Infotheory for Statistics and Learning

Lecture 6

- Basic learning theory [BBL:1–3,HDGR:1,XR]
- Generalization error [HDGR:1,XR]
- Information bounds on generalization error [HDGR:2–4,XR]
- Complexity, information and generalization [BBL:4,HDGR:7]

Learning an Estimator

Consider the general setup: $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, we observe $X \sim P_\theta$ and want to estimate $T(\theta)$ as $\hat{T}$

The decision rule is a kernel $P_{\hat{T}|X=x}$ and the risk is

$$R_\theta(\hat{T}) = \int \left\{ \int \ell(T(\theta), \hat{T}) dP_{\hat{T}|X=x} \right\} dP_\theta$$

Let $Z = (X, T(\theta))$ for $X \sim P_\theta$, that is, knowing Z we know both X and the correct value of $T(\theta)$

For θ deterministic Z is described by P_θ, and with a prior π we have $P_Z = (P_\theta \otimes P_{T|\theta}) \circ \pi$

In either case, let Q be the resulting distribution for Z

Assume $P_{\hat{T}|X}$ is deterministic (for simplicity, can be generalized), and that $s = \hat{T}(x) \in E$ for a (standard Borel) space $(E, \mathcal{E})$

Let $\ell(s, z)$ be the associated cost; e.g. if $T(\theta) = \theta$ and $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, then $\ell(s, z) = \ell(\hat{\theta}(x), (x, \theta)) = (\theta - \hat{\theta}(x))^2$

Define

$$L_Q(s) = E_Q[\ell(s, Z)] = \int \ell(s, z) Q(dz)$$

- The true risk (knowing Q) when using $\hat{T}(x) = s$

Assume $\ell(s, z)$ is chosen such that $L_Q(s) = R_\theta(s)$ for θ deterministic and $L_Q(s) = R_\pi(s)$ with a prior $\theta \sim \pi$, e.g.

$$\ell(s, z) = (\theta - \hat{\theta}(x))^2 \Rightarrow$$
$$L_Q(s) = E_\theta[(\theta - \hat{\theta}(X))^2] \text{ or } E_\pi[E_\theta[(\theta - \hat{\theta}(X))^2]]$$

Assume now that Q is unknown but we have access to Z_i iid $\sim Q$ for $i = 1, \ldots, n$, the training samples

Let $Z^n = (Z_1, \ldots, Z_n) \sim P_{Z^n} = Q^{\otimes n}$ (n-fold product) and consider a kernel $P_{S|Z^n}$, randomly assigning $\hat{T}(x) \in E$ for $Z^n = z^n$

For a given learning algorithm $P_{S|Z^n}$, the resulting $L_Q(S)$ is a random variable with distribution determined by $P_S = P_{S|Z^n} \circ P_{Z^n}$

The probability space for S is $(E, \mathcal{E}, P_S)$, for each hypothesis $s \in E$

Define the empirical loss for hypothesis s

$$L_{Z^n}(s) = \frac{1}{n} \sum_{i=1}^{n} \ell(s, Z_i)$$

- Goal is to minimize $L_Q(s)$ but we can only compute $L_{Z^n}(s)$

So far $Z_i = (X_i, T(\theta_i))$ (or $Z_i = (X_i, T_i)$) $\Rightarrow$ supervised learning

We can also have $Z_i = X_i \Rightarrow$ unsupervised learning

Example: $\theta \in \mathbb{R}^{p\times 1}$, $x \in \mathbb{R}^{p\times 1}$, $Z_i = (X_i, \theta_i)$ with $X_i \sim P_\theta$, $\theta_i \sim \pi$ and using $\ell(s,z) = \ell(s,(x,\theta)) = \|\theta - \hat\theta(x)\|^2$. We do not know P_θ or π, and cannot compute $R_\pi = E[\|\theta - \hat\theta(X)\|^2]$. Choose instead $\hat\theta(x)$ to minimize

$$L_{Z^n}(s) = \frac{1}{n}\sum_{i=1}^{n} \|\theta_i - \hat\theta(X_i)\|^2$$

over $E = \{\text{linear estimators } \hat\theta = Ax\}$. With

$$R(Z^n) = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T \quad \text{and} \quad F(Z^n) = \frac{1}{n}\sum_{i=1}^{n} \theta_i X_i^T$$

we get

$$L_{Z^n}(s) = \mathrm{Tr}\left\{(A - FR^{-1})R(A-FR^{-1})^T + \frac{1}{n}\sum_{i=1}^{n}\theta_i\theta_i^T - FR^{-1}F^T\right\}$$

