Sample Complexity and Uniform Convergence

Eli Upfal

BROWN

Extracting Information from Data

Data science, machine learning, data mining, pattern recognition, statistical inference, "the scientific method",...

DATA $\Rightarrow$ ALGORITHM $\Rightarrow$ MODEL

Extracting Information from Data

Data science, machine learning, data mining, pattern recognition, statistical inference, "the scientific method",...

DATA $\Rightarrow$ ALGORITHM $\Rightarrow$ MODEL

In CS we focus on the algorithm (efficiency, correctness with respect to the input)

DATA $\Rightarrow$ ALGORITHM $\Rightarrow$ MODEL

Extracting Information from Data

Data science, machine learning, data mining, pattern recognition, statistical inference, "the scientific method",...

$$\text{DATA} \Rightarrow \text{ALGORITHM} \Rightarrow \text{MODEL}$$

In CS we focus on the algorithm (efficiency, correctness with respect to the input)

$$\text{DATA} \Rightarrow \text{ALGORITHM} \Rightarrow \text{MODEL}$$

but $$f(garbage) = garbage$$

In data analysis we need to verify that the information is in the data:

$$\text{DATA} \Rightarrow \text{ALGORITHM} \Rightarrow \text{MODEL}$$

Sample Complexity

Sample Complexity addresses the fundamental questions in data analysis:

- Does the data (training set) contains sufficient information to make a valid predictions (or fix a model)?
- Is the sample sufficiently large?
- How accurate is a prediction (model) inferred from a sample of a given size?

Standard statistics/probabilistic techniques do not give adequate soluions

Outline:

- Motivation: learning a binary classifier, the realizable and non-realizable case.
- Uniform convergence
- Uniform convergence through VC-dimension
- Applications: binary classification learning, data analysis
- Rademacher complexity
- Applications of Rademacher complexity in data analysis

Take home message:

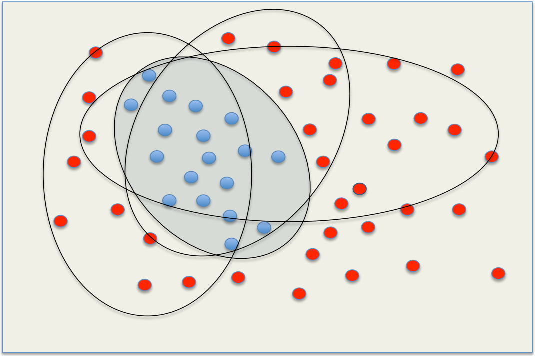
- Beautiful theory
- Not just theory
- Not just machine learning

Learning a Binary Classifier

- An unknown probability distribution $\mathcal{D}$ on a domain $\mathcal{U}$
- An unknown correct classification – a partition c of U to In and Out sets
- Input:
  - Concept class $\mathcal{C}$ – a collection of possible classification rules (partitions of U).
  - A training set $\{(x_i, c(x_i)) \mid i = 1, \ldots, m\}$, where $x_1, \ldots, x_m$ are sampled from $\mathcal{D}$.
- Goal: With probability $1 - \delta$ the algorithm generates a classification that is correct (on items generated from $\mathcal{D}$) with probability $opt(\mathcal{C}) - \epsilon$, where $opt(\mathcal{C})$ is the probability of the best classification in $\mathcal{C}$.

Learning a Binary Classifier

- Out and In items, and a concept class C of possible classification rules



When does the sample identify the correct rule? - The realizable case

- The realizable case - the correct classification $c \in \mathcal{C}$.
- Algorithm: choose $h^* \in \mathcal{C}$ that agrees with all the training set (there must be at least one).
- For any $h \in \mathcal{C}$ let $\Delta(c, h)$ be the set of items on which the two classifiers differ: $\Delta(c, h) = \{x \in U \mid h(x) \neq c(x)\}$
- If the sample (training set) intersects every set in

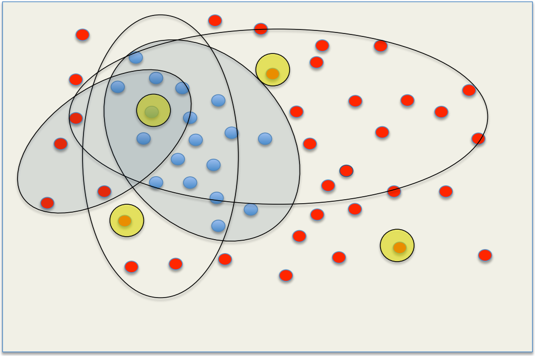
$$\{\Delta(c, h) \mid Pr(\Delta(c, h)) \geq \epsilon\},$$

then

$$Pr(\Delta(c, h^*)) \leq \epsilon.$$

Learning a Binary Classifier

- Red and blue items, possible classification rules, and the sample items 



When does the sample identify the correct rule? The unrealizable (agnostic) case

- The unrealizable case - c may not be in $\mathcal{C}$.
- For any $h \in \mathcal{C}$, let $\Delta(c,h)$ be the set of items on which the two classifiers differ: $\Delta(c,h) = \{x \in U \mid h(x) \neq c(x)\}$
- For the training set $\{(x_i, c(x_i)) \mid i = 1, \ldots, m\}$, let

$$\tilde{Pr}(\Delta(c,h)) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{h(x_i)\neq c(x_i)}$$

