A High-Level Introduction to Concentration Inequalities

Haoen CUI

Uptake Math Club Lightning Talk

July 5, 2019

Outline

Objectives

What are "concentration inequalities" used for?

- To show that some random quantity is close to its mean with high probability

Why does it matter?

- To prove theoretical statistical learning results

Notation

Let P be a probability measure over an instance space $\mathcal{Z}$ and f be a function defined on domain $\mathcal{Z}$, then we write

$$P(f) \stackrel{\text{def}}{=} \mathbb{E}_{Z\sim P}[f(Z)]$$

$$P_n(f) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f(Z_i) = \mathbb{E}_{Z\sim P_n}[f(Z)]$$

where $Z_i \stackrel{\text{iid}}{\sim} P$ are sample observations and hence P_n is the empirical measure,

$$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Z_i \in A\}$$

Desired Concentration Inequalities

In general, we seek for results of the form

$$\mathbb{P}\left(\left|f(Z_1,\ldots,Z_n)-\mathbb{E}[f(Z_1,\ldots,Z_n)]\right|>\epsilon\right)<\delta(\epsilon,n)$$

where $\forall\epsilon>0$, $\delta(\epsilon,n)\to 0$ as $n\to\infty$.
For statistical learning theory, we will need *uniform bounds* of the form

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|f(Z_1,\ldots,Z_n)-\mathbb{E}[f(Z_1,\ldots,Z_n)]\right|>\epsilon\right)<\delta(\epsilon,n)$$

over a class of functions $\mathcal{F}$.

Application: Empirical Risk Minimization in Classification

Consider a binary classification problem. Suppose we have data $\{(X_i, Y_i)\}_{i=1}^n$ where $X_i \in \mathbb{R}^d$ and $Y_i \in \{0, 1\}$. Let $f : \mathbb{R}^d \to \{0, 1\}$ be a classifier, then its

- Training Error

$$\widehat{R}_n(f) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_i \neq f(X_i)\}$$

- Generalization Error:

$$R(f) = \mathbb{P}(Y \neq f(X))$$

Application: ERM in Classification (Cont.)

Let $\widehat{f}_*$ minimize training error $\widehat{R}_n$ and f_* minimize generalization error R. Suppose we have

$$\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)| \leq \epsilon$$

then

$$R(f_*) \leq R(\widehat{f}_*) \leq \widehat{R}_n(\widehat{f}_*) + \epsilon \leq \widehat{R}_n(f_*) + \epsilon \leq R(f_*) + 2\epsilon$$

which translates to

$$|R(\widehat{f}_*) - R(f_*)| \leq 2\epsilon$$

guarantees the consistency of ERM over $\mathcal{F}$.

Outline

Commonly Used Results

Tail Probability of Gaussian Random Variable Decays Exponentially

For $\forall \epsilon > 0$, consider sample average $\overline{X}_n$,

$$X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2) \implies \mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$$

Commonly Used Results (Cont.)

Bounds on Expected Values

Suppose that $\mathbb{P}(X_n \geq 0) = 1$ and $\exists c_1 > \frac{1}{e}, c_2 > 0$ such that $\forall \epsilon > 0$

$$\mathbb{P}(X_n < \epsilon) \leq c_1 \exp(-c_2 n \epsilon^2)$$

then we have, for $C = \frac{1+\ln(c_1)}{c_2}$,

$$\mathbb{E}[X_n] \leq \sqrt{\frac{C}{n}}$$

Commonly Used Results (Cont.)

Bounds on the Maximum of a Set of Random Variables

Let $\{X_i\}_i$ be a set of random variables. Suppose $\exists \sigma > 0$ such that

$$\mathbb{E}[\exp(tX_i)] \leq \exp\left(\frac{t\sigma^2}{2}\right) \qquad \forall t > 0$$

then

$$\mathbb{E}\left[\max_{i\in[n]} X_i\right] \leq \sigma\sqrt{2\ln(n)}$$

Commonly Used Results (Cont.)