$$\Rightarrow P_{S|Z^n} \to s(Z^n) = F(Z^n)(R(Z^n))^{-1}\, x$$

Define the (expected) generalization error

$$G(Q, P_{S|Z^n}) = E[L_Q(S) - L_{Z^n}(S)]$$

That is, on average

$$L_Q(S) \approx L_{Z^n}(S)$$

if $G(Q, P_{S|Z^n})$ is small

Assume that there exists an $s^* \in E$ such that

$$\inf_{s\in E} L_Q(s) = L_Q(s^*)$$

then

$$E[L_Q(S)] - L_Q(s^*) = G(Q, P_{S|Z^n}) + E[L_{Z^n}(S) - L_{Z^n}(s^*)]$$

Thus $E[L_Q(S)] \approx L_Q(s^*)$ if both

$$G(Q, P_{S|Z^n}) \approx 0 \quad \text{and} \quad E[L_{Z^n}(S)] \approx E[L_{Z^n}(s^*)]$$

Sensitivity of $P_{S|Z^n}$ to the training samples:

Assume $Z^n \sim P_{Z^n}$ and $\tilde{Z}^n \sim P_{Z^n}$ independent of Z^n

If S was generated from Z^n (via $P_{S|Z^n}$), then

$$E[\ell(S,\tilde{Z}_i)] = \int E\left[\int \ell(S,z)Q(dz)\middle| Z^n = z^n\right] dP_{Z^n} = E[L_Q(S)]$$

and consequently

$$\begin{aligned} G(Q,P_{S|Z^n}) &= \frac{1}{n}\sum_{i=1}^{n} E[\ell(S,\tilde{Z}_i) - \ell(S,Z_i)] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[\ell(S,\tilde{Z}_i) - \ell(S^{(i)},\tilde{Z}_i)] \end{aligned}$$

where $S^{(i)}$ was generated from $(Z_1,\dots,Z_{i-1},\tilde{Z}_i,Z_{i+1},\dots,Z_n)$

Thus $G(Q,P_{S|Z^n})$ is small if

$$\frac{1}{n}\sum_{i=1}^{n} \ell(S,\tilde{Z}_i) \approx \frac{1}{n}\sum_{i=1}^{n} \ell(S^{(i)},\tilde{Z}_i)$$

That is, if $P_{S|Z^n}$ is stable in the sense that S is not sensitive to local modification of the training samples

The algorithm $P_{S|Z^n}$ is ε stable if

$$D(P_{S|Z^n=z^n} \| P_{S|Z^n=v^n}) \le \varepsilon$$

for all z^n and v^n that differ in one sample; does not depend on Q

The algorithm $P_{S|Z^n}$ is ε information stable w.r.t. Q if

$$I(S;Z^n) \le n\varepsilon$$

ε stable $\Rightarrow$ ε information stable for any Q

Information Bounds on Generalization Error

For a RV X with $m = E[X]$, its logarithmic moment-generating function is

$$\psi(\lambda) = \ln E\left[e^{\lambda(X-m)}\right], \quad \lambda \in \mathbb{R}$$

$\psi(\lambda)$ is convex, $\psi(0) = \psi'(0) = 0$

We have the Chernoff bound, $\Pr(X \geq m + t) \leq e^{-\hat{\psi}(t)}$, for any $t > 0$, where

$$\hat{\psi}(t) = \sup_{\lambda \geq 0} \{\lambda t - \psi(\lambda)\}$$

is the Cramér transform of $\psi(\lambda)$

$\hat{\psi}(\lambda)$ is nonnegative, convex and non-decreasing on $[0, \infty)$

For a general $f(x) : \mathbb{R} \to \mathbb{R}$, we also define the Fenchel–Legendre dual

$$f^*(y) = \sup_x \{yx - f(x)\}, \quad f^*(y) \text{ is convex}$$

Assume we can find convex functions $\psi_1(\lambda)$ and $\psi_2(\lambda)$, with $\psi_i(0) = \psi_i'(0) = 0$ and such that for $\lambda > 0$

$$\sup_{s \in E} E\left[e^{-\lambda(\ell(s,Z) - L_Q(s))}\right] \leq e^{\psi_1(-\lambda)}, \quad \sup_{s \in E} E\left[e^{\lambda(\ell(s,Z) - L_Q(s))}\right] \leq e^{\psi_2(\lambda)}$$

Then[1], for any $P_{S|Z^n}$ such that $I(S; Z^n) < \infty$

$$\hat{\psi}_2^{-1}\left(\frac{1}{n} I(S; Z^n)\right) \leq G(Q, P_{S|Z^n}) \leq \hat{\psi}_1^{-1}\left(\frac{1}{n} I(S; Z^n)\right)$$

