COMP8601: Advanced Topics in Theoretical Computer Science
Lecture 9: ϵ-Nets, ϵ-Samples, VC-dimension (Part 2)
Lecturer: Hubert Chan
Date: 5 Nov 2013

These lecture notes are supplementary materials for the lectures. They are by no means substitutes for attending lectures or replacement for your own notes!

1 Getting More Randomness

Recall that we have a set X some distribution D, and C is a class of boolean functions on X such that the VC-dimension is d. We draw m independent samples from X to form a random subset S, and we wish to find out how large does m have to be in order for S to be an ϵ-net with high probability. This is the result we wish to prove.

Theorem 1.1 (Number of Samples for Class with Bounded VC-Dimension) *Suppose (X, C) has VC-dimension at most d and S is a subset obtained by sampling from X independently m times. If $m \geq \max\{\frac{4}{\epsilon}\log\frac{2}{\delta}, \frac{8d}{\epsilon}\log\frac{16e}{\epsilon}\}$, then with probability at least $1-\delta$, S is an ϵ-net.*

Using VC-dimension, we can bound the number of effective boolean functions on the random subset S. However, after conditioning on S, there is no more randomness left. We see how we can introduce extra artificial randomness in the analysis in order to make the proof works.

Alternative Experiment with More Randomness. We first sample $2m$ points independently from X according to distribution D to form $W \in X^{2m}$. We pick m coordinates uniformly at random from W to form $S \in X^m$. Observe that S has the same distribution as before, but now we have more randomness.

Define A to be the event that there is some $f \in C_\epsilon$ such that for all $x \in S$, $f(x) = 0$.

Define B to be the event that there is some $f \in C_\epsilon$ such that

1. For all $x \in S$, $f(x) = 0$.
2. There exist at least $\frac{\epsilon m}{2}$ points in W such that $f(x) = 1$.

We have $B \subseteq A$. We wish to prove that if $Pr[B]$ is small, then so is $Pr[A]$.

Lemma 1.2 *$Pr[A] \leq 2Pr[B]$, for $m \geq \frac{8\ln 2}{\epsilon}$.*

Proof: It suffices to show that $Pr[\overline{B}|A] \leq \frac{1}{2}$. Then, we would have $Pr[B] = Pr[A \cap B] = Pr[B|A]Pr[A] \geq \frac{1}{2} \cdot Pr[A]$.

We next consider the conditional probability of B given A. Because $Pr[\overline{B}|A] = E[Pr[\overline{B}|A, S]]$, we show an upper bound for the probability conditioned on the random object S

Since A happens, there is some $f_0 \in C_\epsilon$ such that $f_0(x) = 0$ for all $x \in S$. Observe that after we condition on S, all the remaining m points in $W \setminus S$ are still totally random and unknown. Hence, the number Y of points x in $W \setminus S$ such that $f_0(x) = 1$ is a sum of m independent $\{0, 1\}$-random variables, each having expectation at least ϵ.

Event $\overline{B}$ implies that $Y < \frac{\epsilon m}{2} \leq \frac{E[Y]}{2}$. Hence, by Chernoff Bound, $Pr[Y < \frac{1}{2}E[Y]] \leq \exp(-\frac{1}{2} \cdot (\frac{1}{2})^2 E[Y]) \leq \exp(-\frac{1}{8}\epsilon m) \leq \frac{1}{2}$, for $m \geq \frac{8 \ln 2}{\epsilon}$. Hence, it follows that $Pr[\overline{B}|A, S] \leq \frac{1}{2}$. ■

2 Using the Extra Randomness

Lemma 2.1 $Pr[B] \leq (\frac{2em}{d})^d \cdot 2^{-\frac{\epsilon m}{2}}$.

Proof: We next give an upper bound on $Pr[B]$. Using conditional probability, we have $Pr[B] = E[Pr[B|W]]$. Observe that once we fix W, we only need to consider the class $C(W)$ of boolean functions. Since (X, C) has VC-dimension d, it follows that $|C(W)| \leq (\frac{2em}{d})^d$.

