Advanced Machine Learning - Hilary Term 2017
3 : VC Dimension

Lecturer: Varun Kanade

We have seen that a consistent learner can be used to design a PAC-learning algorithm, provided the output hypothesis comes from a class that is not too large, in particular as long as the log of the size of the hypothesis class is polynomial in the required factors. However, when the concept class or hypothesis class is infinite, this result cannot be applied at all. Concept classes that are uncountably infinite are often used in machine learning, linear threshold functions (a.k.a. linear halfspaces) being the most common one. We have already studied the class of axis-aligned rectangles, and proved the correctness of a PAC-learning algorithm for this class using first principles. In this lecture, we'll show that provided a specific *capacity measure* called VC dimenion of a concept class can be bounded, then a consistent learner can be used to design PAC-learning algorithms. In particular, the VC dimension can be finite even for concept classes that are uncountably infinite.

1 VC Dimension

In order to keep the notation cleaner, we will avoid using the subscript n to indicate that we are discussing the learning problem with instances of size n and that the concept class is over X_n. However, it should be clear that the discussion applies to a concept class defined as $\bigcup_{n\geq 1} C_n$. Let $S \subset X$ be a finite set of instances. For a concept $c : X \to \{0,1\}$, we can consider the restriction of c to S, $c|_S : S \to \{0,1\}$, where $c|_S(x) = c(x)$ for $x \in S$. We define the following:

$$\Pi_C(S) = \{c|_S \mid C \in C\}. \tag{1}$$

The set $\Pi_C(S)$ is the number of distinct restrictions of concepts in C defined by the set S. Alternatively, if $S = \{x_1, \ldots, x_m\}$, we can associate each element of $\Pi_C(S)$ with the values at each of the m points,

$$\Pi_C(S) = \{(c(x_1), \ldots, c(x_m)) \mid c \in C\} \tag{2}$$

The set $\Pi_C(S)$ is referred to as the set all possible dichotomies on S induced by C. Clearly for a set S of size m, $|\Pi_C(S)| \leq 2^m$, as C contains boolean functions. If for a set S of size m, $|\Pi_C(S)| = 2^m$, we say that S is *shattered* by C.

Definition 1 (Shattering). *We say that a finite set $S \subset X$ is shattered by C, if $|\Pi_C(S)| = 2^{|S|}$. S is shattered by C if all possible dichotomies over S can be realised by C.*

We can now define the notion of *dimension* for a concept class C, called the Vapnik-Chervonenkis dimension, named after the authors of the seminal paper that introduced this notion to learning theory.

Definition 2 (VC Dimension). *The Vapnik-Chervonenkis dimension of C denoted as $\mathsf{VCD}(C)$ is the cardinality d of the largest set S shattered by C. If C shatters arbitrarily large finite sets, then $\mathsf{VCD}(C) = \infty$.*

1.1 Examples

The language used to define VC-dimension is a bit different from that commonly used in machine learning. Let us use some examples to clarify this idea. The notion of shattering can be phrased as follows, given a set of points $S \subset X$, if we assign labels 0 or 1 (or $+$ or $-$) to the points in S arbitrarily, is there a concept $c \in C$ that is consistent with the labels? If the answer is always yes, then the set S is shattered by C, otherwise it is not.

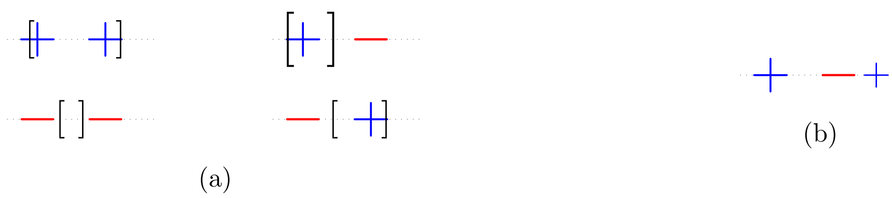

Figure 1: (a) All possible dichotomies on 2 points can be realised using intervals. (b) A dichotomy on three points that cannot be realised by intervals.

