PAC Learning Framework

张超

Dept. of CS&T, Tsinghua University

Questions for Learning Algorithms

- Sample Complexity
  - How many training examples do we need to converge to a successful hypothesis with a high probability?

Questions for Learning Algorithms

- Sample Complexity
    - How many training examples do we need to converge to a successful hypothesis with a high probability?
- Computational Complexity
    - How much computational effort is needed to converge to a successful hypothesis with a high probability?

Questions for Learning Algorithms

- Sample Complexity
  - How many training examples do we need to converge to a successful hypothesis with a high probability?
- Computational Complexity
  - How much computational effort is needed to converge to a successful hypothesis with a high probability?
- Mistake Bound
  - How many training examples will the learner misclassify before converging to a successful hypothesis?

Questions for Learning Algorithms

- What it means for the learner to be "successful"?

Questions for Learning Algorithms

- What it means for the learner to be "successful"?
  - The output hypothesis is identical to the target concept

Questions for Learning Algorithms

- What it means for the learner to be "successful"?
  - The output hypothesis is identical to the target concept
  - The output hypothesis agrees with the target concept most of the time

Questions for Learning Algorithms

- What it means for the learner to be "successful"?
  - The output hypothesis is identical to the target concept
  - The output hypothesis agrees with the target concept most of the time
- The answer to the question denpends on the particular learning model in mind

Introduction

- Probably Approximately Correct Learning

Introduction

- Probably Approximately Correct Learning
- Analysis under the Framework
  - Sample Complexity
  - Computational Complexity

Introduction

- Probably Approximately Correct Learning
- Analysis under the Framework
  - Sample Complexity
  - Computational Complexity
- Restrict the discussion on
  - Learn boolean-valued concept
  - Learn from noise-free training data

Basic Symbols

- Example: The students who passed mid-term exam and hand in all the homework, and passed the final exam or hand in a reading report passed the course Applied Stochastic Process

Basic Symbols

- Example: The students who passed mid-term exam and hand in all the homework, and passed the final exam or hand in a reading report passed the course Applied Stochastic Process
- X: Instance set, $x \in X$

Passed Mid-term Exam	Passed Final Exam	Hand in All The Homework	Hand in Reading Report
1 / 0	1 / 0	1 / 0	1 / 0

$$| X | = 2 \times 2 \times 2 \times 2 = 16$$

Basic Symbols

- Example: The students who passed mid-term exam and hand in all the homework, and passed the final exam or hand in a reading report passed the course Applied Stochastic Process
- X: Instance set, $x \in X$

Passed Mid-term Exam	Passed Final Exam	Hand in All The Homework	Hand in Reading Report
1 / 0	1 / 0	1 / 0	1 / 0

$$| X | = 2 \times 2 \times 2 \times 2 = 16$$

- C: Concept set, $c \in C$

$$C = \{((1,1,1,0),1),((1,0,1,1),1),((1,1,1,1),1)\}$$

Basic Symbols

- Example: The students who passed mid-term exam and hand in all the homework, and passed the final exam or hand in a reading report passed the course Applied Stochastic Process
- X: Instance set, $x \in X$

Passed Mid-term Exam	Passed Final Exam	Hand in All The Homework	Hand in Reading Report
1 / 0	1 / 0	1 / 0	1 / 0

$$|X| = 2 \times 2 \times 2 \times 2 = 16$$

- C: Concept set, $c \in C$

$$C = \{((1,1,1,0),1),((1,0,1,1),1),((1,1,1,1),1)\}$$

- H: All possible hypothesis set, $h \in H$

$$|H| = 1 + 3 \times 3 \times 3 \times 3 = 82$$

Basic Symbols

- Example: The students who passed mid-term exam and hand in all the homework, and passed the final exam or hand in a reading report passed the course Applied Stochastic Process
- X: Instance set, $x \in X$

