Big Data
Solutions Infinite Hypothesis Classes

Exercise 1: Threshold functions

(a) You can set the threshold between any pair of consecutive points, before the smallest point, or after the largest point. That makes $m+1$ different dichotomies in total. So $\tau(m) = m + 1$.

(b) $\tau(1) = 2, \tau(2) = 3, \tau(3) = 4$.

(c) If $\mathcal{H}$ shatters a set of m points, then it can realize all 2^m possible dichotomies on those m points. The VC-dimension of $\mathcal{H}$ is the size of the largest set that $\mathcal{H}$ can shatter. Looking at the growth function, we see that $\tau(1) = 2 = 2^1$. So the VC-dimension is at least 1. Continuing, we see that $\tau(2) = 3 < 2^2 = 4$. So for $m = 2$, the number of dichotomies that $\mathcal{H}$ can realize falls short of 2^m. We conclude that the VC-dimension of $\mathcal{H}$ is 1.

Note that it is not a particularly good idea to determine the VC-dimension via the growth function, because determining the growth function will typically be harder. But the growth function is of interest in its own right, and we wanted to point out the connection with the VC-dimension.

Exercise 2: Intervals

(a) You can choose the two endpoints of the interval in $\binom{m+1}{2}$ ways. This gives

$$\binom{m+1}{2} = \frac{m(m+1)}{2} = \frac{1}{2}m^2 + \frac{1}{2}m$$

In addition, we can choose the two endpoints of the interval in the same line segment between two points. This produces the "all-negative" labeling. So in total we can realize

$$\tau(m) = \frac{1}{2}m^2 + \frac{1}{2}m + 1$$

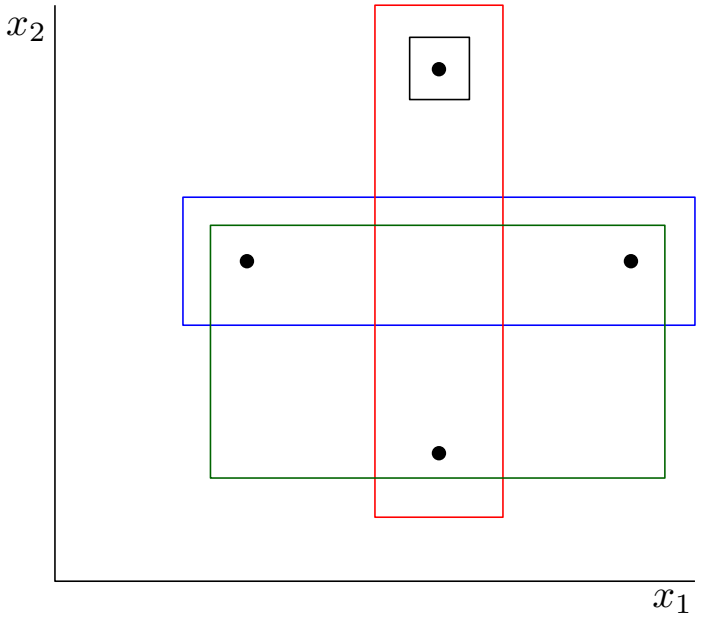
different dichotomies.

(b) $\tau(1) = 2, \tau(2) = 4, \tau(3) = 7$.

(c) For $m = 2$, $\tau(m)$ is equal to 2^m. For $m = 3$, $\tau(m)$ falls short of 2^m. The VC-dimension is 2.

Exercise 3: Axis-aligned rectangles

(a) Arrange the points as the corners of a diamond shape:



I drew in a few dichotomies in different colors. You should be able to complete the picture. Note that you cannot shatter a set of 4 points that are the corners of an axis-aligned rectangle, because then you won't be able to separate the two points on one diagonal from the two points on the other diagonal.

(b) Consider any set of 5 points. Let x_{left} denote the leftmost point, that is the point with the smallest value for x_1. Likewise, let x_{right}, x_{low} and x_{high} denote the rightmost, lowest and highest points respectively. Every rectangle that includes these points includes all 5 points. Hence, the dichotomy that assigns a positive label to x_{left}, x_{right}, x_{low} and x_{high}, and assigns a negative label to the remaining point(s) cannot be realized.