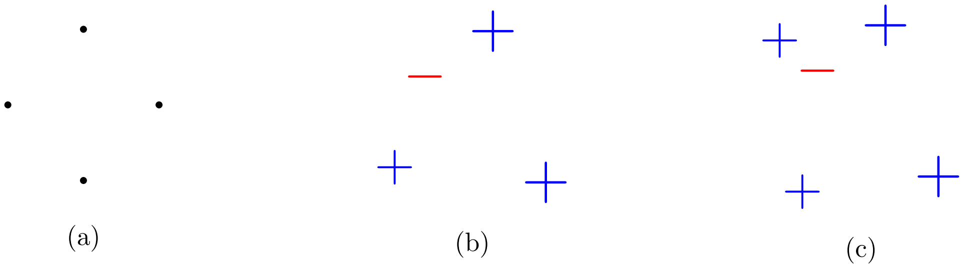

Figure 2: (a) A set of 4 points on which all dichotomies can be realised using rectangles. (b) A set of 4 points with a dichotomy that cannot be realised by rectangles. (c) Any set of 5 points always has a dichotomy that cannot be realised using rectangles.

Intervals in $\mathbb{R}$

Let $X = \mathbb{R}$ and let $C = \{c_{a,b} \mid a, b \in \mathbb{R}, a < b\}$ be the concept class of intervals, where $c_{a,b} : \mathbb{R} \to \{0, 1\}$ is defined as $c_{a,b}(x) = 1$ if $x \in [a, b]$ and 0 otherwise. What is $\mathsf{VCD}(C)$? It is easy to see that any subset $S \subset R$ of size 2 can be shattered by C, but not a set of size 3 as shown in Figure 1. Given a set of size three, if the middle point is labelled negative and the other two positive, there is no interval consistent with the labelling. Thus, $\mathsf{VCD}(C) = 2$.

Rectangles in $\mathbb{R}^2$

Let $X = \mathbb{R}^2$ and let C be the concept class of axis-aligned rectangles. Figure 2(a) shows a set of size 4 that can be shattered, Fig. 2(b) shows a set of size 4 that cannot be shattered, by providing an explicit labelling that cannot be achieved. However, the definition of VC dimension only requires the existence of a set of a certain size that is shattered. It is possible to show that no set of size 5 can be shattered. The reason being that there must be one of the five points that is not the extreme left, right, bottom or top point. If this point is labelled as negative and all the extreme points are labelled as positive, then there is no rectangle that can achieve this dichotomy.

Linear Threshold Functions or Linear Halfspaces

The concept class of linear threshold functions is widely used in machine learning applications. Let us show that the class of linear threshold functions in $\mathbb{R}^2$ has VC-dimension 3. Fig. 3(a) shows a set of size 3 that can be shattered by linear threshold functions; Fig. 3(b) shows a set of size 3 that cannot be shattered by linear threshold functions. No set of size 4 can be shattered by linear threshold functions. There are two possibilities, either the convex hull has four vertices in which case if the opposite ends of the quadrilateral are given the same labels, but adjacent vertices are given opposite ones, then no linear threshold function can achieve this labelling (Fig. 3 (c)). If on the other hand the convex hull only contains three vertices, if the vertices of the convex hull are labelled positive and the point in the interior is labelled negative, this labelling is not consistent with any linear threshold function (see Fig. 3 (d)). The degenerate

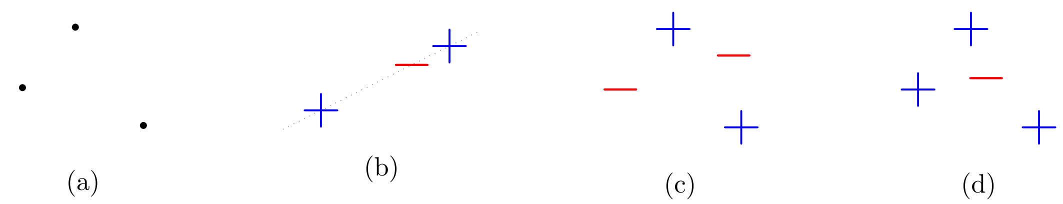

Figure 3: (a) A set of 3 points shattered by linear threshold functions. (b) A dichotomy on a set of 3 points that cannot be realised by linear threshold functions. (c) & (d) No set of 4 points can be shattered by linear threshold functions.

case when all four points lie on a line can be treated easily (*e.g.,* as in the case of Fig. 3(b)).

Exercise: Show that the VC dimension of linear threshold functions in n dimensions is $n+1$.

