不同，概率模型（3.26）允许对 $i \in V$ 采用不同的节点特定参数 $\mathbf{w}^{(i)}$。特别是，汇集所有本地数据集得到的整个数据集并不符合 i.i.d. 假设。

在下文中，我们将重点关注 GTVMin 实例 (3.19)，以学习每个节点 $i \in V$ 的局部线性模型参数 $\mathbf{w}^{(i)}$。对于经验图的合理选择，相连节点的参数 $\mathbf{w}^{(i)}, \mathbf{w}^{(i')}$ $\{i, i'\} \in E$ 应该是相似的。但是，我们不能根据参数 $\mathbf{w}^{(i)}$ 来选择边缘权重，因为它们是未知的。我们只能使用本地数据集中数据点的特征和标签对 $\mathbf{w}^{(i)}$ 的（噪声）估计（见第 7 讲）。

考虑对符合概率模型（3.26）的本地数据集的经验图（及其边权重 $A_{i,i'}$）的给定（设计）选择，其真实基础参数为 $\mathbf{w}^{(i)}$。为便于阐述，我们假设

$$\mathbf{w}^{(i)} = \mathbf{c}，对于某些\ c \in\ R^{d}\ 和所有\ i \in\ V。 \qquad (3.27)$$

为了研究 (3.19) 的解 $\mathbf{w}^{(i)}$ 与真正的 $\hat{}$

基础参数 $\mathbf{w}^{(i)}$，我们将其分解为

$$\mathbf{w}\hat{}^{(i)} =: \tilde{\mathbf{w}}^{(i)} + \hat{}\mathbf{c},\ \hat{}\mathbf{c} := (1/n)\sum_{i'=1}^{n} \mathbf{w}\hat{}^{(i')}。 \qquad (3.28)$$

分量 **c** 在所有节点 $i \in V$ 上都是相同的，并作为正交

$\mathbf{w}^{\wedge} = \{\hat{\mathbf{w}}^{(i)}\}_{i=1}^{n}$ 在子空间 (3.14) 上的投影。分量 $\tilde{\mathbf{w}}^{(i)} := \hat{\mathbf{w}}^{(i)} - (1/n)\sum_{i=1}^{n} \hat{\mathbf{w}}^{(i)}$ 包含每个节点 i 的偏差、$\hat{\mathbf{w}}^{(i)}$ 与所有节点的平均值之间的偏差。实际上，所有节点的偏差 $\tilde{\mathbf{w}}^{(i)}$ 的平均值就是零向量、

$(1/n)\sum_{i=1}^{n} \tilde{\mathbf{w}}^{(i)} = \mathbf{0}$.

分解（3.28）需要对误差 $\mathbf{w}^{(i)} - \mathbf{w}^{(i)}$ 进行类似的（正交）分解。事实上，

对于相同的真实基本模型参数 (3.27)（这使得 **w** 成为子空间 (3.14) 的一个元素），我们有

$$\sum_{i=1}^{n} \|\hat{\mathbf{w}}^{(i)} - \mathbf{w}^{(i)}\|_2^2 \overset{(3.27),(3.28)}{=} \sum_{i=1}^{n} \|\mathbf{c} - \hat{\mathbf{c}}\|_2^2 + \sum_{i=1}^{n} \|\tilde{\mathbf{w}}^{(i)} - \tilde{\mathbf{w}}^{(i)}\|_2^2 \tag{3.29}$$

下面的命题提供了 (3.29) 中第二个误差分量的上限。

命题 3.1. *考虑一个连通的经验图，即 $\lambda_2 > 0$（见 (3.10)），以及本地数据集 (3.26) 的 GTVMin (3.19) 的解 (3.28)。如果 (3.26) 中的真实局部模型参数相同（见 (3.27)），我们可以学习参数的偏差上限 $\tilde{\mathbf{w}}^{(i)} := \hat{\mathbf{w}}^{(i)} - (1/n)\sum_{i=1}^{n} \hat{\mathbf{w}}^{(i)}$ $\hat{\mathbf{w}}^{(i)}$，因为*

$$\sum_{i=1}^{n} \|\tilde{\mathbf{w}}^{(i)}\|_2^2 \leq \frac{1}{\lambda_2} \sum_{i=1}^{n} (1/m_i) \|\boldsymbol{\varepsilon}^{(i)}\|_2^2 \tag{3.30}$$

