Lecture 20

Information-Theoretic Notions of Stability. Differential Privacy.

Sasha Rakhlin

Nov 19, 2018

Generalization

Denote data by $\mathcal{S} = \{Z_1, \ldots, Z_n\}$.

Recall the problem of bounding generalization gap

$$\mathbb{E}[\mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n)]$$

The difficulty is in dependence of $\widehat{f}_n$ on the data on which it is being evaluated. We saw that the difference is generally nonzero (bias). But how strong is the dependence of $\widehat{f}_n$ on data? If not too strong – the empirical loss should be roughly an unbiased estimate of the expected performance.

Today: information-theoretic notions of dependence of an algorithm on data. We have to consider randomized algorithms, however. This approach will not work with deterministic methods.

We shall frame the question of bias in the context of a more general question of bias in adaptive data analysis.

Consider a randomized algorithm $\widehat{f}_n[\mathcal{S}, \xi]$, where ξ is internal randomization of the method. Since the notation becomes cumbersome, let's shorten it to $W = \widehat{f}_n[\mathcal{S}, \xi]$ where W will be treated as a random variable taking values in an abstract set (say, a class of functions).

Let $P_{W|\mathcal{S}}$ denote the distribution of W conditionally on data.

Differential Privacy

Definition: a randomized algorithm is (ϵ, δ)-differentially private if

$$P_{W|\mathcal{S}=\mathbf{s}}(A) \leq e^{\epsilon} \cdot P_{W|\mathcal{S}=\mathbf{s}'}(A) + \delta$$

for any measurable A and any two datasets $\mathbf{s}$ and $\mathbf{s}'$ differing in one example.

Assume loss function is bounded in $[0, 1]$, unless otherwise specified.

Differential Privacy

For a differentially private method, for any given $\mathcal{S}$, z', and z,

$$\mathbb{E}_\xi\left[\ell(\widehat{f}_n[\mathcal{S}^{i,z'},\xi],z)\right] \le e^\epsilon \cdot \mathbb{E}_\xi\left[\ell(\widehat{f}_n[\mathcal{S},\xi],z)\right] + \delta$$

since loss is bounded by 1. Rearranging and using boundedness of loss,

$$\mathbb{E}_\xi\left[\ell(\widehat{f}_n[\mathcal{S}^{i,z'},\xi],z)\right] - \mathbb{E}_\xi\left[\ell(\widehat{f}_n[\mathcal{S},\xi],z)\right] \le (e^\epsilon - 1) + \delta$$

This is uniform stability (extended to randomized algorithms) from previous lecture. Note that $e^\epsilon - 1 \le 2\epsilon$ for $\epsilon \in [0,1]$.

Conclusion: differential privacy is a stronger notion than uniform stability (if loss function is bounded). In particular, differential privacy also yields a bound on generalization gap

$$\mathbb{E}\left[\mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n)\right]$$

Let us discuss a few weaker notions that still yield small generalization gap.

Notions of "distances"

Relative entropy:

$$D(P\|Q) = \mathbb{E}_P \log \frac{dP}{dQ}$$

Total variation distance:

$$d_{TV}(P,Q) = \sup_A |P(A) - Q(A)|$$

Max divergence:

$$D_\infty(P\|Q) = \sup_A \log \frac{P(A)}{Q(A)}$$

Approximate max divergence:

$$D_\infty^\delta(P\|Q) = \sup_{A:P(A)>\delta} \log \frac{P(A)-\delta}{Q(A)}$$

E_γ-divergence ($\gamma \geq 1$):

$$E_\gamma(P\|Q) = \sup_A P(A) - \gamma Q(A)$$

Relations among these "distances"

$$d_{TV}(P,Q) \le \sqrt{\frac{1}{2} D(P\|Q)}$$

$$D(P\|Q) \le D_\infty(P\|Q)$$

$$1 - \gamma(1 - d_{TV}(P,Q)) \le E_\gamma(P\|Q)$$

Corresponding notions of "independence"

(U, V) with joint law P_{UV} and marginals P_U, P_V.

Mutual information

$$I(U;V) = D(P_{UV} || P_U \otimes P_V)$$

T-information

$$T(U;V) = d_{TV}(P_{UV}, P_U \otimes P_V)$$

Max-information

$$I_\infty(U;V) = D_\infty(P_{UV} || P_U \otimes P_V)$$

Approximate max-information

$$I_\infty^\delta(U;V) = D_\infty^\delta(P_{UV} || P_U \otimes P_V)$$

Conditional notions of information are defined as integrated versions

$$I(U;V|Y) = \int P_Y(dy) D(P_{UV|Y=y} || P_{U|Y=y} \otimes P_{V|Y=y})$$

Differential Privacy

(ϵ, δ)-differential privacy can be written as

$$\mathsf{E}_{e^\epsilon}(\mathsf{P}_{W|\mathcal{S}=\mathbf{s}} || \mathsf{P}_{W|\mathcal{S}=\mathbf{s}'}) \leq \delta$$

or as

$$\mathsf{D}_\infty^\delta(\mathsf{P}_{W|\mathcal{S}=\mathbf{s}} || \mathsf{P}_{W|\mathcal{S}=\mathbf{s}'}) \leq \epsilon$$

for $\mathbf{s}, \mathbf{s}'$ differing in one example.