2 Growth Function

Let C be a concept class, we define the *growth function*, as follows:

Definition 3 (Growth Function)**.** *For any natural number m, define,*

$$\Pi_C(m) = \max\{|\Pi_C(S)| \mid |S| = m\}$$

The function $\Pi_C(m)$ indicates the growth of the number of realised dichotomies as the size of the finite set increases. Clearly, for $m \leq d$, where $d = \mathsf{VCD}(C)$, we have $\Pi_C(m) = 2^m$. The question of interest is what happens when $m \geq d$? We show the remarkable result that in fact $\Pi_C(m)$ is bounded by a degree d polynomial in m.

Definition 4. *Define the function $\Phi_d(m)$ as follows: $\Phi_d(0) = 1$ for all $d \in \mathbb{N}$, $\Phi_0(m) = 1$ for all $m \in \mathbb{N}$, and for $d > 0, m > 0$,*

$$\Phi_d(m) = \Phi_d(m-1) + \Phi_{d-1}(m-1)$$

Lemma 1. *Let C be a concept class with $\mathsf{VCD}(C) = d$, then $\Pi_C(m) \leq \Phi_d(m)$.*

Proof. We will prove this by induction simultaneously on d and m. We first check the base cases. If $d = 0$, then no finite set can be shattered, so C only contains one concept. Thus, for all m, $\Pi_C(m) = 1$. If $m = 0$, then clearly $\Pi_C(0) \leq \Phi_d(0)$. Now, suppose that the result holds for all $d' \leq d$ and $m' \leq m$, when at least one of the inequalities is strict.

Let S be any set of size m. Let x be a distinguished point of S. Then,

$$|\Pi_C(S \setminus \{x\})| \leq \Pi_C(m-1) \leq \Phi_d(m-1) \tag{3}$$

Let us look at the difference between $\Pi_C(S)$ and $\Pi_C(S \setminus \{x\})$. Consider the set,

$$C' = \{c \in \Pi_C(S) \mid c(x) = 0 \wedge \exists \widetilde{c} \in \Pi_C(S), \widetilde{c}(x) = 1, \forall z \in S \setminus \{x\}, c(z) = \widetilde{c}(z)\}$$

In words, we look at a dichotomy in $\Pi_C(S \setminus \{x\})$ and see whether this can be extended in two distinct ways in $\Pi_C(S)$, *i.e.,* whether we can keep the assignment on points in $S \setminus \{x\}$ and then still have a choice to label x as either 1 or 0. Then, we have

$$|\Pi_C(S)| = |\Pi_C(S \setminus \{x\})| + |\Pi_{C'}(S \setminus \{x\})| \tag{4}$$

The first term accounts for all the dichotomies on $S \setminus \{x\}$, and the second one accounts for the dichotomies on $S \setminus \{x\}$ that can be extended to two distinct dichotomies on S.

If we show that $\mathsf{VCD}(C') \leq d-1$, we would be done by induction. Let $S' \subseteq S \setminus \{x\}$ be shattered by C'. (Note that x cannot be included in any set shattered by C' since $c'(x) = 0$ for all $c' \in C'$.) Then, by definition $S' \cup \{x\}$ is shattered by C, so it must be the case that $|S'| \leq d-1$. This completes the proof. □

Let us now give a bound on $\Phi_d(m)$ and show that it is bounded by a polynomial of degree d in m.

Lemma 2. *For all m, d,*

$$\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$$

Proof. By induction on d and m. The base cases can easily checked to be true. Suppose this holds for all $d' \leq d$ and $m' \leq m$ as long as at least one inequality is strict.

Then, by definition,

$$\begin{aligned}
\Phi_d(m) &= \Phi_d(m-1) + \Phi_{d-1}(m-1) \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\
&= \binom{m-1}{0} + \sum_{i=1}^{d} \left(\binom{m-1}{i} + \binom{m-1}{i-1} \right) \\
&= \binom{m}{0} + \sum_{i=1}^{d} \binom{m}{i}
\end{aligned}$$

□

Finally, we can use elementary inequalities to give a bound on $\Phi_d(m)$.

