

PS2 due next week (10/15)

- One more P.S. before midterm 11/5

Last time: finished online-to-batch
started probability bounds on large deviations

- Chebyshev's inequality
- log mgf

Today: Finish probability bounds
Consistent Hypothesis (CH) alg.
Empirical Risk Minimization (ERM) alg

Log moment generating function aka cumulant generating function

For r.v. z , log mgf is

$$K_z(\lambda) = \ln E[\exp(\lambda z)] \quad (\text{might be } \infty \text{ at some } \lambda \neq 0)$$

$$K'_z(\lambda) = E[z_\lambda] \quad \text{where } z_\lambda \sim p_z(z) \exp(\lambda z - K_z(\lambda))$$

$$K'_z(0) = E[z]$$

$$K''_z(\lambda) = \text{var}(z_\lambda)$$

$$K''_z(0) = \text{var}(z)$$

$\vdots$

Examples: $z \sim \text{Normal}(\mu, \sigma^2)$: $K_z(\lambda) = \mu\lambda + \frac{\sigma^2 \lambda^2}{2}$

$$z \sim \text{Bernoulli}(\mu) : K_z(\lambda) = \ln(1 + \mu(e^\lambda - 1))$$

$$z \sim \text{Poisson}(\mu) : K_z(\lambda) = \mu(e^\lambda - 1)$$

* If $z = x_1 + \dots + x_n$ is sum of independent r.v.'s $x_1, \dots, x_n$, then:

$$K_z(\lambda) = K_{x_1}(\lambda) + \dots + K_{x_n}(\lambda)$$

Proposition (Cramér - Chernoff method)

For $t > E(z)$,

$$\Pr(z > t) \leq \exp(-\sup_{\lambda > 0} \{t\lambda - K_z(\lambda)\})$$

For $t < E(z)$,

$$\Pr(z < t) \leq \exp(-\sup_{\lambda < 0} \{t\lambda - K_z(\lambda)\})$$

Proof: Markov's inequality!

$$\Pr(\boxed{\exp(\lambda z)} > \exp(\lambda t)) \leq \frac{E[\exp(\lambda z)]}{\exp(\lambda t)}$$

$\vdots$

This $\sup_{\lambda > 0} \{t\lambda - K_z(\lambda)\}$ can be replaced with

$$\sup_{\lambda \in \mathbb{R}} \{t\lambda - K_z(\lambda)\} =: K_z^*(t) \quad \text{"Fenchel conjugate of } K_z(\lambda)\text{"}$$

a.k.a. "rate function"
"Cramér transform"

examples: $N(\mu, \sigma^2)$: $K_z^*(t) = \sup_{\lambda \in \mathbb{R}} \{t\lambda - (\mu\lambda + \frac{\sigma^2 \lambda^2}{2})\}$
 $= \frac{(t - \mu)^2}{2\sigma^2}$

$$K_{z-E(z)}^*(t) = \frac{t^2}{2\sigma^2}$$

$$\text{Bernoulli}(\mu) : K_{z-E(z)}^*(t) = RE(\text{Bern}(\mu+t), \text{Bern}(\mu))$$

$$\geq 2(\mu+t-\mu)^2 = 2t^2 \quad \text{by Pinsker's inequality}$$

$$\text{Poisson}(\mu) : K_{z-E(z)}^*(t) = \mu \cdot h_B\left(\frac{t}{\mu}\right)$$

$$h_B(a) = (1+a) \ln(1+a) - a$$

Ex $z = x_1 + \dots + x_n$ sum of n iid rv's $E(x_i) = 0$

$$K_z(\lambda) = \sum_{i=1}^n K_{x_i}(\lambda)$$

$$= n \cdot K_{x_1}(\lambda)$$

$$K_z^*(t) = \sup_{\lambda \in \mathbb{R}} \{t\lambda - n K_{x_1}(\lambda)\}$$

$$= n \cdot \sup_{\lambda \in \mathbb{R}} \left\{ \frac{t}{n} \cdot \lambda - K_{x_1}(\lambda) \right\}$$

$$= n \cdot K_{x_1}^*\left(\frac{t}{n}\right)$$

$$\Pr\left(\frac{z}{n} \geq \underline{\varepsilon}\right) \leq \exp(-K_z^*(\varepsilon n)) \\ = \exp(-n \cdot K_{x_1}^*(\varepsilon))$$

$$x_1 + \mu \sim \text{Bern}(\mu)$$

$$\Rightarrow K_{x_1}^*(\varepsilon) \geq 2\varepsilon^2$$

$z = \text{Sum of } n \text{ iid Bern}(\mu) \text{ rv's}$

$$\Pr(z \geq (\mu + \varepsilon)n) \leq \exp(-2n\varepsilon^2)$$

With prob. $\geq 1 - \delta$,

$$\frac{z}{n} \leq \mu + \sqrt{\frac{\ln(1/\delta)}{2n}}$$

Poisson heuristic

$$z \sim \text{Binomial}(n, p) \xrightarrow{n \rightarrow \infty} N(np, np(1-p))$$

$$z \sim \text{Binomial}(n, p) \longrightarrow \text{Poisson}(np) \\ p = \Theta\left(\frac{1}{n}\right)$$

