Learning From Data Problems: Chapter II

J. David Giese

Exercise 2.1

The breaking point for (1) is $N = 2$ because $(1, -1) \notin \mathcal{H}(\mathbf{x}_1, \mathbf{x}_2)$.

The breaking point for (2) is $N = 3$ because $(1, -1, 1) \notin \mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$.

There is no breaking point for (3) because every dichotomy can be generated by $\mathcal{H}$.

Exercise 2.2

a) For (1), we have $m_{\mathcal{H}}(N) \leq \binom{N}{1} + \binom{N}{0} = N + 1$, which is true.

For (2), we have $m_{\mathcal{H}}(N) \leq \binom{N}{2} + \binom{N}{1} + \binom{N}{0} = N^2/2 + N/2 + 1$, which is true.

There is no bound for (3) as there is no break point.

b) No, because if $m_{\mathcal{H}}(N) < 2^N$ then there must be a break point, however if there is a break point it will be polynomial bounded.

Exercise 2.3

The Vapnik-Chervonenkis dimension is 1, 2, and ∞ respectively.

Exercise 2.4

a) First we select $\mathcal{D}$ such that the samples, when placed in the rows of a matrix form:

$$G = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{1} & \mathbf{I}_d \end{bmatrix}$$

where $\mathbf{I}_d$ is the $d \times d$ identity matrix. We can see by inspection that G is invertible, however for fun we can use a matrix identity to prove it:

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(A)\det(D - CA^{-1}B) \quad \Longrightarrow$$
$$\det(G) = \det(1)\det(I_d - \mathbf{1} \times \mathbf{0}) = 1.$$

We can thus solve $\mathbf{b} = G\mathbf{w}$ given a weight vector $\mathbf{w}$, and can generate each dichotomy by selecting the sign of each dimension of $\mathbf{b}$.

b) Select our first $d+1$ points as in (a). Clearly the $d+2$ point will be a linear combination of the first points (because we have already spanned the space). This means we we no longer have enough free parameters to vary $\mathbf{b}$ and generate each dichotomy.

Exercise 2.5

$$\begin{aligned} \delta &= 4m_{\mathcal{H}}(2N)\exp(-N\epsilon^2/8) \\ &\leq 4((2N)^{d_{\text{vc}}} + 1)\exp(-N\epsilon^2/8) \\ &= 4((2 \cdot 100) + 1)\exp(-100(0.1)^2/8) \\ &\approx 709. \end{aligned}$$

Clearly, although δ is a probability, the bound we have set for it can be much greater than 1, and thus useless.

Exercise 2.6

a) The Hoeffding Inequality on the test data gives the bound $\epsilon = \sqrt{\frac{1}{2 \cdot 200} \ln \frac{2}{0.05}} = 0.096$. We can also use the Hoeffding Inequality for the trained bound, because the hypothesis set is finite; in other words, we don't need to resort to the VC bound. We have $\epsilon = \sqrt{\frac{1}{2 \cdot 400} \ln \frac{2 \cdot 1000}{0.05}} = 0.115$. The bound provided by the test data is clearly better.

b) This is a subtle question. Note we are not changing $\mathcal{H}$ so it is not a trade off between model complexity and in-sample error. Looking out the simple generalization bound with $M = 1$

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{1}{2N} \ln \frac{2}{\delta}}$$

we see that for a given single hypothesis g, using a larger test sample size further tightens the bound. One may then jump to the conclusion that your in-sample error will take a penalty, however this is not necessarily the case. In the extreme case (e.g.

when you have a single training sample) your E_{in} is likely to be perfect, because your hypothesis has enough free dimensions to fit the data. Clearly this is still not a good thing, although our mathematical analysis presented in this chapter insufficient to account for it.

This appears to exemplify how over-fitting is a separate issue that must also be accounted for. Clearly, by taking too many samples from our training-set, we will be unable to select the proper g, even if the in-sample error is low.

Exercise 2.7

a) The squared distance between 0 and 1 is 1, hence the pointwise mean-squared error is equivalent "binary error-measure" used in Chapter 1.

b) In this case, the squared distance between -1 and 1 is 4, hence you will need to normalize by 4 to make the pointwise measures equivalent.

Exercise 2.8

a) The expectation operator is linear, hence if $\mathcal{H}$ is closed under linear combinations, then $\bar{g} \in \mathcal{H}$.

b) If we let $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and $\mathcal{H} = 1, 0$ then, unless one of the hypothesis occurs with probability 0, $\bar{g}$ will not be in $\mathcal{H}$.

c) No.