VIDY Summarizer:
UML Chapter 30:Compression Bounds & Stronger Generalization Bounds for Deep Nets via a Compression Approach

Jianpeng Chen

6/19/2024

Summarize

Tighter generalization bound via compression

– **1. Data compression (**UML 30**)**

$$L_D(h_T) \le L_V(h_T) + \sqrt{\frac{2L_V(h_T)\log(1/\delta)}{|V|}} + \frac{4\log(1/\delta)}{|V|}$$

Data compression $\Big\downarrow$ Sample a subset of the dataset ***I*** *.

$$P\left(L_D(h_{I^*}) \le L_V(h_{I^*}) + \sqrt{\frac{4kL_V(h_{I^*})\log(m/\delta')}{m}} + \frac{8k\log(m/\delta')}{m}\right) \ge 1-\delta'$$

– **Examples:**

- Axis Aligned Rectangles
- Halfspaces
- Separating Polynomials
- Separation with Margin

Summarize

Tighter generalization bound via compression

– **2. Model compression (**Stronger generalization bounds for deep nets via a compression approach**)**
  - **Tight the generalization bound not of** f **but of its compression** g.

Theorem 2.2. *((Neyshabur et al., 2017a)) For any deep net with layers* $A^1, A^2, \dots A^d$ *and output margin* γ *on a training set* S*, the generalization error can be bounded by*

$$\tilde{O}\left(\sqrt{\frac{hd^2 \max_{x\in S}\|x\| \prod_{i=1}^{d}\|A^i\|_2^2 \sum_{i=1}^{d}\frac{\|A^i\|_F^2}{\|A^i\|_2^2}}{\gamma^2 m}}\right).$$

Example 1. MLP:

Theorem 4.1. *For any fully connected network* f_A *with* $\rho_\delta \geq 3d$*, any probability* $0 < \delta \leq 1$ *and any margin* γ*, Algorithm 1 generates weights* $\tilde{A}$ *for the network* $f_{\tilde{A}}$ *such that with probability* $1-\delta$ *over the training set and* $f_{\tilde{A}}$*, the expected error* $L_0(f_{\tilde{A}})$ *is bounded by*

$$\hat{L}_\gamma(f_A) + \tilde{O}\left(\sqrt{\frac{c^2 d^2 \max_{x\in S}\|f_A(x)\|_2^2 \sum_{i=1}^{d}\frac{1}{\mu_i^2\mu_{i\rightarrow}^2}}{\gamma^2 m}}\right)$$

Example 2. CNN:

Theorem 5.1. *For any convolutional neural network* f_A *with* $\rho_\delta \geq 3d$*, any probability* $0 < \delta \leq 1$ *and any margin* γ*, Algorithm 4 generates weights* $\tilde{A}$ *for the network* $f_{\tilde{A}}$ *such that with probability* $1-\delta$ *over the training set and* $f_{\tilde{A}}$*:*

$$L_0(f_{\tilde{A}}) \leq \hat{L}_\gamma(f_A) + \tilde{O}\left(\sqrt{\frac{c^2 d^2 \max_{x\in S}\|f_A(x)\|_2^2 \sum_{i=1}^{d}\frac{\beta^2(\lceil \kappa_i/s_i\rceil)^2}{\mu_i^2\mu_{i\rightarrow}^2}}{\gamma^2 m}}\right)$$