Lemma 3. *For $m \geq d$,*

$$\Phi_d(m) \leq \left(\frac{em}{d}\right)^d$$

Proof.

$$\begin{aligned}
\Phi_d(m) &= \sum_{i=0}^{d} \binom{m}{i} \\
&= \left(\frac{m}{d}\right)^d \cdot \left(\sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^d \right) \\
&\leq \left(\frac{m}{d}\right)^d \cdot \left(\sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^i \right) && \text{As } \frac{d}{m} \leq 1 \\
&\leq \left(\frac{m}{d}\right)^d \cdot \left(\sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i \right) \\
&= \left(\frac{m}{d}\right)^d \cdot \left(1 + \frac{d}{m}\right)^m \leq \left(\frac{m}{d}\right)^d \cdot e^d
\end{aligned}$$

□

3 Sample Complexity Upper Bound

In this section, we'll prove that the VC dimension plays a role analogous to $\log |H_n|$ in the case of finite hypothesis classes. Provided the learning algorithm outputs a consistent learner from come class H which has bounded VC dimension, say d, and the sample size is modestly large as a function of the d, $1/\epsilon$ and $1/\delta$, then this yields a PAC-learning algorithm.

Theorem 5. *Let C be a concept class with VC-dimension d. Let L be a consistent learner for C that outputs a hypothesis $h \in C$. Then L is a PAC-learning algorithm for C provided it is given as input a random sample of size m drawn from $\mathsf{EX}(c, D)$ and the following bound on m holds*

$$m \geq \kappa_0 \left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{d}{\epsilon} \log \frac{1}{\epsilon} \right)$$

for some universal constant κ_0.

Proof. Let c and c' be two concepts in C. Denote by $c' \oplus c$ the concept defined as follows:

$$(c' \oplus c)(x) = \begin{cases} 1 & \text{if } c'(x) \neq c(x) \\ 0 & \text{if } c'(x) = c(x) \end{cases}$$

Note that if D is the target distribution and c is the target concept, then $\mathrm{err}(c') = \mathbb{P}_{x \sim D}\left[(c \oplus c')(x) = 1\right]$.

Let $\Delta_c(C) = \{c' \oplus c \mid c' \in C\}$. The first thing to show is that $\mathsf{VCD}(\Delta_c(C)) = \mathsf{VCD}(C)$ (left as an exercise to the reader). Furthermore, we define the class,

$$\Delta_{c,\epsilon}(C) = \{\widetilde{c} \in \Delta_c(C) \mid \mathbb{P}_{x \sim D}\left[\widetilde{c}(x) = 1\right] \geq \epsilon\}$$

Thus, any $c' \in C$ such that $c' \oplus c \in \Delta_{c,\epsilon}(C)$ is potentially problematic as $\mathrm{err}(c') \geq \epsilon$.

We say that a set S is an ϵ-net for $\Delta_c(C)$ if for every $\widetilde{c} \in \Delta_{c,\epsilon}(C)$, there exists $x \in S$, such that $\widetilde{c}(x) = 1$. Thus, our main goal is to bound the probability that a set S of size m drawn from $\mathsf{EX}(c, D)$ fails to be an ϵ-net for $\Delta_c(C)$. If S is an ϵ-net, then any $c' \in C$ that is consistent with S is not in $\Delta_{c,\epsilon}(C)$ and hence satisfies, $\mathrm{err}(c') \leq \epsilon$.

We will draw a sample S in two phases. First draw a sample S_1 of size m from $\mathsf{EX}(c, D)$. Let A be the event that S_1 is not an ϵ-net for $\Delta_c(C)$. Now, suppose the event A occurs, then there exists $\widetilde{c} \in \Delta_{c,\epsilon}(C)$ such that $\widetilde{c}(x) = 0$ for all $x \in S_1$. Fix such a $\widetilde{c}$ and draw a second sample S_2 of size m. Now, let us obtain a *lower* bound on the number of elements x in S_2 that satisfy $\widetilde{c}(x) = 1$. Let X_i denote the random variable that takes value 1 if the i^{th} element of S_2 satisfies $\widetilde{c}(x) = 1$ and X_i takes value 0 otherwise. Thus, if $X = \sum_{i=1}^{m} X_i$, then X is the (random variable) number of such points in S_2. Note that, $\mathbb{E}[X] \geq \epsilon m$, so by using a Chernoff bound (see Appendix B),

$$\mathbb{P}\left[X < \epsilon m/2\right] \leq \mathbb{P}\left[|X - \mathbb{E}[X]| > \frac{\mathbb{E}[X]}{2}\right] \leq 2\exp\left(-\frac{\epsilon m}{12}\right)$$

Provided $\epsilon m \geq 24$ (which our final bound will ensure), the probability that at least $\epsilon m/2$ points in S_2 satisfy $\widetilde{c} = 1$ is at least $1/2$.

