Aalto University
School of Science

CS-E4070 — Computational learning theory

Slide set 08 : the Vapnik-Chervonenkis dimension II

Cigdem Aslay and Aris Gionis

Aalto University

spring 2019

reading material

- K&V, chapter 3
- SS&BD, chapter 6

the VC dimension – reminder

- a set A of m instances is shattered by $\mathcal{H}$ iff there exist hypotheses in $\mathcal{H}$ that label A in all possible 2^m ways
- $\Pi_{\mathcal{H}}(A)$: restriction of $\mathcal{H}$ to A

$$\Pi_{\mathcal{H}}(A) = \{(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_m)) : h \in \mathcal{H}\}$$

- $\mathcal{H}$ shatters A iff

$$\Pi_{\mathcal{H}}(A) = \{0, 1\}^m$$

the VC dimension – reminder

- equivalent set-theoretic definitions:
  - restriction of $\mathcal{H}$ to A

$$\Pi_{\mathcal{H}}(A) = \{h \cap A : h \in \mathcal{H}\}$$

  - $\mathcal{H}$ shatters A iff

$$\Pi_{\mathcal{H}}(A) = 2^{A}$$

the VC dimension – reminder

- the VC dimension, $VCD(\mathcal{H})$, of a hypothesis class $\mathcal{H}$ is the cardinality of the largest finite subset of X shattered by $\mathcal{H}$.

$$VCD(\mathcal{H}) = sup\{|A| : \mathcal{H} \text{ shatters } A\}$$

- If $\mathcal{H}$ can shatter arbitrarily large finite sets, then

$$VCD(\mathcal{H}) = \infty$$

the VC dimension – reminder

- to show that $VCD(\mathcal{H})$ is d we need to show that:
  - there exists a set of size d which is shattered by $\mathcal{H}$
  - no set of size $d+1$ can be shattered by $\mathcal{H}$

growth function

- the VC dimension only looks at the largest set that $\mathcal{H}$ can shatter
- the growth function $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ gives the number of ways that m instances can be labeled by $\mathcal{H}$

$$\Pi_{\mathcal{H}}(m) = \max_{A \subset X, |A|=m} |\Pi_{\mathcal{H}}(A)|$$

- that is how many different dichotomies that $\mathcal{H}$ can produce maximally

$$\Pi_{\mathcal{H}}(m) = \max_{A \subset X, |A|=m} |\{(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_m)) : h \in \mathcal{H}\}|$$

growth function

- the growth function further characterizes complexity of $\mathcal{H}$: the faster the growth, more dichotomies with increasing m
- clearly, if $\mathcal{H}$ does not have finite VC dimension, then

$$\Pi_{\mathcal{H}}(m) = 2^m, \forall m$$

- if $VCD(\mathcal{H}) = d$ and $m \leq d$, then $\Pi_{\mathcal{H}}(m) = 2^m$
  - if there is a d sized set that $\mathcal{H}$ can shatter, for each integer $k < d$, there is also a set of size k that $\mathcal{H}$ can shatter

growth function

- what about $m > d$?
- the fact that $\mathcal{H}$ cannot shatter a set of size m doesn't mean that it is completely useless for sets of size m
  - it might label almost all m instances correctly, or
  - might do a horrible labeling for any m instances
- Sauer-Shelah-Perles lemma tells us what to expect when $m > d$

a polynomial bound on $\Pi_{\mathcal{H}}(m)$

- Sauer-Shelah-Perles lemma: let $\mathcal{H}$ be a hypothesis class with $VCD(\mathcal{H}) \leq d < \infty$. then, $\forall m$:

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

  and, if $m > d$ then

$$\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^{d} = \mathcal{O}(m^{d})$$

a polynomial bound on $\Pi_{\mathcal{H}}(m)$

- Sauer-Shelah-Perles lemma shows that

  when m becomes larger than d, the growth function

  increases polynomially rather than exponentially

  with sample size m

- to prove Sauer-Shelah-Perles lemma, we first need Pajor's lemma

- Pajor's lemma: for any A, the cardinality of $\Pi_{\mathcal{H}}(A)$ is

  bounded by the number of subsets of A that $\mathcal{H}$ shatters

Pajor's lemma

- lemma: let $\mathcal{H}$ be any hypothesis class with $VCD(\mathcal{H}) = d$.

