Topics in Learning Theory

Lecture 3: Concentration and Rademacher Complexity

Topics

- General Exponential inequalities for sum of independent variables
- Concentration inequality: McDiarmid's inequality
- Symmetrization and Rademacher Complexity

Independent Variables and Concentration

- Let $\xi_1, \ldots, \xi_n$ be iid random variables.
- Law of large numbers: empirical mean converges to the true mean in probability

$$\frac{1}{n}\sum_{i=1}^{n} \xi_i \to_P \mathbf{E}\xi_1.$$

  – the empirical mean is concenrated around the true mean when n is large.

- Given deviation $\epsilon > 0$, degree of concentration: measured by tail inequality for

$$P\left[\frac{1}{n}\sum_{i=1}^{n} \xi_i - \mathbf{E}\xi_1 \geq \epsilon\right] \qquad \text{one sided}$$

or

$$P\left[|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mathbf{E}\xi_1| \geq \epsilon\right] \qquad \text{two sided}$$

- Markov inequlaity (two sided): if $f(u) \geq 0$, then:

$$P\left[|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mathbf{E}\xi_1| \geq \epsilon\right] \leq \frac{Ef(|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mathbf{E}\xi_1|)}{\inf_{u\geq\epsilon} f(u)}.$$

- Chebyshev's inequality:

$$P\left[|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mathbf{E}\xi_1| \geq \epsilon\right] \leq \frac{1}{n\epsilon^2}Var(\xi_1)$$

can be proven using $f(u) = u^2$ (second moment)

– decays slowly when is large.

- Hoeffding's inequality ($_1 \in [0;1]$):

$$\mathsf{P}[\frac{1}{\mathsf{n}}\sum_{i=1}^{n} \ _i - \mathbf{E} \ _1 \geq \ ] \leq \exp(-2\mathsf{n} \ ^2)$$

$$\mathsf{P}[\frac{1}{\mathsf{n}}\sum_{i=1}^{n} \ _i - \mathbf{E} \ _1 \leq - \ ] \leq \exp(-2\mathsf{n} \ ^2)$$

can be proven using $\mathsf{f}(\mathsf{u}) = \mathsf{e}^{\lambda u}$

McDiarmid's inequality

Theorem 1. *Let $\xi_1, \ldots, \xi_n$ be independent random variables. Consider a real-valued function $g(\xi_1, \ldots, \xi_n)$ such that $\forall i = 1, \ldots, n$:*

$$\sup_{\xi_1, \ldots, \xi_n, \xi_i'} |g(\xi_1, \ldots, \xi_n) - g(\xi_1, \ldots, \xi_{i-1}, \xi_i', \xi_{i+1}, \ldots, \xi_n)| \leq c_i.$$

Then for all $\eta \geq 0$, with probability of at least $1 - \eta$,

$$g(\xi_1, \ldots, \xi_n) \leq \mathbf{E} g(\xi_1, \ldots, \xi_n) + \sqrt{\frac{\ln(1/\eta)}{2} \sum_{i=1}^{n} c_i^2}.$$

Remark: generalize Hoeffding's inequality

- proof is analogous
- obtain Hoeffding's inequality for $\xi_i \in [0,1]$ and $g(\xi_1, \ldots, \xi_n) = \frac{1}{n} \sum_{i=1}^{n} \xi_i$

Application to Learning

- We let

$$g(S_n) = \sup_{f \in \mathcal{H}} [E\phi(f(X), Y) - \frac{1}{n} \sum_{i=1}^{n} \phi(f(X_i), Y_i)]$$

- If $\phi(f(X), Y) \in [0, 1]$, then can take $c_i = 1$ in McDiarmid's inequality: with probability $1 - \eta$:

$$[E_{X,Y}\phi(f(X), Y) - \frac{1}{n} \sum_{i=1}^{n} \phi(f(X_i), Y_i)]$$

$$\leq E_{S_n} \sup [E_{X,Y}\phi(f(X), Y) - \frac{1}{n} \sum_{i=1}^{n} \phi(f(X_i), Y_i)] + \sqrt{\frac{\ln(1/\eta)}{2n}}.$$

- McDiarmid gives $\sqrt{1/n}$ rate at best
  - similar to additive form of Hoeffding's inequality
- In order to get fast $1/n$ rate, one requires more sophisticated concentration inequality
  - bernstein-type
  - also requires localization techniques (localized Rademacher complexity) — important developments in recent years

