Chapter 24

VC Dimension, ε-nets and ε-approximation

By Sariel Har-Peled, November 20, 2008[1]

> "I've never touched the hard stuff, only smoked grass a few times with the boys to be polite, and that's all, though ten is the age when the big guys come around teaching you all sorts to things. But happiness doesn't mean much to me, I still think life is better. Happiness is a mean son of a bitch and needs to be put in his place. Him and me aren't on the same team, and I'm cutting him dead. I've never gone in for politics, because somebody always stand to gain by it, but happiness is an even crummier racket, and their ought to be laws to put it out of business."
> – Emile Ajar, Momo.

In this lecture, we would be interested in using sampling to capture or learn a concept. For example, consider an algorithm that tries to learn a classifier, that given positive and negative examples, construct a model of the universe. For example, the inputs are records of clients, and we would like to predict whether or not one should give them a loan.

Clearly, we are trying to approximate a function. The natural question to ask, is how many samples one needs to learn a concept reliably? It turns out that this very fundamental question has a (partial) answer, which is very useful in the developement of algorithms.

24.1 VC Dimension

Definition 24.1.1 A *range space* S is a pair $(X, \mathcal{R})$, where X is a (finite or infinite) set and $\mathcal{R}$ is a (finite or infinite) family of subsets of X. The elements of X are *points* and the elements of $\mathcal{R}$ are *ranges*. For $A \subseteq X$, $P_{\mathcal{R}}(A) = \left\{ r \cap A \;\middle|\; r \in \mathcal{R} \right\}$ is the *projection* of $\mathcal{R}$ on A.

If $P_{\mathcal{R}}(A)$ contains all subsets of A (i.e., if A is finite, we have $|P_{\mathcal{R}}(A)| = 2^{|A|}$) then A is *shattered* by $\mathcal{R}$.

The *Vapnik-Chervonenkis* dimension (or VC-dimension) of S, denoted by $\mathsf{VC}(S)$, is the maximum cardinality of a shattered subset of X. It there are arbitrarily large shattered subsets then $\mathsf{VC}(S) = \infty$.

24.1.1 Examples

Example. Let $X = \mathbb{R}^2$, and let $\mathcal{R}$ be the set of disks in the plane. Clearly, for three points in the plane $1, 2, 3$, one can find 8 disks that realize all possible 2^3 different subsets.

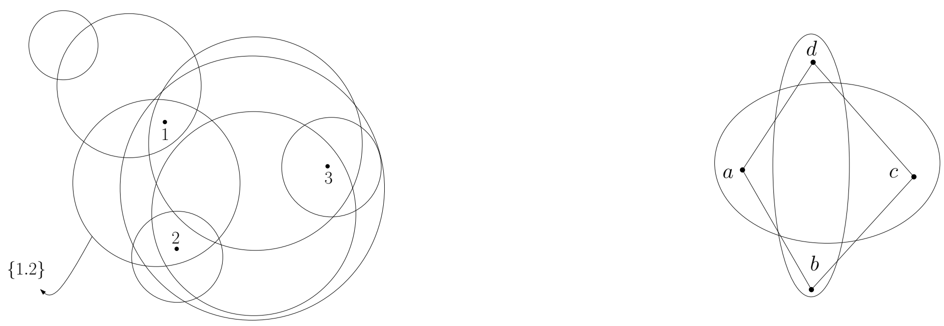
But can disks shatter a set with four points? Consider such a set P of four points, and there are two possible options. Either the convex-hull of P has three points on its boundary, and in this case, the subset having those vertices in the subset but not including the middle point is impossible, by convexity. Alternatively, if all four points are vertices of the convex hull, and they are p_1, p_2, p_3, p_4 along the boundary of the convex hull, either the set $\{p_1, p_3\}$ or the set $\{p_2, p_4\}$ is not realizable. Indeed, if both options are realizable, then consider the two disks D_1, D_2 the realizes those assignments. Clearly, D_1 and D_2 must intersect in four points, but this is not possible, since two disks have at most two intersection points. See Figure 24.1 (b).