- Algorithm: choose $h^* = \arg\min_{h\in\mathcal{C}} \tilde{Pr}(\Delta(c,h))$.
- If for every set $\Delta(c,h)$,

$$|Pr(\Delta(c,h)) - \tilde{Pr}(\Delta(c,h))| \leq \epsilon,$$

then

$$Pr(\Delta(c,h^*)) \leq opt(\mathcal{C}) + 2\epsilon.$$

Uniform Convergence [Vapnik – Chervonenkis 1971]

Definition

A set of functions $\mathcal{F}$ has the *uniform convergence* property with respect to a domain Z if there is a function $m_{\mathcal{F}}(\epsilon, \delta)$ such that

- for any $\epsilon, \delta > 0$, $m(\epsilon, \delta) < \infty$
- for any distribution D on Z, and a sample $z_1, \dots, z_m$ of size $m = m_{\mathcal{F}}(\epsilon, \delta)$,

$$Pr(\sup_{f \in \mathcal{F}} |\frac{1}{m}\sum_{i=1}^{m} f(z_i) - E_{\mathcal{D}}[f]| \leq \epsilon) \geq 1 - \delta.$$

Let $f_E(z) = \mathbf{1}_{z \in E}$ then $\mathbf{E}[f_E(z)] = Pr(E)$.

Uniform Convergence and Learning

Definition

A set of functions $\mathcal{F}$ has the *uniform convergence* property with respect to a domain Z if there is a function $m_{\mathcal{F}}(\epsilon, \delta)$ such that

- for any $\epsilon, \delta > 0$, $m(\epsilon, \delta) < \infty$
- for any distribution D on Z, and a sample $z_1, \ldots, z_m$ of size $m = m_{\mathcal{F}}(\epsilon, \delta)$,

$$Pr(\sup_{f \in \mathcal{F}} |\frac{1}{m} \sum_{i=1}^{m} f(z_i) - E_{\mathcal{D}}[f]| \leq \epsilon) \geq 1 - \delta.$$

- Let $\mathcal{F}_{\mathcal{H}} = \{f_h \mid h \in H\}$, where f_h is the loss function for hypothesis h.
- $\mathcal{F}_H$ has the uniform convergence property $\Rightarrow$ an ERM (Empirical Risk Minimization) algorithm "learns" $\mathcal{H}$.
- The *sample complexity* of learning $\mathcal{H}$ is bounded by $m_{\mathcal{F}_{\mathcal{H}}}(\epsilon, \delta)$

Uniform Convergence - 1971, PAC Learning - 1984

Definition

A set of functions $\mathcal{F}$ has the *uniform convergence* property with respect to a domain Z if there is a function $m_{\mathcal{F}}(\epsilon, \delta)$ such that

- for any $\epsilon, \delta > 0$, $m(\epsilon, \delta) < \infty$
- for any distribution D on Z, and a sample $z_1, \ldots, z_m$ of size $m = m_{\mathcal{F}}(\epsilon, \delta)$,

$$Pr(\sup_{f \in \mathcal{F}} |\frac{1}{m} \sum_{i=1}^{m} f(z_i) - E_{\mathcal{D}}[f]| \leq \epsilon) \geq 1 - \delta.$$

- Let $\mathcal{F}_{\mathcal{H}} = \{f_h \mid h \in H\}$, where f_h is the loss function for hypothesis h.
- $\mathcal{F}_H$ has the uniform convergence property $\Rightarrow$ an ERM (Empirical Risk Minimization) algorithm "learns" $\mathcal{H}$. PAC efficiently learnable if there a polynomial time ϵ, δ-approximation for minimum ERM.