Symmetrization

- Purpose: derive complexity measure using empirical quantities
  - such as empirical covering number or empirical Rademacher complexity
- $S'_n = \{(X'_i, Y'_i)\}$: iid copies of $S_n = \{(X_i, Y_i)\}$
- σ_i: iid Bernoulli with equal probability ± 1

$$E_{S_n} \sup[E_{X,Y}\phi(f(X),Y) - \frac{1}{n}\sum_{i=1}^{n}\phi(f(X_i),Y_i)]$$

$$= E_{S_n} \sup[E_{S'_n}\frac{1}{n}\sum_{i=1}^{n}\phi(f(X'_i),Y'_i) - \frac{1}{n}\sum_{i=1}^{n}\phi(f(X_i),Y_i)]$$

$$\leq E_{S_n,S'_n} \sup[\frac{1}{n}\sum_{i=1}^{n}\phi(f(X'_i),Y'_i) - \frac{1}{n}\sum_{i=1}^{n}\phi(f(X_i),Y_i)]$$

$$= E_{S_n,S'_n} \sup[\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi(f(X'_i),Y'_i) - \frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi(f(X_i),Y_i)]$$

$$\leq 2E_{S_n} \sup \frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi(f(X'_i),Y'_i).$$

Rademacher Complexity

$$R(\phi(\mathcal{H})) = E_{S_n} R(\phi(\mathcal{H})|S_n)$$

and

$$R(\phi(\mathcal{H})|S_n) = E_\sigma \sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \phi(f(X_i'), Y_i').$$

- Learning Bound: if $\phi \in [0, 1]$, then (McDiarmid)

$$E_{X,Y} \phi(f(X), Y) \leq \frac{1}{n} \sum_{i=1}^{n} \phi(f(X_i), Y_i) + 2R(\phi(\mathcal{H})) + \sqrt{\frac{\ln(1/\eta)}{2n}}.$$

Estimating Rademacher complexity: L_2 covering number and chaining

- Given S_n, L_2 empirical covering is the cover of $\phi(cH)$ with respect to metric:

$$d(f, f'|S_n) = [\frac{1}{n}\sum_{i=1}^{n}[\phi(f(X_i), Y_i) - \phi(f(X_i)', Y_i')]^2]^{1/2},$$

  which we denote as $N_2(\phi(\mathcal{H}), \epsilon|S_n)$

- Empirical L_2 covering number is smaller than empirical L_∞ covering number.

- Chaining bound for Rademacher complexity: there is universal $C > 0$:

$$R(\mathcal{H}|S_n) \leq 0.5C \inf_{\epsilon_0} \left[\epsilon_0 + \int_{\epsilon > \epsilon_0} \sqrt{\frac{1}{n} \ln(1 + N_2(\mathcal{H}, \epsilon|S_n))} d\epsilon \right]$$

- Learning bound using empirical L_2 covering number: if $\phi \in [0, 1]$

$$\begin{aligned} E_{X,Y}\phi(f(X), Y) \leq & \frac{1}{n} \sum_{i=1}^{n} \phi(f(X_i), Y_i) + \sqrt{\frac{\ln(1/\eta)}{2n}} \\ & + C E_{S_n} \inf_{\epsilon_0} \left[\epsilon_0 + \int_{\epsilon > \epsilon_0} \sqrt{\frac{1}{n} \ln(1 + N_2(\phi(\mathcal{H}), \epsilon|S_n))} d\epsilon \right]. \end{aligned}$$

- Consequence of chaining bound:

– if covering number $N_2(\phi(\mathcal{H}), \epsilon|S_n) = O(1/\epsilon^p)$ for $p < 2$, then

$$R(\mathcal{H}|S_n) = O(1/\sqrt{n})$$

– sharper than bound not based on the chaining argument: complexity in the last lecture

$$\inf\{\epsilon : \epsilon \geq b\sqrt{\ln[aN_\infty(\phi(\mathcal{H}), \epsilon/4)/\eta]/n}\}$$

* $N_2(\phi(\mathcal{H}), \epsilon|S_n) \leq N_\infty(\phi(\mathcal{H}), \epsilon)$

References

- General Exponential inequality: follows

  T. Zhang. Data dependent concentration bounds for sequential prediction algorithms. In *COLT 05*, 2005.

- Simple Concentration Inequality/Rademacher Complexity:

  P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. In *COLT 01*, pages 224–240, 2001.

- Chaining:

  A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

- Localized Rademacher Complexity:

  P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.