Passed Mid-term Exam	Passed Final Exam	Hand in All The Homework	Hand in Reading Report
1 / 0	1 / 0	1 / 0	1 / 0

$$|X| = 2 \times 2 \times 2 \times 2 = 16$$

- C: Concept set, $c \in C$

$$C = \{((1,1,1,0),1),((1,0,1,1),1),((1,1,1,1),1)\}$$

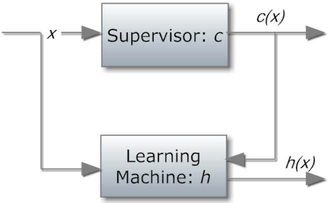
- H: All possible hypothesis set, $h \in H$

$$|H| = 1 + 3 \times 3 \times 3 \times 3 = 82$$

- $\mathscr{D}$: Instance Distribution; D: Training set

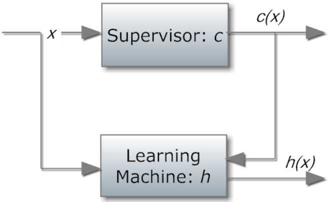
Basic Symbols

- Model of learning from examples



Basic Symbols

- Model of learning from examples



- x is drawn randomly from X according to $\mathscr{D}$

Error of a hypothesis

- Ture Error

Definition

True Error of hypothesis h with respect to target concept c and distribution $\mathscr{D}$ is the probability that h will misclassify an instance drawn randomly according to $\mathscr{D}$, i.e.

$$\text{error}_{\mathscr{D}}(h) \equiv P(c \Delta h)$$

$A \Delta B$ is symmetric difference, that $A \Delta B = (A \cup B) \setminus (A \cap B)$

Error of a hypothesis

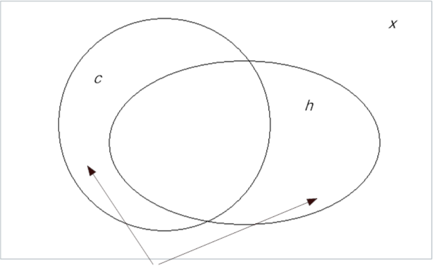
- Ture Error

Definition

True Error of hypothesis h with respect to target concept c and distribution $\mathscr{D}$ is the probability that h will misclassify an instance drawn randomly according to $\mathscr{D}$, i.e.

$$\text{error}_{\mathscr{D}}(h) \equiv P(c\Delta h)$$

$A\Delta B$ is symmetric difference, that $A\Delta B = (A \cup B) \setminus (A \cap B)$



where $c(x)$ not equals $h(x)$

PAC Learnability

- Learnabilty("Successful")

PAC Learnability

- Learnabilty("Successful")
- $\text{error}_{\mathscr{D}}(h) = 0$ is futile

PAC Learnability

- Learnabilty("Successful")
- $\text{error}_{\mathscr{D}}(h) = 0$ is futile
  - We need to provide training examples corresponding to every possible instance in X

PAC Learnability

- Learnabilty("Successful")
- $\text{error}_{\mathscr{D}}(h) = 0$ is futile
  - We need to provide training examples corresponding to every possible instance in X
  - Randomly drawned training examples will always be some nonzero probabilty to be misleading

PAC Learnability

- Learnabilty("Successful")
- $\text{error}_{\mathscr{D}}(h) = 0$ is futile
  - We need to provide training examples corresponding to every possible instance in X
  - Randomly drawned training examples will always be some nonzero probabilty to be misleading
- PAC Learnability

PAC Learnability

- Learnabilty("Successful")
- $\text{error}_{\mathscr{D}}(h) = 0$ is futile
  - We need to provide training examples corresponding to every possible instance in X
  - Randomly drawned training examples will always be some nonzero probabilty to be misleading
- PAC Learnability
  - We will require only that h's true error be bounded by some constant, ϵ, that can be made arbitrarily small