Uniform Convergence

Definition

A set of functions $\mathcal{F}$ has the *uniform convergence* property with respect to a domain Z if there is a function $m_{\mathcal{F}}(\epsilon, \delta)$ such that

- for any $\epsilon, \delta > 0$, $m(\epsilon, \delta) < \infty$
- for any distribution D on Z, and a sample $z_1, \ldots, z_m$ of size $m = m_{\mathcal{F}}(\epsilon, \delta)$,

$$Pr(\sup_{f \in \mathcal{F}} |\frac{1}{m}\sum_{i=1}^{m} f(z_i) - E_{\mathcal{D}}[f]| \leq \epsilon) \geq 1 - \delta.$$

VC-dimension and Rademacher complexity are the two major techniques to

- prove that a set of functions $\mathcal{F}$ has the uniform convergence property
- charaterize the function $m_{\mathcal{F}}(\epsilon, \delta)$

Some Background

- Let $z_1, \ldots, z_m$ i.i.d. observation from distribution $F(x)$, and $F_m(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{z_i \le x}$ (empirical distribution function)
- Strong Law of Large Numbers: for a given x,

$$F_m(x) \rightarrow_{a.s} F(x) = Pr(z \le x).$$

- Glivenko-Cantelli Theorem (uniform convergence of $\{F(x) \mid x \in \mathbf{R}\}$):

$$\sup_{x \in \mathbf{R}} |F_m(x) - F(x)| \rightarrow_{a.s} 0.$$

- Dvoretzky-Keifer-Wolfowitz Inequality (Kolmogorov-Smirnov ditribution)

$$Pr(\sup_{x \in \mathbf{R}} |F_m(x) - F(x)| \ge \epsilon) \le 2e^{-2m\epsilon^2}.$$

- VC-dimension characterizes uniform convergence property for arbitrary sets of events.

Simplest Uniform Convergence - Union Bound

Theorem

In the realizable case, any concept class $\mathcal{C}$ can be learned with $m = \frac{1}{\epsilon}(\ln |\mathcal{C}| + \ln \frac{1}{\delta})$ samples.

Proof.

We need a sample that intersects every set in the family of sets

$$\{\Delta(c, c') \mid Pr(\Delta(c, c')) \geq \epsilon\}.$$

There are at most $|\mathcal{C}|$ such sets, and the probability that a sample is chosen inside a set is $\geq \epsilon$.
The probability that m random samples did not intersect with at least one of the sets is bounded by

$$|\mathcal{C}|(1-\epsilon)^m \leq |\mathcal{C}|e^{-\epsilon m} \leq |\mathcal{C}|e^{-(\ln |\mathcal{C}| + \ln \frac{1}{\delta})} \leq \delta.$$

$\square$

How Good is This Bound? - Learning an Interval

- A distribution $\mathcal{D}$ is defined on universe that is an interval $[A, B]$.
- The true classification rule is defined by a sub-interval $[a, b] \subseteq [A, B]$.
- The concept class $\mathcal{C}$ is the collection of all intervals,

$$\mathcal{C} = \{[c, d] \mid [c, d] \subseteq [A, B]\}$$

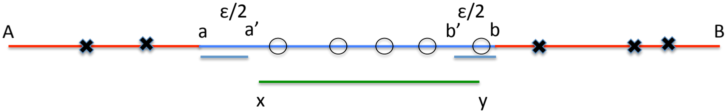
Theorem

There is a learning algorithm that given a sample from $\mathcal{D}$ *of size* $m = \frac{2}{\epsilon} \ln \frac{2}{\delta}$, *with probability* $1 - \delta$, *returns a classification rule (interval)* $[x, y]$ *that is correct with probability* $1 - \epsilon$.

Note that the sample size is independent of the size of the concept class $|\mathcal{C}|$, which is infinite.

Learning an Interval

- If the classification error is ≥ ε then the sample missed at least one of the the intervals [a,a’] or [b’,b] each of probability ≥ ε/2



Each sample excludes many possible intervals.
The union bound sums over overlapping hypothesis.
Need better characterization of concept's complexity!

Proof.

Algorithm: Choose the smallest interval $[x, y]$ that includes all the "In" sample points.

- Clearly $a \leq x < y \leq b$, and the algorithm can only err in classifying "In" points as "Out" points.
- Fix $a < a'$ and $b' < b$ such that $Pr([a, a']) = \epsilon/2$ and $Pr([b, b']) = \epsilon/2$.
- If the probability of error when using the classification $[x, y]$ is $\geq \epsilon$ then either $a' \leq x$ or $y \leq b'$ or both.
- The probability that the sample of size $m = \frac{2}{\epsilon} \ln \frac{2}{\delta}$ did not intersect with one of these intervals is bounded by

$$2(1 - \frac{\epsilon}{2})^m \leq e^{-\frac{\epsilon m}{2} + \ln 2} \leq \delta$$

□

- The union bound is far too loose for our applications. It sums over overlapping hypothesis.
- Each sample excludes many possible intervals.
- Need better characterization of concept's complexity!