From differential Privacy to max-information

$(\epsilon, 0)$-differential privacy:

$$\log \frac{P_{W|\mathcal{S}=s}(A)}{P_{W|\mathcal{S}=s'}(A)} \leq \epsilon$$

where s, s' differ in one coordinate. Applying repeatedly,

$$\log \frac{P_{W|\mathcal{S}=s}(A)}{P_{W|\mathcal{S}=s''}(A)} \leq n\epsilon$$

for any two datasets s, s''. Then, trivially,

$$I_\infty(W; \mathcal{S}) \leq n\epsilon \cdot \log(e).$$

From differential Privacy to max-information

It is also possible to show (Dwork et al '15) a stronger result by giving up in δ:

$$I_\infty^\delta(W;\mathcal{S}) \leq O(n\epsilon^2 + \epsilon\sqrt{n\log(1/\delta)}).$$

Furthermore, (ϵ,δ)-differential privacy implies (Rogers et al '16)

$$I_\infty^\beta(W;\mathcal{S}) \leq O(n\epsilon^2 + n\sqrt{\delta/\epsilon}), \qquad \beta = O(n\sqrt{\delta/\epsilon})$$

Advantage of I_∞ over mutual info or other measures: possible to prove high probability bounds – see (Dwork et al '15).

Incremental notions of independence

Define *erasure mutual information*

$$\mathrm{I}^{-}(W;\mathcal{S}) = \sum_{i=1}^{n} \mathrm{I}(W;Z_i|\mathcal{S}^{-i})$$

where $\mathcal{S} = \{Z_1, \ldots, Z_n\}$ and $\mathcal{S}^{-i}$ is with Z_i removed.

Similarly, *erasure* T*-information* is

$$\mathsf{T}^{-}(W;\mathcal{S}) = \sum_{i=1}^{n} \mathsf{T}(W;Z_i|\mathcal{S}^{-i})$$

Upper bounds on generalization gap

If loss is bounded $0 \le \ell \le 1$,

$$\mathbb{E}\left[\mathbf{L}(W) - \widehat{\mathbf{L}}(W)\right] \le \frac{1}{n} T^{-}(W;\mathcal{S})$$

If loss $\ell(w, Z)$ is σ^2-subgaussian for any w (loss can be unbounded),

$$\mathbb{E}\left[\mathbf{L}(W) - \widehat{\mathbf{L}}(W)\right] \le \sqrt{\frac{2\sigma^2}{n} I(W;\mathcal{S})}$$

If data in $\mathcal{S}$ are assumed to be independent,

$$I(W;\mathcal{S}) \le I^{-}(W;\mathcal{S}).$$

If for any $\mathbf{s}, \mathbf{s}'$ differing in one example,

- $D(P_{W|\mathcal{S}=\mathbf{s}} \| P_{W|\mathcal{S}=\mathbf{s}'}) \le \epsilon$, then $I^{-}(W;\mathcal{S}) \le n\epsilon$
- $\frac{1}{n}\sum_{i=1}^{n} D(P_{W|\mathcal{S}=\mathbf{s}} \| P_{W|\mathcal{S}=\mathbf{s}^{-i}}) \le \epsilon$, then $I^{-}(W;\mathcal{S}) \le n\epsilon$
- $d_{TV}(P_{W|\mathcal{S}=\mathbf{s}} \| P_{W|\mathcal{S}=\mathbf{s}'}) \le \epsilon$, then $T^{-}(W;\mathcal{S}) \le n\epsilon$

Example: Gibbs algorithm

Suppose algorithm takes values in a set $\mathcal{F}$ and loss $0 \le \ell \le 1$. Let Q be some prior distribution on $\mathcal{F}$. For a fixed dataset $\mathbf{s}$, define the Gibbs measure

$$P_{W|\mathcal{S}=\mathbf{s}}(dw) \propto \exp\left\{-\beta \widehat{\mathbf{L}}_{\mathbf{s}}(w)\right\} Q(dw)$$

Large β means distribution is focused on minimum loss (ERM). On the other hand $\beta = 0$ means no dependence on data. We should expect bounds on generalization gap to be vacuous when β is too large, and 0 when $\beta = 0$.

A direct computation shows that

$$D(P_{W|\mathcal{S}=\mathbf{s}} || P_{W|\mathcal{S}=\mathbf{s}'}) \le \frac{\beta^2}{2n^2}$$

whenever $\mathbf{s}, \mathbf{s}'$ differ in one example. Hence,

$$I^{-}(W;\mathcal{S}) \le \frac{\beta^2}{2n}$$

Since subgaussianity parameter $\sigma^2 = 1/4$, we get

$$\mathbb{E}\left[\mathbf{L}(W) - \widehat{\mathbf{L}}(W)\right] \le \sqrt{\frac{2\sigma^2\beta^2}{2n^2}} = \frac{\beta}{2n}$$

Decoupling lemma

Mutual information upper bound on generalization gap is a consequence of the following general lemma.

