Machine Learning **Spring 2021**

Lecture 8: Algorithm Stability and Generalization, Clustering

Lecturer: Liwei Wang *Scribe: Jianing Lou, Xuanyu Peng, Yuedi Chen, Zhao Zhang*

8.1 Term Project 3

Deep learning has a large number of parameters, which often exceed the number of data. Therefore, the generalization of deep learning is a problem worth studying. Read the paper *Uniform Convergence May be unable to Explain the Generalization in deep learning*. If agree with the article, trying to construct general case; Give reasons for disagreeing.

8.2 Algorithmic Stability and Generalization

Definition 8.1 (Uniform Stability) *Let $\mathcal{A}$ be a learning algorithm. $S = (z_1, \cdots, z_n)$ be a training dataset. Let $S^i = (z_1, \cdots, z_{i-1}, z'_i, z_{i+1}, \cdots, z_n)$ denote a neighboring dataset. Let $\mathcal{A}(S)$ denote a classifier learned by $\mathcal{A}$ from S. Let $\ell(\cdot,\cdot)$ be a loss function.*

A learning algorithm $\mathcal{A}$ is said to have uniform stability β with respect to loss $\ell(\cdot,\cdot)$, if $\forall S, \forall i, \forall S^i, \forall z$,

$$|\ell(\mathcal{A}(S), z) - \ell(\mathcal{A}(S^i), z)| \leq \beta$$

Theorem 8.2 (Uniform stability implies generalization) *Define the risk (similar to test error) as follows,*

$$R(\mathcal{A}(S)) = \mathbb{E}_{z\sim D}[\ell(\mathcal{A}(S), z)]$$

And define the empirical risk (similar to training error) as follows,

$$R_{emp}(\mathcal{A}(S)) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}(S), z_i)$$

Then assume $|\ell(\cdot,\cdot)| \leq M$, we have

$$\mathbb{P}(R(\mathcal{A}(S)) - R_{emp}(\mathcal{A}(S)) \geq \beta + \epsilon) \leq \exp\left(\frac{-n\epsilon^2}{2(n\beta + M)^2}\right)$$

The proof of 8.2 is based on the following lemmas.

Lemma 8.3 *Suppose $\mathcal{A}$ is symmetric with respect to $(z_1, \cdots, z_n)$,* i.e. *for any permutation σ, $\mathcal{A}(z_1, \cdots, z_n) = \mathcal{A}(\sigma(z_1, \cdots, z_n))$, then*

$$\mathbb{E}_S[R(\mathcal{A}(S)) - R_{emp}(\mathcal{A}(S))] \leq \beta \tag{8.1}$$

Proof: *On the one hand,*

$$\begin{aligned}\mathbb{E}_S[R_{emp}(\mathcal{A}(S))] &= \mathbb{E}_S[\frac{1}{n}\sum_{i=1}^{n} l(\mathcal{A}(S), z_i)] \\ &= \mathbb{E}_S[l(\mathcal{A}(S), z_1)]\end{aligned}$$

That is because $l(\mathcal{A}(z_1, \cdots, z_i, \cdots, z_n), z_i) = l(\mathcal{A}(z_i, \cdots, z_1, \cdots, z_n), z_1) = l(\mathcal{A}(S), z_1)$*, according to the symmetry of* $\mathcal{A}$.
On the other hand,

$$\begin{aligned}\mathbb{E}_S[R(\mathcal{A}(S))] &= \mathbb{E}_S\mathbb{E}_z[l(\mathcal{A}(S), z)] \\ &= \mathbb{E}_{z_1,\cdots,z_n,z}[l(\mathcal{A}(S), z)]\end{aligned}$$

That means the expected loss on the random data $z_1, \cdots, z_n, z$*. Switch* z *and* z_1*, we have*

$$\mathbb{E}_S[R(\mathcal{A}(S))] = \mathbb{E}_S[l(\mathcal{A}(S'), z_1)]$$

where S' *denotes* $(z, z_2, \cdots, z_n)$.
According to the definition of β,

$$\begin{aligned}\mathbb{E}_S[R(\mathcal{A}(S)) - R_{emp}(\mathcal{A}(S))] &= \mathbb{E}_S[l(\mathcal{A}(S), z_1) - l(\mathcal{A}(S'), z_1)] \\ &\leq \beta\end{aligned}$$

■

Lemma 8.4 (McDiarmid's Inequality) *Suppose* $|f(x_1, \cdots, x_n) - f(x_1, \cdots, x_i', \cdots, x_n)| \leq c_i, \forall i \in [n], \forall x_1, \cdots x_i, x_i', \cdots, x_n$. *Then*

$$\mathbb{P}(f(X_1, \cdots, X_n) - \mathbb{E}f(X_1, \cdots, X_n) \geq \epsilon) \leq \exp\{-\frac{2\epsilon^2}{\sum \mathrm{c_i^2}}\} \tag{8.2}$$

