CSE 513T Theory of Artificial Intelligence & Machine Learning Sep 29, 2016

Lecture 10

Lecturer: Brendan Juba *Scribe: Shali Jiang (VCD), Muhan Zhang (NFLT)*

1 VC Dimension

1.1 Motivations

In Occam's Razor, we gave sample complexity bound using size of the representations $size(c)$ or log size of the hypothesis class $\log|\mathcal{H}|$. In this chapter, we'll see VC dimension is a lower bound of $size(c)$ or $\log|\mathcal{H}|$, and we can get tighter bounds using VC dimension, as in the following theorem:

Theorem 1 *For a class $\mathcal{C}$ of VC dimension d, with probability $1-\delta$, any $h \in \mathcal{C}$ that is consistent with $m \geq c_0(\frac{1}{\epsilon}(d\log\frac{1}{\epsilon}+\log\frac{1}{\delta}))$ examples drawn from D and labeled by $c \in \mathcal{C}$ will satisfy $Pr_{x\in D}[h(x)=c(x)] \geq 1-\epsilon$, for some constant c_0. (see page 61 of [1])*

This theorem upper bounds (i.e. we need at most this many examples) the number of examples needed for PAC learning with VC dimension, rather than size of representation or size of the concept classes (hypothesis class).

We can also give a lower bound on the number of examples needed for PAC learning with VC dimension:

Theorem 2 *If there is a PAC-learning algorithm for a class $\mathcal{C}$ of VC dimension d, the algorithm requires $\Omega(\frac{d}{\epsilon})$ examples. (see section 3.6 of [1])*

Also, when dealing with representions involved with real numbers, the size of representations can grow with higher precision of the real numbers, but that doesn't necessarily imply that we need more examples; in such cases, VC dimension can capture more essential characteristics of a representation class.

1.2 Definitions

(Also see section 3.2 of [1])

Definition 3 *For any concept class $\mathcal{C}$ over X, and any $S = \{x_1, \ldots, x_m\} \subset X$, all the dichotomies realized by $\mathcal{C}$, is defined as*

$$\Pi_{\mathcal{C}}(S) = \{(c(x_1), \ldots, c(x_m)) : c \in \mathcal{C}\}.$$

Note $\Pi_{\mathcal{C}}(S)$ is a subset of $\{0,1\}^m$,

Definition 4 *If $\Pi_{\mathcal{C}}(S) = \{0,1\}^m$ (where $m = |S|$), then we say that S is shattered by $\mathcal{C}$. Thus, S is shattered by $\mathcal{C}$ if $\mathcal{C}$ realizes all possible dichotomies of S.*

Definition 5 *The **Vapnik-Chervonenkis (VC) dimension** of $\mathcal{C}$, denoted as $VCD(\mathcal{C})$, is the cardinality d of the largest set S shattered by $\mathcal{C}$. If arbitrarily large finite sets can be shattered by $\mathcal{C}$, then $VCD(\mathcal{C}) = \infty$.*

Note by this definition, for $VCD(\mathcal{C})$ to be d, we don't need $\mathcal{C}$ to shatter every sample S with size d. The existence of one such set S suffices.

1.3 Examples

(Also see section 3.3 of [1])

- *Intervals of the Real Line.*
  Let $c \in \mathcal{C}$ be a concept induced by $[a,b]$ such that $c(x) = 1$ if $x \in [a,b]$; $c(x) = 0$ otherwise. It's easy to see that any two different points can be shattered by a interval. So the VCD is at least two. But for any three points on the real line, there is a dichotomy that is not realizable by intervals, as shown in Figure 1. Hence the VCD is two.



Figure 1: A dichotomy unrealizable by intervals

- *Linear halfspaces in the Plane.*
  Let $c \in \mathcal{C}$ the a concept induced by a hyperplane $w^T x + b = 0$ such that $c(x) = 1$ if $w^T x + b > 0$; $c(x) = 0$ otherwise. Here we consider two dimensional space, i.e. $w, x \in R^2, b \in R$. It's easy to see that any three points not colinear can be shattered by halfspaces. So the VCD is at least three. To show no four points can be shattered by halfspaces, we consider all possible cases in terms of the number of points lying on the convex hull of the four points. If the convex hull is a segment (i.e. the four points are colinear), then obviously they cannot be shattered by halfspaces. If the convex hull is not a segment. Then two cases can exist. First, all four points lie on the convex hull, as shown in the middle of Figure 2, then there is a dichotomy that can't be realized by halfspaces. Second, only three points lie on the convex hull; that means there must be a point inside the convex hull; we label that point negative, and the other three positive; this labeling can't be realized by halfspaces. We have explored all possible cases of how four points can be positioned on a plane, and for each case there is always some labeling that can't be realized by halfspaces, hence the VCD of halfspaces is three. In general, we can show VCD of halfspaces in R^n is $n + 1$.

