Learning Theory

Machine Learning II (2023-2024)
UMONS

1 Exercise 1

Consider a sample of 10 marbles drawn independently from a bin that holds red and green marbles. The probability of a red marble is μ. For $\mu = 0.05$, $\mu = 0.5$, and $\mu = 0.8$, compute the probability of getting no red marbles ($\nu = 0$) in the following cases.

(a) We draw only one such sample. Compute the probability that $\nu = 0$.

(b) We draw 1,000 independent samples. Compute the probability that (at least) one of the samples has $\nu = 0$.

(c) Repeat (b) for 1,000,000 independent samples.

Solution

(a)

$$\mathbb{P}(\text{green}) = 1 - \mu$$

Now,

$$\begin{aligned}\mathbb{P}(\nu = 0) &= \mathbb{P}(\text{First green} \cap \text{Second green} \cap \cdots \cap \text{Tenth green}) \\ &= \mathbb{P}(\text{First green}) \times \mathbb{P}(\text{Second green}) \times \cdots \times \mathbb{P}(\text{Tenth green}) \\ &= (1-\mu)^{10}\end{aligned}$$

- For $\mu = 0.05$, $\mathbb{P}(\nu = 0) = 0.599$
- For $\mu = 0.5$, $\mathbb{P}(\nu = 0) = 0.00098$
- For $\mu = 0.8$, $\mathbb{P}(\nu = 0) = 1.024 \times 10^{-7}$

(b)

$$\begin{aligned}&\mathbb{P}(\text{AT LEAST one of the 1000 independent sample has } \nu = 0) \\ &= 1 - \mathbb{P}(\text{NO sample has } \nu = 0) \\ &= 1 - \mathbb{P}(\text{ALL sample has } \nu \neq 0) \\ &= 1 - [\mathbb{P}(\text{ONE sample has } \nu \neq 0)]^{1000} \\ &= 1 - [1 - \mathbb{P}(\text{ONE sample has } \nu = 0)]^{1000} \\ &= 1 - [1 - (1-\mu)^{10}]^{1000}\end{aligned}$$

- For $\mu = 0.05$, $\mathbb{P}(\text{AT LEAST one of the 1000 independent sample has } \nu = 0) \approx 1$
- For $\mu = 0.5$, $\mathbb{P}(\text{AT LEAST one of the 1000 independent sample has } \nu = 0) \approx 0.62$
- For $\mu = 0.8$, $\mathbb{P}(\text{AT LEAST one of the 1000 independent sample has } \nu = 0) \approx 0.0001$

(c)

$$\mathbb{P}(\text{AT LEAST one of the 1000000 independent sample has } \nu = 0) = 1 - [1 - (1-\mu)^{10}]^{1000000}$$

- For $\mu = 0.05$, $\mathbb{P}$(AT LEAST one of the 1000000 independent sample has $\nu = 0) \approx 1$
- For $\mu = 0.5$, $\mathbb{P}$(AT LEAST one of the 1000000 independent sample has $\nu = 0) \approx 1$
- For $\mu = 0.8$, $\mathbb{P}$(AT LEAST one of the 1000000 independent sample has $\nu = 0) \approx 0.097$

2 Exercise 2

Here is an experiment that illustrates the difference between a single bin and multiple bins. Run a computer simulation for flipping 1,000 fair coins. Flip each coin independently 10 times. Let's focus on 3 coins as follows: c_1 is the first coin flipped; c_{rand} is a coin you choose at random; c_{min} is the coin that had the minimum frequency of heads (pick the earlier one in case of a tie). Let ν_1, ν_{rand} and ν_{min} be the fraction of heads you obtain for the respective three coins.

(a) What is μ for the three coins selected?

(b) Repeat this entire experiment a large number of times (e.g., 100,000 runs of the entire experiment) to get several instances of ν_1, ν_{rand} and ν_{min} and plot the histograms of the distributions of ν_1, ν_{rand} and ν_{min}. Notice that which coins end up being c_{rand} and c_{min} may differ from one run to another.

(c) Using (b), plot estimates for $\mathbb{P}[|\nu - \mu| > \epsilon]$ as a function of ϵ, together with the Hoeffding bound $2e^{-2\epsilon^2 N}$ (on the same graph) .

(d) Which coins obey the Hoeffding bound, and which ones do not? Explain why.

(e) Relate part (d) to the multiple bins in Figure 1.

