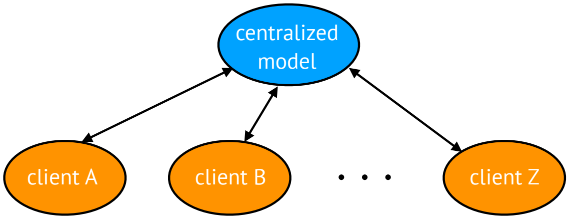
Agnostic federated learning

Mehryar Mohri[1,2], Gary Sivek[1], Ananda Theertha Suresh[1]

[1]Google Research, [2]Courant Institute

June 11, 2019

Federated learning scenario [McMahan et al., '17]



- Data from large number of clients (phones, sensors)
- Data remains distributed over clients
- Centralized model trained based on data

What is the loss function?

Standard federated learning

Setting

- Merge samples from all clients and minimize loss
- Domains: clusters of clients
- Clients belong to p domains: $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_p$

Training procedure

- $\hat{\mathcal{D}}_k$: empirical distribution of $\mathcal{D}_k$ with m_k samples
- $\hat{\mathcal{U}}$: uniform distribution over all observed samples

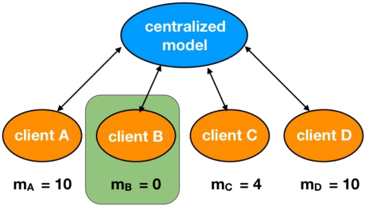
$$\hat{\mathcal{U}} = \sum_{k=1}^{p} \frac{m_k}{\sum_{i=1}^{p} m_i} \hat{\mathcal{D}}_k$$

- Minimize loss over uniform distribution

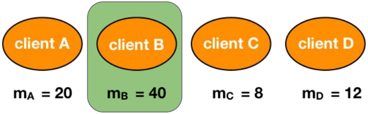
$$\min_{h \in \mathcal{H}} \mathcal{L}_{\hat{\mathcal{U}}}(h)$$

Inference distribution

Training



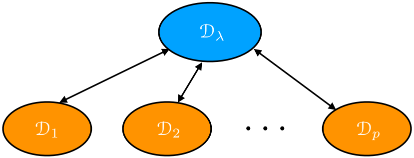
Inference



Inference distribution is not same as the training distribution

Permissions, hardware compatibility, network constraints

Agnostic federated learning



- Learn model that performs well over any mixture of domains
- $\overline{\mathcal{D}}_\lambda = \sum_{k=1}^{p} \lambda_k \cdot \hat{D}_k$
- λ is unknown and belongs to $\Lambda \subseteq \Delta_p$
- Minimize the agnostic loss

$$\min_{h \in \mathcal{H}} \max_{\lambda \in \Lambda} \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h)$$

- Fairness implications

Theoretical results

Generalization bound

Asume $\mathcal{L}$ is bounded by M. For any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in \mathcal{H}$ and $\lambda \in \Lambda$,

$$\mathcal{L}_{\mathcal{D}_\lambda}(h) \leq \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h) + 2\mathfrak{R}_{\mathbf{m}}(\mathcal{G}, \lambda) + M\epsilon + M\sqrt{\frac{\mathfrak{s}(\lambda \,\|\, \overline{\mathbf{m}})}{2m} \log \frac{|\Lambda_\epsilon|}{\delta}}$$

- $\mathfrak{R}_{\mathbf{m}}(\mathcal{G}, \lambda)$: weighted Rademacher complexity
- $\mathfrak{s}(\lambda \,\|\, \overline{\mathbf{m}})$: skewness parameter $1 + \chi^2(\lambda, \mathbf{m})$
- Regularization based on generalization bound

Efficient algorithms?

Stochastic optimization as a two player game

Algorithm $\textsc{Stochastic-AFL}$

Initialization: $w_0 \in \mathcal{W}$ and $\lambda_0 \in \Lambda$.

Parameters: step size $\gamma_w > 0$ and $\gamma_\lambda > 0$.

For $t = 1$ to T:

1. Stochastic gradients: $\delta_w \mathsf{L}(w_{t-1}, \lambda_{t-1})$ and $\delta_\lambda \mathsf{L}(w_{t-1}, \lambda_{t-1})$
2. $w_t = \textsc{Project}(w_{t-1} - \gamma_w \delta_w \mathsf{L}(w_{t-1}, \lambda_{t-1}), \mathcal{W})$
3. $\lambda_t = \textsc{Project}(\lambda_{t-1} + \gamma_\lambda \delta_\lambda \mathsf{L}(w_{t-1}, \lambda_{t-1}), \Lambda)$

Output: $w^A = \frac{1}{T} \sum_{t=1}^{T} w_t$ and $\lambda^A = \frac{1}{T} \sum_{t=1}^{T} \lambda_t$

Results

- $1/\sqrt{T}$ convergence
- Extensions to stochastic mirror descent
- Experimental validation of the above results

Thank you!, more at poster #172