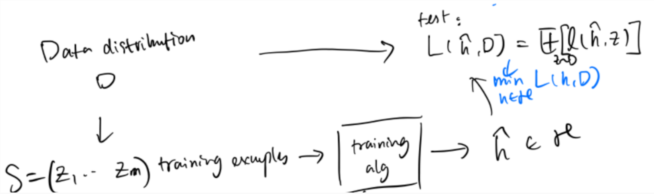
Statistical learning via loss minimization

Chicheng Zhang

CSC 588, University of Arizona

The statistical learning pipeline



Goal: design good training algorithm that can output model $\hat{h}$, that approximately minimizes excess loss

$$L(\hat{h}, D) - \min_{h \in \mathcal{H}} L(h, D)$$

Loss minimization: examples

- Classification: $z = (x, y) \in \mathcal{X} \times \{\pm 1\}$, $\ell(h, z) = I(h(x) \neq y)$
  - e.g. $\mathcal{H} = \left\{ h_w(x) := \mathbf{sign}(\langle w, x \rangle) : w \in \mathbb{R}^d \right\}$
- Regression: $z = (x, y) \in \mathcal{X} \times \mathbb{R}$, $\ell(h, z) = |h(x) - y|^p$ $(p > 0)$
  - $p = 2$: least squares regression; $p = 1$: least absolute deviation regression
  - e.g. $\mathcal{H} = \{h_w(x) = \langle w, x \rangle : \|w\|_2 \leq B\}$
- Density estimation: $z \in \mathbb{R}$, $\ell(h, z) = \ln \frac{1}{h(z)}$
  - e.g. $\mathcal{H} = \left\{ h_\mu(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}} : \mu \in \mathbb{R} \right\}$
- Ranking from pairwise comparisons: $z = (x, x', b) \in \mathcal{X} \times \mathcal{X} \times \{\pm 1\}$, $\ell(h, z) = I(b(h(x) - h(x')) \leq 0)$
  - e.g. $\mathcal{H} = \{h_w(x) = \langle w, x \rangle : \|w\|_2 \leq B\}$

Sample complexity analysis

Question

Can we analyze the sample complexities of these learning problems in a unified framework?

- We will analyze ERM: $\hat{h} = \operatorname{argmin}_{h \in \mathcal{H}} L(h, S)$
- Main tool: Rademacher complexity

A general theorem on ERM

Theorem (ERM's excess loss)
Suppose $\forall z \in \mathcal{Z}, h \in \mathcal{H}$, $|\ell(h, z)| \leq M$. *Then with probability* $1 - \delta$,

$$L(\hat{h}, D) - \min_{h \in \mathcal{H}} L(h, D) \leq 4M\sqrt{\frac{\ln \frac{4}{\delta}}{2m}} + 4\,\mathrm{Rad}_m(\mathcal{F}),$$

where $\mathcal{F} = \{f_h(z) := \ell(h, z) : h \in \mathcal{H}\}$

Remarks:

- We have seen a special form of this bound in the classification setting
- To obtain interpretable excess loss bounds, it requires further work to bound $\mathrm{Rad}_m(\mathcal{F})$ concretely

Proof sketch

Step 1: With probability $1-\delta/2$:

$$
\begin{aligned}
&\sup_{h\in\mathcal{H}} \left(L(h,D)-L(h,S)\right) \\
\leq &\mathbb{E}\left[\sup_{h\in\mathcal{H}} \left(L(h,D)-L(h,S)\right)\right] + 2M\sqrt{\frac{\ln\frac{4}{\delta}}{2m}} \text{ (McDiarmid's Inequality)} \\
\leq &2\operatorname{Rad}_m(\mathcal{F}) + 2M\sqrt{\frac{\ln\frac{4}{\delta}}{2m}} =: \mu \text{ (Symmetrization Lemma)}
\end{aligned}
$$