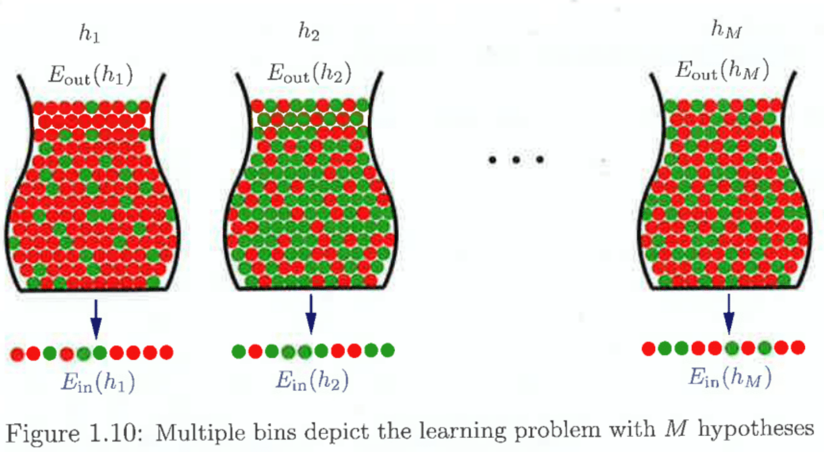

Figure 1: Source: Abu-Mostafa et al. Learning from data. AMLbook.

See this Python notebook

https://colab.research.google.com/drive/1CB3s2RpkfdU9y7tx7RLT5hS_bGNZoM_s?usp=sharing

3 Exercise 3

The Hoeffding Inequality is one form of the law of large numbers. One of the simplest forms of that law is the Chebyshev Inequality, which you will prove here.

(a) If t is a non-negative random variable, prove that for any $\alpha > 0$, $\mathbb{P}[t \geq \alpha] \leq \mathbb{E}(t)/\alpha$.

(b) If u is any random variable with mean μ and variance σ^2, prove that for any $\alpha > 0$, $\mathbb{P}[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{\alpha}$. [**Hint:** Use (a)]

(c) If $u_1, \cdots, u_N$are iid random variables, each with mean μ and variance σ^2, and $u = \frac{1}{N}\sum_{n=1}^{N} u_n$, prove that for any $\alpha > 0$,

$$\mathbb{P}[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{N\alpha}.$$

Notice that the RHS of this Chebyshev Inequality goes down linearly in N, while the counterpart in Hoeffding's Inequality goes down exponentially. In Exercise 5, we develop an exponential bound using a similar approach.

Solution

(a) Let's assume t is a non-negetive random variable i.e $t \geq 0$ and $\alpha > 0$. So, we can write,

$$\begin{aligned}
\alpha I(t \geq \alpha) &\leq t && [I \text{ is the indicator which returns 1 when } t \geq \alpha \text{ else } 0] \\
\mathbb{E}[\alpha I(t \geq \alpha)] &\leq \mathbb{E}[t] && [\text{Taking expected value on either side}] \\
\alpha \mathbb{E}[I(t \geq \alpha)] &\leq \mathbb{E}[t] \\
\mathbb{E}[I(t \geq \alpha)] &\leq \frac{\mathbb{E}[t]}{\alpha} \\
\mathbb{P}(t \geq \alpha) &\leq \frac{\mathbb{E}[t]}{\alpha} && [\text{Markov inequality}]
\end{aligned}$$

(b) Let's consider a random variable u with mean μ, variance σ^2 and $\alpha > 0$. Therefore, we have,

$$\begin{aligned}
\mathbb{E}[u] &= \mu \\
Var(u) &= \sigma^2 \\
\mathbb{P}(u \geq \alpha) &\leq \frac{\mathbb{E}[u]}{\alpha} \text{ for } u \geq 0
\end{aligned}$$

Now, if we consider the random variable $(u-\mu)^2 \geq 0$, we obtain,

$$\begin{aligned}
\mathbb{P}[(u-\mu)^2 \geq \alpha] &\leq \frac{\mathbb{E}[(u-\mu)^2]}{\alpha} \\
&= \frac{Var(u)}{\alpha} \\
&= \frac{\sigma^2}{\alpha} && [\text{Chebyshev inequality}]
\end{aligned}$$

(c) $u_1, u_2, ...u_N$ are iid random variable with mean $\mathbb{E}[u_i] = \mu$ and variance $Var(u_i) = \sigma^2$. Now, considering $u = \frac{1}{N}\sum_{n=1}^{N} u_n$ and $\alpha > 0$ we obtain,

$$\begin{aligned}
mean(u) &= \mathbb{E}[u] = \frac{1}{N}\sum_{n=1}^{N}\mu = \mu \\
Var(u) &= Var(\frac{1}{N}\sum_{n=1}^{N} u_n) \\
&= \frac{Var(\sum_{n=1}^{N} u_n)}{N^2} \qquad [\text{As } Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]] \\
&= \frac{\sum_{n=1}^{N} Var(u_n)}{N^2} = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}
\end{aligned}$$

