Sample Complexity for Function Approximation. Model Selection.

Maria-Florina (Nina) Balcan

09/24/2018
Two Core Aspects of Machine Learning

Algorithm Design. How to optimize?

Automatically generate rules that do well on observed data.

- E.g.: logistic regression, SVM, Adaboost, etc.

Confidence Bounds, Generalization

Confidence for rule effectiveness on future data.
PAC/SLT models for Supervised Classification

Learning Algorithm

Data Source

Distribution \( D \) on \( X \)

Labeled Examples

(\( x_1, c^*(x_1) \), ..., \( x_m, c^*(x_m) \))

Expert / Oracle

Alg.outputs

\( h : X \rightarrow Y \)

\( c^* : X \rightarrow Y \)

\( x_1 > 5 \)

\( x_6 > 2 \)

+1

-1

+1

+1

-
PAC/SLT models for Supervised Learning

• **X** - feature/instance space; distribution **D** over **X**
  
e.g., \( X = \mathbb{R}^d \) or \( X = \{0,1\}^d \)

• Algo sees training sample **S**: \( (x_1, c^*(x_1)), \ldots, (x_m, c^*(x_m)) \), \( x_i \) i.i.d. from **D**
  - labeled examples - drawn i.i.d. from **D** and labeled by target \( c^* \)
  - labels \( \in \{-1,1\} \) - binary classification

• Algo does optimization over **S**, find hypothesis \( h \).

• Goal: \( h \) has small error over **D**.
  \[
  err_D(h) = \Pr_{x \sim D} (h(x) \neq c^*(x))
  \]

• Fix hypothesis space **H**  [whose complexity is not too large]
  • Realizable: \( c^* \in H \).
  • Agnostic: \( c^* \) “close to” **H**.
Sample Complexity for Supervised Learning

Realizable Case

**Consistent Learner**

- **Input:** \( S: (x_1, c^*(x_1)), \ldots, (x_m, c^*(x_m)) \)
- **Output:** Find \( h \) in \( H \) consistent with \( S \) (if one exits).

**Theorem**

\[
m \geq \frac{1}{\varepsilon} \left[ \ln(|H|) + \ln\left(\frac{1}{\delta}\right) \right]
\]

labeled examples are sufficient so that with prob. \( 1 - \delta \), all \( h \in H \) with \( err_D(h) \geq \varepsilon \) have \( err_S(h) > 0 \).

**Linear in \( 1/\varepsilon \)**

**Theorem**

\[
m = \mathcal{O}\left(\frac{1}{\varepsilon} \left[ VCDim(H) \log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right) \right] \right)
\]

labeled examples are sufficient so that with probab. \( 1 - \delta \), all \( h \in H \) with \( err_D(h) \geq \varepsilon \) have \( err_S(h) > 0 \).
Sample Complexity: Infinite Hypothesis Spaces

Realizable Case

**Theorem**

\[ m = \mathcal{O}\left(\frac{1}{\epsilon} \left[ VCdim(H) \log \left(\frac{1}{\epsilon}\right) + \log \left(\frac{1}{\delta}\right) \right] \right) \]

Labeled examples are sufficient so that with probab. \(1 - \delta\), all \(h \in H\) with \(err_D(h) \geq \epsilon\) have \(err_S(h) > 0\).

E.g., \(H=\) linear separators in \(\mathbb{R}^d\)

\(VCdim(H)= d+1\)

\[ m = \mathcal{O}\left(\frac{1}{\epsilon} \left[ d \log \left(\frac{1}{\epsilon}\right) + \log \left(\frac{1}{\delta}\right) \right] \right) \]

Sample complexity linear in \(d\)

So, if double the number of features, then I only need roughly twice the number of samples to do well.
Empirical Risk Minimization (ERM)

- **Input**: \( S: (x_1, c^*(x_1)), \ldots, (x_m, c^*(x_m)) \)
- **Output**: Find \( h \) in \( H \) with smallest \( \text{err}_S(h) \)