PAC Learnability

- Learnabilty("Successful")
- $\text{error}_{\mathscr{D}}(h) = 0$ is futile
  - We need to provide training examples corresponding to every possible instance in X
  - Randomly drawned training examples will always be some nonzero probabilty to be misleading
- PAC Learnability
  - We will require only that h's true error be bounded by some constant, ϵ, that can be made arbitrarily small
  - We will require only that h's probability of failure be bounded by some constant, δ, that can be made arbitrarily small

PAC Learnability

- Learnabilty("Successful")
- $\text{error}_{\mathscr{D}}(h) = 0$ is futile
  - We need to provide training examples corresponding to every possible instance in X
  - Randomly drawned training examples will always be some nonzero probabilty to be misleading
- PAC Learnability
  - We will require only that h's true error be bounded by some constant, ϵ, that can be made arbitrarily small
  - We will require only that h's probability of failure be bounded by some constant, δ, that can be made arbitrarily small
  - $(1 - \delta) \rightarrow$ probably; $\epsilon \rightarrow$, approximately correct
  - We only require learner probably learn a hypothesis that is approximately correct, which is PAC

PAC Learnability

- PAC Learnability

Definition

Consider a concept class C defined over a set of instances X of length n and a learner L using hypothesis space H. C is PAC-learnable by L using H if for all $c \in C$, distributions $\mathscr{D}$ over X, ϵ such that $0 < \epsilon < \frac{1}{2}$, and δ such that $0 < \delta < \frac{1}{2}$,
learner L will with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $\text{error}_{\mathscr{D}}(\text{h}) \leq \epsilon$,
in time that is polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$, n, and size(c).

PAC Learnability

- PAC Learnability

Definition

Consider a concept class C defined over a set of instances X of length n and a learner L using hypothesis space H. C is PAC-learnable by L using H if for all $c \in C$, distributions $\mathscr{D}$ over X, ϵ such that $0 < \epsilon < \frac{1}{2}$, and δ such that $0 < \delta < \frac{1}{2}$,
learner L will with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $\text{error}_{\mathscr{D}}(\text{h}) \leq \epsilon$,
in time that is polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$, n, and size(c).

- $L \rightarrow (1 - \delta), \epsilon,$
- $L \rightarrow$Polynomial time, Polynomial samples

PAC Learnability

- PAC Learnability

Definition

Consider a concept class C defined over a set of instances X of length n and a learner L using hypothesis space H. C is PAC-learnable by L using H if for all $c \in C$, distributions $\mathscr{D}$ over X, ϵ such that $0 < \epsilon < \frac{1}{2}$, and δ such that $0 < \delta < \frac{1}{2}$,
learner L will with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $\text{error}_{\mathscr{D}}(\text{h}) \leq \epsilon$,
in time that is polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$, n, and $\text{size}(c)$.

- $L \rightarrow (1 - \delta), \epsilon,$
- $L \rightarrow$Polynomial time, Polynomial samples
- $X \rightarrow n$
- $C \rightarrow \text{size}(c)$

Sample Complexity for Finite Hypothesis Space

- Sample Complexity
  - The growth in the number of required training examples with problem size, called the sample complexity of the learning problem

Sample Complexity for Finite Hypothesis Space

- Sample Complexity
  - The growth in the number of required training examples with problem size, called the sample complexity of the learning problem
- Consistent Learner

Definition

A hypothesis is consistent with set D, if and only if $h(x) = c(x)$, i.e.,

$$\text{Consistent}(h, D) \equiv (\forall < x, c(x) > \in D), h(x) = c(x)$$

A learner is consistent if it outputs hypotheses that perfectly fit the training data

Consistent Learner

- Version Space

Definition

The Version Space, denoted $\mathrm{VS}_{H,D} \equiv \{h \in H \mid Consistent(h, D)\}$

$$C = \{(1,1,1,0),(1,0,1,1),(1,1,1,1)\}, c = (1,0,1,1)$$
$$D = \{((1,0,1,1),1),((1,0,0,1),0)\}$$
$$\Rightarrow VS_{H,D} = \{(?,?,1,?),...,(1,0,1,1),...,(1,1,1,0),...,(1,1,1,1)\}$$