Bounds on Moment Generating Functions

Suppose $\mathbb{P}(|X| \leq c) = 1$ and $\mathbb{E}[X] = 0$. Let $\sigma^2 \stackrel{\text{def}}{=} \text{Var}[X]$. Then

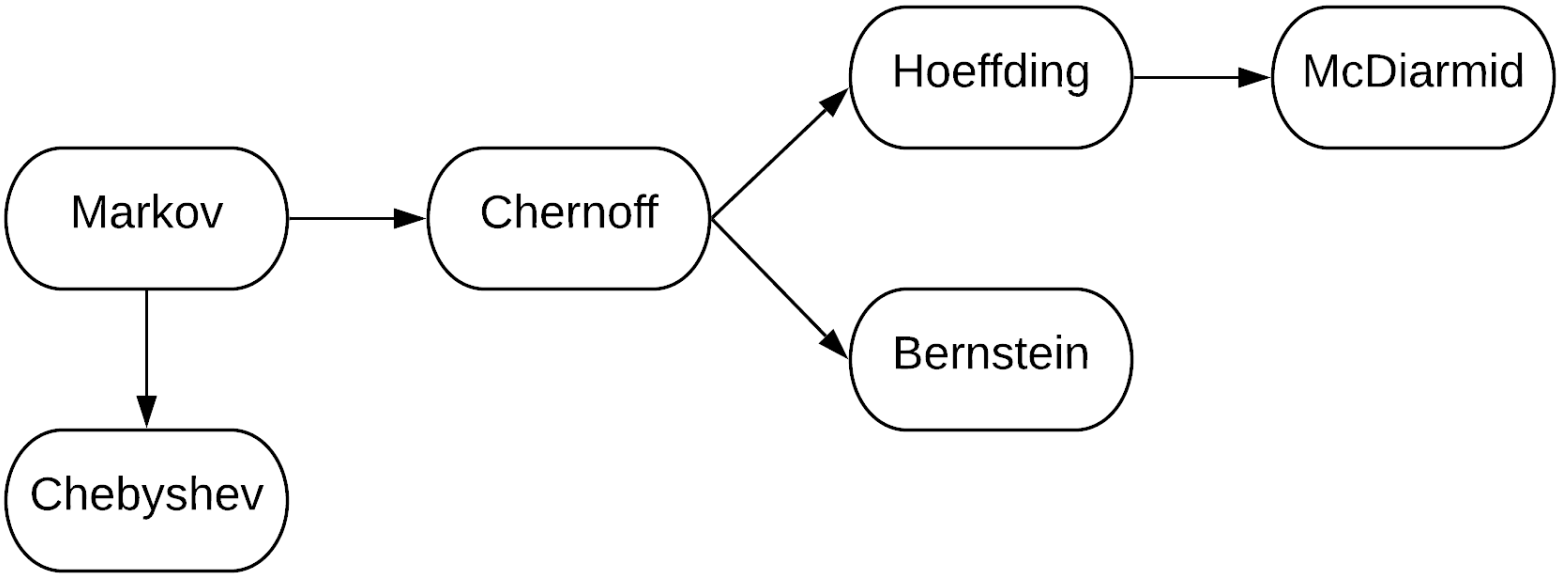
$$\mathbb{E}[e^{tX}] \leq \exp\left\{t^2\sigma^2\left(\frac{e^{tc} - 1 - tc}{(tc)^2}\right)\right\} \qquad \forall t > 0$$

Another Bound on Moment Generating Functions Useful for Proof of Hoeffding's Inequality

Suppose $\mathbb{P}(a \leq X \leq b) = 1$ and let $\mu \stackrel{\text{def}}{=} \mathbb{E}[X]$, then

$$\mathbb{E}[e^{tX}] \leq \exp\left(t\mu + \frac{(b-a)^2}{8}t^2\right) \qquad \forall t \in \mathbb{R}$$

Roadmap


Hoeffding
McDiarmid
Markov
Chernoff
Bernstein
Chebyshev

Markov's Inequality and Chebyshev's Inequality

Markov's Inequality

Suppose that $\mathbb{P}(X \geq 0) = 1$ and $\mathbb{E}[X] < \infty$, then for $\forall \epsilon > 0$,

$$\mathbb{P}(X > \epsilon) \leq \frac{E[X]}{\epsilon}$$

Chebyshev's Inequality (Corollary of Markov's Inequality)

Suppose that $\mathbb{E}[X] < \infty$ and $\mathbb{E}[X^2] < \infty$, then for $\forall \epsilon > 0$,

$$\mathbb{P}(|X - \mathbb{E}[X]| > \epsilon) \leq \frac{\mathrm{Var}[X]}{\epsilon^2}$$

Chernoff Bounding Trick

Consider applying *Chebyshev's Inequality* to sample average $\overline{X}_n$ of iid random variables with mean μ and variance σ^2,

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

does not decay exponentially fast as sample size n increases.