Now consider the event B defined as follows: A sample $S = S_1 \cup S_2$ of size $2m$ with $|S_1| = |S_2| = m$ is drawn from $\mathsf{EX}(c, D)$, there exists a $\widetilde{c} \in \Pi_{\Delta_{c,\epsilon}(C)}(S)$, such that $|\{x \in S \mid \widetilde{c}(x) = 1\}| \geq \epsilon m/2$ and $\widetilde{c}(x) = 0$ for all $x \in S_1$. Note that $\mathbb{P}[B] \geq \frac{1}{2}\mathbb{P}[A]$, since if S_1 fails to be an ϵ-net for $\Delta_c(C)$, then the probability of there being a $\widetilde{c} \in \Delta_{c,\epsilon}(C)$ such that $|\{x \in S_2 \mid \widetilde{c}(x) = 1\}| \geq \epsilon m/2$ is at least $1/2$. Thus, $\mathbb{P}[A] \leq 2\mathbb{P}[B]$.

We will bound $\mathbb{P}[B]$, which is a purely combinatorial problem. The question is the following: We are given $2m$ balls out of which $r \geq \epsilon m/2$ are red and the remaining are black. If we divided them into two sets of size m each, without seeing the colours, what is the probability that the first set has no red balls and the second set has all of them? This probability is simply given by $\binom{m}{r}/\binom{2m}{r}$. We can bound this as follows:

$$\frac{\binom{m}{r}}{\binom{2m}{r}} = \prod_{i=0}^{r-1} \frac{m-i}{2m-i} \le \frac{1}{2^r}$$

Thus, we have

$$\mathbb{P}[A] \le 2 \cdot \mathbb{P}[B] \le 2 \cdot \left|\Pi_{\Delta_{c,\epsilon}(C)}(S)\right| 2^{-\frac{\epsilon m}{2}} \qquad \text{By the union bound over all } \widetilde{c} \in \Pi_{\Delta_{c,\epsilon}(C)}(S).$$

$$\le 2 \cdot \left|\Pi_{\Delta_{c}(C)}(S)\right| 2^{-\frac{\epsilon m}{2}} \le 2 \cdot \left(\frac{2em}{d}\right)^d 2^{-\frac{\epsilon m}{2}}$$

Thus, there exists a universal constant, κ_0, such that provided m is larger than the bound given in the statement of the theorem, $\mathbb{P}[A] \le \delta$. This completes the proof. □

In our final definition of PAC learning, we allowed the output hypothesis to be from a different (and typically larger) class. Theorem 5 still applies, but d needs to be the VC-dimension of the hypothesis class H. Provided $C \subseteq H$, it holds that $\mathsf{VCD}(C) \le \mathsf{VCD}(H)$. This shows that using a hypothesis class that is larger comes at a cost of increased sample complexity; however, as we've seen it may allow us to solve problems that may otherwise be computationally intractable.

4 Sample Complexity Lower Bounds

The notion of VC dimension also allows us to show sample complexity lower bounds. These lower bounds apply no matter what algorithm we use and hold even for algorithms using unbounded computation. These lower bounds are purely information-theoretic in nature.

Theorem 6. *Let C be a concept class with $\mathsf{VCD}(C) = d$, with $d \ge 25$.*[1] *Then any PAC-learning algorithm for C requires at least $\frac{d-1}{32\epsilon}$ examples.*

Proof. (Note: We will be a little hand-wavy and only convey the essential ideas of the proof. For a fully formal proof, please refer to the book by Mohri et al. (2012).)

Let S be a set of size d that is shattered by C. Suppose $S = \{x_1, x_2, \ldots, x_d\}$. Let D be a distribution defined as follows: $D(x_1) = 1 - 8\epsilon$, and $D(x_j) = 8\epsilon/(d-1)$ for $j = 2, \ldots, d$. Suppose the learning algorithm receives $m = (d-1)/(32\epsilon)$ examples drawn according to D. We claim that in fact it receives very few examples from the set $\{x_2, \ldots, x_d\}$. Let Z_i be the random variable that is 1 if the i^{th} example drawn from D is in the set $S \setminus \{x_1\}$ and 0 otherwise. Then $Z_i = 1$ with probability 8ϵ and 0 with probability $1 - 8\epsilon$. Let $Z = \sum_{i=1}^m Z_i$ be the number of examples seen from the set $S \setminus \{x_1\}$ (possibly with repetitions). Then $\mathbb{E}[Z] = \frac{d-1}{4}$ and using a Chernoff bound,

$$\mathbb{P}\left[Z \ge \frac{d-1}{2}\right] \le \mathbb{P}\left[|Z - \mathbb{E}[Z]| \ge \mathbb{E}[Z]\right] \le 2\exp\left(-\frac{d-1}{12}\right)$$