Exercise 4: Bounding the growth function

We first verify the base cases $k = 0$ and $k = 1$:

$$\sum_{i=0}^{0} \binom{m}{i} = \binom{m}{0} = 1 \leq m^0 + 1$$
$$\sum_{i=0}^{1} \binom{m}{i} = \binom{m}{0} + \binom{m}{1} = 1 + m \leq m^1 + 1.$$

So far, so good. Next we show that if the claim is true for k, then it is also true for $k + 1$. Assume (induction hypothesis):

$$\sum_{i=0}^{k} \binom{m}{i} \leq m^k + 1.$$

To prove:

$$\sum_{i=0}^{k+1} \binom{m}{i} \leq m^{k+1} + 1.$$

Proof:

$$\begin{aligned}
\sum_{i=0}^{k+1} \binom{m}{i} &= \sum_{i=0}^{k} \binom{m}{i} + \binom{m}{k+1} && \text{(use induction hypothesis)} \\
&\leq m^k + 1 + \frac{m!}{(m-k-1)!(k+1)!} \\
&= m^k + 1 + \frac{m(m-1)\cdots(m-k)}{(k+1)!} \\
& && ((k+1)! > 1\text{; lift out the } (m-1) \text{ term}) \\
&\leq m^k + 1 + (m-1)m^k. \\
&= m^{k+1} + 1.
\end{aligned}$$

Exercise 5: Application of the bound on the growth function

(a) At exercise 1 we found that for threshold functions, $\tau(m) = m + 1$ and the VC-dimension was 1. Filling in $d = 1$ in the bound given by Sauer's Lemma, we obtain

$$\tau(m) \leq \binom{m}{0} + \binom{m}{1} = 1 + m.$$

So in this case the bound gives the exact growth function.

(b) At exercise 2, we found that $\tau(m) = \frac{1}{2}m^2 + \frac{1}{2}m + 1$, and the VC dimension was $d = 2$. Filling in $d = 2$ in the bound given by Sauer's Lemma, we obtain

$$\tau(m) \leq \binom{m}{0} + \binom{m}{1} + \binom{m}{2} = 1 + m + \frac{m(m-1)}{2} = \frac{1}{2}m^2 + \frac{1}{2}m + 1.$$

So in this case the bound gives the exact growth function. Note that this is not true in general.

(c) For $m = 10$:

$$\binom{10}{0} + \binom{10}{1} + \binom{10}{2} + \binom{10}{3} + \binom{10}{4} = 386.$$

For $m = 20$:

$$\binom{20}{0} + \binom{20}{1} + \binom{20}{2} + \binom{20}{3} + \binom{20}{4} = 6,196.$$

The total number is $2^{10} = 1,024$ respectively $2^{20} = 1,048,576$.

Exercise 6: Linear classifiers with a single predictor variable

(a) A "brute force" approach informs us that it is best to put the threshold between 19 and 21. The corresponding classifier makes 5 errors on the training set: the three positive examples to the left of the threshold, and the two negative examples to the right. The training error therefore is:

$$L_D = \frac{5}{25} = 0.2$$

Any threshold in the interval $(19, 21]$ produces the same labeling of the training set, so the ERM solution is not unique.

(b) Choose for example $w_0 = -20$ and $w_1 = 1$. Then we predict class $+1$ if

$$\begin{aligned} -20 + x &\geq 0 \\ x &\geq 20 \end{aligned}$$

This gives us a threshold in the interval $(19, 21]$, producing 5 errors and a training error of 20%. Again, the ERM solution is not unique. We can dream up infinitely many weight combinations that result in a threshold in the interval $(19, 21]$.

(c) The linear classifier predicts class +1 if

$$\begin{aligned} w_0 + w_1 x &\geq 0 \\ w_1 x &\geq -w_0 \\ x &\geq -\frac{w_0}{w_1} && \text{(if } w_1 \text{ is positive)} \\ x &\leq -\frac{w_0}{w_1} && \text{(if } w_1 \text{ is negative)} \end{aligned}$$

So the linear classifier is more powerful because we can also predict the positive class to the left of the threshold $-w_0/w_1$, namely if $w_1 < 0$.

