Lec 24-25 Concentration "crash course"

"generalization"

Why?

$R(\hat{f})$ ← alg

Concentation:

well-behaved functions of r.v.'s behave as some simple statistic.

best reference:
lecture notes
ramon van handel
APC550

E.g. Lipschitz function, gaussian r.v.'s,
→ mean

Lipschitz + convex ← false without.
function → mean

bounded r.v.'s

λ_{max}(Gaussian matrix) concentrates

Question:

* Classic
(always fa

$\hat{R}(\hat{f}$

7.7K

h course"

Why?

$$R(\hat{f}) - R(g) = \boxed{R(\hat{f}) - \hat{R}(\hat{f})} + \hat{R}(\hat{f}) - \hat{R}(f) + \hat{R}(f) - R(f) + R(f) - R(g)$$

$\hat{f}$: alg

$\hat{R}(\hat{f}) - \hat{R}(f)$ ✓ opt

$\hat{R}(f) - R(f)$: reference in ? 于

$R(f) - R(g)$ ✓ apx

Question: why $R(\hat{f}) - \hat{R}(\hat{f})$ small?

* Classical view: (always false)

R & $\hat{R}$ are pop/emp risks

$R = \mathbb{E}\, \ell(y, \hat{f}(x))$

$\hat{R} = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \hat{f}(x_i))$

iid, same dist.

$\hat{R}(\hat{f}) \xrightarrow{n\to\infty} R(\hat{f})$ via "super-powered LLN"

* Modern perspec

* Interpolation

Motivation: DL e

[all prominent s

unclear if curren

* OOD (out

test dist ≠ trai

(e.g., alphafold)

* Classical view is

(e.g., "rethinking g

* Transfer learning

$(\hat{f})$?

$f)$ ✓ opt

reference in

$f)$? ✗

) ✓ apx

nall?

pop/emp ratio

$\ell(y_i, \hat{f}(x_i))$

powered

N "

* Modern perspectives:

* Interpolation $\hat{R}(\hat{f}) \approx 0 \ll R(\hat{f}) \approx \inf_g R(g)$

motivation: DL experiments (Srebro et al '14, Zhang '16)

[all prominent stat/theorists working on this; unclear if currently true.]

* OOD (out-of-domain/distributive) — test dist ≠ train dist (e.g., alphafold) — Ristoski, Zhao

* Classical view is broken (e.g., "rethinking generalization" (Zhang '16))

* Transfer learning — Raghunathan

Classical view

General form:

$R(f)$

Rem. * We'll d

* $\frac{1}{n}$ ro

* "Uniform deviation

R(g)

)

eski

Classical view

"uniform term"
modern/popular view:
this is an inf

General form: w/ pr $\geq 1-\delta$, $\forall f \in \mathcal{F}$

$$R(f) \leq \hat{R}(f) + \sqrt{\frac{1 \cdots \delta}{n}} + \sqrt{\frac{\text{comp}(\mathcal{F})}{n}}$$

Rem. * We'll define comp($\mathcal{F}$) later

* $\frac{1}{n}$ not $\frac{1}{\sqrt{n}}$ is possible, not really relevant to us.

* "Uniform deviation" because uniform if f (we'll explain later)

Theorem (Hoeffding).

Suppose $(Z_1 \dots Z_n)$ ind, $Z_i \in [a_i, b_i]$ a.s.

Then $\Pr\left[\frac{1}{n}\sum (Z_i - \mathbb{E}Z_i) \geq \varepsilon\right] \leq \exp\left(-\frac{2n^2\varepsilon^2}{\sum_i (b_i - a_i)^2}\right)$ (2)

alternatively, w/ prob $\geq 1-\delta$, $\frac{1}{n}\sum_i \mathbb{E}Z_i \leq \frac{1}{n}\sum_i Z_i + \sqrt{\frac{\sum_i (a_i - b_i)^2}{2n^2} \ln\frac{1}{\delta}}$

$\Rightarrow$

Plan: (a) use it for $\hat{R}(\hat{f}) - R(\hat{f})$

(b) sketch proof

$R(\hat{f}) - R(f^*)$ (1) Given fixed f, $\underset{\text{i.i.d.}}{\text{data}} \; (x_i, y_i)_{i=1}^n$

define $Z_i := \ell(y_i, f(x_i))$ assume $\in [a,b]$

$\Rightarrow$ w/ prob $\geq 1-\delta$, $R(f) = \frac{1}{n}\sum \mathbb{E}Z_i \leq \frac{1}{n}\sum_i Z_i + \sqrt{\frac{n(b-a)^2}{2n^2} \ln\frac{1}{\delta}}$

$= \hat{R}(f) + (b-a)\sqrt{\frac{\ln 1/\delta}{2n}}$.