For each $f \in C(W)$, we let B_f to be the event that

1. For all $x \in S$, $f(x) = 0$.
2. There exist at least $\frac{\epsilon m}{2}$ points in W such that $f(x) = 1$.

Then, we have $Pr[B|W] \leq Pr[\cup_{f \in C(W)} B_f|W] \leq \sum_{f \in C(W)} Pr[B_f|W]$. Hence, it suffices to obtain a uniform bound on $Pr[B_f|W]$, for each $f \in C(W)$.

Observe that once W and f are both fixed, we exactly know which of the $2m$ points are marked 1 and which are marked 0. The only randomness left is how we pick m random points to form S. At this point, if we see that the number of points in W marked 1 is less than $\frac{\epsilon m}{2}$, then we have $Pr[B_f|W] = 0$. Also, if the number of points W marked 1 is more than m, then we know that there must be a point in S that would be marked 1, and so in this case we also have $Pr[B_f|W] = 0$.

We are left with the case when the number of points marked 1 is $L \geq \frac{\epsilon m}{2}$. The number of ways to choose S such that none of the L points are contained is $\binom{2m-L}{m}$. It follows that

$$Pr[B_f|W] \leq \frac{\binom{2m-L}{m}}{\binom{2m}{m}} = \frac{m}{2m} \cdot \frac{m-1}{2m-1} \cdot \cdots \cdot \frac{m-L+1}{2m-L+1} \leq \frac{1}{2^L} \leq 2^{-\frac{\epsilon m}{2}}.$$

Hence, we have $Pr[B|W] \leq |C(W)| \cdot 2^{-\frac{\epsilon m}{2}} \leq (\frac{2em}{d})^d \cdot 2^{-\frac{\epsilon m}{2}}$. Taking expectation again, we have the required upper bound on $Pr[B]$. ■

2.1 Choosing the Right Value for m

It follows that we have $Pr[A] \leq 2(\frac{2em}{d})^d \cdot 2^{-\frac{\epsilon m}{2}}$. We next show that this is at most δ when $m \geq \max\{\frac{4}{\epsilon} \log \frac{2}{\delta}, \frac{8d}{\epsilon} \log \frac{16e}{\epsilon}\}$.

Observe that the required result is equivalent to

$\frac{\epsilon m}{2} \geq d \log \frac{2em}{d} + \log \frac{2}{\delta}$.

From the choice of m, we have $\frac{\epsilon m}{4} \geq \log \frac{2}{\delta}$.

It suffices to check that $\frac{\epsilon m}{4} \geq d \log \frac{2em}{d}$. Putting $m = \frac{8d}{\epsilon} \log \frac{16e}{\epsilon}$, this is equivalent to $\frac{16e}{\epsilon} \geq \log \frac{16e}{\epsilon}$, which is certainly true since $\frac{16e}{\epsilon} \geq 16$.

3 ϵ-Sample

We have seen that an ϵ-net for X under some class C is some subset $S \subseteq X$ such that if a function $f \in C$ marks at least an ϵ fraction of points in X positive, then f also marks at least 1 point in S positive.

We next consider S such that for each function $f \in C$, about the same fraction of points in S and X are marked positive by f. But now, instead of being a set, S can have repetitions too. Given a bag (multi-set) S of points in X and a function $f \in C$, we denote by $\mathsf{Avg}_S[f]$ the fraction of points in S marked 1 by f.

Observe that $E_X[f]$ depends on the distribution D on X, while $\mathsf{Avg}_S[f]$ assumes that each copy in S has the same weight. Of course, if a point appears more times in S, it would have higher weight.

Definition 3.1 (ϵ-Sample) *An ϵ-sample S for a set X with distribution D under a class C of boolean functions on X is a bag (multi-set) of points from X satisfying the following:*

For each $f \in C$, $|E_X[f] - \mathsf{Avg}_S[f]| \leq \epsilon$.

Example

Suppose X are points in the plane $\mathbb{R}^2$ with some distribution, and C is the class of functions, each of which corresponds to an axis-aligned rectangle that marks the points inside 1 and 0 otherwise. Then, an ϵ-sample S is a bag of points such if the fraction (weighted according to D) of points in X contained inside a rectangle is p, then the fraction of points in S contained in the same rectangle is $p \pm \epsilon$.

