泛化理论

第五章 Information-based Bound

§5.1.1 Information-based Bound

@ 滕佳烨

[ref] Russo, D., & Zou, J. (2016, May). Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics* (pp. 1232-1240). PMLR.
[ref] Xu, A., & Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, *30*.

Recall:

Uniform Convergence: $L(\hat{f}) - \hat{L}(\hat{f}) \le \sup_{f \in \mathcal{F}} \left| L(f) - \hat{L}(f) \right|.$

Stability-based bound: similar dataset returns similar models

PAC-Bayesian bound: Bayesian training (parameter following a distribution), $\sqrt{\frac{KL(P||Q)}{n}}$ (P: prior distribution, Q: posterior distribution)

Today's topic:

Information-based bound: another approach to generalization (measuring the dependency)

Information-based bound

One sentence: bound the generalization directly using its mutual information.

Similar to PAC-Bayes, we still consider stochastic parameter.
We want to bound the generalization gap

$$\mathbb{E}\ell\left(\bar{S}, W(S)\right) - \mathbb{E}\ell\left(S, W(S)\right).$$

A direct intuition: when $W(S)$ has large **dependency** on S, the gap might be large.
But how to **measure** the **dependency**? $\rightarrow$ **mutual information**

Mutual Information:
For random variable X and Y, we define its mutual information as

$$I(X, Y) = E_{XY} \log \frac{p(x, y)}{p(x)p(y)},$$

where $I(X, Y) = 0$ if X ind Y.

Information-based bound

One sentence: if W(S) and S has small mutual information, the bound is better.

Theorem (Mutual Information). Suppose the loss function $\ell(w,z)$ is σ-subGaussian for all w (where the probability is on sample z), then

$$\mathbb{E}\ell(\bar{S}, W(S)) - \mathbb{E}\ell(S, W(S)) \leq \sqrt{\frac{2\sigma^2}{n} I(S, W(S))}.$$

Remark: Different paper may use different information form. Here we use the version in Xu & Raginsky (2017) with the mutual information between the training set and the trained parameter.

The bound is still $\sqrt{n}$ convergence rate.

Information-based bound (proof)

Theorem (Mutual Information). Suppose the loss function $\ell(w,z)$ is σ-subGaussian for all w (where the probability is on sample z), then

$$\mathbb{E}\ell\left(\bar{S},W(S)\right)-\mathbb{E}\ell\left(S,W(S)\right)\leq\sqrt{\frac{2\sigma^2}{n}I\left(S,W(S)\right)}.$$

Lemma: for random variable $(X,Y)\sim P_X\times P_{Y|X}$ with independent copy $\left(\tilde{X},\tilde{Y}\right)\sim P_{\tilde{X}}\times P_{\tilde{Y}}$, where $P_X=P_{\tilde{X}}$ and $P_Y=P_{\tilde{Y}}$, if $f\left(\tilde{X},\tilde{Y}\right)$ is subGaussian, we have

$$\left|\mathbb{E}f(X,Y)-\mathbb{E}f\left(\tilde{X},\tilde{Y}\right)\right|\leq\sqrt{2\sigma^2 I(X,Y)}.$$

Therefore, by setting $f\left(\tilde{X},\tilde{Y}\right)$ as the test loss $\ell\left(\bar{S},W(S)\right)$, it is $\sigma/\sqrt{n}$-subGaussian (by concentration inequality on $\bar{S}$ with n samples). We can derive the theorem.

Proof of the Lemma (one can check 4.1.2 for more details).

$$KL(P_{XY}|P_X\times P_Y)\underbrace{\geq}_{\text{Jenson's}}\mathbb{E}[\lambda f(X,Y)]-\log\mathbb{E}\exp\left[\lambda f\left(\tilde{X},\tilde{Y}\right)\right]\underbrace{\geq}_{\text{subGaussian}}\lambda\left[\mathbb{E}f(X,Y)-\mathbb{E}f\left(\tilde{X},\tilde{Y}\right)\right]-\frac{\lambda^2\sigma^2}{2}.$$

Take-away messages

(a) Mutual Information

(b) Information-based bound:

If $W(S)$ does not depend on S much, the bound is small.

For σ-subGaussian loss, we have

$$\mathbb{E}\ell(\bar{S}, W(S)) - \mathbb{E}\ell(S, W(S)) \leq \sqrt{\frac{2\sigma^2}{n} I(S, W(S))}.$$

All the slides will be available at www.tengjiaye.com/generalization soon.

@ 滕佳烨

Thanks!