$\implies$ With probability $1-\delta$:

$$\sup_{h\in\mathcal{H}} |L(h,D)-L(h,S)| \leq \mu$$

Step 2: $\implies$ With probability $1-\delta$: $L(\hat{h},D)-\min_{h\in\mathcal{H}} L(h,D) \leq 2\mu$

Application: least-squares regression

- $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\|_2 \le R\right\}$, $\mathcal{H} = \{h_w(x) := \langle w, x\rangle : \|w\|_2 \le B\}$
- $\mathcal{Y} = [-Y, Y]$, $\ell(h, (x, y)) = (y - h(x))^2$.
- ERM:

$$\hat{h} = h_{\hat{w}}, \quad \text{where } \hat{w} = \operatorname*{argmin}_{w:\|w\|_2 \le B} \frac{1}{m} \sum_{i=1}^{m} (\langle w, x_i\rangle - y_i)^2,$$

- Can we give an (interpretable) high-probability upper bound on $L(\hat{h}, D) - \min_{h \in \mathcal{H}} L(h, D)$?

Application: least-squares regression (cont'd)

Applying Theorem:

Step 1: choose M, a bound on pointwise losses.

- Recall: $\ell(h,(x,y)) = (h(x)-y)^2$
- $\forall x \in \mathcal{X}, h_w \in \mathcal{H}$, $h_w(x) = \langle w, x\rangle \in [-BR, BR]$.
- Therefore, $h_w(x) - y \in [-(BR+Y), BR+Y]$; can choose $M := (BR+Y)^2$.

Step 2: bound the Rademacher complexity of the loss class.

- $\mathrm{Rad}_m(\mathcal{F}) = \mathbb{E}\,\mathrm{Rad}_S(\mathcal{F})$, where

$$\begin{aligned}\mathrm{Rad}_S(\mathcal{F}) &= \frac{1}{m}\mathbb{E}_{\sigma\sim U(\pm 1)^m}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^m \sigma_i f(z_i)\right]\\ &= \frac{1}{m}\mathbb{E}_{\sigma\sim U(\pm 1)^m}\left[\sup_{w:\|w\|_2\le B}\sum_{i=1}^m \sigma_i (y_i - \langle w, x_i\rangle)^2\right]\end{aligned}$$

How to control this?

The contraction inequality for Rademacher complexities

Lemma (Contraction inequality)
Suppose $S = (z_1, \ldots, z_m)$ is a sample, $\mathcal{G} : \mathcal{Z} \to [a, b]$ is a function class, $\phi : [a, b] \to \mathbb{R}$ is a L_ϕ-Lipschitz function. $\mathcal{F} = \{\phi \circ g : g \in \mathcal{G}\}$. Then,

$$\mathrm{Rad}_S(\mathcal{F}) \leq L_\phi \, \mathrm{Rad}_S(\mathcal{G}).$$

Interpretation:

- Equivalently,

$$\frac{1}{m}\mathbb{E}_\sigma\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^m \sigma_i \phi(g(z_i))\right] \leq L_\phi \cdot \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^m \sigma_i g(z_i)\right].$$

- Removing the ϕ function, at the price of introducing a L_ϕ factor

Lipschitz function

Definition
$\phi : [a, b] \to \mathbb{R}$ is said to be L-Lipschitz, if for any $u, v \in [a, b]$,

$$|\phi(u) - \phi(v)| \leq L\,|u - v|\,.$$

Theorem (Practical Lipschitzness criterion)
If ϕ is differentiable, then

$$\phi \text{ is } L\text{-Lipschitz} \Leftrightarrow \max_{v \in [a,b]} |\phi'(v)| \leq L.$$

Proof.
$\Leftarrow$: Lagrange mean value theorem
$\Rightarrow$: definition of derivative $\square$

Application: least-squares regression (cont'd)

$$\mathrm{Rad}_S(\mathcal{F}) = \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{w:\|w\|_2\le B}\sum_{i=1}^m \sigma_i(y_i - \langle w, x_i\rangle)^2\right]$$