Example. Consider the range space $S = (\mathbb{R}^2, \mathcal{R})$, where $\mathcal{R}$ is the set of all (closed) convex sets in the plane. We claim that the $\mathsf{VC}(S) = \infty$. Indeed, consider a set U of n points $p_1, \ldots, p_n$ all lying on the boundary of the unit circle in the plane. Let V be any subset of U, and consider the convex-hull $\mathcal{CH}(V)$. Clearly, $\mathcal{CH}(V) \in \mathcal{R}$, and furthermore, $\mathcal{CH}(V) \cap U = V$. Namely, any subset of U is realizable by S. Thus, S can shatter sets of arbitrary size, and its VC dimension is unbounded.

Example 24.1.2 Let $S = (X, R)$, where $X = \mathbb{R}^d$ and R is the set of all (closed) halfspaces in $\mathbb{R}^d$. To see what is the VC dimension of S, we nee the following result of Radon:

> **Theorem 24.1.3 (Randon's Lemma)** *Let A be a set of $d + 2$ points in $\mathbb{R}^d$. Then, there exists two disjoint subsets C, D of A, such that $\mathcal{CH}(C) \cap \mathcal{CH}(D) \neq \emptyset$.*
>
> *Proof:* The points $p_1, \ldots, p_{d+2}$ of A are linearly dependent. As such, there exists $\beta_1, \ldots, \beta_{d+2}$, not all of them zero, such that $\sum_i \beta_i p_i = 0$ and $\sum_i \beta_i = 0$ (to see that, remember that the affine subspace spanned by $p_1, \ldots, p_{d+2}$ is induced by all points that can be represented as $p_1 + \sum_{i=2}^{d+2} \alpha_i (p_i - p_1)$ where $\sum_i \alpha_i = 0$). Assume, for the sake of simplicity of exposition, that the $\beta_1, \ldots \beta_k \geq 0$ and $\beta_{k+1}, \ldots, \beta_{d+2} < 0$. Furthermore, let $\mu = \sum_{i=1}^{k} \beta_i$. We have that
>
> $$\sum_{i=0}^{k} \beta_i p_i = -\sum_{i=k+1}^{d+2} \beta_i p_i.$$
>
> In particular, $v = \sum_{i=0}^{k} (\beta_i/\mu) p_i$ is a point in the $\mathcal{CH}(\{p_1, \ldots, p_k\})$ and $\sum_{i=k+1}^{d+2} -(\beta_i/\mu) p_i \in \mathcal{CH}(\{p_{k+1}, \ldots, p_{d+2}\})$. We conclude that v is in the intersection of the two convex hulls, as required. ■

In particular, this implies that if a set Q of $d+2$ points is being shattered by S, we can partition this set Q into two disjoint sets A and B such that $\mathcal{CH}(A) \cap \mathcal{CH}(B) \neq \emptyset$. It should now be clear that any halfspace h^+ containing all the points of A, must also contain a point of the $\mathcal{CH}(B)$. But this implies that a point of B must be in h^+. Namely, the subset A can not be realized by a halfspace, which implies that Q can not be shattered. Thus $\mathsf{VC}(S) < d + 2$. It is also easy to verify that the regular simplex with $d + 1$ vertices is being shattered by S. Thus, $\mathsf{VC}(S) = d + 1$.



(a) (b)

Figure 24.1: Disks in the plane can shatter three points, but not four.

24.2 VC-Dimensions and the number of different ranges

Let

$$g(d,n) = \sum_{i=0}^{d} \binom{n}{i}.$$

Note that for all $n, d \geq 1$, $g(d,n) = g(d, n-1) + g(d-1, n-1)$

Lemma 24.2.1 (Sauer's Lemma) *If (X, R) is a range space of* VC*-dimension d with $|X| = n$ points then $|R| \leq g(d,n)$.*

Proof: The claim trivially holds for $d = 0$ or $n = 0$.

