CEU "Topics in Combinatorics" course

Topics covered in the March 11, 2014 class

Let us fix a set system $\mathcal{R}$. We call the sets in $\mathcal{R}$ *ranges*. Typical examples: circular disks is the plane, triangles in the plane, etc. The restriction of $\mathcal{R}$ to a finite set H is $\mathcal{R}(H) = \{R \cap H \mid R \in \mathcal{R}\}$.

Recall from last class that a set S is called *shattered* by $\mathcal{R}$ if $\mathcal{R}(H)$ contains every subset of H, i.e., $|\mathcal{R}(H)| = 2^{|H|}$.

The Vapnik-Chervonenkis dimension (VC-dimension) of $\mathcal{R}$ is the maximal size of a shattered set. This is interesting if finite.

Sauer Lemma (from last week): If $|H| = n$ and the VC-dimension of $\mathcal{R}$ is d, then $|\mathcal{R}(H)| \leq \sum_{i=0}^{d} \binom{n}{i}$.

We can choose a small subset S of size k of a larger set H of size n. This subset represents H with respect to a range R within error ϵ if $\left| \frac{|R\cap H|}{n} - \frac{|R\cap S|}{k} \right| \leq \epsilon$. Take a random k-subset S of H. The probability that S fails to represent H with respect to a fixed range R within error ϵ can be bounded by the Chernoff bound: this probability is at most $2e^{-\epsilon^2 k/2}$. Note that the error with which S represents H with respect to R depends only on $R \cap H$. Thus, if $|\mathcal{R}(H)| \leq n^d$ (in particular, if the VC-dimension is d), we can choose $k = 2d \log n/\epsilon^2$ and the union bound gives that the random k-subset S of H represents H within error ϵ with respect to *all ranges* $R \in \mathcal{R}$ with positive probability. Vapnik and Chervonenkis showed that if $\mathcal{R}$ has VC-dimension d, then much smaller sets (namely of size $k = O(\frac{d}{\epsilon^2} \log \frac{d}{\epsilon^2})$) also exist that represent H within error ϵ with respect to all ranges $R \in \mathcal{R}$. Note that here k does not depend on n.

We prove a similar statement for so called ϵ-nets (see Haussler-Welzl theorem below). $S \subseteq H$ is said to be an *ϵ-net* with respect to $\mathcal{R}$ if for every $R \in \mathcal{R}$ with $|R \cap H| \geq \epsilon n$ we have $R \cap S \neq \emptyset$. This is a weaker property than the one considered above, and smaller sizes are enough. A random subset S of size k fails this property with regard to a single fixed range R with $|R \cap H| \geq \epsilon n$ with probability at most $(1-\epsilon)^k < e^{-\epsilon k}$. If $|\mathcal{R}(H)| \leq n^d$ a random k-subset is an ϵ-net with respect to $\mathcal{R}$ with probability at least $1 - n^d e^{-\epsilon k}$. This shows the existence of ϵ-nets of size $(d/\epsilon)\log n$. As above, the challenge is to get rid of the dependence on n. The next theorem replaces the $\log n$ term with the typically much smaller $\log(d/\epsilon)$ term in case the VC-dimension of $\mathcal{R}$ id d.

Haussler-Welzl theorem (1987): Let $\mathcal{R}$ be a family of ranges with VC-dimension d, H be a set of size n and $0 < \epsilon < 1/100$ a parameter. There exists an ϵ-net for H with respect to $\mathcal{R}$ of size $k = 6\frac{d}{\epsilon} \log \frac{d}{\epsilon}$. In fact, a random k-subset of H works with probability over $1/2$.

This is the high level description of the proof. We choose a random $2k$ element subset Z first and then select our final set S as a random size k subset of Z in a second step. The probability that a range was reasonably well represented in Z but we loose all its elements in the second step can be well estimated and is small. When using the union bound it is enough to consider the restrictions of the ranges to Z and this is a small number independent of n: $|\mathcal{R}(Z)| < |Z|^d$. On the other hand if S is not an ϵ-net, namely there is a range $R \in \mathcal{R}$ with $|R \cap H| \geq \epsilon n$ but $R \cap S = \emptyset$, then most likely we "lost R" in the second step, $Z \setminus S$ probably has decent size intersection with R.

Proof: Take two uniform random k-subsets S and T of H and let Z be their union. The following procedure produces the same (uniform) distribution on S and T. First take a uniform random $2k$-subset Z of H, then uniformly randomly split Z into two equal parts: S and T. We will use both views.

It is technically simpler to choose subsets with repetitions. That is, instead of a k element subset S we choose a sequence $s_1, \ldots, s_k$ each element uniformly picked from H, but some of the elements may coincide. The same works for the elements of T and for Z. This is a technical thing to make the calculation of probabilities simpler. For example we do not have to worry about choosing S and T to be disjoint.

Let A be the event that S is not an ϵ-net for H with respect to $\mathcal{R}$. Let B be the event that there exists a range $R \in \mathcal{R}$ with $R \cap S = \emptyset$ and $|R \cap T| \geq l := \epsilon k/2$. Technical note: here T is a sequence $t_1, \ldots, t_k$ and $|R \cap T|$ really means $|\{i \mid t_i \in R\}|$, that is, we count elements with their multiplicities..

We claim two statements:

(i) $Pr[B] \geq Pr[A]/2$,

(ii) $Pr[B] \leq (2k)^d/2^l$.

Clearly, these claims together imply that $Pr[A] \leq 2(2k)^d/2^l$. The proof is finished by simply evaluating this expression for the known values of k and l and realizing that it is less than 1/2 as the theorem claims. (Do the calculation.) Technical note: the set $S = \{s_1, \ldots, s_k\}$ may actually be smaller than size k if there are some coincidences $s_i = s_j$. In this case simply extend S to a size k set randomly, this will not ruin its property of being an ϵ-net.