(d) To determine the growth function, we can reason as follows. Pick any m *distinct* points, and consider them in their sorted order. We can choose the weights to produce a threshold between any pair of consecutive points, which gives $m - 1$ possibilities. For each threshold we can predict $+$ to the right of the threshold and $-$ to the left, or vice versa; this gives a total of $2(m-1)$ possible labelings. In addition, we can also produce the "all positive" and "all negative" labeling by choosing the weights to give a threshold to the left of the smallest point, or to the right of the largest point. Hence, the growth function is $\tau(m) = 2m$. We evaluate the growth function until it falls short of 2^m:

$$\begin{aligned} \tau(1) &= 2 = 2^1 \\ \tau(2) &= 4 = 2^2 \\ \tau(3) &= 6 < 2^3 \end{aligned}$$

We conclude that the size of the largest set that is shattered is 2, that is, the VC-dimension is 2.

Exercise 7: The VC-dimension of linear classifiers

We discuss the general case, where we have to shatter $d + 1$ points in $\mathbb{R}^d$ (that is, there are d predictor variables). Choose the points as follows (recall we add $x_0 \equiv 1$ for w_0):

	x_0	x_1	x_2	$\dots$	x_d
$\mathbf{x}_0$	1	0	0	$\dots$	0
$\mathbf{x}_1$	1	1	0	$\dots$	0
$\mathbf{x}_2$	1	0	1	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathbf{x}_d$	1	0	0	$\dots$	1

The first column consists entirely of 1's, and the main diagonal consists entirely of 1's as well. All remaining entries are 0. Using results from linear algebra we can reason as follows.

Let's call the given data matrix $\mathbf{X}$. To produce any given labeling $\mathbf{y} = (y_0, y_1, \ldots, y_d)$, we must choose weight values such that

$$\mathbf{X}\mathbf{w} = \mathbf{y}$$

Since $\mathbf{X}$ is invertible, the unique solution is:

$$\mathbf{w} = \mathbf{X}^{-1}\mathbf{y}.$$

We could stop here, but this argument is not convincing unless you are familiar with the results from linear algebra that we used. So let's reason from first principles. To produce any given labeling, we must choose $\mathbf{w}$ such that:

$$y_i = \mathbf{w} \cdot \mathbf{x}_i, \qquad i = 0, \ldots, d.$$

Of the weight vector $\mathbf{w}$, only w_0 and w_i are relevant for producing y_i, because the remaining entries of $\mathbf{x}_i$ are all 0. We see immediately that if $y_0 = +1$, we must choose $w_0 = +1$ as well. We continue on the premiss that $y_0 = +1$. For $i > 0$, if $y_i = +1$, choose $w_i = 0$, and if $y_i = -1$, choose $w_i = -2$. The case for $y_0 = -1$ is completely analogous, and is left as an exercise.

We have shown that we can shatter a set of $d + 1$ points. To finish the proof that the VC-dimension is $d + 1$, we must show that no set of size $d + 2$ is shattered.

As given in the hint, we can write one of the rows as a linear combination of the other rows:

$$\mathbf{x}_j = \sum_{i \neq j} c_i \mathbf{x}_i, \tag{1}$$

where not all c_i are zero because the first component of every vector has the value 1.
Now assign the label -1 to row j, that is, $y_j = -1$. To produce this label, we must have:

$$\mathbf{w} \cdot \mathbf{x}_j < 0.$$

Furthermore, for $c_i \neq 0$, set $y_i = \text{sign}(c_i)$. This requires that $\text{sign}(\mathbf{w} \cdot \mathbf{x}_i) = \text{sign}(c_i)$.
From equation(1) it follows that:

$$\mathbf{w} \cdot \mathbf{x}_j = \mathbf{w} \cdot \sum_{i \neq j} c_i \mathbf{x}_i = \sum_{i \neq j} c_i (\mathbf{w} \cdot \mathbf{x}_i). \tag{2}$$

For all i, $c_i(\mathbf{w} \cdot \mathbf{x}_i) \geq 0$ since $\text{sign}(\mathbf{w} \cdot \mathbf{x}_i) = \text{sign}(c_i)$. This implies that $\mathbf{w} \cdot \mathbf{x}_j \geq 0$. This leads to a contradiction with the requirement that $\mathbf{w} \cdot \mathbf{x}_j < 0$. Hence, it is not possible to produce the requested labeling.