Let x be any element of X, and consider the sets

$$R_x = \left\{ r \setminus \{x\} \;\middle|\; x \in r, r \in R, r \setminus \{x\} \in R \right\}$$

and

$$R \setminus x = \left\{ r \setminus \{x\} \;\middle|\; r \in R \right\}.$$

Observe that $|R| = |R_x| + |R \setminus x|$ (Indeed, if r does not contain x than it is counted in R_x, if does contain x but $r \setminus x \notin R$, then it is also counted in R_x. The only remaining case is when both $r \setminus \{x\}$ and $r \cup \{x\}$ are in R, but then it is being counted once in R_x and once in $R \setminus x$.)

Observe that R_x has VC dimension $d-1$, as the largest set that can be shattered is of size $d-1$. Indeed, any set $A \subset X$ shattered by R_x, implies that $A \cup \{x\}$ is shattered in R.

Thus,

$$|R| = |R_x| + |R \setminus x| = g(n-1, d-1) + g(n-1, d) = g(d,n),$$

by induction. ∎

By applying Lemma 24.2.1, to a finite subset of X, we get:

Corollary 24.2.2 *If $(X, \mathcal{R})$ is a range space of* VC*-dimension d then for every finite subset A of X, we have $|P_{\mathcal{R}}(A)| \leq g(d, |A|)$.*

Lemma 24.2.3 *Let $S = (X, \mathcal{R})$ and $S' = (X, \mathcal{R}')$ be two range spaces of dimension d and d', respectively, where $d, d' > 1$. Let $\widehat{\mathcal{R}} = \left\{ r \cup r' \,\middle|\, r \in \mathcal{R}, r' \in \mathcal{R}' \right\}$. Then, for the range space $\widehat{S} = (X, \widehat{\mathcal{R}})$, we have that* $\mathsf{VC}\left(\widehat{S}\right) = O((d + d') \log(d + d'))$

Proof: Let A be a set of n points in X that are being shattered by $\widehat{S}$. There are $g(n, d)$ and $g(n, d')$ different assignments for the elements of A by ranges of $\mathcal{R}$ and $\mathcal{R}'$, respectively. Every subset C of A realized by $\widehat{r} \in \widehat{\mathcal{R}}$, is a union of two subsets $A \cap r$ and $A \cap r'$ where $r \in \mathcal{R}$ and $r' \in \mathcal{R}'$. Thus, the number of different subsets of A realized by $\widehat{S}$ is bounded by $g(n, d)g(n, d')$. Thus, $2^n \leq n^d n^{d'}$, for $d, d' > 1$. We conclude $n \leq (d + d') \lg n$, which implies that $n \leq O((d + d') \log(d + d'))$. ∎

24.3 On ε-nets and ε-sampling

Definition 24.3.1 Let (X, R) be a range space, and let A be a finite subset of X. For $0 \leq \varepsilon \leq 1$, a subset $B \subseteq A$, is an *ε-sample* for A if for any range $r \in R$, we have

$$\left| \frac{|A \cap r|}{|A|} - \frac{|B \cap r|}{|B|} \right| \leq \varepsilon.$$

Similarly, $N \subseteq A$ is an *ε-net* for A, if for any range $r \in R$, if $|r \cap A| \geq \varepsilon |A|$ implies that r contains at least one point of N (i.e., $r \cap N \neq \emptyset$).

Theorem 24.3.2 *There is a positive constant c such that if (X, R) is any range space of* VC-*dimension at most d, $A \subseteq X$ is a finite subset and $\varepsilon, \delta > 0$, then a random subset B of cardinality s of A where s is at least the minimum between $|A|$ and*

$$\frac{c}{\varepsilon^2}\left(d \log \frac{d}{\varepsilon} + \log \frac{1}{\delta}\right)$$

is an ε-sample for A with probability at least $1 - \delta$.

Theorem 24.3.3 (ε-net Theorem) *Let (X, R) be a range space of* VC-*dimension d, let A be a finite subset of X and suppose $0 < \varepsilon, \delta < 1$. Let N be a set obtained by m random independent draws from A, where*

$$m \geq \max\left(\frac{4}{\varepsilon} \log \frac{2}{\delta}, \frac{8d}{\varepsilon} \log \frac{8d}{\varepsilon}\right). \tag{24.1}$$

Then N is an ε-net for A with probability at least $1 - \delta$.