  For any $A = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset X$

  $$|\Pi_{\mathcal{H}}(A)| \leq |\{B \subseteq A : \mathcal{H} \textit{ shatters } B\}|$$

- proof (sketch): by induction. for $m = 1$, either both sides are equal to 1 or are equal to 2
  - empty set is always considered to be shattered by $\mathcal{H}$
- now assume that the inequality holds for all $k < m$
- let $A' = A \setminus \{\mathbf{x}_1\}$

Pajor's lemma

- (proof cont'd.) define two sets Y_0 and Y_1:

$$Y_0 = \{(y_2, \cdots, y_m) : (0, y_2, \cdots, y_m) \in \Pi_{\mathcal{H}}(A) \ \vee \\ (1, y_2, \cdots, y_m) \in \Pi_{\mathcal{H}}(A)\}$$

and

$$Y_1 = \{(y_2, \cdots, y_m) : (0, y_2, \cdots, y_m) \in \Pi_{\mathcal{H}}(A) \ \wedge \\ (1, y_2, \cdots, y_m) \in \Pi_{\mathcal{H}}(A)\}$$

- Notice that $|\Pi_{\mathcal{H}}(A)| = |Y_0| + |Y_1|$

Pajor's lemma

- (proof cont'd.) since $Y_0 = \Pi_{\mathcal{H}}(A')$, by the induction assumption (applied on $\mathcal{H}$ and A'), we have:

$$\begin{aligned} |Y_0| = |\Pi_{\mathcal{H}}(A')| &\leq |\{B \subseteq A' : \mathcal{H} \textit{ shatters } B\}| \\ &= |\{B \subseteq A : \mathbf{x}_1 \notin B \ \wedge \ \mathcal{H} \textit{ shatters } B\}| \end{aligned}$$

- let $\mathcal{H}' \subseteq \mathcal{H}$ contain the pairs of hypotheses that agree on A' but disagree on $\mathbf{x}_1$

$$\begin{aligned} \mathcal{H}' = \{h \in \mathcal{H} : \exists h' \in \mathcal{H} \ s.t. \ h(\mathbf{x}_1) \neq h'(\mathbf{x}_1) \text{ and} \\ h(\mathbf{x}_i) = h'(\mathbf{x}_i), i = 2, \cdots, m\} \end{aligned}$$

- notice that, if $\mathcal{H}'$ shatters a set $B \subseteq A'$ it also shatters the set $B \cup \{\mathbf{x}_1\}$ and vice versa.

Pajor's lemma

- (proof cont'd.) notice also that $Y_1 = \Pi_{\mathcal{H}'}(A')$
- so by induction (applied on $\mathcal{H}$' and A') we obtain

$$\begin{aligned}
|Y_1| = |\Pi_{\mathcal{H}'}(A')| &\leq |\{B \subseteq A' : \mathcal{H}' \textit{ shatters } B\}| \\
&= |\{B \subseteq A' : \mathcal{H}' \textit{ shatters } B \cup \{\mathbf{x}_1\}\}| \\
&= |\{B \subseteq A : \mathbf{x}_1 \in B \ \wedge \ \mathcal{H}' \textit{ shatters } B\}| \\
&\leq |\{B \subseteq A : \mathbf{x}_1 \in B \ \wedge \ \mathcal{H} \textit{ shatters } B\}|
\end{aligned}$$

- Hence, we have:

$$\begin{aligned}
|\Pi_{\mathcal{H}}(A)| &= |Y_0| + |Y_1| \\
&\leq \{B \subseteq A : \mathbf{x}_1 \notin B \ \wedge \ \mathcal{H} \textit{ shatters } B\} \\
&+ |\{B \subseteq A : \mathbf{x}_1 \in B \ \wedge \ \mathcal{H} \textit{ shatters } B\}| \\
&= |\{B \subseteq A : \mathcal{H} \textit{ shatters } B\}|
\end{aligned}$$

Sauer-Shelah-Perles lemma

- Sauer-Shelah-Perles lemma: let $\mathcal{H}$ be a hypothesis class with $VCD(\mathcal{H}) \leq d < \infty$. then, $\forall m$:

  $$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

  and, if $m > d$ then

  $$\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^{d} = \mathcal{O}(m^{d})$$