Chernoff Bounding Trick

$$\mathbb{P}(X > \epsilon) \leq \inf_{t \geq 0} e^{-t\epsilon}\mathbb{E}[e^{tX}] \qquad \forall \epsilon \in \mathbb{R}$$

Hoeffding's Inequality

Hoeffding's Inequality

If $X_1, X_2, \cdots, X_n$ are independent (but not necessarily identical) random variables with

$$\mathbb{P}(a_i \leq X_i \leq b_i) = 1$$

and common mean $\mathbb{E}[X_i] = \mu$, then for $\forall \epsilon > 0$

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq 2 \exp\left(- \frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2} \right)$$

McDiarmid's Inequality

McDiarmid's Inequality relaxes bounded random variable assumption in *Hoeffding's Inequality* to *bounded difference*.

McDiarmid's Inequality

If $X_1, X_2, \cdots, X_n$ are independent random variables with

$$\sup_{z_1,\cdots,z_n,z_i'} \left| f(\cdots, z_{i-1}, z_i, z_{i+1}, \cdots) - f(\cdots, z_{i-1}, z_i', z_{i+1}, \cdots) \right| \leq c_i$$

for some constant vector $(c_1, \cdots, c_n)$, then

$$\mathbb{P}(|P_n(f) - P(f)| > \epsilon) \leq 2\exp\left(- \frac{2\epsilon^2}{\sum_{i=1}^n {c_i}^2} \right) \qquad \forall \epsilon > 0$$

Bernstein's Inequality

Hoeffdings inequality does not use any information about the random variables except the fact that they are bounded. If the variance of X_i is small, then we can get a sharper inequality from *Bernsteins inequality*.

Bernsteins Inequality

Suppose $\mathbb{P}(|X_i| \leq c) = 1$ and $\mathbb{E}[X_i] = \mu$, then

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq 2 \exp\left\{ - \frac{n\epsilon^2}{2\sigma^2 + \frac{2}{3}c\epsilon} \right\} \qquad \forall \epsilon > 0$$

where $\sigma^2 \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \text{Var}[X_i]$.

Outline

Shattering Number

Let $\mathcal{F}$ be a class of binary functions $f : \mathcal{Z} \to \{0, 1\}$ over instance space $\mathcal{Z}$. For any finite set $S = \{z_1, z_2, \cdots, z_n\} \subset \mathcal{Z}$, define *projection of* $\mathcal{F}$ *onto* S as

$$\mathcal{F}_S \stackrel{\text{def}}{=} \left\{ \big(f(z_1), \cdots, f(z_n)\big) : f \in \mathcal{F} \right\}$$

Note that $\mathcal{F}_S$ is a finite collection of binary vectors and $|\mathcal{F}_S| < 2^n$.

Shattering Number

$$s(\mathcal{F}, n) \stackrel{\text{def}}{=} \sup_{z_1, z_2, \cdots, z_n \in \mathcal{Z}} |\mathcal{F}_{z_1, z_2, \cdots, z_n}|$$

Shattering Number (Cont.)

Example

Consider instance space $\mathcal{Z} = \mathbb{R}$ and function class be positive rays $\mathcal{F} = \{\mathbb{1}\{z > t\} : \mathcal{Z} \mapsto \{0, 1\} | t \in \mathbb{R}\}$. Pick three real numbers $z_1 < z_2 < z_3$ and let $S = \{z_1, z_2, z_3\}$. Then

$$\mathcal{F}_S = \{(0, 0, 0), (0, 0, 1), (0, 1, 1), (1, 1, 1)\}$$

Vapnik-Chervonenkis (VC) Dimension

Vapnik-Chervonenkis (VC) Dimension

The *Vapnik-Chervonenkis (VC) Dimension* of a class of binary function $\mathcal{F}$ is defined as

$$d_{\mathrm{VC}}(\mathcal{F}) \stackrel{\mathrm{def}}{=} \sup\left\{n \in \mathbb{N} : s(\mathcal{F}, n) = 2^n\right\}$$

VC dimension can be think of as

- binary degrees of freedom, or
- estimate of effective number of parameters.