24.4 Proof of the ε-net Theorem

Let (X, R) be a range space of VC-dimension d, and let A be a subset of X of cardinality n. Suppose that m satisfiers Eq. (24.1). Let $N = (x_1, \ldots, x_m)$ be the sample obtained by m independent samples from A (the elements of N are not necessarily distinct, and thats why we treat N as a ordered set). Let E_1 be the probability that N fails to be an ε-net. Namely,

$$E_1 = \left\{ \exists r \in R \,\middle|\, |r \cap A| \geq \varepsilon n, r \cap N = \emptyset \right\}.$$

(Namely, there exists a "heavy" range r that does not contain any point of N.) To complete the proof, we must show that $\mathbf{Pr}[E_1] \le \delta$. Let $T = (y_1, \ldots, y_m)$ be another random sample generated in a similar fashion to N. Let E_2 be the event that N fails, but T "works", formally

$$E_2 = \left\{ \exists r \in R \;\middle|\; |r \cap A| \ge \varepsilon n, r \cap N = \emptyset, |r \cap T| \ge \frac{\varepsilon m}{2} \right\}.$$

Intuitively, since $E_T\Big[|r \cap T|\Big] \ge \varepsilon m$, then for the range r that N fails for, we have with "good" probability that $|r \cap T| \ge \frac{\varepsilon n}{2}$. Namely, E_1 and E_2 have more or less the same probability.

Claim 24.4.1 $\mathbf{Pr}[E_2] \le \mathbf{Pr}[E_1] \le 2\,\mathbf{Pr}[E_2]$.

Proof: Clearly, $E_2 \subseteq E_1$, and thus $\mathbf{Pr}[E_2] \le \mathbf{Pr}[E_1]$. As for the other part, note that $\mathbf{Pr}\Big[E_2 \;\Big|\; E_1\Big] = \mathbf{Pr}[E_2 \cap E_1] \,/\, \mathbf{Pr}[E_1] = \mathbf{Pr}[E_2] \,/\, \mathbf{Pr}[E_1]$. It is thus enough to show that $\mathbf{Pr}\Big[E_2 \;\Big|\; E_1\Big] \ge 1/2$.

Assume that E_1 occur. There is $r \in R$, such that $|r \cap A| > \varepsilon n$ and $r \cap N = \emptyset$. The required probability is at least the probability that for this specific r, we have $|r \cap T| \ge \frac{\varepsilon n}{2}$. However, $|r \cap T|$ is a binomial variable with expectation εm, and variance $\varepsilon(1-\varepsilon)m \le \varepsilon m$. Thus, by Chebychev inequality (Theorem 24.7.1),

$$\mathbf{Pr}\left[|r \cap T| < \frac{\varepsilon m}{2}\right] \le \mathbf{Pr}\left[\Big||r \cap T| - \varepsilon m\Big| > \frac{\varepsilon m}{2}\right] \mathbf{Pr}\left[\Big||r \cap T| - \varepsilon m\Big| > \frac{\sqrt{\varepsilon m}}{2}\sqrt{\varepsilon m}\right] \le \frac{4}{\varepsilon m} \le \frac{1}{2},$$

by Eq. (24.1). Thus, $\mathbf{Pr}[E_2] \,/\, \mathbf{Pr}[E_1] = \mathbf{Pr}\Big[|r \cap T| \ge \frac{\varepsilon n}{2}\Big] = 1 - \mathbf{Pr}\Big[|r \cap T| < \frac{\varepsilon m}{2}\Big] \ge \frac{1}{2}$. ∎

Thus, it is enough to bound the probability of E_2. Let

$$E_2' = \left\{ \exists r \in R \;\middle|\; r \cap N = \emptyset, |r \cap T| \ge \frac{\varepsilon m}{2} \right\},$$

Clearly, $E_2 \subseteq E_2'$. Thus, bounding the probability of E_2' is enough to prove the theorem. Note however, that a shocking thing happened! We no longer have A as participating in our event. Namely, we turned bounding an event that depends on a global quantity, into bounding a quantity that depends only on local quantity/experiment. This is the crucial idea in this proof.