It remains only to prove the two statements of the claim.

For the proof of (i) we take the first view of our random process and fix the choice of S so that A is satisfied, i.e., S is no ϵ-net. We still choose T randomly from $H \setminus S$ and show that the probability that B holds is at least 1/2. Let us fix a range $R \in \mathcal{R}$ showing that S is no ϵ-net, namely $|R \cap H| \geq \epsilon n$, but $R \cap S = \emptyset$. We choose all elements t_i of T uniformly randomly, so $Pr[t_i \in R] = |R \cap H|/n \geq \epsilon$. We choose k elements t_i independently, each falls in R with probability at least ϵ, so the expected number of them is $E[|T \cap R|] \geq \epsilon k = 2l$. From the expectation being $2l$ it is easy to conclude that the probability of reaching l (thus making event B hold) is large. In class I applied a version of the Chernoff bound that was not quite enough, then improvised a direct calculation. But the simplest is to invoke a stronger form the Chernoff bound. See both forms in the appendix. The second form implies that after fixing S such that A holds B will hold with probability at least $1 - e^{-l/4}$, that is, almost certainly, clearly with probability way above 1/2.

For the proof of (ii) we take the second view of our random process and fix Z first. In order for B to happen we need to have $R \in \mathcal{R}$ with $m := |R \cap Z| \geq l$ and we have to split Z such that all elements in $R \cap Z$ end up in T and none in S. Each element has probability 1/2 ending up in T, so all would end up in T with probability exactly 2^{-m} if these choices were independent. They are not entirely independent, but this only decreases the probability. The exact probability is $\binom{2k-m}{k}/\binom{2k}{k} \leq 2^{-m} \leq 2^{-l}$. This is the bound that a given range R causes B to happen. We apply the union bound. As B does not depend on anything outside the fixed set Z we need to multiply our probability bound with the number of ranges in $\mathcal{R}(Z)$. This is bounded by the VC-dimension and the Sauer lemma finishing the proof of (ii). Q.E.D.

It is clear that for "reasonable" families $\mathcal{R}$ the ϵ-nets for certain sets need to be of size at least $\lfloor 1/\epsilon \rfloor$. For this it is enough to choose this many pairwise disjoint ranges and making H consist of an equal number of points in each. Any ϵ-net for H must then contain a point in each of the ranges. (Note: this can be done for disks, triangles, etc., but not for half-planes, as no three half-planes in the plane are pairwise disjoint. Still, one cannot get away with fewer points in an ϵ-net with respect to half-planes. Why?)

For many types of ranges $O(1/\epsilon)$ size ϵ-nets exists for any set. These include disks and half-planes in the plane (see assignment). Noga Alon proved that this is not the case for triangles, rectangles, or even lines in the plane.

Theorem (Alon, 2010): $O(1/\epsilon)$ size ϵ-nets do not always exist for planar point sets with respect to any of (a) triangles, (b) rectangles, (c) lines.

Proof: The proof is based on the density Hales Jewett theorem of Fürstenberg and Katznelson. The asymptotic bound $O(1/\epsilon)$ would mean the existence of a constant c such that ϵ-nets always exist with size at most c/ϵ. We construct a counterexample for this claim. Let $k > 2c$ be an integer. Using DHJ, choose m such that any set $T \subseteq [k]^m$ with $|T| \geq k^m/2$ contains a combinatorial line. Set $\epsilon = k^{1-m}$. The set $H_0 = [k]^m$ has size $n = k^m$ and any ϵ-net for H_0 with respect to the combinatorial lines has size over $n/2$. Indeed, if $|S| \leq n/2$, then its complement $[k]^m \setminus S$ contains a combinatorial line of size $k = \epsilon n$ and S is disjoint from this line. Thus any ϵ-net for H_0 has size over $n/2 > c/\epsilon$.

We can consider H_0 to be a set in the m-dimensional Euclidean space, where combinatorial lines are of the form $L \cap H_0$ for a real line L. To turn this construction to a planar one we take an orthogonal projection of H_0 to the plane. Let us choose a projection ϕ such that for a real line L defining a combinatorial line $L \cap H_0$ and a point $x \in [k]^m \setminus L$ we have $\phi(x) \notin \phi(L)$. A generic (or random) projection will do. We have to satisfy finitely many conditions, each violated if our projection satisfies a linear constraint.

The planar set $H = \phi(H_0)$ will "behave as" H_0, namely the images of a combinatorial lines will show up among the restrictions of lines to H. Thus no ϵ-net for H with respect to lines exists with size at most $\frac{c/}{\epsilon}$.

The same set H works also for triangles and rectangles, simply because among the restrictions of the triangle or rectangle ranges to H one finds the lines too (and also other sets). Q.E.D.

Appendix

Two forms of the Chernoff inequality

Let X_i be independent 0-1 valued random variables (Bernoulli random variables) with $Pr[X_i = 1] = p_i$. Let $S = \sum_{i=1}^n X_i$. Its expectation is $\mu := E[S] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n p_i$. For any $\delta \geq 0$ we have

$$Pr[X \leq \mu - \delta n] \leq e^{-\delta^2 n/2},$$

$$Pr[X \leq (1-\delta)\mu] \leq e^{-\mu\delta^2/2}.$$

In words: You make n independent coin flips, possibly with coins of varying biases. The bias of coin i (the probability of falling on head) is p_i (in our applications all p_i are the same). S is the total number of heads. Its expectation is $\mu := \sum_{i=1}^n p_i$.

The first form compares the distance of S from its expectation to n (absolute error), the second form compares it to μ (relative error). In case $\mu << n$ the second form is stronger.