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 $\mathcal{F}$:**

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

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

PAC-bayesian bounds

Consider stochastic learning algorithm: $\hat{f}_n = \mathcal{A}(S_n) \sim Q|S_n$.

Corresponding bound:

$$\mathbb{E}_{Q|S_n}\left(R(\hat{f}_n) - \hat{R}_n(\hat{f}_n)\right) \leq \mathrm{bound}(N(Q|S_n), n, \delta) \quad \text{w.p. } \geq 1 - \delta \text{ over } S_n.$$

PAC-bayesian bound (McAllester, 1999)[1]:

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

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