Remark

* w
(
Alg
$\hat{f}(x$

$\exp\left(-\frac{2n^2\varepsilon}{\sum_i (b_i-a_i)^2}\right)$

$\ldots Z_i + \sqrt{\frac{\sum_i (a_i-b_i)^2}{2n^2} \ln\frac{1}{\delta}}$

$\ldots (x_i, y_i))_{i=1}^n$ assume $\in [a,b]$

$\ldots \sum_i Z_i + \sqrt{\frac{n(b-a)^2}{2n^2} \ln\frac{1}{\delta}}$

$\ldots -a)\sqrt{\frac{\ln 1/\delta}{2n}}$.

---

② Given $\mathcal{F}$ with $|\mathcal{F}| < \infty$,

define $\delta_f := \frac{\delta}{|\mathcal{F}|}$

for any $f \in \mathcal{F}$ w/ $pr \geq 1-\delta_f$, $R(f) \leq \hat{R}(f) + (b-a)\sqrt{\frac{\ln|\mathcal{F}| + \ln\frac{1}{\delta_f}}{2n}}$

$\Rightarrow$ via union bound

w/ $pr \geq 1-\delta = 1-\sum_{f\in\mathcal{F}} \delta_f$, $\forall f \in \mathcal{F}$,

$R(f) \leq \hat{R}(f) + (b-a)\sqrt{\frac{\ln|\mathcal{F}| + \ln 1/\delta}{2n}}$.

Remark * $\ln|\mathcal{F}|$ was our $\text{comp}(\mathcal{F})$

* Why did need union bound?

(Why not apply Hoeffding to output f_S?)

Alg output depends data $\Rightarrow$ $(Z_1, \ldots Z_n)$ might be dependent

$$\hat{f}(x) = \begin{cases} y_i & \text{if } x = x_i \\ 0 & \text{o.w.} \end{cases}$$

Can construct joint pr on $X \times Y$ s.t. $\hat{R}(f) \xrightarrow{n\to\infty} R(f)$, but also $\hat{R}(f) = 0 \neq 1 = R(f)$.

---

* Fixes:

(a) use cross-valid

(b) alg-specific an

(c) Brute force / uniform control all poss

How to prove Hoeffding

Definition. rv Z is σ^2-

(sub-Gaussian with vari

if $\forall t \in \mathbb{R}$, $\mathbb{E}\exp($

Rem. * $\exists$ 4 equivalent definitio

(vershynin hdp

* $\exists$ other examples of "

(su

$\ln \frac{1}{\delta_f}$

$\frac{\ln|\mathcal{F}| + \ln 1/\delta}{2n}$

$\sqrt{\frac{\ln|\mathcal{F}| + \ln 1/\delta}{2n}}$

* Fixes

(a) use cross-validation error

(b) alg-specific analysis (typically ...)

(c) Brute force / uniform deviation: control all possible outputs of alg.

How to prove Hoeffding

<u>Definition.</u> rv Z is σ^2-sub-Gaussian (sub-Gaussian with variance proxy σ^2)

if $\forall t \in \mathbb{R}$. $\mathbb{E}\exp(t(X - \mathbb{E}X)) \le \exp\left(\frac{t^2\sigma^2}{2}\right)$

<u>Rem.</u> * $\exists$ 4 equivalent definitions (vershynin hdp book)

* $\exists$ other examples of "well concentrated r.v.s" (sub-exponential)

...mples.

* $\mathcal{N}(\mu, \sigma^2)$

* $X \in [a,b]$ a.s.

One subcase.

$\mathbb{E}\exp(tX) = \frac{1}{2}$

$= \frac{1}{2}$

* If $(X_1, \dots, X_n)$

$\Rightarrow \frac{1}{n}\sum X_i$ is

...?)

... be dependent

joint pr on $\mathcal{X} \times \mathcal{Y}$

... $\hat{R}(f) = 0 \neq 1 = R(f)$.

Examples.

* $\mathcal{N}(\mu, \sigma^2)$ is σ^2 - sub-Gaussian (complete the square)

* $X \in [a,b]$ a.s. then X is $\left(\frac{b-a}{2}\right)^2$ - sub-Gaussian.

  One subcase. if $X \sim \mathrm{Unif}\{-1,+1\}$

  $\mathbb{E}\exp(tX) = \frac{1}{2}\left[\exp(-t) + \exp(t)\right]$

  $= \frac{1}{2}\left[\sum \frac{(-t)^i}{i!} + \sum \frac{(t)^i}{i!}\right] = \frac{1}{2}\cdot 2 \cdot \sum_{i\geq 0} \frac{t^{2i}}{(2i)!}$

  $\leq \sum \frac{(t^2)^i}{2^i \cdot i!} = \sum_{i\geq 0} \frac{(t^2/2)^i}{i!}$

  $= \exp(t^2/2)$. /

* If $(X_1, \ldots, X_n)$ are σ^2-subgaussian & independent

  $\Rightarrow \frac{1}{n}\sum X_i$ is $\frac{\sum \sigma_i^2}{n^2}$-sub-Gaussian

  (behaves like $\frac{1}{n}$)

"Concentration phenomenon"

$\exp\left(\frac{t^2\sigma^2}{2}\right)$.

one p...

...tly.

...ntal r.v.s"