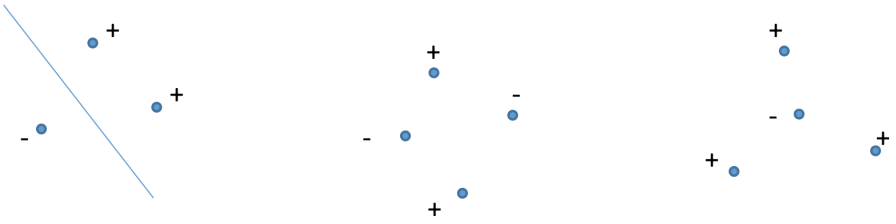

Figure 2: Left: A dichotomy and its realization by a halfspace, with the shaded region indicating the positive side. Middle and right: Dichotomies unrealizable by halfspaces.

- *Axis-Aligned Rectangles in the Plane.*
  Let $c \in \mathcal{C}$ be a concept induced by an axis-aligned rectangle in the plane such that $c(x) = 1$ if x is inside the rectangle, and 0 otherwise. It's easy to have four points shattered by rectangles, as shown in left of Figure 3. So the VCD is at least four. Note it's also easy to find four points that is not shattered by rectangles, as shown in the middle of Figure 3, but existence of one set of four points that can be shattered is enough to show the VCD is at least four. To show no five points can be shatterred by rectangles, we can reason in this way: For any five points, suppose A is the smallest rectangle that contains the five points, then there must be at least two points lying on the same edge, or there must be at least one point lying strictly inside A. For the first case, we label the two points differently; for the second case, we label the point inside negative, and the other four points positive. We can see, this labeling can't be realized by rectangles. Hence, the VCD of rectangles is four.

2 No Free Lunch Theorem

Theorem 6 *No free lunch thoerem: Let $\mathcal{C}$ be a concept class of VC dimension d. Then any algorithm that with probability $> \frac{1}{7}$ produces h s.t. $Pr_{x \in D}[h(x) = c(x)] \geq 1 - \epsilon$ must use $\Omega(\frac{d}{\epsilon})$ exmaples.*



Figure 3: Left: A dichotomy and its realization by an axis-aligned rectangle. Middle and right: Dichotomies unrealizable by axis-aligned rectangles.

Proof *Intuitively, the more complex your concept class is (i.e., the larger d is), the more examples you need to learn a good hypothesis (because complex classes mean that data can be more irregularly distributed and can still be shattered, then you need more observations to explore this irregular distribution.).*

Because the theorem does not specify the distribution D, it must hold for any unknown distribution D. We prove the theorem by looking at a worst-case data distribution D. But before that, let's first consider a better distribution – D is a uniform distribution over d points $\{x^1, ..., x^d\}$. *Because* $VCD(\mathcal{C}) = d$, *there must exist a set of d points* $\{x^1, ..., x^d\}$ *such that no matter what labels they have, there will exist a* $c \in \mathcal{C}$ *assigning the same labels as their true labels. This means that all* 2^d *possible label combinations are realizable by* $\mathcal{C}$.