Claim 24.4.2 $\mathbf{Pr}[E_2] \le \mathbf{Pr}\Big[E_2'\Big] \le g(d, 2m)2^{-em/2}$.

Proof: We imagine that we sample the elements of $N \cup T$ together, by picking a set $Z = (z_1, \ldots, z_{2m})$ from A, by picking each element independently from A. Next, we randomly decide which of the m elements of Z form N, and remaining elements from T. Clearly,

$$\mathbf{Pr}[E_2'] = \sum_Z \mathbf{Pr}\Big[E_2' \;\Big|\; Z\Big] \mathbf{Pr}[Z].$$

Thus, from this point on, we fix the set Z, and we bound $\mathbf{Pr}\Big[E_2' \;\Big|\; Z\Big]$.

It is now enough to consider the ranges in the projection space $P_R(Z)$. By Lemma 24.2.1, we have $|P_R(Z)| \le g(d, 2m)$.

Let us fix any $r \in P_R(Z)$, and consider the event

$$E_r = \left\{ r \cap N = \emptyset \text{ and } |r \cap T| > \frac{\varepsilon m}{2} \right\}.$$

For $k = |r \cap (N \cup T)|$, we have

$$\begin{aligned}
\mathbf{Pr}[E_r] &\le \mathbf{Pr}\left[r \cap N = \emptyset \;\middle|\; |r \cap (N \cup T)| > \frac{\varepsilon m}{2}\right] = \frac{\binom{2m-k}{m}}{\binom{2m}{m}} \\
&= \frac{(2m-k)(2m-k-1)\cdots(m-k+1)}{2m(2m-1)\cdots(m+1)} \\
&= \frac{m(m-1)\cdots(m-k+1)}{2m(2m-1)\cdots(2m-k+1)} \le 2^{-k} \le 2^{-\varepsilon m/2}.
\end{aligned}$$

Thus,

$$\mathbf{Pr}\left[E_2' \;\middle|\; Z\right] \le \sum_{r \in P_R(Z)} \mathbf{Pr}[E_r] \le |P_R(Z)|\, 2^{-\varepsilon m/2} = g(d, 2m) 2^{-\varepsilon m/2},$$

implying that $\mathbf{Pr}\left[E_2'\right] \le g(d, 2m) 2^{-\varepsilon m/2}$. ■

Proof of Theorem 24.3.3: By Lemma 24.4.1 and Lemma 24.4.2, we have $\mathbf{Pr}[E_1] \le 2g(d, 2m)2^{-\varepsilon m/2}$. It is thus remains to verify that if m satisfies Eq. (24.1), then $2g(d, 2m)2^{-\varepsilon m/2} \le \delta$. One can verify that this inequality is implied by Eq. (24.1).

Indeed, we know that $2m \ge 8d$ and as such $g(d, 2m) = \sum_{i=0}^{d} \binom{2m}{i} \le \sum_{i=0}^{d} \frac{(2m)^i}{i!} \le (2m)^d$, for $d > 1$. Thus, it is sufficient to show that the inequality $2(2m)^d 2^{-\varepsilon m/2} \le \delta$ holds. By taking lg of both sides and rearranging, we have that this is equivalent to

$$\frac{\varepsilon m}{2} \ge d \lg(2m) + \lg \frac{2}{\delta}.$$

By our choice of m (see Eq. (24.1)), we have that $\varepsilon m/4 \ge \lg(2/\delta)$. Thus, we need to show that

$$\frac{\varepsilon m}{4} \ge d \lg(2m).$$

We verify this inequality for $m = \frac{8d}{\varepsilon} \lg \frac{8d}{\varepsilon}$, indeed

$$2d \lg \frac{8d}{\varepsilon} \ge d \lg\left(\frac{16d}{\varepsilon} \lg \frac{8d}{\varepsilon}\right).$$

This is equivalent to $\left(\frac{8d}{\varepsilon}\right)^2 \ge \frac{16d}{\varepsilon} \lg \frac{8d}{\varepsilon}$. Which is equivalent to $\frac{4d}{\varepsilon} \ge \lg \frac{8d}{\varepsilon}$, which is certainly true for $0 \le \varepsilon \le 1$ and $d > 1$. Note that it is easy to verify that the inequality holds for $m \ge \frac{8d}{\varepsilon} \lg \frac{8d}{\varepsilon}$, by deriving both sides of the inequality.

