A simplified proof of Haussler's packing Theorem

Nikita Zhivotovskiy[1]

[1]Technion

Based on https://arxiv.org/abs/1711.10414

VC dimension

Let $V \subseteq \{0,1\}^n$.

For $I = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$ denote the projection

$$V|_I = \{(v_{i_1}, \ldots, v_{i_k}) : v \in V\}.$$

Definition: Vapnik-Chervonenkis (VC) dimension of V

VC dimension of V is the largest d such that there is $I \subset \{1, \ldots, n\}, |I| = d$ *with the following property*

$$|V|_I| = 2^d.$$

0	0	1	1	0
0	1	1	1	0
1	0	0	1	0
1	1	1	0	0
0	0	1	1	1
0	1	1	1	1
1	0	1	0	0
1	1	0	1	0

Lemma: V-C'68, Sauer'71, Shelah'72

For $V \subset \{0, 1\}^n$ with VC dimension d

$$|V| \leq \sum_{i=0}^{d} \binom{n}{i}.$$

Note that for $n \geq d$

$$\left(\frac{n}{d}\right)^d \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d$$

In many applications we also need to understand the covering and packing properties of V (when VC dimension is bounded by d).

For $v, u \in V$ let $\rho_H(v, u)$ denote the Hamming distance between v and u.

Question

Assume that $V \subset \{0,1\}^n$ has VC dimension d and for any two distinct $u, v \in V$ we have $\rho_H(u, v) \geq k$.

What can we say about $|V|$ in this case?

History: R. Dudley, Ann. of Probability, 1978

$$|V| \leq C_d \left(\frac{n}{k}\right)^d \log^d \left(\frac{n}{k}\right),$$

where C_d depends only on d.

D. Haussler, JoCT, Ser. A, 1995 (submitted 91)

$$|V| \leq e(2d+1)\left(\frac{2en}{k}\right)^d,$$

The proof was simplified by Chazzele in 1992.

In the book of Jiri Matousek (Geometric discrepancy, 1999) the proof of Haussler is described "a probabilistic argument which looks like a magician's trick".

If we consider the 'normalized' distance $\rho = \rho_H/n$ and consider ε-separated subsets of V in ρ then the result of Haussler implies:

$$|V| \leq \left(\frac{10}{\varepsilon}\right)^d.$$

Up to constant factors this coincides with the packing number of the unit sphere in $\mathbb{R}^d$ — the maximal number of $\varepsilon/2$-balls one can pack in the unit ball.

The proof

Up to some point the proof follows the lines of the original proof of Haussler. We need the following definition.

Definition: Unit distance graph

For $V \subset \{0,1\}^n$ define the following graph:

- *set of vertexes is V;*
- *set of edges: any two $v, u \in V$ are connected iff $\rho_H(u, v) = 1$.*

Lemma: Haussler

If $V \subset \{0,1\}^n$ has VC dimension d then it is possible to orient the unit distance graph of V in a way such that the out-degree of each vertex is at most d.

Shifting

The proof is very instructive: For a column i, change each 1 to a 0, unless it would lead to a row that is already in the table.
Shifting *all* the columns from left to right gives:

0	1	0	1	1
1	0	0	1	1
1	1	1	0	1
0	1	1	0	0
0	0	0	1	0

$\longrightarrow$

0	1	0	0	0
0	0	0	1	1
0	0	0	0	1
0	0	0	0	0
0	0	0	1	0

It is easy to check that when *all* the columns are shifted from left to right the resulting set V^* will have the following properties:

- $|V| = |V^*|$,
- $VCdim(V^*) \leq VCdim(V)$,
- If (V, E) is a unit-distance graph of V and (V^*, E^*) is a unit-distance graph of V^* then $|E^*| \geq |E|$.
- All the vectors in V^* have at most d ones (this implies the VC lemma). Therefore, the edge density $|E^*|/|V^*| < d$. In particular, $|E|/|V| < d$.

To prove the orientation result we need the following result (based on the application of Hall's theorem)

Theorem: Alon, Tarsi 1992

If the graph and all of its subgraphs have the edge density bounded by k then we may orient the graph in a way such that the out-degree of each vertex is at most k.

Prediction problem

From here we choose a path which differs from the original argument.

- Our opponent chooses $v^* \in V$, which we do not know.
- We know V and observe both I and $v^*|_I$, where I is a set obtained by uniform sampling from $\{1, \ldots, n\}$ exactly m times (we may have copies of the same element, so that $|I| < m$).
- Our aim is to construct an estimate $\hat{v}$ (based on what we observe) such that

$$\mathbb{E}\, \rho_H(\hat{v}, v^*)/n \quad \text{is small},$$

We need the following algorithm, which takes its roots in the paper of Haussler, Littlestone and Warmuth, 1988.

Given $V \subset \{0,1\}^n$ for all $M \subseteq \{1,\ldots,n\}$ orient the one-distance graph corresponding to V in a way such that the max out-degree is at most d. This provides a deterministic family of orientations.