The function $\hat{\psi}_i^{-1}$ is concave. If it is also non-decreasing, then if $P_{S|Z^n}$ is ε stable, and/or $P_{S|Z^n}$ is ε information stable for Q,

$$G(Q, P_{S|Z^n}) \leq \hat{\psi}_1^{-1}(\varepsilon)$$

[1] [J-H-W] Jiao, Han and Weissman, *IEEE ISIT 2017*

The proof (see [XR] and [J-H-W] for details) relies on the following lemma, of general value:

Lemma: For X and Y with joint distribution P_{XY} and marginals P_X and P_Y, and $f(X,Y)$ real-valued such that

$$\sup_x \ln E\left[e^{\lambda(f(x,Y)-E[f(x,Y)])}\right] \leq \psi_2(\lambda), \quad \lambda > 0$$
$$\sup_x \ln E\left[e^{\lambda(f(x,Y)-E[f(x,Y)])}\right] \leq \psi_1(\lambda), \quad \lambda < 0$$

for $\psi_i(\lambda)$ convex and $\psi(0) = \psi'(0) = 0$, then

$$E[f(X,Y)] \leq \int f(x,y)d(P_X \otimes P_Y) + \hat{\psi}_2^{-1}(I(X;Y))$$
$$E[f(X,Y)] \geq \int f(x,y)d(P_X \otimes P_Y) - \hat{\psi}_1^{-1}(I(X;Y))$$

The proof of the lemma relies on the following observations:

For any convex $\phi(x)$ with $\phi(0) = \phi'(0) = 0$, the transform $\hat{\phi}(y) = \sup_{x\geq 0}(yx - \phi(x))$ has an inverse $\hat{\phi}^{-1}(y)$ that can be written as

$$\hat{\phi}^{-1}(y) = \inf_{\lambda>0} \frac{y + \phi(\lambda)}{\lambda}$$

From the Donsker–Varadhan Lemma (more about this next lecture), we have $D(P_{Y|X=x}\|P_Y) \geq \lambda E[f(x,Y)|X=x] - \ln E[e^{\lambda f(x,Y)}]$ and by assumption $\ln E[e^{\lambda f(x,Y)}] \leq \psi(\lambda) + \lambda E[f(x,Y)]$

Hence

$$E[f(x,Y)|X=x] - E[f(x,Y)] \leq \inf_{\lambda>0} \frac{D(\cdot\|\cdot) + \psi(\lambda)}{\lambda} = \hat{\psi}^{-1}(D(\cdot\|\cdot))$$

and consequently $\int (E[f(x,Y)|X=x] - E[f(x,Y)])dP_X$

$$\leq \int \hat{\psi}^{-1}(D(P_{Y|X=x}\|P_Y))dP_X \leq \hat{\psi}^{-1}(I(X;Y))$$

For $P_X = \mathcal{N}(0, \sigma^2)$ we get $\psi(\lambda) = (\lambda\sigma)^2/2$

$\Rightarrow$ any P_X such that $m = E[X] < \infty$ and

$$\psi(\lambda) \leq \frac{(\lambda\sigma)^2}{2}$$

is called σ^2-sub-Gaussian $\Rightarrow \hat{\psi}(t) \geq \frac{t^2}{2\sigma^2}$ and thus

$$\Pr(X + m \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}$$

In general, if $\Pr(X \in [a, b]) = 1$ for $-\infty < a \leq b < \infty$ then

$$\psi(\lambda) \leq \frac{(\lambda(b-a))^2}{8}$$

$\Rightarrow$ all such X are σ^2-sub-Gaussian with $\sigma^2 = (b-a)^2/4$

If $\ell(s, Z)$ is σ^2-sub-Gaussian for all $s \in E$, we can use

$$\psi_i(\lambda) = \frac{(\lambda\sigma)^2}{2}, \quad \hat{\psi}_i^{-1}(r) = \sqrt{2r\sigma^2}$$

Then for any $P_{S|Z^n}$ we get

$$|G(Q, P_{S|Z^n})| \leq \sqrt{\frac{2\sigma^2}{n} I(S; Z^n)}$$

and if $P_{S|Z^n}$ is ε stable and/or ε information stable w.r.t. Q

$$|G(Q, P_{S|Z^n})| \leq \sqrt{2\sigma^2\varepsilon}$$

Complexity and Generalization

Consider binary classification, that is $T(\theta) = \theta \in \{0, 1\}$, and samples $Z_i = (X_i, \theta_i)$ where $\theta_i =$ "X_i belongs to class θ_i"

Assume $\ell(s, Z) = \mathbb{1}(\{\hat{\theta}(X) \neq \theta\}) \Rightarrow L_Q(s) = \Pr(\hat{\theta}(X) \neq \theta)$