This completes the proof of the theorem. ■

24.5 Exercises

Exercise 24.5.1 (Flip and Flop.) (A) **[5 Points]** Let $b_1, \ldots, b_{2m}$ be m binary bits. Let Ψ be the set of all permutations of $1, \ldots, 2m$, such that for any $\sigma \in \Psi$, we have $\sigma(i) = i$ or $\sigma(i) = m + i$, for $1 \leq i \leq m$, and similarly, $\sigma(m+i) = i$ or $\sigma(m+i) = m+i$. Namely, $\sigma \in \Psi$ either leave the pair $i, i+m$ in their positions, or it exchange them, for $1 \leq i \leq m$. As such $|\Psi| = 2^m$.

Prove that for a random $\sigma \in \Psi$, we have

$$\mathbf{Pr}\left[\left|\frac{\sum_{i=1}^m b_{\sigma(i)}}{m} - \frac{\sum_{i=1}^m b_{\sigma(i+m)}}{m}\right| \geq \varepsilon\right] \leq 2e^{-\varepsilon^2 m/2}.$$

(B) **[5 Points]** Let Ψ' be the set of all permutations of $1, \ldots, 2m$. Prove that for a random $\sigma \in \Psi'$, we have

$$\mathbf{Pr}\left[\left|\frac{\sum_{i=1}^m b_{\sigma(i)}}{m} - \frac{\sum_{i=1}^m b_{\sigma(i+m)}}{m}\right| \geq \varepsilon\right] \leq 2e^{-C\varepsilon^2 m/2},$$

where C is an appropriate constant. [**Hint:** Use (A), but be careful.]

(C) **[10 Points]** Prove Theorem 24.3.2 using (B).

Exercise 24.5.2 (Dual VC dimension.) Let $(X, \mathcal{R})$ be a range space with VC dimension d, and let $A \subseteq X$ be a finite set. Consider the induced range space $\mathsf{S} = (A, P_R(A))$.

Next, for a point $\mathsf{q} \in A$, let $\mathcal{R}(\mathsf{q})$ denote the set of all the ranges of $P_R(A)$ that contains is, and consider the *dual range space* $\mathsf{D} = \left(P_R(A), \left\{\mathcal{R}(\mathsf{q}) \;\middle|\; \mathsf{q} \in A\right\}\right)$.

Prove that the VC dimension of D is at most 2^d.

Exercise 24.5.3 (On VC dimension.) (A) Prove directly a bound on the VC dimension of the range space of ellipses in two dimensions (i.e., the ranges are the interior of ellipses). Show a matching lower bound (or as matching as you can).

(B) Prove that the VC dimension of regions defined by a polynomial of degree at most s in d dimensions is bounded. Such an inequality might be for example $ax^2 + bxy + y^3 - x^2y^2 \leq 3$ ($s = 2+2 = 4$ in this example), and the region it defines is all the points that comply with this inequality.

[**Hint:** Consider a mapping of $\mathbb{R}^d$ into $\mathbb{R}^k$, such that all polynomials of degree s correspond to linear inequalities.]

Exercise 24.5.4 (Dual VC dimension.) Let $(X, \mathcal{R})$ be a range space with VC dimension d, and let $A \subseteq X$ be a finite set. Consider the induced range space $\mathsf{S} = (A, P_R(A))$.

Next, for a point $\mathsf{q} \in A$, let $\mathcal{R}(\mathsf{q})$ denote the set of all the ranges of $P_R(A)$ that contains is, and consider the *dual range space* $\mathsf{D} = \left(P_R(A), \left\{\mathcal{R}(\mathsf{q}) \;\middle|\; \mathsf{q} \in A\right\}\right)$.

Prove that the VC dimension of D is at most 2^d.