3.1 ϵ-Sample for Finite C

Similar to the case for ϵ-net, we can bound the number of independent samples to form an ϵ-sample, when the class C of boolean functions is finite.

Theorem 3.2 *Suppose C is finite and S is a subset obtained by sampling from X independently m times. If $m \geq \frac{1}{2\epsilon^2} \ln \frac{2|C|}{\delta}$, then with probability at least $1-\delta$, S is an ϵ-sample.*

Proof: Fix a function $f \in C$. Let $p := E_X[f]$. Suppose x_i is a point drawn from distribution D on X, and define $Z_i := f(x_i)$. Then, it follows that $E[Z_i] = p$. Let S be a bag formed from the x_i's, $i \in [m]$. Then, we have $\mathsf{Avg}_S[f] = \frac{1}{m}\sum_i Z_i$.

Hence, by Hoeffding's Inequality,

$Pr[|E_X[f(x)] - \mathsf{Avg}_S[f]| > \epsilon] \leq 2\exp(-2m\epsilon^2)$.

By the union bound, the probability that S fails for some function $f \in C$ is at most $|C| \cdot 2\exp(-2m\epsilon^2)$, which is at most δ, for $m \geq \frac{1}{2\epsilon^2} \ln \frac{2|C|}{\delta}$. ■

3.2 ϵ-Sample for Infinite C

For the case of infinite C, we use the same approach as that for ϵ-net. We assume that (X, C) has VC-dimension d and use similar techniques to obtain the following result.

Theorem 3.3 *Suppose (X, C) has VC-dimension at most d. Then, suppose S is a bag of points in*

X obtained by sampling from X under distribution D independently m times. If $m \geq \Omega(\frac{1}{\epsilon^2}(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}))$, then with probability at least $1 - \delta$, S is an ϵ-sample.

We shall prove this result in a homework question.

4 Homework Preview

1. **ϵ-Sample for (X, C) with VC-dimension d.** Suppose X is a set and C is a collection of boolean functions such that (X, C) has VC-dimension d. In this question, we derive a sufficient number m of independent random samples from X with distribution D such that the resulting bag S is an ϵ-sample under class C of boolean functions with probability at least $1 - \delta$.

   (a) **Introducing Extra Randomness.** Suppose we sample $2m$ copies independently from X to form the bag W. Then, we pick m copies out of W at random to form S. In other words, W can be view as a tuple in X^{2m}, and we pick m distinct coordinates at random and use them to form S.

   Let A be the event that there exists some $f \in C$ such that $|E_X[f] - \mathsf{Avg}_S[f]| > \epsilon$.

   Let B be the event that there exists some $f \in C$ such that $|E_X[f] - \mathsf{Avg}_S[f]| > \epsilon$ and $|\mathsf{Avg}_W[f] - \mathsf{Avg}_S[f]| > \frac{\epsilon}{4}$.

   Prove that $Pr[A] \leq 2Pr[B]$.

   (Hint: Show that $Pr[\overline{B}|A] \leq \frac{1}{2}$.

   Observe that given A, the event $\overline{B}$ implies that there is some $f_0 \in C$ such that $|E_X[f_0] - \mathsf{Avg}_S[f_0]| > \epsilon$ and $|\mathsf{Avg}_W[f_0] - \mathsf{Avg}_S[f_0]| \leq \frac{\epsilon}{4}$. This means that $|E_X[f_0] - \mathsf{Avg}_{W \setminus S}[f_0]| > \frac{\epsilon}{2}$.

   Use Hoeffding's Inequality and you may assume $m \geq \frac{2 \ln 4}{\epsilon^2}$.)

   (b) **Conditional Probability.** For $f \in C$, define B_f to be the event that $|E_X[f] - \mathsf{Avg}_S[f]| > \epsilon$ and $|\mathsf{Avg}_W[f] - \mathsf{Avg}_S[f]| > \frac{\epsilon}{4}$. (Hence, $B = \cup_f B_f$.)