For a fixed s we have

$$\Pr(L_{Z^n}(s) \leq L_Q(s) - t) \leq \exp(-2nt^2)$$

($L_{Z^n}(s)$ is $1/(4n)$ sub-Gaussian for all $s \in E$ and $E[L_{Z^n}] = L_Q$)

Thus with probability at least $1 - \delta$,

$$L_Q(s) \leq L_{Z^n}(s) + \sqrt{\frac{\ln(1/\delta)}{2n}}$$

Again, the bound is for a *fixed* $s \in E$. Assume $E = \{s_1, \ldots, s_M\}$ is finite, then

$$\Pr(\text{ there is an } s \in E \text{ such that } L_{Z^n}(s) \leq L_Q(s) - t)$$
$$\leq \sum_{i=1}^{M} \Pr(L_{Z^n}(s_i) \leq L_Q(s_i) - t) \leq M e^{-2nt^2}$$

Hence, for *all* $s \in E$ and with probability at least $1 - \delta$

$$L_Q(s) \leq L_{Z^n}(s) + \sqrt{\frac{\ln M + \ln(1/\delta)}{2n}}$$

How do we handle the case when E is infinite/uncountable?
Can we replace the $\ln M$ term with something finite?

Even if E is infinite, the set

$$T(z^n) = \{(\ell(s, z_1), \ldots, \ell(s, z_n)) : s \in E\}$$

is finite

Let $S(n) = \sup_{z^n} |T(z^n)|$, then with probability at least $1 - \delta$

$$L_Q(s) \leq L_{Z^n}(s) + 2\sqrt{2\,\frac{\ln S(2n) + \ln(2/\delta)}{n}}$$

A classifier $s = \hat{\theta}(x) \in E$ shatters the samples $\{(x_i, \theta_i)\}_{i=1}^n$ if $\hat{\theta}(x_i) = \theta_i$, $i = 1, \ldots, n$

The Vapnik–Chervonenkis (VC) dimension d of the set E of classifiers = the largest n for which there is a set $\{x_i\}$ such that for any $\{\theta_i\}$ there is an $s \in E$ that shatters $\{(x_i, \theta_i)\}_{i=1}^n$

That is, $d =$ the largest n such that $S(n) = 2^n$

Example: For $E = \{\text{mappings } \mathbb{1}(\{x \in [a, b]\}),\ -\infty < a < b < \infty\}$ we get $d = 2$, since for any three points $x < y < z$ the set $\{(x, 1), (y, 0), (z, 1)\}$ cannot be shattered

In general, for all $n \geq d$ it can be shown that

$$S(n) \leq \left(\frac{en}{d}\right)^d$$

which implies the bound

$$L_Q(s) \leq L_{Z^n}(s) + 2\sqrt{2\frac{d\ln(2en/d) + \ln(2/\delta)}{n}}$$

for a class E with VC dimension d

The VC dimension measures the *complexity* in learning hypotheses from the class E

Universal versus Algorithm-dependent

The bound

$$L_Q(s) \le L_{Z^n}(s) + 2\sqrt{2\frac{d\ln(2en/d) + \ln(2/\delta)}{n}}$$

holds with probability $1-\beta$ for *all* $s \in E$, i.e. *uniformly* over E

Our previous bound

$$|G(Q, P_{S|Z^n})| \le \sqrt{\frac{2\sigma^2}{n} I(S;Z^n)}$$

is valid for the *expected* generalization error $E[L_Q(S) - L_{Z^n}(S)]$ for a *given* Q and algorithm $P_{S|Z^n}$

$I(S;Z^n)$ characterizes the complexity in producing S from Z^n

In modern (deep) learning, the VC dimension can be enormous, resulting in the universal bound becoming vacuous

It has been argued[2] that distribution Q and algorithm $P_{S|Z^n}$ dependent bounds are necessary to characterize deep learning, resulting in more useful complexity metrics than VC dimension

As an example, a recent[3] high-probability bound reads: Let R be any distribution on E, then with probability not smaller than $1-\beta$

$$E[L_Q(S) - L_{Z^n}(S)|Z^n] \le \hat{\psi}^{-1}\left(\frac{D(P_{S|Z^n}\|R) + \ln(n/\beta) + 1}{n}\right)$$

(where $\psi(\lambda)$ is such that $\ln E[\exp(-\lambda(\ell(s,Z) + L_Q(s)))] \le \psi(\lambda)$)

Note that choosing $R = P_S$ gives $E[D(P_{S|Z^n}\|P_S)] = I(S;Z^n)$

[2] C. Zhang et al., "Understanding deep learning requires re-thinking generalization," 2017

[3] Rodríguez–Gálvez, Thobaben and Skoglund, "More PAC Bayes bounds," 2023