Now, if we replace the value of $Var(u)$ in the equation (b) we obtain,

$$\mathbb{P}[(u - \mu)^2 \geq \alpha] \leq \frac{\sigma^2}{N\alpha}$$

4 Background

The moment generating function (MGF) of a random variable X is given by:

$$M_X(s) = \mathbb{E}[e^{Xs}].$$

We called it the moment generating function because its derivatives evaluated at 0 provides the moments of X. In fact,

$$M_X'(0) = \left[\frac{d}{ds}\mathbb{E}[e^{Xs}]\right]_{s=0} = \mathbb{E}\left[\frac{d}{ds}e^{Xs}\right]_{s=0} = \mathbb{E}\left[Xe^{Xs}\right]_{s=0} = \mathbb{E}[X].$$

More generally, we have

$$M_X^{(k)}(0) = \mathbb{E}[X^k],$$

for $k = 1, 2, \ldots$.

There are two important properties of MGFs:

- *Sums of independnet random variables*: If we have random variables $X_1, X_2, \ldots, X_N$, which are independent, and $Y = \sum_{n=1}^{N} X_n$, then

$$M_Y(s) = \prod_{n=1}^{N} M_{X_n}(s).$$

  Basically, this allows us to calculate effectively every moment of a sum of independent random variables.

- *Equality of MGFs*: If the MGF of X and Y exist, and are equal, then X and Y have the same distribution.

5 Exercise 4

In this problem, we derive a form of the law of large numbers that has an exponential bound, called the Chernoff bound. We focus on the simple case of flipping a fair coin, and use an approach similar to Exercise 3.

(a) Let t be a (finite) random variable, a be a positive constant, and s be a positive parameter. If $T(s) = \mathbb{E}_t(e^{st})$, prove that

$$\mathbb{P}[t \geq \alpha] \leq e^{-s\alpha}T(s).$$

[**Hint:** e^{st} is monotonically increasing in t]

(b) Let $u_1, \cdots, u_N$ be iid random variables, and let $u = \frac{1}{N}\sum_{n=1}^{N} u_n$. If $U(s) = \mathbb{E}_{u_n}(e^{su_n}$ (for any n), prove that

$$\mathbb{P}[u \geq \alpha] \leq (e^{-s\alpha}U(s))^N.$$

(c) Suppose $\mathbb{P}[u_n = 0] = \mathbb{P}[u_n = 1] = \frac{1}{2}$ (fair coin). Evaluate $U(s)$ a a function of s, and minimize $e^{s\alpha}U(s)$ with respect to s for fixed α, $0 < \alpha < 1$.

(d) Conclude in (c) that, for $0 < \epsilon < \frac{1}{2}$,

$$\mathbb{P}[u \geq \mathbb{E}(u) + \epsilon] \leq 2^{-\beta N},$$

where $\beta = 1 + (\frac{1}{2} + \epsilon)\log_2(\frac{1}{2} + \epsilon) + (\frac{1}{2} - \epsilon)\log_2(\frac{1}{2} - \epsilon)$ and $\mathbb{E}(u) = \frac{1}{2}$. Notice that this bound is exponentially decreasing in N.

Solution

(a) Assume t to be a random variable, $\alpha > 0$ and $s \geq 0$. We have $T(s) = \mathbb{E}_t(e^{st})$

$$\begin{aligned}
\mathbb{P}(t \geq \alpha) &= \mathbb{P}(st \geq s\alpha) \\
&= \mathbb{P}(e^{st} \geq e^{s\alpha}) \qquad \text{[Using Hint]} \\
&\leq \frac{\mathbb{E}[e^{st}]}{e^{s\alpha}} \qquad \text{[Using Markov inequality]} \\
&\leq e^{-s\alpha}T(s)
\end{aligned}$$

(b) Let $u_1, u_2, ...u_N$ be iid random variable and let $u = \frac{1}{N}\sum_{n=1}^{N} u_n$. Let's consider $U(s) = \mathbb{E}_{u_n}[e^{su_n}]$ for any n.

$$\begin{aligned}
\mathbb{P}[u \geq \alpha] &= \mathbb{P}[Nu \geq N\alpha] \\
&\leq e^{-sN\alpha}\mathbb{E}[e^{sNu}] \qquad \text{[Using } \textbf{(a)}\text{]} \\
&\leq e^{-sN\alpha}\mathbb{E}[e^{sN\frac{1}{N}\sum_{n=1}^{N} u_n}] \\
&\leq e^{-sN\alpha}\mathbb{E}[e^{\sum_{n=1}^{N} su_n}] \\
&\leq e^{-sN\alpha}\prod_{n=1}^{N}\mathbb{E}[e^{su_n}] \\
&\leq e^{-sN\alpha}\prod_{n=1}^{N}U(s) \\
&\leq e^{-sN\alpha}[U(s)]^N \\
&\leq [e^{-s\alpha}U(s)]^N
\end{aligned}$$

