

Exercise 6.1

Let $X_n = \{-1, 1\}^n, H_n : X_n \rightarrow Y, Y = \{-1, 1\}$ and $VCDim(H_n) = d$.

”if”. $VCDim(H_n) \subseteq Poly(n) \Rightarrow \log |H_n| \subseteq Poly(n)$?

The set D of all boolean functions in $\{-1, 1\}$ is finite. Restrict the function class H to this set, choose $m = 2^n$ (want to shatter all possible boolean functions) and use Sauer’s lemma.

$$S_H(m) = \max_{|D|=m} S_H(D) = |H_D| \leq \left(\frac{em}{d}\right)^d = \left(\frac{e2^n}{d}\right)^d \quad (67)$$

where $S_H()$ is the shattering coefficient, that is, the maximum number of ways H is able to split some set of size m (or a given set D). Applying log on both sides shows

$$\log |H_n| \leq d \log \frac{e2^n}{d}, \quad (68)$$

which is polynomial in n .

”only”. Shattering d examples clearly requires atleast 2^d functions. Hence

$$|H_n| \geq 2^d \Leftrightarrow \log |H_n| \geq d \log 2 \approx d. \quad (69)$$

Thus, if d is not polynomial, neither is $\log |H_n|$.

□

Exercise 6.2

a) Lets handle $k = 1$. Suppose $VCDim(H_2) = d+2$. Drop arbitrary hypothesis from H_2 , i.e.

$$H_1 = H_2 \setminus \{h\} \Leftrightarrow H_2 = H_1 \cup h. \quad (70)$$

Doing this, you lose the ability to classify 1 example. This is because H_2 is able to make all $2^{VCDim(H_2)}$ dichotomies for a set of size $VCDim(H_2)$, there must be one or more hypotheses in H_2 that differ from the removed hypothesis h only by the label of 1 example. Thus H_1 is able to make atleast $2^{VCDim(H_2)-1}$ dichotomies for a set of size $VCDim(H_2) - 1$ and it follows that $VCDim(H_1) \geq d + 1$.

General k inductively. Result: the claim is true.

b) Take two hypothesis classes, $H_1 = \{1\}$ and $H_2 = \{-1\}$. That is, both classes contain just one constant function. As neither function class can shatter any set of one point (for example $X = \{1\}$) arbitrarily, $VCDim(H_1) = VCDim(H_2) = 0$. However, let $H = H_1 \cup H_2$. Now the combined hypothesis class can shatter a set of single point. Thus,

$$1 = VCDim(H) > VCDim(H_1) + VCDim(H_2) = 0. \quad (71)$$

Result: the claim is false.

□

Exercise 6.3

With our hypothesis class, we can set upper and lower limit separately for each coordinate axis. So, in each dimension, we can label a set of 2 well-chosen points arbitrarily, by selecting either of the points, both points, or no points at all. Hence, with n dimensions, we can label at least $2n$ points arbitrarily if they're well chosen. Then $VCDim(H) \geq 2n$.

The following dataset illustrates a suitable configuration.

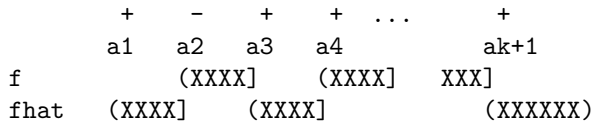
```
-1 0 0 ... 0
 1 0 0 ... 0
 0 1 0 ... 0
 0 -1 0 ... 0
    ...
 0 0 0 ... 1
 0 0 0 ... -1
```

However, trying to add a $2n+1$ th point will necessarily result in a set containing at least three examples with each having a nonzero value in a single dimension. Our hypothesis class is clearly not expressive enough to label three such points arbitrarily. Geometrically, the example added after $2n$ will result in some point being necessarily inside the box and thus can't be labeled in both ways.

□

Exercise 6.4

Graphically, the two interval classifiers can be illustrated roughly as follows.



Given a set of labeled examples on the real axis, the task of learning an interval classifier is to find k split locations a_1, \dots, a_k to classify the data "as well as possible". The other way to think about is that you have a set of intervals and you wish to choose their endpoints so, that each interval labels the examples ending up inside it as positive while minimizing the total classification error. The number of intervals given designate the maximum complexity we allow to the classifier (clearly selecting number of splits on the same order as the number of examples allow us to overfit randomly labeled data quite well, unless some labels of otherwise identical points differ).