Consistent Learner

- Version Space

Definition

The Version Space, denoted $\mathrm{VS}_{H,D} \equiv \{h \in H \mid Consistent(h, D)\}$

$$C = \{(1,1,1,0),(1,0,1,1),(1,1,1,1)\}, c = (1,0,1,1)$$
$$D = \{((1,0,1,1),1),((1,0,0,1),0)\}$$
$$\Rightarrow VS_{H,D} = \{(?,?,1,?),...,(1,0,1,1),...,(1,1,1,0),...,(1,1,1,1)\}$$

- $\epsilon-$exhausted

  To bound the number of examples needed by any consistent learner, we need only to bound the number of examples needed to assure that the version space contains no unacceptable hypotheses

Definition

$VS_{H,D}$, is said to be ϵ-exhausted with respect to c and $\mathscr{D}$, if every hypothesis h in $VS_{H,D}$ has error less than ϵ with respect to c and $\mathscr{D}$, i.e.

$$(\forall h \in VS_{H,D})\mathrm{error}_{\mathscr{D}}(h) < \epsilon$$

Consistent Learner

- $\epsilon-$Exhausting The Version Space

  A probabilistic way to probabilistic to bound the probability that the version space will be $\epsilon-$exhausted after a given number of training examples, without knowing the identity of the target concept or the distribution from which training examples are drawn

Definition

If the hypothesis space H is finite, and D is a sequence of $m \geq 1$ independent randomly drawn examples of some target concept c, then for any $0 \leq \epsilon \leq 1$, the probability that the version space $VS_{H,D}$ is not $\epsilon-$exhausted is less than or equal to,

$$| H | e^{-\epsilon m}$$

Consistent Learner

- $\epsilon-$Exhausting The Version Space
  - proof:
    Assume $h_1, h_2, \ldots, h_k$ be all the hypotheses in H that have true error greater than ϵ with respect to c. The probability one such hypothesis is consistent with m randomly drawn samples is $(1-\epsilon)^m$, then the probability that there's at least one such hypothesis in $VS_{H,D}$ is,

$$k(1-\epsilon)^m \tag{1}$$

$$k(1-\epsilon)^m \leq \mid H \mid (1-\epsilon)^m \leq \mid H \mid e^{-\epsilon m} \tag{2}$$

Consistent Learner

- $\epsilon-$Exhausting The Version Space
  - proof:
    Assume $h_1, h_2, \ldots, h_\mathrm{k}$ be all the hypotheses in H that have true error greater than ϵ with respect to c. The probability one such hypothesis is consistent with m randomly drawn samples is $(1-\epsilon)^m$, then the probability that there's at least one such hypothesis in $VS_{H,D}$ is,

$$k(1-\epsilon)^m \tag{1}$$

$$k(1-\epsilon)^m \leq \mid H \mid (1-\epsilon)^m \leq \mid H \mid e^{-\epsilon m} \tag{2}$$

- Get m

$$\mid H \mid e^{-\epsilon m} \leq \delta \tag{3}$$

$$m \geq \frac{1}{\epsilon}(\ln \mid H \mid + \ln(\frac{1}{\delta})) \tag{4}$$

Consistent Learner

- Conjunctions of Boolean Literals
  - $n = 4, \mid H \mid = 3^4, \epsilon = 0.1, \delta = 0.05$,then,

$$m \geq \frac{1}{\epsilon}(n \ln 3 + \ln(\frac{1}{\delta})) = 10 \times (4 \ln 3 + \ln 20) \approx 74$$

Consistent Learner

- Conjunctions of Boolean Literals
  - $n = 4, \mid H \mid = 3^4, \epsilon = 0.1, \delta = 0.05$,then,

$$m \geq \frac{1}{\epsilon}(n \ln 3 + \ln(\frac{1}{\delta})) = 10 \times (4 \ln 3 + \ln 20) \approx 74$$