   Fix $f \in C$. Define H_f to be the event that $|\mathsf{Avg}_W[f] - \mathsf{Avg}_S[f]| > \frac{\epsilon}{4}$. Then, clearly $B_f \subseteq H_f$, and so $Pr[B_f|W] \leq Pr[H_f|W]$. We analyze $Pr[H_f|W]$

   Suppose $P_{max} := \max_{f \in C} Pr[H_f|W]$. Prove that $Pr[B] \leq (\frac{2em}{d})^d \cdot P_{max}$.

   (Hint: Recall that (X, C) has VC-dimension d. After conditioning on W which has only $2m$ points, how many boolean functions can the class C induce on W?)

   (c) **Bounding P_{max}.** This is the most technical part of the proof and this part differs the most from the proof for ϵ-net.

   After W and f are fixed, we know precisely how many copies in W are marked 1 by f. Let this number be L. The only randomness left is the choice of S among W. Recall that S is formed from W by choosing m copies from the $2m$ copies in W.

   We can order the objects in W in an arbitrary list, and assign one by one whether each object is in S in the following way: suppose when object a is considered, there are already x objects assigned to S and y objects assigned to $W \setminus S$. Then, object a is assigned to S with probability $\frac{m-x}{(m-x)+(m-y)}$ and to $W \setminus S$ with probability $\frac{m-y}{(m-x)+(m-y)}$.

i. Suppose the L objects marked 1 are being considered first. For $1 \leq i \leq L$, let u_i be the variable that takes value 1 if the ith object is assigned to S and -1 if it is assigned to $W \setminus S$.
Define $U_i := \sum_{j=1}^{i} u_j$. Compute the probability that the $(i+1)$st object is assigned to S in terms of i and U_i.
What does it mean when $U_i > 0$? When $U_i > 0$, what happens to this probability? Are the u_i's independent?

ii. Find an expression β in terms of ϵ and m such that $|\mathsf{Avg}_W[f] - \mathsf{Avg}_S[f]| > \frac{\epsilon}{4}$ *iff* $U_L^2 > \beta$.
(We want to obtain an upper bound for $Pr[U_L^2 > \beta]$.)

iii. We saw that the u_i's are not independent. This makes the analysis difficult. Hence, we would like to compare the u_i's with another collection of independent random variables. For each $1 \leq i \leq L$, we define independent random variable γ_i that takes values in $\{-1, 1\}$ uniformly, i.e., each value with probability $\frac{1}{2}$. Define $Y_i := \sum_{1 \leq j \leq i} \gamma_j$.
Observe that we would like U_L^2 to be small. Can you explain intuitively why Y_L^2 is more likely to be larger than U_L^2?
Prove that $E[U_L^2] \leq E[Y_L^2]$.
(Hint: Prove by induction on i that $E[U_i^2] \leq E[Y_i^2]$. In the inductive step, you might find considering the conditional probability $Pr[U_i u_{i+1} | U_i]$ useful.)
(Optional: Prove that for all non-negative integers r, $E[U_L^{2r}] \leq E[Y_L^{2r}]$. You may use this result for later parts of the question.)

iv. Let t be a real number. Prove that $E[\exp(tU_L^2)] \leq E[\exp(tY_L^2)]$.
(Hint: Recall the Taylor expansion $\exp(y) := \sum_{r \geq 0} \frac{y^r}{r!}$.)

v. By considering moment generating functions, prove an upper bound for $Pr[U_L^2 > \beta]$, and conclude that $P_{max} \leq 2\exp(-\frac{\epsilon^2 m}{32})$.
(Hint: Recall from the lecture on Johnson-Lindenstrauss Lemma, we have $E[\exp(tY_L^2)] \leq (1 - 2tL)^{-1/2}$, for $t < \frac{1}{2L}$.)

(d) **Wrapping Everything Up.** Prove that if $m \geq \max\{\frac{64}{\epsilon^2} \ln \frac{4}{\delta}, \frac{256d}{\epsilon^2} \ln \frac{16e}{\epsilon}\}$, then with probability at least $1 - \delta$, the bag S is an ϵ-sample for X under class C.