Sauer-Shelah-Perles lemma

- proof: since $VCD(\mathcal{H}) \leq d$, no set with size larger than d is shattered by $\mathcal{H}$. let $A_m = \underset{A \subset X, |A|=m}{\arg\max} |\Pi_{\mathcal{H}}(A)|$
- then by Pajor's lemma it follows that for any m:

$$\Pi_{\mathcal{H}}(m) \leq |\{B \subseteq A_m : \mathcal{H} \textit{ shatters } B\}| \leq \sum_{i=0}^{d} \binom{m}{i}$$

- and when $m > d$:

$$\sum_{i=0}^{d} \binom{m}{i} < \left(\frac{em}{d}\right)^d$$

- (verify the above inequality, see Lemma A.5 in SS&BD if you need help)

polynomial sample complexity of PAC learning

- previously: finite hypothesis classes are PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$$

- if a finite hypothesis class $\mathcal{H}$ shatters a finite set A then

$$|\mathcal{H}| \geq |\Pi_{\mathcal{H}}(A)| = 2^{|A|}$$

- this immediately implies that $VCD(\mathcal{H}) \leq \log|\mathcal{H}|$
- the difference between $VCD(\mathcal{H})$ and $|\log \mathcal{H}|$ can be arbitrarily large

sample complexity upper bound

- theorem 1: let $\mathcal{C}$ be a concept class with VC dimension d. Let L be any algorithm that takes as input a set S of m labeled examples of a concept in $\mathcal{C}$ and outputs a hypothesis $h \in \mathcal{C}$ that is consistent with S.

  Then, L is a PAC learning algorithm for $\mathcal{C}$ provided that it is given a random sample of m examples from $EX(\mathcal{D}, c)$ where m satisfies

  $$m \geq a_0 \left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{d}{\epsilon} \log \frac{1}{\epsilon} \right)$$

  for some constant $a_0 > 0$.

sample complexity upper bound

- theorem 2: let $\mathcal{C}$ be any concept class. let $\mathcal{H}$ be any representation class of VC dimension d. Let L be any algorithm that takes as input a set S of m labeled examples of a concept in $\mathcal{C}$ and outputs a hypothesis $h \in \mathcal{H}$ that is consistent with S.

  Then, L is a PAC learning algorithm for $\mathcal{C}$ using $\mathcal{H}$ provided that it is given a random sample of m examples from $EX(\mathcal{D}, c)$ where m satisfies

  $$m \geq a_0 \left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{d}{\epsilon} \log \frac{1}{\epsilon} \right)$$

  for some constant $a_0 > 0$.

sample complexity upper bound - proof (sketch)

- let c denote the target concept
- denote by $c \oplus h$ the hypothesis defined as

$$(c \oplus h)(\mathbf{x}) = \begin{cases} 1 & \text{if } c(\mathbf{x}) \neq h(\mathbf{x}) \\ 0 & \text{if } c(\mathbf{x}) = h(\mathbf{x}) \end{cases}$$

- notice that $error_{\mathcal{D}}(h) = \mathbf{Pr}_{\mathbf{x} \sim \mathcal{D}}[(c \oplus h)(\mathbf{x}) = 1]$
- define the class of error regions w.r.t c and $\mathcal{H}$ as follows

$$\Delta(c) = \{c \oplus h : h \in \mathcal{H}\}$$

- notice that $VCD(\mathcal{H}) = VCD(\Delta(c))$
  - for any set S, we can map each element $h \in \Pi_{\mathcal{H}}(S)$ to a $\tilde{h} \in \Pi_{\Delta(c)}(S)$. this mapping is bijective.

sample complexity upper bound - proof (sketch)

- refine $\Delta(c)$ to consider only error regions with weight at least ϵ under $\mathcal{D}$

$$\Delta_\epsilon(c) = \{\tilde{h} \in \Delta(c) : \mathbf{Pr}_{\mathbf{x} \sim \mathcal{D}}[\tilde{h}(\mathbf{x}) = 1] \geq \epsilon\}$$