- $\mid X \mid = 16 < m_0^{'} = 74$
  - The weakness of this bound is mainly due to the $\mid H \mid$ term, which arises in the proof when summing the probability that a single hypothesis could be unacceptable, over all possible hypotheses

Agnostic Learning and Inconsistent Hypotheses

- Agnostic Learner if H does not contain the target concept c, then a zero-error hypothesis cannot always be found

Definition

A learner that makes no assumption that the target concept is representable by H and that simply finds the hypothesis with minimum training error, is often called an agnostic learner

Agnostic Learning and Inconsistent Hypotheses

- Agnostic Learner if H does not contain the target concept c, then a zero-error hypothesis cannot always be found

Definition

A learner that makes no assumption that the target concept is representable by H and that simply finds the hypothesis with minimum training error, is often called an agnostic learner

- Inconsistent Hypothesis

Agnostic Learning and Inconsistent Hypotheses

- Agnostic Learner if H does not contain the target concept c, then a zero-error hypothesis cannot always be found

Definition

A learner that makes no assumption that the target concept is representable by H and that simply finds the hypothesis with minimum training error, is often called an agnostic learner

- Inconsistent Hypothesis
  - $\text{error}_D(h)$ denote the training error of hypothesis h, h_{best} denote the hypothesis from H having lowest training error over training set

Agnostic Learning and Inconsistent Hypotheses

- Agnostic Learner if H does not contain the target concept c, then a zero-error hypothesis cannot always be found

Definition

A learner that makes no assumption that the target concept is representable by H and that simply finds the hypothesis with minimum training error, is often called an agnostic learner

- Inconsistent Hypothesis
  - $\text{error}_D(h)$ denote the training error of hypothesis h, h_{best} denote the hypothesis from H having lowest training error over training set
  - Using Hoeffding bound which characterize the deviation between the true probability of some event and its observed frequency over m independent trials
    To ensure the true error $\text{error}_{\mathscr{D}}(h_{\text{best}}) \leq \epsilon + \text{error}_D(h_{\text{best}})$

$$P((\exists h \in H)(\text{error}_{\mathscr{D}}(h_{\text{best}}))\text{error}_D(h_{\text{best}}) + \epsilon) \leq | H | e^{-2m\epsilon^2}$$

$$m \geq \frac{1}{2\epsilon^2}(\ln | H | + \ln \frac{1}{\delta})$$

PAC-Learnability of Other Concept Classes

- Unbiased Learner
  - The unbiased concept class C that contains every teachable concept relative to X, suppose that instances in X are defined by n boolean features, then $\mid C \mid = 2^{|X|} = 2^{2^n}$,

$$m \geq \frac{1}{\epsilon}(2^n \ln 2 + \ln \frac{1}{\delta})$$

is not PAC learnable

PAC-Learnability of Other Concept Classes

- Unbiased Learner
  - The unbiased concept class C that contains every teachable concept relative to X, suppose that instances in X are defined by n boolean features, then $\mid C \mid = 2^{|X|} = 2^{2^n}$,

$$m \geq \frac{1}{\epsilon}(2^n \ln 2 + \ln \frac{1}{\delta})$$

    is not PAC learnable
- K-Term DNF And K-CNF Concepts
  - K-Term DNF has polynomial sample complexity, it does not have polynomial computational complexity for a learner using $H = C$
  - K-CNF polynomial computation complexity per example and still has polynomial sample complexity
  - K-Term DNF $\subseteq K - CNF$

Sample Complexity for Infinite Hypothesis Spaces

- Shatter

Definition

A set of instances S is shattered by hypothesis space H if and only if for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy.

Sample Complexity for Infinite Hypothesis Spaces

- Shatter

Definition

A set of instances S is shattered by hypothesis space H if and only if for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy.