**Theorem**

\[
m \geq \frac{1}{2\varepsilon^2} \left[ \ln(|H|) + \ln \left( \frac{2}{\delta} \right) \right]
\]

labeled examples are sufficient s.t. with probab. \( \geq 1 - \delta \), all \( h \in H \)

**Theorem**

\[
m = \mathcal{O} \left( \frac{1}{\varepsilon^2} \left[ VCdim(H) + \log \left( \frac{1}{\delta} \right) \right] \right)
\]

labeled examples are sufficient so that with probab. \( 1 - \delta \), all \( h \in H \)

with \( |err_D(h) - err_S(h)| \leq \varepsilon \).

1/\( \varepsilon^2 \) dependence [as opposed to \( 1/\varepsilon \) for realizable]
Sample Complexity: Finite Hypothesis Spaces

Agnostic Case

1) How many examples suffice to get UC whp (so success for ERM).

Theorem

\[ m \geq \frac{1}{2\epsilon^2} \left[ \ln(|H|) + \ln \left( \frac{2}{\delta} \right) \right] \]

labeled examples are sufficient s.t. with probab. \( \geq 1 - \delta \), all \( h \in H \)
have \( |err_D(h) - err_S(h)| < \epsilon \).

2) Statistical Learning Theory style:

With prob. at least \( 1 - \delta \), for all \( h \in H \):

\[ err_D(h) \leq err_S(h) + \sqrt{\frac{1}{2m} \left( \ln (2|H|) + \ln \left( \frac{1}{\delta} \right) \right)}. \]

\( 1/\epsilon^2 \) dependence [as opposed to \( 1/\epsilon \) for realizable], but get for something stronger.

\( \sqrt{\frac{1}{m}} \) as opposed to \( \frac{1}{m} \) for realizable
Sample Complexity: Infinite Hypothesis Spaces

Agnostic Case

1) How many examples suffice to get UC whp (so success for ERM).

Theorem

\[ m = \mathcal{O}\left(\frac{1}{\varepsilon^2} \left[ VCdim(H) + \log\left(\frac{1}{\delta}\right)\right]\right) \]

labeled examples are sufficient so that with probab. $1 - \delta$, all $h \in H$ with $|err_D(h) - err_S(h)| \leq \varepsilon$.

2) Statistical Learning Theory style:

With prob. at least $1 - \delta$, for all $h \in H$:

\[ err_D(h) \leq err_S(h) + \Theta\left(\frac{1}{2m} \left( VCdim(H) \ln\left(\frac{em}{VCdim(H)}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right) \]
VC dimension Generalization Bounds

E.g., \( \text{err}_D(h) \leq \text{err}_S(h) + O\left(\sqrt{\frac{1}{2m} \left( \text{VCdim}(H) \ln \left( \frac{\text{emp}}{\text{VCdim}(H)} \right) + \ln \left( \frac{1}{\delta} \right) \right)} \right) \).

VC bounds: distribution independent bounds

- **Generic**: hold for any concept class and any distribution.  
  [nearly tight in the WC over choice of D]

- Might be very loose specific distr. that are more benign than the worst case....

- Hold only for binary classification; we want bounds for functions approximation in general (e.g., multiclass classification and regression).
Rademacher Complexity Bounds

[Koltchinskii & Panchenko 2002]

• Distribution/data dependent. Tighter for nice distributions.
• Apply to general classes of real valued functions & can be used to recover the VC bounds for supervised classification.
• Prominent technique for generalization bounds in last decade.

See “Introduction to Statistical Learning Theory”
O. Bousquet, S. Boucheron, and G. Lugosi.
Rademacher Complexity

Problem Setup

• A space $Z$ and a distr. $D|Z$
• $F$ be a class of functions from $Z$ to $[0,1]$
• $S = \{z_1, ..., z_m\}$ be i.i.d. from $D|Z$

Want a high prob. uniform convergence bound, all $f \in F$ satisfy:

$$E_D[f(z)] \leq E_S[f(z)] + \text{term(complexity of } F, \text{ niceness of } D/S)$$

What measure of complexity?