证明。 参见第 3.5.1 节。

^{2} $\overline{\lambda_2 \alpha}$ $_{i=1}$ ^{i} ^{2}

□

上限 (3.30) 包括三个部分：

- 通过 (3.26) 中的噪声项 $\boldsymbol{\varepsilon}^{(i)}$ 得出本地数据集的属性、

- 经验图通过特征值 $\lambda L_2^{(G)}$ （见 (3.10)）、
- GTVMin 参数 α。

根据 (3.30)，我们可以确保$^{(i)}$ 的误差分量很小。

换句话说，当 $\tilde{}$足够大时，GTVMin 的解就会出现。

在大 α 和连通的经验图中（其中 $\lambda_2 > 0$），GTVMin $\mathbf{w}^{(i)}$ 的解对于所有节点 $i \in \hat{V}$ 都大致相同。这对于某些 FL 应用可能是可取的，因为这些应用的目标是为所有节点训练一个共同的模型[13]。然而，有些 FL 应用涉及异构本地数据集，对这些数据集来说，强迫所有节点就一个共同模型达成一致是不利的（见第 6 讲）。

3.4 编码作业概述

Python 笔记本。FLDesignPrinciple_CodingAssignment.ipynb

数据文件。作业_MLBasicsData.csv

This assignment revolves around a collection of temperature measurements that we store in the empirical graph $G^{(\mathrm{FMI})}$. Each node $i \in V$ represents a FMI weather station at latitude $\mathrm{lat}^{(i)}$ and longitude $\mathrm{lon}^{(i)}$, which we stack into the vector $\mathbf{z}^{(i)} := \mathrm{lat}^{(i)}, \mathrm{lon}^{(i)}{}^{T} \in \mathbb{R}^2$. The local dataset $D^{(i)}$ contains m_i temperature measurements $y^{(i,1)}, \ldots, y^{(i,m_i)}$ at the station i.

$G^{(\mathrm{FMI})}$ 的边由 Python 函数 add_edges() 获得。利用相应向量 $\mathbf{z}^{(i)}$, $\mathbf{z}^{(i')}$ 之间的欧氏距离，将每个 FMI 站点 i 与其最近的邻居 i' 连接起来。邻居的数量由输入参数 numneighbors 控制。 所有边 $\{i, i'\} \in E$ 具有相同的边重 $A_{i,i'} = 1$。

对于每个站点 $i \in V$，您需要学习预测温度的假设 $h(x) = w^{(i)}$ 的单参数 $w^{(i)} \in \mathbb{R}$。我们用平均平方误差损失 $L_i\, w^{(i)} =$ 来衡量假设的质量。 $(1/m_i) \sum_{r=1}^{m_i} y^{(i,r)} - w^{(i)2}$ 。您应该通过平衡来学习参数 $w^{(i)}$ 局部损失与 w 的总变化$^{(i)}$、

$$\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}{}^{T} \in \operatorname*{argmin}_{w^{(1)},\ldots,w^{(n)}{}^{T}} \sum_{i=1}^{n} L\, w_i{}^{(i)} + \alpha \sum_{\{i,\, i'\} \in E} {}^{(i)} - w^{(i')\,2}. \quad (3.31)$$

您的任务包括

- 使用合适的特征选择 $\mathbf{x}^{(i,r)}$ ，将 (3.31) 重译成 (3.19)。请注意，这种

选择必须不同于 "多项式运算基础 "作业中使用的原始特征（经度、纬度、年、月、日、时、分）。

- 通过 (3.25) 并使用 Python 函数 numpy.linalg.inv 计算 (3.31) 的解 $w^{(i)}$。为此，您应该根据本地数据集 $D^{(i)}$ 和经验图 G 的拉普拉斯矩阵 $\mathbf{L}^{(\mathrm{FMI})}$ 确定矩阵 $\mathbf{Q}$ 和向量 $\mathbf{q}$。$^{(\mathrm{FMI})}$