- this means that, any $h \in \mathcal{H}$ such that $c \oplus h \in \Delta_\epsilon(c)$ is potentially problematic as $error_{\mathcal{D}}(h) \geq \epsilon$
- definition: for any $\epsilon > 0$, a set S is an ϵ-net for $\Delta(c)$ if, for every $\tilde{h} \in \Delta_\epsilon(c)$, there exists $\mathbf{x} \in S$ such that $\tilde{h}(\mathbf{x}) = 1$

sample complexity upper bound - proof (sketch)

- main idea: if S is an ϵ-net for $\Delta(c)$, and L outputs $h \in \mathcal{H}$ that is consistent with S, then it must be that $error_{\mathcal{D}}(h) \leq \epsilon$
  - any $h \in \mathcal{H}$ consistent with S cannot be in $\Delta_\epsilon(c)$
- main goal: if we can bound the probability that a set S of m random examples fails to be an ϵ-net for $\Delta(c)$, then we have bounded the probability that h consistent with S has error greater than ϵ

sample complexity upper bound - proof (sketch)

- notice that for finite $\mathcal{H}$, we bound this probability by $|\mathcal{H}|(1-\epsilon)^m$
- we want to show that if we draw a small set of instances from $EX(\mathcal{D}, c)$, then they form an ϵ-net with high probability
- also we want to show that the sample size required for this depends on $VCD(\mathcal{H})$, ϵ, and δ (independent of $|\mathcal{H}|$ and $|X|$)

sample complexity upper bound - proof (sketch)

- draw a multiset S_1 of m random examples from $\mathcal{D}$
- let $\mathcal{A}$ be the event that elements of S_1 fail to form an ϵ-net for $\Delta(c)$
- suppose that $\mathcal{A}$ occurs, then there exists $\tilde{h} \in \Delta_\epsilon(c)$ such that $\tilde{h}(\mathbf{x}) = 0, \forall \mathbf{x} \in S_1$
- now, fix this $\tilde{h}$ and draw a second sample S_2 of size m
- our goal is to upper bound the probability of $\mathcal{A}$
- we will do so by obtaining a lower bound on the number of instances $\mathbf{x}$ in S_2 that satisfy $\tilde{h}(\mathbf{x}) = 1$

sample complexity upper bound - proof (sketch)

- let $\mathcal{Z}_i$ denote the random variable that takes value 1 if the i-th element $\mathbf{x}_i$ of S_2 satisfies $\tilde{h}(\mathbf{x}_i) = 1$ and 0 otherwise
- let $\mathcal{Z} = \sum_{i=1}^{m} \mathcal{Z}_i$ be the number of such instances in S_2
- notice that $\mathbf{E}[\mathcal{Z}] \geq \epsilon m$, because each element of S_2 has probability at least ϵ to hit an error region

sample complexity upper bound - proof (sketch)

- using Markov's inequality, we get

$$\mathbf{Pr}\left[\mathcal{Z} < \frac{\epsilon m}{2}\right] \leq \mathbf{Pr}\left[|\mathcal{Z} - \mathbf{E}[\mathcal{Z}]| > \frac{\mathbf{E}[\mathcal{Z}]}{2}\right] \leq 2\exp\left(-\frac{\epsilon m}{2}\right)$$

- the probability that at least $\epsilon m/2$ instances in S_2 satisfy $\tilde{h}(\mathbf{x}) = 1$ is at least $1/2$ (for $\epsilon m \geq 24$)
- let $\mathcal{B}$ be the combined event over the random draws of S_1 and S_2 that $\mathcal{A}$ occurs on the draw of S_1 (i.e., S_1 is not an ϵ-net) and S_2 has at least $\epsilon m/2$ hits in a region of $\Delta_\epsilon(c)$ that is missed by S_1

sample complexity upper bound - proof (sketch)

- the definition of $\mathcal{B}$ requires that $\mathcal{A}$ occurs on S_1
- we have shown in previous slide that $\mathbf{Pr}[\mathcal{B} \mid \mathcal{A}] \geq 1/2$
- then we have $\mathbf{Pr}[\mathcal{B}] = \mathbf{Pr}[\mathcal{B} \mid \mathcal{A}]\,\mathbf{Pr}[\mathcal{A}] \geq 1/2\mathbf{Pr}[\mathcal{A}]$
- so our goal of bounding $\mathbf{Pr}[\mathcal{A}]$ is equivalent to finding δ such that

$$\mathbf{Pr}[\mathcal{B}] \leq \frac{\delta}{2}$$

  because this would imply

$$\mathbf{Pr}[\mathcal{A}] \leq \delta$$

sample complexity upper bound - proof (sketch)