Lemma 8.5 *Assume* $|l(\cdot, \cdot)| \leq M$,

$$\big|[R(\mathcal{A}(S)) - R_{emp}(\mathcal{A}(S))] - [R(\mathcal{A}(S^i)) - R_{emp}(\mathcal{A}(S^i))]\big| \leq 2(\beta + \frac{M}{n}) \tag{8.3}$$

Proof:

$$\begin{aligned}
& \big|[R(\mathcal{A}(S)) - R_{emp}(\mathcal{A}(S))] - [R(\mathcal{A}(S^i)) - R_{emp}(\mathcal{A}(S^i))]\big| \\
\leq\ & \big|R_{emp}(\mathcal{A}(S)) - R_{emp}(\mathcal{A}(S^1))\big| + \big|R(\mathcal{A}(S)) - R(\mathcal{A}(S^1))\big| \\
\leq\ & \frac{1}{n}\big|l(\mathcal{A}(S), z_1) - l(\mathcal{A}(S^1), z_1')\big| + \\
& \frac{1}{n}\sum_{i=2}^{n}\big|l(\mathcal{A}(S), z_i) - l(\mathcal{A}(S^1), z_i)\big| + \\
& \mathbb{E}_z[l(\mathcal{A}(S), z) - l(\mathcal{A}(S^1), z)] \\
\leq\ & \frac{1}{n}(\big|l(\mathcal{A}(S), z_1) - l(\mathcal{A}(S^1), z_1)\big| + \big|l(\mathcal{A}(S^1), z_1)\big| + \big|l(\mathcal{A}(S^1), z_1')\big|) + \frac{n-1}{n}\beta + \beta \\
\leq\ & 2(\beta + \frac{M}{n})
\end{aligned}$$

■

Though loss functions are usually unbounded, 8.3 still holds. That's because the proof only uses $\left|l(\mathcal{A}(S^1), z_1)\right| \leq M$ and $\left|l(\mathcal{A}(S^1), z_1')\right| \leq M$, which are ensured by the bounded data.

Finally, we conclude the proof of Theorem 8.2:

Proof: Denote $\Phi(S) = R(\mathcal{A}(S)) - R_{emp}(\mathcal{A}(S))$. According to Lemma 8.3, we have

$$\mathbb{P}(\Phi(S) \geq \beta + \epsilon) \leq \mathbb{P}(\Phi(S) - \mathbb{E}_S[\Phi(S))] \geq \epsilon)$$

Lemma 8.5 means that $\Phi(S)$ is a stable function, where $c_i = 2(\beta + \frac{M}{n})$, then we can used McDiarmid's Inequality to get the result,

$$\mathbb{P}(\Phi(S) - \mathbb{E}_S[\Phi(S))] \geq \epsilon) \leq \exp\left(\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right) = \exp\left(\frac{-n\epsilon^2}{2(n\beta + M)^2}\right)$$

■

8.3 Clustering

Clustering is an unsupervised learning task, and is described as follows:

Given a set of datas $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ and a non-zero integer $k \leq n$, where each data is a d-dimensional real vector, clustering(k-means clustering) aims to partition these n datas into k sets $S_1, S_2, \ldots, S_k$ so as to minimize the following loss function

$$\phi = \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in S_i} ||\boldsymbol{x} - \boldsymbol{\mu}_i||^2$$

where $\boldsymbol{\mu}_i$ is the cluster center of S_i.

The most common algorithm is "k-means algorithm".

Algorithm 8.3.1: k-means algorithm

Initialize: choose k points randomly as the cluster centers $m_1, \ldots, m_k$;
do
  Assign each data to the cluster center with the nearest mean;
  $S_i \leftarrow \{x_j : x_j \text{ is assigned to } m_i\}, \forall i$;
  $m_i \leftarrow$ the mean of points in $S_i, \forall i$;
while k *cluster centers changes*;
return $m_1, \ldots, m_k$;

However, this naive algorithm is only guaranteed to find a local optimum.

Improvement: k-means++

We can optimize the "initialize" step in line 1 as follows:

Algorithm 8.3.2: Improved initialization

1 Choose one center uniformly at random among the data points;
2 **for** $i : 2 \rightarrow k$ **do**
3 Choose one new data point at random as a new center, a point $\boldsymbol{x}$ is chosen with probability proportional to $||\boldsymbol{x} - m^*||^2$, where $m^* \in \{m_1, \ldots, m_{i-1}\}$ and is nearest to $\boldsymbol{x}$.
4 **end**

Letting ϕ_{OPT} denote the global optimal loss, it has been proved by Arthur and Vassilvitskii[1] that after choosing centers in this way, we have

$$\mathbb{E}[\phi] \leq 8(\ln k + 2)\phi_{OPT}$$

References

[1] Arthur, D.; Vassilvitskii, S. (2007). "k-means++: the advantages of careful seeding". *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms.* Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035.