- 研究（3.31）中 α 值的变化对相应解的局部损失和总体变化的影响 $\hat{w}^{(i)}$。

3.5 证明

3.5.1 命题 3.1 的证明

Let us introduce the shorthand $f\left(\mathbf{w}^{(i)}\right)$ for the objective function of the GTVMin instance (3.19). We verify the bound (3.30) by showing that if it does not hold, the choice for the local model parameters $\mathbf{w}^{(i)} := \overline{\mathbf{w}}^{(i)}$ (see (3.26)) results in a smaller objective function value, $f\left(\mathbf{w}^{(i)}\right) < f\left(\overline{\mathbf{w}}^{(i)}\right)$. This would contradict the fact that $\mathbf{w}^{(i)}$ is a solution to (3.19).

首先，请注意

$$
\begin{aligned}
f\left(\overline{\mathbf{w}}^{(i)}\right) &= \sum_{i \in \mathcal{V}} (1/m_i) \left\| \mathbf{y}^{(i)} - \mathbf{X}^{(i)} \overline{\mathbf{w}}^{(i)} \right\|_2^2 + \alpha \sum_{\{i,i'\} \in \mathcal{E}} A_{i,i'} \left\| \overline{\mathbf{w}}^{(i)} - \overline{\mathbf{w}}^{(i')} \right\|_2^2 \\
&\overset{(3.27)}{=} \sum_{i \in \mathcal{V}} (1/m_i) \left\| \mathbf{y}^{(i)} - \mathbf{X}^{(i)} \mathbf{w}^{(i)} \right\|_2^2 \\
&\overset{(3.26)}{=} \sum_{i \in \mathcal{V}} (1/m_i) \left\| \mathbf{X}^{(i)} \overline{\mathbf{w}}^{(i)} + \boldsymbol{\varepsilon}^{(i)} - \mathbf{X}^{(i)} \overline{\mathbf{w}}^{(i)} \right\|_2^2 \\
&= \sum_{i \in \mathcal{V}} (1/m_i) \left\| \boldsymbol{\varepsilon}^{(i)} \right\|_2^2 \qquad (3.32)
\end{aligned}
$$

将 (3.28) 插入 (3.19)、

$$
\begin{aligned}
f\left(\hat{\mathbf{w}}^{(i)}\right) &= \sum_{i \in \mathcal{V}} (1/m_i) \underbrace{\left\| \mathbf{y}^{(i)} - \mathbf{X}^{(i)} \hat{\mathbf{w}}^{(i)} \right\|_2^2}_{\geq 0} + \alpha \sum_{\{i,i'\} \in \mathcal{E}} A_{i,i'} \underbrace{\left\| \hat{\mathbf{w}}^{(i)} - \hat{\mathbf{w}}^{(i')} \right\|_2^2}_{\overset{(3.28)}{=} \left\| \tilde{\mathbf{w}}^{(i)} - \tilde{\mathbf{w}}^{(i')} \right\|_2^2} \\
&\geq \alpha \sum_{\{i,i'\} \in \mathcal{E}} A_{i,i'} \left\| \tilde{\mathbf{w}}^{(i)} - \tilde{\mathbf{w}}^{(i')} \right\|_2^2 \\
&\overset{(3.13)}{\geq} \alpha \lambda_2 \sum_{i=1} \left\| \tilde{\mathbf{w}}^{(i)} \right\|_2^2 . \qquad (3.33)
\end{aligned}
$$

如果约束条件 (3.30) 不成立，那么根据 (3.33) 和 (3.32) 我们可以得到

$f\,\mathrm{w}\hat{}^{(i)} > f\,\overline{\mathbf{w}^{(i)}}$。这与 $\mathrm{w}\hat{}^{(i)}$ 解 (3.19) 的事实相矛盾。