The interval classifier can also be seen as a special case of the one-dimensional segmentation problem (assume signal that comes from source 1 or source -1. Segment the signal to segments denoting either origin 1 or -1).

a) Note that having k splits means that we can make all dichotomies where the sign is changed at most k times (consider reading a label stream $+ + + + \dots$ from left to right). Now k examples can change sign $k - 1$ times. Hence we can split $k + 1$ examples with k splits. $k + 2$ examples can no longer be splitted, we are out of splits. It follows that $VCDim(H) = k + 1$.

b) Dynamic programming can be used. First proceed by sorting the one-dimensional data of n points to get $(x'_1, x'_2, \dots, x'_n)$. Now create a dynamic programming matrix M

	x1'	x2'	...	xn'
1				
2	0			
.				
k	0	0		

The idea is to fill the matrix from top-down, left-right so that we consider putting the first split to data point x'_1 and so on. While progressing, we sum up the errors made by the previous choices and add the error caused by the current choice. At $M(i, j)$ we use the information from $M(i - 1, j)$ and $M(i - 1, j - 1)$ and add the local increase. We might have to keep a separate matrix for f_a and $f_{\hat{a}}$. The final hypothesis is found by backtracking the path that resulted in the smallest total error (the empirical risk minimizer). The time-complexity of this solution is $O(kn)$ for filling the matrix plus $O(n \log n)$ for sorting the data.

□

Exercise 6.5

The lecture notes presented how empirical risk minimization can be used to estimate the Rademacher complexity. This solution is a slight modification of the technique shown in the slides following theorem 3.23.

Theorem 3.23 the second part states that given fixed $S \in Z^m$, we have

$$R_m(F) \leq \sup_{f \in F} \left| \frac{2}{m} \sum_{i=1}^m r_i f(z_i) \right| + \sqrt{\frac{8}{m} \ln \frac{1}{\delta}} \quad (72)$$

with prob. atleast $1 - \delta$ over random choice of r and S . Now we are interested in estimating the first term on the right.

Let $F = L_{0-1}(H)$ be the discrete loss class for some H .

Assume $h \in H \Rightarrow -h \notin H$ ($-h = -h$).

Let $S = ((x_1, y_1), \dots, (x_m, y_m)) \in (X \times Y)^m, r \in \{-1, 1\}^m$ fixed,

$S' = ((x_1, r_1 y_1), \dots, (x_m, r_m y_m))$. $L_{0-1}(h, x, y) = \frac{1}{2}(1 - yh(x))$. Now

$$2 \sum r_i f(z_i) = 2 \sum r_i L(h, x_i, y_i) \quad (73)$$

$$= 2 \sum r_i \frac{1}{2} (1 - y_i h(x_i)) \quad (74)$$

$$= \sum r_i - \sum y_i r_i h(x_i) \quad (75)$$

$$= \sum r_i - \sum (1 - 2L(h, x_i, r_i y_i)) \quad (76)$$

$$= \sum r_i - m + 2 \sum L(h, x_i, r_i y_i) \quad (77)$$

$$= \sum r_i - m + 2m \widehat{err}(h, S'). \quad (78)$$

similarly $-2 \sum r_i f(z_i) = -\sum(r_i) + m - 2m \widehat{err}(h, S')$. The assumption that the hypothesis class is not closed under complementation means must look at the "flipped hypotheses" separately, unlike what could be done in the lecture notes case.

Notice that $\sup |x| = \max(\sup(x), \sup(-x))$. We can write

$$\sup \left| \frac{2}{m} \sum r_i f(z_i) \right| = \max(\sup(\frac{2}{m} \sum r_i f(z_i)), \sup(-\frac{2}{m} \sum r_i f(z_i))), \quad (79)$$

which is

$$\max(\sup \frac{1}{m} (\sum r_i - m + 2m \widehat{err}_1(h, S')), \sup \frac{1}{m} (-\sum r_i + m - 2m \widehat{err}_2(h, S'))).$$

Hence, maximize \widehat{err}_1 and minimize \widehat{err}_2 . Calculate $\sum r_i$ and plug in with the error estimates. Choose the max of the two.

□