Assume that D is a uniform distribution over these d points $\{x^1, ..., x^d\}$. *We also assume that the true labels of these d points are all decided by tossing a fair coin independently (label = 1 if head, label = 0 if tail). Then each of the* 2^d *possible label combinations has equal probability* $\frac{1}{2^d}$. *In addition, knowing some of the labels will not help us guess other unseen labels better because the coins are tossed independently, which means that a hypothesis produced from observed points will disagree with the true label of an unseen point with probability 1/2.*

Consider the following experiment: fix any PAC-learning algorithm for $\mathcal{C}$ *and let it observe d/2 different points from* $\{x^1, ..., x^d\}$ *to produce a hypothesis h. Then for any* x^i *not observed in the algorithm's set of examples, the hypothesis disagrees with its label with probability 1/2. (Choosing d/2 does not mean that we give the algorithm restrictions. We will later show that to observe* $\theta(d/2)$ *different points the algorithm has to sample many more points.) The expected error of h on D is:* $\sum_{i=1}^{d} Pr(x^i) \times (\text{expected error on } x^i) \geq \frac{d}{2} \times \frac{1}{d} \times (\text{expected error on unobserved } x^i) = \frac{d}{2} \times \frac{1}{d} \times \frac{1}{2} = \frac{1}{4}$. *(d/2 is the number of unobserved points under D, 1/d is the probability of each point.)*

However, $1/4$ *is just a lower bound of the expected error of h which normalizes all randomness. We want to know with what probability h can achieve even smaller error than* $1/4$. *Now we define p as the probability that the algorithm produces h with error* $< 1/8$. *We have:* $1/4 \leq$ *(expected error of h) = (probability that error* $< 1/8$*) * (the error, which is* $< 1/8$*) + (probability that error* $\geq 1/8$*) * (the error, which is between* $1/8$ *and* 1*)* $< p * 1/8 + (1 - p) * 1$.

Solving the above we have $p < 6/7$, *which means w.p.* $> 1/7$ *the h will have error* $> 1/8$ *if h uses* $d/2$ *different points. So to ensure that h has error* $< 1/8$ *w.p.* $> 1/7$, *we have to use* $\Omega(d/2)$ *examples.*

Now consider a worse case: D is no longer uniform distribution over $\{x^1, ..., x^d\}$ *but is extremely uneven. Then it will be more difficult to sample points that can explore enough the distribution. An extreme case is that D gives* x^1 *a probability* $1 - 8\epsilon$, *and gives the other* $d - 1$ *points probability* $\frac{8\epsilon}{d-1}$ *each. And we let the label distribution be unchanged (still* $1/2^d$ *for each combination).*

If the algorithm misses $\geq \frac{d-1}{2}$ of the $d-1$ rare examples, then h will on expectation be wrong on $\geq \frac{d-1}{2}\frac{1}{2}$ of these rare examples. To calculate the probability that h has a small error, we let p be the probability that h is wrong on $< \frac{1}{8}$ of these $d-1$ rare examples. Then $\frac{d-1}{2}\frac{1}{2} \leq$ (the expected number that h is wrong on the $d-1$ rare points) $< p\frac{d-1}{8} + (1-p)(d-1)$. Solving it we get $p < 6/7$ again.

So w.p. $> 1/7$, h is wrong on more than $(d-1)/8$ rare examples. Then its error $> \frac{d-1}{8}\frac{8\epsilon}{d-1} = \epsilon$. To show how many samples from D is neccesary to observe at least $(d-1)/2$ rare examples, consider tossing an unfair coin m times with prob. 8ϵ for head (which means we sample a rare example), and $1-8\epsilon$ for tail (which means we sample x^1). Assume we are lucky enough so that every rare example we sample is different. Then to get $(d-1)/2$ different rare examples we have Chernoff bound as follows:

$$\sum_{i=1}^{m} X_i = \text{number of heads } (X_i = 1 \text{ if head})$$

$$E[\sum_{i=1}^{m} X_i] = 8\epsilon m$$

$$Pr(\sum_{i=1}^{m} X_i \geq (1+\gamma)(8\epsilon)m) \leq e^{-m(8\epsilon)\gamma^2/3}$$

Let $(1+\gamma)(8\epsilon)m = \frac{d-1}{2}$, we have $m = \Theta(\frac{d}{\epsilon})$. This means that if we sample $\Theta(\frac{d}{\epsilon})$ examples and we are lucky enough (the rare examples are all different), our probability of observing more than $(d-1)/2$ different rare examples is still less than $e^{-m(8\epsilon)\gamma^2/3} = e^{-\Theta(d)}$. So to ensure the probability of observing $(d-1)/2$ different rare examples is larger than some constant ($1/7$ in our case), we need $m = \Omega(\frac{d}{\epsilon})$.

■

References

[1] Michael J. Kearns and Umesh V. Vazirani, An introduction to computational learning theory.. MIT Press, 1994.