General discrete $Y$

E.g., $Z = X \times Y$, $Y = \{-1,1\}$, $H = \{h: X \rightarrow Y\}$ hyp. space (e.g., lin. sep)

$$F = L(H) = \{l_h: X \times Y \rightarrow [0,1]\}, \text{ where } l_h(z = (x,y)) = 1_{\{h(x) \neq y\}}$$

Then $E_{z \sim D}[l_h(z)] = \text{err}_D(h)$ and $E_S[l_h(z)] = \text{err}_S(h)$. [Loss fnc induced by $h$ and 0/1 loss]

$$\text{err}_D[h] \leq \text{err}_S[h] + \text{term(complexity of } H, \text{ niceness of } D/S)$$
Space $Z$ and a distr. $D|z$; $F$ be a class of functions from $Z$ to $[0,1]$

Let $S = \{z_1, \ldots, z_m\}$ be i.i.d from $D|z$.

The empirical Rademacher complexity of $F$ is:

$$\hat{R}_m(F) = E_{\sigma_1, \ldots, \sigma_m} \left[ \sup_{f \in F} \frac{1}{m} \sum_{i} \sigma_i f(z_i) \right]$$

where $\sigma_i$ are i.i.d. Rademacher variables chosen uniformly from $\{-1,1\}$.

The Rademacher complexity of $F$ is: $R_m(F) = E_S[\hat{R}_m(F)]$

$\sup$ measures for any given set $S$ and Rademacher vector $\sigma$, the max correlation between $f(z_i)$ and $\sigma_i$ for all $f \in F$

So, taking the expectation over $\sigma$ this measures the ability of class $F$ to fit random noise.
The empirical Rademacher complexity of $F$ is:

$$\hat{R}_m(F) = \mathbb{E}_{\sigma_1, \ldots, \sigma_m} \left[ \sup_{f \in F} \frac{1}{m} \sum_{i} \sigma_i f(z_i) \right]$$

where $\sigma_i$ are i.i.d. Rademacher variables chosen uniformly from $\{-1, 1\}$.

The Rademacher complexity of $F$ is:

$$R_m(F) = \mathbb{E}_S[\hat{R}_m(F)]$$

**Theorem:** Whp all $f \in F$ satisfy:

$$\mathbb{E}_D[f(z)] \leq \mathbb{E}_S[f(z)] + 2R_m(F) + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

$$\mathbb{E}_D[f(z)] \leq \mathbb{E}_S[f(z)] + 2\hat{R}_m(F) + 3\sqrt{\frac{\ln(1/\delta)}{m}}$$
Rademacher Complexity

Space \( Z \) and a distr. \( D_{|Z} \); let \( F \) be a class of functions from \( Z \) to \([0,1]\). Let \( S = \{z_1, ..., z_m\} \) be i.i.d from \( D_{|Z} \).

The empirical Rademacher complexity of \( F \) is:

\[
\hat{R}_m(F) = E_{\sigma_1, ..., \sigma_m} \left[ \sup_{f \in F} \frac{1}{m} \sum_i \sigma_i f(z_i) \right]
\]

where \( \sigma_i \) are i.i.d. Rademacher variables chosen uniformly from \( \{-1,1\} \).

The Rademacher complexity of \( F \) is: \( R_m(F) = E_S[\hat{R}_m(F)] \)

E.g.,

1) \( F = \{f\} \), then \( \hat{R}_m(F) = 0 \)  
   [Linearity of expectation: each \( \sigma_i f(z_i) \) individually has expectation 0.]