- Applying contraction inequality with $\phi(v) := v^2$, $\phi : [-(BR+Y), BR+Y] \to \mathbb{R}$. How to choose L_ϕ?
- Choose $L_\phi = \max_{v\in[-(BR+Y),BR+Y]} |\phi'(v)| = 2(BR+Y)$
- Contraction inequality $\Rightarrow$

$$\begin{aligned}\mathrm{Rad}_S(\mathcal{F}) \le& 2(BR+Y)\cdot\frac{1}{m}\mathbb{E}_\sigma\left[\sup_{w:\|w\|_2\le B}\sum_{i=1}^m \sigma_i(y_i - \langle w, x_i\rangle)\right]\\ =& 2(BR+Y)\cdot\frac{1}{m}\mathbb{E}_\sigma\left[\sup_{w:\|w\|_2\le B}\sum_{i=1}^m (-\sigma_i)\langle w, x_i\rangle)\right]\\ =& 2(BR+Y)\cdot \mathrm{Rad}_S(\mathcal{H})\end{aligned}$$

(Recall $\mathcal{H} = \{h_w(x) := \langle w, x\rangle : \|w\|_2 \le B\}$)

Rademacher complexity of linear predictor classes

Theorem
Under the above settings of $\mathcal{X}$ and $\mathcal{H}$, for any S of size m,
$\mathrm{Rad}_S(\mathcal{H}) \leq BR\sqrt{\frac{1}{m}}$.

Proof.

$$
\begin{aligned}
\mathrm{Rad}_S(\mathcal{H}) &= \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{w:\|w\|_2\leq B}\left\langle w, \sum_{i=1}^m \sigma_i x_i\right\rangle\right] \\
&= \frac{B}{m}\mathbb{E}_\sigma\left[\|\sum_{i=1}^m \sigma_i x_i\|_2\right] && \left(\sup_{u:\|u\|_2\leq 1}\langle u, v\rangle = \|v\|_2\right) \\
&\leq \frac{B}{m}\sqrt{\mathbb{E}_\sigma\left[\|\sum_{i=1}^m \sigma_i x_i\|_2^2\right]} && \left(\mathbb{E}[Z] \leq \sqrt{\mathbb{E}[Z^2]}\right) \\
&= \frac{B}{m}\sqrt{\mathbb{E}_\sigma\left[\sum_{i=1}^m\sum_{j=1}^m \sigma_i\sigma_j\langle x_i, x_j\rangle\right]} \\
&\leq \frac{B}{m}\sqrt{mR^2} = BR\sqrt{\frac{1}{m}}.
\end{aligned}
$$

Application: least-squares regression (cont'd)

Putting everything together:

- $\mathrm{Rad}_S(\mathcal{F}) \leq 2(BR + Y)\,\mathrm{Rad}_S(\mathcal{H}) = 2(BR + Y)BR\sqrt{\frac{1}{m}}$
- Also recall: $M = (BR + Y)^2$

ERM excess loss theorem $\implies$ with probability $1 - \delta$,

$$\begin{aligned} L(\hat{h}, D) - \min_{h \in \mathcal{H}} L(h, D) \leq& 4M\sqrt{\frac{\ln \frac{4}{\delta}}{2m}} + 4\,\mathrm{Rad}_m(\mathcal{F}) \\ \leq& \mathrm{const} \cdot (BR + Y)^2 \sqrt{\frac{\ln \frac{4}{\delta}}{2m}}. \end{aligned}$$

Remark: this is a dimension-free bound – excess loss guarantee only depends on norms of examples and predictors