PAC learning a concept class C by minimizing training error rate

Key: Union bound

$$\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$$

$$\begin{aligned}\Pr(\text{neither } A \text{ or } B \text{ happen}) &= 1 - \Pr(A \cup B) \\ &\geq 1 - (\Pr(A) + \Pr(B))\end{aligned}$$

Consistent hypothesis (CH) $H \subseteq Y^X$

- Given: $((x_1, y_1), \dots, (x_T, y_T)) \in (X \times Y)^T$ training data
- Return $\hat{h} \in H$ that satisfies $\hat{h}(x_t) = y_t \quad \forall t$
consistent w/ training data
or $\perp$ if no such $\hat{h}$ exists

Thm: If $T \geq \frac{\ln(|H|/\delta)}{\epsilon}$ and $C \subseteq H$

then CH (ϵ, δ)-PAC learns C .

Ex MCONJ_n - can use AE.

Proof: For each $h \in H$ s.t. $\text{err}(h; P) > \epsilon$,

B_h is the event that h is consistent w/ training data.

$$\begin{aligned}\Pr(B_h) &= (1 - \text{err}(h; P))^T \leq \exp(-\text{err}(h; P) \cdot T) \quad 1 - x \leq e^{-x} \\ &\leq \exp(-\epsilon T)\end{aligned}$$

$$\Pr\left(\bigcup_{\substack{h \text{ s.t.} \\ \text{err}(h; P) > \epsilon}} B_h\right) \leq \sum_{\substack{h \text{ s.t.} \\ \text{err}(h; P) > \epsilon}} \Pr(B_h) \leq |H| \cdot \exp(-\epsilon T) \leq \delta$$

$\Rightarrow \text{b/c } T \geq \frac{\ln(|H|/\delta)}{\epsilon}$

□

EMPIRICAL RISK MINIMIZATION (ERM)

Return $\hat{h} \in \mathcal{H}$ that has smallest training error rate

$$\text{err}(h; S) := \frac{1}{|S|} \sum_{(x, y) \in S} \mathbb{1}\{h(x) \neq y\}$$

Thm If $S \sim P^T$ for some dist P over $\mathcal{X} \times \mathcal{Y}$
then with prob. $\geq 1 - \delta$,

$$\text{err}(\hat{h}; P) \leq \min_{h \in \mathcal{H}} \text{err}(h; P) + \sqrt{\frac{2 \ln(2|\mathcal{H}|/\delta)}{T}}$$

eg. suppose $\epsilon \leq \epsilon/2$

and suppose T is large enough so $\epsilon \leq \epsilon/2$

then: ERM is (ϵ, δ) -PAC learner

Proof: For any $h \in \mathcal{H}$

$$\Pr \left(\left| T \cdot \text{err}(h; S) - T \cdot \text{err}(h; P) \right| \leq \sqrt{\frac{T \ln(1/\delta)}{2}} \right) \geq 1 - 2\delta$$

w.p. $\geq 1 - 2\delta$

$$\forall h \in \mathcal{H}, \quad |\text{err}(h; S) - \text{err}(h; P)| \leq \sqrt{\frac{\ln(1/\delta)}{2T}}$$

$$\text{err}(\hat{h}; P) - \text{err}(h^*; P) = \text{err}(\hat{h}; P) - \text{err}(\hat{h}; S)$$

$$+ \text{err}(\hat{h}; S) - \text{err}(h^*; S)$$

$$+ \text{err}(h^*; S) - \text{err}(h^*; P)$$

$$\leq 2 \cdot \sqrt{\frac{\ln(1/\delta)}{2T}} = \sqrt{\frac{2 \ln(2/\delta)}{T}}$$

Occam's Razor

Countable hypothesis classes

$h_1, h_2, \dots$

Choose the earliest hypothesis on this list that's consistent w/ training data.

with prob $\geq 1 - \delta$

$$\text{err}(h_i | S) = 0 \Rightarrow \text{err}(h_i | P) \leq \sqrt{\frac{\ln(i(i+1)/\delta)}{T}}$$