2) \( F = \{\text{all 0/1 fnc}\} \), then \( \hat{R}_m(F) = 1/2 \)  
   [To maximize set \( f(z_i) = 1 \) when \( \sigma_i = 1 \) and \( f(z_i) = 0 \) when \( \sigma_i = -1 \). Then quantity inside expectation is \#1's \( \in \sigma \), which is \( m/2 \) by linearity of expectation.]
Space $Z$ and a distr. $D_{|Z}$; $F$ be a class of functions from $Z$ to $[0,1]$

Let $S = \{z_1, ..., z_m\}$ be i.i.d from $D_{|Z}$.

The empirical Rademacher complexity of $F$ is:

$$\hat{R}_m(F) = E_{\sigma_1, ..., \sigma_m} \left[ \sup_{f \in F} \frac{1}{m} \sum_0^m \sigma_i f(z_i) \right]$$

where $\sigma_i$ are i.i.d. Rademacher variables chosen uniformly from $\{-1, 1\}$.

The Rademacher complexity of $F$ is: $R_m(F) = E_S[\hat{R}_m(F)]$

E.g.:
1) $F=\{f\}$, then $\hat{R}_m(F) = 0$

2) $F=\{\text{all 0/1 fnc}\}$, then $\hat{R}_m(F) = 1/2$

3) $F=L(H)$, $H=$binary classifiers then: $R_S(F) \leq \sqrt{\frac{\ln(2|H[S]|)}{m}}$

$H$ finite: $R_S(F) \leq \sqrt{\frac{\ln(2|H|)}{m}}$
Rademacher Complexity Bounds

Space \( Z \) and a distr. \( D_{|Z} \); \( F \) be a class of functions from \( Z \) to \([0,1] \).
Let \( S = \{z_1, ..., z_m\} \) be i.i.d from \( D_{|Z} \).

The empirical Rademacher complexity of \( F \) is:

\[
\hat{R}_m(F) = E_{\sigma_1, ..., \sigma_m} \left[ \sup_{f \in F} \frac{1}{m} \sum_{i=0}^{m} \sigma_i f(z_i) \right]
\]
where \( \sigma_i \) are i.i.d. Rademacher variables chosen uniformly from \( \{-1,1\} \).

The Rademacher complexity of \( F \) is: \( R_m(F) = E_S[\hat{R}_m(F)] \)

**Theorem:** Whp all \( f \in F \) satisfy: 

\[
E_D[f(z)] \leq E_S[f(z)] + 2R_m(F) + \sqrt{\frac{\ln(2/\delta)}{2m}}
\]

bound expectation of each \( f \) in terms of its empirical average & the RC of \( F \)

\[
E_D[f(z)] \leq E_S[f(z)] + 2 \hat{R}_m(F) + 3 \sqrt{\frac{\ln(1/\delta)}{m}}
\]

Data dependent bound! Proof uses Symmetrization and Ghost Sample Tricks! (same as for VC bound)
Rademacher Complex: Binary classification

Fact: \( H = \{ h: X \rightarrow Y \} \) hyp. space (e.g., lin. sep) \( F = \mathcal{L}(H), \ d = \text{VCdim}(H) \):

\[
R_S(F) \leq \sqrt{\frac{\ln(2|H[S]|)}{m}}
\]

So, by Sauer's lemma, \( R_S(F) \leq \sqrt{\frac{2d\ln\left(\frac{em}{d}\right)}{m}} \)

Theorem: For any \( H \), any distr. \( D \), w.h.p. \( \geq 1 - \delta \) all \( h \in H \) satisfy:

\[
err_D(h) \leq err_S(h) + R_m(H) + 3 \sqrt{\frac{\ln(2/\delta)}{2m}}.
\]

\[
err_D(h) \leq err_S(h) + \sqrt{\frac{2d\ln\left(\frac{em}{d}\right)}{m}} + 3 \sqrt{\frac{\ln(2/\delta)}{2m}}.
\]

generalization bound

Many more uses!!! Margin bounds for SVM, boosting, regression bounds, deep nets bounds etc.
What you should know

• Notion of sample complexity.

• Shattering, VC dimension as measure of complexity, Sauer’s lemma, form of the VC bounds

• Rademacher Complexity.