Proof of contraction inequality

High-level idea:

$$
\begin{aligned}
m \operatorname{Rad}_S(\mathcal{F}) &= \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \left(\sum_{i=1}^{m} \sigma_i \phi(g(z_i)) \right) \right] \\
&\overset{(1)}{\leq} \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \left(L_\phi \sigma_1 g(z_1) + \sum_{i=2}^{m} \sigma_i \phi(g(z_i)) \right) \right] \\
&\overset{(2)}{\leq} \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \left(L_\phi \sigma_1 g(z_1) + L_\phi \sigma_2 g(z_2) + \sum_{i=3}^{m} \sigma_i \phi(g(z_i)) \right) \right] \\
&\leq \ldots \\
&\overset{(m)}{\leq} \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \left(L_\phi \sigma_1 g(z_1) + L_\phi \sigma_2 g(z_2) + \ldots + L_\phi \sigma_m g(z_m) \right) \right] \\
&= m L_\phi \operatorname{Rad}_S(\mathcal{G}).
\end{aligned}
$$

We will focus on the proof of inequality (1), as the rest are similar.

Proof of contraction inequality (cont'd)

Proof of (1):

$$\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\left(\sum_{i=1}^{m}\sigma_i\phi(g(z_i))\right)\right] = \mathbb{E}_{\sigma_{2:n}}\mathbb{E}_{\sigma_1}\left[\sup_{g\in\mathcal{G}}\left(\sigma_1\phi(g(z_i)) + \sum_{i=2}^{m}\sigma_i\phi(g(z_i))\right)\right]$$

For any fixed $\sigma_{2:n}$, denote by $F(g) := \sum_{i=2}^{m}\sigma_i\phi(g(z_i))$.

It suffices to show

$$\mathbb{E}_{\sigma_1}\left[\sup_{g\in\mathcal{G}}\left(\sigma_1\phi(g(z_1)) + F(g)\right)\right] \leq \mathbb{E}_{\sigma_1}\left[\sup_{g\in\mathcal{G}}\left(L_{\phi}\sigma_1 g(z_1) + F(g)\right)\right]$$

Proof of contraction inequality (cont'd)

$$
\begin{aligned}
&\mathbb{E}_{\sigma_1}\left[\sup_{g\in\mathcal{G}}\left(\sigma_1\phi(g(z_1))+F(g)\right)\right]\\
=&\frac{1}{2}\left(\sup_{g\in\mathcal{G}}\left(\phi(g(z_1))+F(g)\right)+\sup_{g'\in\mathcal{G}}\left(-\phi(g'(z_1))+F(g')\right)\right)\\
=&\frac{1}{2}\left(\sup_{g,g'\in\mathcal{G}}\left(\phi(g(z_1))-\phi(g'(z_1))+F(g)+F(g')\right)\right)\\
\le&\frac{1}{2}\left(\sup_{g,g'\in\mathcal{G}}\left(L_\phi\left|g(z_1)-g'(z_1)\right|+F(g)+F(g')\right)\right)\\
\le&\frac{1}{2}\left(\sup_{g,g'\in\mathcal{G}:g(z_1)\ge g'(z_1)}\left(L_\phi\left|g(z_1)-g'(z_1)\right|+F(g)+F(g')\right)\right)\\
=&\frac{1}{2}\left(\sup_{g,g'\in\mathcal{G}:g(z_1)\ge g'(z_1)}\left(L_\phi g(z_1)-L_\phi g'(z_1)+F(g)+F(g')\right)\right)\\
\le&\frac{1}{2}\left(\sup_{g,g'\in\mathcal{G}}\left(L_\phi g(z_1)-L_\phi g'(z_1)+F(g)+F(g')\right)\right)\\
=&\mathbb{E}_{\sigma_1}\left[\sup_{g\in\mathcal{G}}\left(L_\phi\sigma_1 g(z_1)+F(g)\right)\right]\qquad\square
\end{aligned}
$$

What have we learned?

- Loss minimization appears in many statistical learning problems
- We established general excess loss upper bound of ERM in terms of Rademacher complexity
- Contraction inequality: useful tool for bounding the Rademacher complexity of function classes
- Rademacher complexity bound for ℓ_2 bounded linear function classes