Given $I \subset \{1,\ldots,n\}$ and $v^*|_I$ consider the following vector $\hat{v}_I$ (for a vector $v \in \{0,1\}^n$ let $v(i)$ is its i-th coordinate)

- For all $i \in I$ set $\hat{v}_I(i) = v^*(i)$.
- For $i \notin I$ if all vectors $u \in V$ such that $v^*|_I = u|_I$ have the same coordinate $u(i)$, then set $\hat{v}_I(i) = u(i)$.
- For $i \notin I$ if there are $u, w \in V$ such that $v^*|_I = u|_I = w|_I$ but $u(i) \neq w(i)$ set $v^*(i)$ according to the direction of the edge in the orientation of the graph corresponding to $V|_{I \cup i}$:
  if the edge goes to $w(i)$ to $u(i)$ then set $\hat{v}_I(i) = u(i)$, otherwise $\hat{v}_I(i) = w(i)$.

A simple computation shows that for $\hat{v}_I$ constructed this way the following inequality holds

$$\mathbb{E}\,\frac{\rho_H(\hat{v}_I, v^*)}{n} \le \frac{d}{m+1}.$$

Indeed, let $M = \{M_1, \ldots, M_{m+1}\} \subset \{1, \ldots, n\}$ of size $m+1$. Denote $M^{\setminus i} = M \setminus \{M_i\}$. Observe that the following holds:

$$\frac{1}{m+1}\sum_{i=1}^{m+1} \mathbb{1}\{\hat{v}_{M^{\setminus i}}(i) \ne v^*(i)\} \le \frac{\text{outdegree of } v^*}{m+1} \le \frac{d}{m+1}.$$

At the same time, since all the summands have the same distribution if elements of M were sampled uniformly from $\{1, \ldots, M\}$ we have

$$\mathbb{E}\,\frac{1}{m+1}\sum_{i=1}^{m+1} \mathbb{1}\{\hat{v}_{M^{\setminus i}}(i) \ne v^*(i)\}$$

$$= Pr\{\hat{v}_{M^{\setminus 1}}(1) \ne v^*(1)\} = \mathbb{E}\,\frac{\rho_H(\hat{v}, v^*)}{n}.$$

Some trivial computations

Recall $\mathbb{E}\,\frac{\rho_H(\hat{v}_I, v^*)}{n} \leq \frac{d}{m+1}$.

Using Markov's inequality we have for any $\varepsilon \geq 0$

$$Pr\left\{\frac{\rho_H(\hat{v}_I, v^*)}{n} \geq \frac{\varepsilon}{2}\right\} < \frac{2d}{m\varepsilon},$$

therefore, for $\delta \in [0, 1]$ if $m = \frac{2d}{\varepsilon\delta}$ then

$$1 - \delta \leq Pr\left\{\frac{\rho_H(\hat{v}_I, v^*)}{n} < \frac{\varepsilon}{2}\right\}.$$

Recall that we want to understand the size of V under the assumption that V has VC dimension d and for any two distinct $u, v \in V$ it holds $\frac{\rho(u,v)}{n} \geq \frac{k}{n} = \varepsilon$.

Now we proceed with the lower bound argument taking its roots in the paper of Benedek and Itai, 1991.

We slightly abuse the notation: when v^* is a 'target' and I is a set of observations denote $\hat{v}_{v^*} := \hat{v}_I$.

Observe that when for $u, w \in V$ it holds $u\big|_I = w\big|_I$ we have

$$\hat{v}_u = \hat{v}_w.$$

However, in this case since for any two distinct $u, w \in V$ we have $\frac{\rho_H(u,w)}{n} \geq \varepsilon$ it may not happen that simultaneously

$$\frac{\rho_H(\hat{v}_u, u)}{n} < \varepsilon/2 \quad \text{and} \quad \frac{\rho_H(\hat{v}_w, w)}{n} < \varepsilon/2$$

Just because of the contradiction with the triangle inequality.

Finally, using the previous slide together with the VC lemma in the last line we have for $m = \frac{2d}{\varepsilon\delta}$ that ($\mathbb{E}$ is with respect to the choice of I)

$$\begin{aligned} 1 - \delta &\leq \frac{1}{|V|} \sum_{v \in V} Pr \left\{ \frac{\rho_H(\hat{v}_v, v)}{n} < \frac{\varepsilon}{2} \right\} \\ &= \frac{1}{|V|} \mathbb{E} \sum_{v \in V} \mathbb{1} \left\{ \frac{\rho_H(\hat{v}_v, v)}{n} < \frac{\varepsilon}{2} \right\} \\ &\leq \frac{1}{|V|} \mathbb{E} \, |V|_I| \leq \frac{1}{|V|} \left(\frac{em}{d} \right)^d = \frac{1}{|V|} \left(\frac{2e}{\varepsilon\delta} \right)^d . \end{aligned}$$

Therefore,

$$|V| \leq \inf_{\delta \in (0,1)} \frac{1}{1 - \delta} \left(\frac{2e}{\varepsilon\delta} \right)^d \leq e(d+1) \left(\frac{2e}{\varepsilon} \right)^d .$$