- VC Dimension

Definition

$VC(H)$, of hypothesis space H defined over instance space X is the size of the largest finite subset of X shattered by H. If arbitrarily large finite sets of X can be shattered by H, then $VC(H) =$

$$VC(H) = d = \log_2 2^d \leq \log_2 \mid H \mid$$

Sample Complexity and the VC Dimension

- Upper Bound to PAC Learn Any Target in C

$$m \geq \frac{1}{\epsilon}(4\log_2\frac{2}{\delta} + 8VC(H)\log_2\frac{13}{\epsilon})$$

Sample Complexity and the VC Dimension

- Upper Bound to PAC Learn Any Target in C

$$m \geq \frac{1}{\epsilon}(4 \log_2 \frac{2}{\delta} + 8VC(H) \log_2 \frac{13}{\epsilon})$$

- Lower Bound

$$m \geq \max\{\frac{1}{\epsilon} \log \frac{1}{\delta}, \frac{VC(C) - 1}{32\epsilon}\}$$

Some More General Scenario

- Learning Real-Valued Target Functions

Definition

A concept class is a nonempty set $C \subseteq 2^X$ of concepts. Assume X is a fixed set, either finite, countably infinite, $[0, l]^n$, or E^n for some $n \geq 1$. In the latter cases, we assume that each $c \in C$ is a Borel set. For $\overline{x} = (x_1, \ldots, x_\mathrm{m}) \in X^m$, $m \geq 1$, the m-sample of $c \in C$ generated by $\overline{x}$ is given by $\mathrm{sam}_\mathrm{c}(x) = (< x_1, I_\mathrm{c}(x_1) >, \ldots, < x_\mathrm{m}, I_\mathrm{c}(x_\mathrm{m}) >)$. The sample space of C, denoted S_C, is the set of all m-samples over all $c \in C$ and all $\overline{x} \in X^m$, for all $m \geq 1$. $\cdots$

Blumer, A., Ehrenfeucht, A., Haussler, D., &Warmuth, M.
Learnability and the Vapnik-Chervonenkis dimension.
Journal of the ACM, 34(4) (October), 929-965. 1989.

Natarajan, B. K.
Machine Learning: A theoretical approach.
San Mateo, CA: Morgan Kaufmann, 1991.

Some More General Scenario

- Learning from Certain Types of Noisy Data

Laird, P.
Learning from good and bad data.
Dordrecht: Kluwer Academic Publishers, 1988.

Kearns, M. J., Vazirani, U. V.
An introduction to computational learning theory.
Cambridge, MA: MIT Press, 1994.

Papers in Recent Years

Auer, P., Ortner, R.
A new PAC bound for intersection-closed concept classes
Machine learning, 2007.

Critisms of The PAC Model

- The worst-case emphasis in the model makes it unusable in practice

Critisms of The PAC Model

- The worst-case emphasis in the model makes it unusable in practice
- The notions of target concept and noise-free training data are unrealistic in practice

Critisms of The PAC Model

- The worst-case emphasis in the model makes it unusable in practice
- The notions of target concept and noise-free training data are unrealistic in practice
- Without its computational part, the result are contained by Statistical Learning Theory

Reference

Valiant, L.
A theory of the learnable.
Communications of the ACM, 27, 1984.

Blumer, A., Ehrenfeucht, A., Haussler, D., &Warmuth, M.
Learnability and the Vapnik-Chervonenkis dimension.
Journal of the ACM, 34(4) (October), 929-965. 1989.

Haussler, D.
Overview of the Probably Approximately Correct (PAC) Learning Framework.
An introduction to the topic

Reference

- Vidyasagar, M.
  A theory of learning and generalization : with applications to neural networks and control systems.
  London ; New York : Springer, c1997
- Vapnik，V. N.
  Statistical learning theory.
  New York : Wiley, c1998.
- Mitchell, T. M.著, 曾华军, 张银奎等译
  机器学习.
  北京: 机械工业出版社, 2003.

Thanks!