Exercise 24.5.5 (Improved Hitting Set.) Let $(X, \mathcal{R})$ be a range space with constant VC dimension d. Furthermore, assume that you have access to an oracle, such that given a finite set $A \subseteq X$ of n elements, it computes the range space $\mathsf{S} = (A, P_{\mathcal{R}}(A))$ in time $O(|A| + |P_{\mathcal{R}}(A)|)$.

(A) Assume, that ever element of $\mathsf{q} \in A$ has an associated weight w_{q}, where the weight is a positive integer number. Show, how to compute ε-net efficiently so that it is an ε-net for the weighted points.

(B) In fact, the computation in the previous part would be slow if the weights are very large integers. To make things easier, assume very weight w_{q} is of the form 2^j, where j is a non-negative integer bounded by a parameter M. Show how to compute efficiently an ε-net in this case. (You can assume that computations on integers smaller than $M^{O}(1)$ can be performed in constant time.)

(C) Prove the following theorem:

Theorem 24.5.6 *Let $(X, \mathcal{R})$ be a range space with constant VC dimension d. Let A be subset of X with n elements. Furthermore, assume that there is a* hitting set *$H \subseteq A$ of size k for $(A, P_{\mathcal{R}}(A))$. Namely, any range $\mathbf{r}$ of $P_{\mathcal{R}}(A)$ contains a point of H.*

Then one can compute in polynomial time, a set U of $O(dk \log(dk))$ points of X, such that U is a hitting set for $\mathsf{S} = (A, P_{\mathcal{R}}(A))$.

To this end, assign weight 1 to all the points of A. Next, consider an δ-net for S, for the appropriate δ. If it is the required hitting set, then we are done. Otherwise, consider a "light" range (which is not being hit) and double the weight of its elements. Repeat. Argue that this algorithm terminates (by comparing the weight of H to the weight of the whole set A). What is the number of iterations of the algorithm being performed? What is the required value of δ? What is the exact size of the generated hitting set.

(D) Show a polynomial time algorithm that compute a hitting set of the range space $\mathsf{S} = (A, P_{\mathcal{R}}(A))$, of size $O(kd \log(kd))$, where d is the VC dimension of S, $n = |A|$, and k is the smallest hitting set of S. What is the expected running time of your algorithm?

(This is interesting because in general the smallest hitting set of a range space can not be approximated within a factor better than $\Omega(\log n)$ unless $P = NP$.)

24.6 Bibliographical notes

The exposition here is based on [AS00]. The usual exposition of the ε-net/ε-sample tend to be long and tedious in the learning literature. The proof of the ε-net theorem is due Haussler and Welzl [HW87]. The proof of the ε-sample theorem is due to Vapnik and Chervonenkis [VC71]. However, the importance of Vapnik and Chervonenkis result was not realized at the time, and only in the late eighties the strong connection to learning was established.

An alternative proof of both theorems exists via the usage of discrepancy. Using discrepancy, one can compute ε-samples and ε-nets deterministically. In fact, in some geometric cases, discrepancy yields better results than the ε-net and ε-sample theorem. See [Mat99, Cha01] for more details.

Exercise 24.5.1 is from Anthony and Bartlett [AB99].

24.7 From previous lectures

Theorem 24.7.1 (Chebychev inequality) *Let X be a random variable with $\mu_x = \mathbf{E}[X]$ and σ_x be the standard deviation of X. That is $\sigma_X^2 = \mathbf{E}\left[(X - \mu_x)^2\right]$. Then, $\mathbf{Pr}[|X - \mu_X| \geq t\sigma_X] \leq \dfrac{1}{t^2}$.*

Bibliography

[AB99] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge, 1999.

[AS00] N. Alon and J. H. Spencer. *The probabilistic method*. Wiley Inter-Science, 2nd edition, 2000.

[Cha01] B. Chazelle. *The Discrepancy Method: Randomness and Complexity*. Cambridge University Press, New York, 2001.

[HW87] D. Haussler and E. Welzl. ε-nets and simplex range queries. *Discrete Comput. Geom.*, 2:127–151, 1987.

[Mat99] J. Matoušek. *Geometric Discrepancy*. Springer, 1999.

[VC71] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.