We can simulate the example oracle by drawing examples from D and assigning a random label (out of 0 or 1) to any newly seen example. For examples, that were seen previously, we retain the label initially given to them. Since S is shattered by C, this labelling is consistent with some $c \in C$. Thus, any hypothesis output by the algorithm errs with probability at least 1/2 on any example that it has not seen. Thus, with probability at least $2\exp\left(-\frac{d-1}{12}\right) \ge 1/2$ (provided $d \ge 25$), the error of any hypothesis output by the algorithm is at least 2ϵ, as it has not seen at least half the examples from the set $\{x_2, \ldots, x_d\}$ which has a total probability mass of 8ϵ (equally distributed). □

[1] The condition $d \ge 25$ is not really necessary, with a slightly improved argument this can be shown for any $d \ge 2$.

References

Michel Goemans. Chernoff bounds, and some applications. Available online at `http://math.mit.edu/~goemans/18310S15/chernoff-notes.pdf`, 2015. Lecture Notes.

Michael J. Kearns and Umesh K. Vazirani. *An Introduction to Computational Learning Theory.* The MIT Press, 1994.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning.* MIT Press, 2012.

A Consistent Learner for Linear Threshold Functions

Let us define the class of linear threshold functions over $\mathbb{R}^n$,

$$\mathsf{LTF}_n = \{x \mapsto \mathbb{1}_{\geq 0}(w \cdot x + w_0) \mid w \in \mathbb{R}^n, \|w\|_2 = 1, w_0 \in \mathbb{R}\}, \tag{5}$$

where $\mathbb{1}_{\geq 0}(z) = 1$ if $z \geq 0$ and 0 otherwise.

We would like to design a consistent learner for the class of linear threshold functions. The problem is the following: Given $(x_1, y_1), \ldots, (x_m, y_m)$, where $x_i \in \mathbb{R}^n$ and $y_i \in \{0, 1\}$, such that there exists $w^* \in \mathbb{R}^n$, $\|w^*\|_2 = 1$ and $w_0^* \in \mathbb{R}$ such that $y_i = \mathbb{1}_{\geq 0}(w^* \cdot x_i + w_0^*)$, find some w, w_0, such that $y_i = \mathbb{1}_{\geq 0}(w \cdot x_i + w_0)$.

We consider the following linear program with variables, $w_0, w_1, \ldots, w_n$. The objective function is constant, so we are in fact only looking for a feasible point. The constraints are given by:

$$w_0 + w_1 x_{i1} + w_2 x_{i2} + \cdots w_n x_{in} \geq 0 \qquad \text{For all } i \text{ such that } y_i = 1 \tag{6}$$

$$-w_0 - w_1 x_{i1} - w_2 x_{i2} - \cdots w_n x_{in} \geq 1 \qquad \text{For all } i \text{ such that } y_i = 0 \tag{7}$$

Let us first discuss the second inequality (7). An example is classified as negative if $w^* \cdot x + w_0^* < 0$; however, strict inequalities cannot be given as part of the linear program. The choice of 1 is arbitrary, we could have used any strictly positive real number. Let us show that the above linear program has a feasible solution; given that a feasible solution exists, there are known polynomial time algorithms to find one.

Let the target linear threshold function be defined by w^*, w_0^*. Let $\alpha = \min\{-(w^* \cdot x_i + w_0^*) \mid y_i = 0\}$; we know that $\alpha > 0$. Consider $\frac{w^*}{\alpha} \in \mathbb{R}^n$, $\frac{w_0}{\alpha} \in \mathbb{R}$; it can be checked that this is a feasible solution to the constraints defined by (6) and (7).

B Chernoff Bounds

We'll use the following version of the Chernoff bound frequently. For other versions and a proof, please refer to the notes by Goemans (2015).

Theorem 7. *Let $X_1, \ldots, X_n$ be independent random variables, where $X_i = 1$ with probability p_i and $X_i = 0$ with probability $1 - p_i$. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[X] = \sum_{i=1}^n p_i$. Then for $\alpha \in [0, 1]$,*

$$\mathbb{P}\left[|X - \mu| > \alpha\mu\right] \leq 2 \exp\left(-\frac{\mu\alpha^2}{3}\right)$$