VC Dimension: Collection of Results

Binary Function Class $\mathcal{F}$	$d_{\mathrm{VC}}(\mathcal{F})$
Finite set with N elements	$\log_2(N)$
Intervals $[a, b] \subset \mathbb{R}$	2
Discs in $\mathbb{R}^2$	3
Closed balls in $\mathbb{R}^d$	$d + 2$
Rectangles in $\mathbb{R}^d$	$2d$
Half-spaces in $\mathbb{R}^d$	$d + 1$
Convex polygons in $\mathbb{R}^d$	∞

VC Dimension: Example

VC dimension works like this:

- You choose the points,
- then the adversary chooses the labeling such that
- there always exists a function f that can correctly classify the specified labeling of those points.

If you are able to succeed for all labelings of the adversary, we say that the VC dimension of binary function class $\mathcal{F}$ is at least the number of points you were able to choose.

VC Dimension of Intervals $[a, b]$ on $\mathbb{R}$ is 2

- Any two-point set can be shattered
- No three-point set can be shattered (Consider $\{(a, 1), (b, 0), (c, 1)\}$ with $a < b < c$)

Finite VC Dimensional Function Classes Has Bounded Shattering Number

If the VC dimension is finite, then the growth function cannot grow too quickly. In fact, there is a phase transition:

- for $n < d$, $s(\mathcal{F}, n) = 2^n$, and
- then the growth switches to polynomial $s(\mathcal{F}, n) \sim \mathcal{O}(n^d)$

Sauer-Shelah Lemma

Suppose $d_{\text{VC}}(\mathcal{F}) = d$, then

$$s(\mathcal{F}, n) \leq \sum_{i=0}^{d} \binom{n}{i}$$

In particular, when $n \geq d$, $s(\mathcal{F}, n) \leq \left(\frac{en}{d}\right)^d$.

Application: Bound Generalization Error

For a binary function class $\mathcal{F}$ that is not too complex, consistency of generalization is guaranteed. One can think of it as a bias-variance trade-off:

- Bias: training error $P_n(f)$
- Variance: "confidence internal" on the right-hand side of the inequality

Suppose $d_{\mathrm{VC}}(\mathcal{F}) < \infty$, then with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \leq \sqrt{\frac{8}{n}\left(\ln\left(\frac{4}{\delta}\right) + d_{\mathrm{VC}}(\mathcal{F})\left(1 + \ln\left(\frac{n}{d_{\mathrm{VC}}(\mathcal{F})}\right)\right)\right)}$$

Rademacher Complexity

Rademacher Complexity of $\mathcal{F}$

Consider *Rademacher random variables* $\sigma_i \overset{\text{iid}}{\sim} \text{Uniform}(\{-1, 1\})$, define

$$\text{Rad}_n(\mathcal{F}) \overset{\text{def}}{=} \mathbb{E}_Z \left[\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \right| \right] \right]$$

Rademacher complexity generalizes correlation. Notice that

$$\widehat{R}_n(f) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_i \neq f(X_i)\} = \frac{1}{2} - \frac{1}{2n} \sum_{i=1}^{n} Y_i f(X_i)$$

This gives us Rademacher correlation: what's the best that an algorithm can do on random labels.

Rademacher Complexity Extends Shattering Number and VC Dimension

Rademacher Complexity is defined on a class of bounded functions.

In the special case where $\mathcal{F}$ is a class of binary functions,

$$\text{Rad}_n(\mathcal{F}) \leq \sqrt{\frac{2}{n} \ln\left(s(\mathcal{F}, n)\right)} \qquad \forall n \in \mathbb{N}$$

Devroye and Lugosi (2001)

Suppose $d_{\text{VC}}(\mathcal{F}) < \infty$, then $\exists$ universal constant $C > 0$,

$$\text{Rad}_n(\mathcal{F}) \leq C\sqrt{\frac{d_{\text{VC}}(\mathcal{F})}{n}}$$

Application: Bound Generalization Error

With probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| \leq 2\mathrm{Rad}_n(\mathcal{F}) + \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\delta}\right)}$$

Combining everything together, we have showed a generalization error bound for any learning algorithm $\mathcal{F}$ using ERM

With probability at least $1-\delta$,

$$R(\widehat{f}_*) \leq R(f_*) + 4C\sqrt{\frac{d_{\mathrm{VC}}(\mathcal{F})}{n}} + \sqrt{\frac{2}{n}\ln\left(\frac{2}{\delta}\right)}$$

Concentration of measure is a vast and still growing area. Literature contains many refinements and extensions of these results

Questions?