- bounding $\mathbf{Pr}[\mathcal{B}]$ is a purely combinatorial problem
- we are given $2m$ balls out of which $r \geq \epsilon m/2$ are red and the remaining are black. if we divided them into two sets of size m, without seeing the colors, what is the probability that the first set has no red balls and the second set has all of them?
- this probability is simply given by

$$\frac{\binom{m}{r}}{\binom{2m}{r}} \leq \frac{1}{2^r}$$

sample complexity upper bound - proof (sketch)

- thus we have, by the union bound over all $\tilde{h} \in \Pi_{\Delta_\epsilon(c)}(S)$

$$\begin{aligned}
\mathbf{Pr}[\mathcal{A}] \leq 2 \cdot \mathbf{Pr}[\mathcal{B}] &\leq 2 \cdot |\Pi_{\Delta_\epsilon(c)}(S)| \cdot 2^{-\frac{\epsilon m}{2}} \\
&\leq 2 \cdot |\Pi_{\Delta(c)}(S)| \cdot 2^{-\frac{\epsilon m}{2}} \\
&\leq 2 \cdot \left(\frac{2em}{d}\right)^d \cdot 2^{-\frac{\epsilon m}{2}}
\end{aligned}$$

sample complexity lower bound

- theorem: any algorithm for PAC learning a hypothesis class $\mathcal{H}$ with VC dimension d must use $\Omega(d/\epsilon)$ examples in the worst case.

sample complexity lower bound – proof (main ideas)

- let $S = \{|\mathbf{x}_1, \cdots, \mathbf{x}_d\}$ be a set of size d shattered by $\mathcal{H}$
- let $\mathcal{D}$ be a distribution defined as follows
  - $\mathcal{D}(\mathbf{x}_1) = 1 - 8\epsilon$
  - $\mathcal{D}(\mathbf{x}_j) = 8\epsilon/(d-1)$, for $j = 2, \cdots, d$
- suppose the learning algorithm L receives

$$m = \frac{d-1}{32\epsilon}$$

  examples drawn from $\mathcal{D}$

sample complexity lower bound – proof (main ideas)

- claim: L receives very few examples from the set $S \setminus \{\mathbf{x}_1\}$
- let $\mathcal{Z}_i$ be the random variable that equals 1 if the i-th example drawn from $\mathcal{D}$ is in the set $S \setminus \{\mathbf{x}_1\}$ and 0 otherwise
- then $\mathcal{Z}_i = 1$ with probability 8ϵ and $\mathcal{Z}_i = 0$ with probability $1 - 8\epsilon$

sample complexity lower bound – proof (main ideas)

- let $\mathcal{Z} = \sum_{i=1}^{m} \mathcal{Z}_i$ be the number of examples seen from the set $S \setminus \{\mathbf{x}_1\}$ (possibly with repetitions)
- $\mathbf{E}[\mathcal{Z}] = \frac{d-1}{4}$
- using Markov's inequality

$$\mathbf{Pr}\left[\mathcal{Z} \geq \frac{d-1}{2}\right] \leq \mathbf{Pr}\left[|\mathcal{Z} - \mathbf{E}[\mathcal{Z}]| \geq \mathbf{E}[\mathcal{Z}]\right] \leq 2\exp\left(-\frac{d-1}{12}\right)$$

sample complexity lower bound – proof (main ideas)

- we can simulate the example oracle by drawing examples from $\mathcal{D}$ and assigning a random label by coin tosses to any newly seen example
- for the previously seen examples, retain the labelings initially given
- since S is shattered by $\mathcal{H}$, the labeling is consistent with some $h \in \mathcal{H}$

sample complexity lower bound – proof (main ideas)

- thus any h output by L errs with probability at least $1/2$ on any example it has not seen
- hence with probability at least $2\exp\left(-\frac{d-1}{12}\right) \geq 1/2$, the error of h output by L is at least 2ϵ, as it has not seen at least half the examples from $S \setminus \{\mathbf{x}_1\}$ which has total probability mass of 8ϵ (equally distributed)