Let U, V be random vars with joint P_{UV}. Let $\bar{U}, \bar{V} \sim P_U \otimes P_V$ be independent copies from marginals. Assume $f(u, V)$ is σ^2-subgaussian for any u. Then

$$|\mathbb{E}f(U, V) - \mathbb{E}f(\bar{U}, \bar{V})| \leq \sqrt{2\sigma^2 I(U; V)}$$

This is a consequence of Donsker-Varadhan.

Adaptive composition

Suppose we execute m algorithm sequentially, with input to each algorithm $\mathcal{A}_i$ being the output of all previous $\mathcal{A}_j$, $j \leq i-1$, and the dataset $\mathcal{S}$.

If each algorithm $\mathcal{A}_i$ is (ϵ_i, δ_i)-differentially private, then the composition is $(\sum \epsilon_i, \sum \delta_i)$-differentially private. This is "linear" composition.

Advanced composition: m-fold composition of (ϵ, δ) diff private mechanisms, enjoys $(\epsilon', m\delta + \delta')$ differential privacy for $\epsilon' = \sqrt{2m \ln(1/\delta')}\epsilon + m\epsilon(e^{\epsilon} - 1)$. That is, *sublinear* composition for small ϵ.

Composition (of the linear type) also holds for max-information $I_\infty(\,)$, erasure mutual information $I^-(\,)$, and other KL-based notions of stability from previous slide.

See the survey by (Dwork and Roth) for more information (and other refs at the end of these slides).

Adaptive data analysis

Decoupling lemma can also be used in a slightly different context of adaptive data analysis. Imagine we compute various statistics

$$\Phi = (\phi_1(\mathcal{S}), \ldots, \phi_m(\mathcal{S}))$$

(not necessarily average loss). Let W be a (possibly randomized) selection rule in values in $\{1, \ldots, m\}$. Then

$$|\mathbb{E}_{(W,\mathcal{S})}[\phi_W(\mathcal{S})] - \mathbb{E}_W[\mu_W]|$$

is the bias of our procedure, where $\mu_j = \mathbb{E}\phi_j(\mathcal{S})$.

- The quantity $\mathbb{E}[\phi_W(\mathcal{S})]$ is the expected value of the statistic chosen after seeing the data $\mathcal{S}$. Expectation is over joint $(W, \mathcal{S})$. This can introduce bias.
- $\mathbb{E}_W[\mu_W]$ is the decoupled version – "future performance" of the chosen statistic. Imagine we had a fresh dataset $\bar{\mathcal{S}}$. Then $\phi_W(\bar{\mathcal{S}})$ would an unbiased estimate of $\mathbb{E}_W[\mu_W]$.

Data Split: Train-Validate-Test

One often trains multiple models on a given training dataset and evaluates each model on a validation set (this is the collection of ϕ_i's). The model with best performance is then reported. The result is precisely the bias described above.

To report an unbiased estimate of true expected performance, one needs to leave out a (third) test dataset and evaluate on this fresh dataset only once (e.g. 5 min before the paper deadline). Of course, it is difficult to police this practice.

Adaptive data analysis

Lemma: if $\phi_i - \mu_i$ is σ^2-subgaussian,

$$|\mathbb{E}[\phi_W(\mathcal{S})] - \mathbb{E}_W[\mu_W]| \le \sqrt{2\sigma^2 I(W;\Phi)}$$

Note that

$$I(W;\Phi) \le I(W;\mathcal{S})$$

by data-processing inequality.

Proof of Lemma: use Decoupling Lemma with $U = W$, $V = \mathcal{S}$, and $f(W,\mathcal{S}) = \phi_W(\mathcal{S})$. Then

$$\mathbb{E}[f(W,\mathcal{S})] = \mathbb{E}\phi_W(\mathcal{S}), \quad \mathbb{E}[f(\bar{W},\bar{\mathcal{S}})] = \mathbb{E}\phi_{\bar{W}}(\bar{\mathcal{S}}) = \mathbb{E}\mu_W$$

References

The Algorithmic Foundations of Differential Privacy by Dwork and Roth, 2014.

Generalization in Adaptive Data Analysis and Holdout Reuse by Dwork et al, 2015

Calibrating Noise to Variance in Adaptive Data Analysis by Feldman and Steinke, 2018

Max-Information, Differential Privacy, and Post-Selection Hypothesis Testing by Rogers et al, 2016

Controlling Bias in Adaptive Data Analysis Using Information Theory by Russo and Zou, 2016

Overview chapter *Information-Theoretic Stability and Generalization* by Raginsky et al, 2018