Sample Complexity for Function Approximation. Model Selection.

Maria-Florina (Nina) Balcan

03/10/2018
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.

Computation

(Labeled) Data
PAC/SLT models for Supervised Classification

Data Source

Distribution D on $X$

Labeled Examples

$(x_1, c^*(x_1)), \ldots, (x_m, c^*(x_m))$

Expert / Oracle

$c^*: X \rightarrow Y$

Learning Algorithm

Algorithm outputs

$h: X \rightarrow Y$

$x_1 > 5$

$x_6 > 2$

$+1$

$-1$

+ + - - - -
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 $D$.
  \[
  \text{err}_D(h) = \Pr_{x \sim D} (h(x) \neq c^*(x))
  \]
- **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:** $S: (x_1, c^*(x_1)), ..., (x_m, c^*(x_m))$
- **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**

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

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

**Prob. over different samples of $m$ training examples**
Sample Complexity: Infinite Hypothesis Spaces

Realizable Case

**Theorem**

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

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

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

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

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

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.
Sample Complexity: Uniform Convergence

Agnostic Case

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. \( 1 - \delta \), all \( h \in H \) have \( |\text{err}_D(h) - \text{err}_S(h)| < \varepsilon \).

**Theorem**

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

labeled examples are sufficient so that with probab. \( 1 - \delta \), all \( h \in H \) with \( |\text{err}_D(h) - \text{err}_S(h)| \leq \varepsilon \).
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 \( h \in H \):
\[ err_D(h) \leq err_S(h) + \sqrt{\frac{1}{2m} \left( \ln(2|H|) + \ln \left( \frac{1}{\delta} \right) \right)}. \]
Sample Complexity: Infinite Hypothesis Spaces

Agnostic Case

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

Theorem

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

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 \( h \in H \):

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

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

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 fns approximation in general (e.g., multiclass classification and regression).
Rademacher Complex: Binary classification

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

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

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:

\[
\text{err}_D(h) \leq \text{err}_S(h) + R_m(H) + 3 \sqrt{\frac{\ln(2/\delta)}{2m}}.
\]

\[
\text{err}_D(h) \leq \text{err}_S(h) + \sqrt{\frac{2d\ln\left(\frac{em}{d}\right)}{m}} + 3 \sqrt{\frac{\ln(2/\delta)}{2m}}
\]

generalization bound

Many more uses!!! Margin bounds for SVM, boosting, regression bounds, etc.
Can we use our bounds for model selection?
**True Error, Training Error, Overfitting**

Model selection: trade-off between decreasing training error and keeping $H$ simple.

$$\text{err}_D(h) \leq \text{err}_S(h) + R_m(H) + \ldots$$
Structural Risk Minimization (SRM)

\[ H_1 \subseteq H_2 \subseteq H_3 \subseteq \cdots \subseteq H_i \subseteq \cdots \]

\[ \text{error rate} \]

\[ \text{Hypothesis complexity} \]

\[ \text{Upper bound on true error} \]

\[ \text{True error} \]

\[ \text{overfitting} \]

\[ \text{empirical error} \]
What happens if we increase $m$?

Black curve will stay close to the red curve for longer, everything shift to the right...
Structural Risk Minimization (SRM)

\[ H_1 \subseteq H_2 \subseteq H_3 \subseteq \ldots \subseteq H_i \subseteq \ldots \]
Structural Risk Minimization (SRM)

- \( H_1 \subseteq H_2 \subseteq H_3 \subseteq \cdots \subseteq H_i \subseteq \cdots \)
- \( \hat{h}_k = \arg\min_{h \in H_k} \{ \text{err}_S(h) \} \)
  
  As \( k \) increases, \( \text{err}_S(\hat{h}_k) \) goes down but complex. term goes up.

- \( \hat{k} = \arg\min_{k \geq 1} \{ \text{err}_S(\hat{h}_k) + \text{complexity}(H_k) \} \)
  
  Output \( \hat{h} = \hat{h}_{\hat{k}} \)

Claim: W.h.p., \( \text{err}_D(\hat{h}) \leq \min_k \min_{h^* \in H_k} [\text{err}_D(h^*) + 2\text{complexity}(H_k^*)] \)

Proof:
- We chose \( \hat{h} \) s.t. \( \text{err}_S(\hat{h}) + \text{complexity}(H_{\hat{k}}) \leq \text{err}_S(h^*) + \text{complexity}(H_{k^*}) \).
- Whp, \( \text{err}_D(\hat{h}) \leq \text{err}_S(\hat{h}) + \text{complexity}(H_{\hat{k}}) \).
- Whp, \( \text{err}_S(h^*) \leq \text{err}_D(h^*) + \text{complexity}(H_{k^*}) \).
Techniques to Handle Overfitting

• **Structural Risk Minimization (SRM).** \[ H_1 \subseteq H_2 \subseteq \cdots \subseteq H_i \subseteq \cdots \]  
  Minimize gener. bound: \[ \hat{h} = \arg\min_{k \geq 1} \{ \text{err}_S(\hat{h}_k) + \text{complexity}(H_k) \} \]

• Often computationally hard....

• Nice case where it is possible: M. Kearns, Y. Mansour, ICML’98, “A Fast, Bottom-Up Decision Tree Pruning Algorithm with Near-Optimal Generalization”

• **Regularization:** general family closely related to SRM
  
  • E.g., SVM, regularized logistic regression, etc.
  
  • minimizes expressions of the form: \[ \text{err}_S(h) + \lambda \| h \|^2 \]

• **Cross Validation:** Picked through cross validation
  
  • Hold out part of the training data and use it as a proxy for the generalization error
What you should know

• Notion of sample complexity.

• Understand reasoning behind the simple sample complexity bound for finite $H$ [exam question!].

• Shattering, VC dimension as measure of complexity, Sauer’s lemma, form of the VC bounds (upper and lower bounds).

• Rademacher Complexity.

• Model Selection, Structural Risk Minimization.