PAC-bayesian bounds

Theoretical Deep Learning #2: generalization ability

Eugene Golikov

MIPT, fall 2019

Neural Networks and Deep Learning Lab., MIPT

Notation and goal

- Data distribution: $\mathcal{D}$;
- Dataset: $S_n = \{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$, where all $y_i \in \{-1, 1\}$, all $x_i \in X$;
- Model: $f : X \to \mathbb{R}$;
- Loss function $l(y, f(x))$;
- Risk: $R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}} l(y, f(x))$;
- Empirical risk: $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^n l(y_i, f(x_i))$;
- Result of learning on dataset S_n: $\hat{f}_n = \mathcal{A}(S_n) \in \mathcal{F}$.

Our goal is to bound the risk difference:

$$R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) \leq \mathrm{bound}(N(\hat{f}_n), n, \delta) \quad \text{w.p. } \geq 1 - \delta \text{ over } S_n.$$

PAC-bayesian bounds

Bounds for deterministic $\mathcal{A}$:

- **Finite $\mathcal{F}$:**

$$R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) \leq \sqrt{\frac{1}{2n}\left(\log\frac{1}{\delta} + \log|\mathcal{F}|\right)} \quad \text{w.p. } \geq 1-\delta \text{ over } S_n.$$

- **At most countable $\mathcal{F}$ (McAllester, 1998)**[1]**:**

$$R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) \leq \sqrt{\frac{1}{2n}\left(\log\frac{1}{\delta} + \log\frac{1}{P(\hat{f}_n)}\right)} \quad \text{w.p. } \geq 1-\delta \text{ over } S_n,$$

where P is a distribution over $\mathcal{F}$ (**prior**).

[1] Preliminary theorem 2 in `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.1745&rep=rep1&type=pdf`

PAC-bayesian bounds

Consider stochastic learning algorithm: $\hat{f}_n = \mathcal{A}(S_n) \sim Q|S_n$.
Define $R(Q) := \mathbb{E}_{f\sim Q} R(f), \quad \hat{R}_n(Q) := \mathbb{E}_{f\sim Q}\hat{R}_n(f)$.

Corresponding bound:

$$R(Q|S_n) - \hat{R}_n(Q|S_n) \le \mathrm{bound}(N(Q|S_n), n, \delta) \quad \text{w.p. } \ge 1-\delta \text{ over } S_n.$$

PAC-bayesian bound (McAllester, 1999)[2]:

$$R(Q|S_n) - \hat{R}_n(Q|S_n) \le \sqrt{\frac{1}{2n-1}\left(\log\frac{4n}{\delta} + KL(Q|S_n \,\|\, P)\right)} \quad \text{w.p. } \ge 1-\delta$$

for any distribution P on $\mathcal{F}$.

[2] Theorem 2 in `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.1908&rep=rep1&type=pdf`

PAC-bayesian bounds

Define: $\Delta_n(f) := |R(f) - \hat{R}_n(f)|$.

Lemma (McAllester, 1999)[3]:

$$\mathbb{E}_{f\sim P} e^{(2n-1)\Delta_n(f)^2} \leq \frac{4n}{\delta} \quad \text{w.p. } \geq 1-\delta \text{ over } S_n$$

for any distribution P on $\mathcal{F}$.

Lemma (Donsker & Varadhan):
Let P and Q be distributions on X. Then:

$$KL(P \,\|\, Q) = \sup_{h\colon X \to \mathbb{R}} \left(\mathbb{E}_{x\sim P} h(x) - \log \mathbb{E}_{x\sim Q} e^{h(x)} \right).$$

[3] Lemma 17 in `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.1908&rep=rep1&type=pdf`

PAC-bayesian bounds

Lemma (Langford & Seeger, 2001)[4]:

$$\mathbb{E}_{f\sim P} e^{(n-1)KL(\hat{R}_n(f) \,\|\, R(f))} \leq \frac{2n}{\delta} \quad \text{w.p. } \geq 1-\delta \text{ over } S_n$$

for any distribution P on $\mathcal{F}$.

Theorem (Langford & Seeger, 2001)[5]:

$$KL(\hat{R}_n(Q|S_n) \,\|\, R(Q|S_n)) \leq \frac{1}{n-1}\left(\log\frac{2n}{\delta} + KL(Q|S_n \,\|\, P)\right) \quad \text{w.p. } \geq 1-\delta$$

for any distribution P on $\mathcal{F}$.

---

[4] Lemma 2 in `http://hunch.net/~jl/projects/prediction_bounds/averaging/averaging_tech.pdf`

[5] Theorem 3 there.

PAC-bayesian bounds

Let $X_{1:n}$ be i.i.d., $X_i \sim \mathcal{B}(p)\ \forall i$.

Hoeffding's inequality:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq p + \epsilon\right) \leq e^{-2n\epsilon^2}; \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq p - \epsilon\right) \leq e^{-2n\epsilon^2}.$$

Chernoff-Hoeffding's inequality:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq p + \epsilon\right) \leq e^{-nKL(p+\epsilon \,\|\, p)};$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq p - \epsilon\right) \leq e^{-nKL(p-\epsilon \,\|\, p)}.$$