(c) Suppose we have a fair coin, $\mathbb{P}[u_n = 0] = \mathbb{P}[u_n = 1] = \frac{1}{2}$.

$$\begin{aligned}
U(s) &= \mathbb{E}_{u_n}(e^{su_n}) \\
&= \mathbb{P}[u_n = 0]e^{s\cdot 0} + \mathbb{P}[u_n = 1]e^{s\cdot 1} \\
&= \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot e^s \\
&= \frac{1}{2}(1 + e^s)
\end{aligned}$$

Let's consider

$$\begin{aligned}
f(s) &= e^{-s\alpha}U(s) \\
&= e^{-s\alpha}\frac{1}{2}(1 + e^s) \\
f'(s) &= \frac{e^{-s\alpha}}{2}[(1-\alpha)e^s - \alpha]
\end{aligned}$$

Now if we solve $f'(s) = 0$, we get a root $s^* = \ln(\frac{\alpha}{1-\alpha})$

$$\begin{aligned}
f''(s) &= \frac{e^{-s\alpha}}{2}[(\alpha - 1)^2 e^s + \alpha^2] \\
f''(s^*) &> 0
\end{aligned}$$

Therefore, s^* is minimum of $f(s)$.

(d)

$$\mathbb{E}(u) = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}(u_n) = \frac{1}{N}\sum_{n=1}^{N}(0\cdot\frac{1}{2} + 1\cdot\frac{1}{2}) = \frac{1}{2}$$

Now, for $0 < \epsilon < 1$, we can write

$$\begin{aligned}
\mathbb{P}[u \geq \mathbb{E}(u) + \epsilon] &= \mathbb{P}[u \geq \frac{1}{2} + \epsilon] \\
&\leq [e^{-s(\frac{1}{2}+\epsilon)}U(s)]^N \qquad \text{[Using (b)]} \\
&\leq \min_s [e^{-s(\frac{1}{2}+\epsilon)}U(s)]^N \\
&\text{[If the condition hold for any } s\text{, it will hold for the minimum value of } s \text{]} \\
&\leq [e^{-s^*(\frac{1}{2}+\epsilon)}U(s^*)]^N \\
&\leq [e^{-\ln(\frac{\alpha}{1-\alpha})(\frac{1}{2}+\epsilon)}U(\ln(\frac{\alpha}{1-\alpha}))]^N \\
&\leq \left[e^{-\ln\frac{\frac{1}{2}+\epsilon}{1-\frac{1}{2}-\epsilon}(\frac{1}{2}+\epsilon)}U\Big(\ln\frac{\frac{1}{2}+\epsilon}{1-\frac{1}{2}-\epsilon}\Big)\right]^N \\
&\leq \left[e^{-\ln\frac{\frac{1}{2}+\epsilon}{\frac{1}{2}-\epsilon}(\frac{1}{2}+\epsilon)}U\Big(\ln\frac{\frac{1}{2}+\epsilon}{\frac{1}{2}-\epsilon}\Big)\right]^N \\
&\leq \left[\Big(\frac{\frac{1}{2}+\epsilon}{\frac{1}{2}-\epsilon}\Big)^{(\frac{1}{2}+\epsilon)}\Big(\frac{1}{2}\Big(1 + e^{\ln\frac{\frac{1}{2}+\epsilon}{\frac{1}{2}-\epsilon}}\Big)\Big)\right]^N \qquad \text{[Using (c)]} \\
&\leq \left[\frac{1}{2}\Big(\frac{\frac{1}{2}+\epsilon}{\frac{1}{2}-\epsilon}\Big)^{(\frac{1}{2}+\epsilon)}\Big(1 + \frac{\frac{1}{2}+\epsilon}{\frac{1}{2}-\epsilon}\Big)\right]^N
\end{aligned}$$

$$
\begin{aligned}
&\leq \Big[\frac{1}{2}\Big(\frac{\frac{1}{2}+\epsilon}{\frac{1}{2}-\epsilon}\Big)^{(\frac{1}{2}+\epsilon)}\Big(\frac{1}{2}-\epsilon\Big)^{-1}\Big]^N \\
&\leq \Big[2^{-1}\Big(\frac{1}{(1/2+\epsilon)^{(1/2+\epsilon)}}\Big)\Big(\frac{1}{(1/2-\epsilon)^{(1/2-\epsilon)}}\Big)\Big]^N \\
&\leq \Big[2^{-1-\log_2(1/2+\epsilon)^{(1/2+\epsilon)}-\log_2(1/2-\epsilon)^{(1/2-\epsilon)}}\Big]^N \\
&\leq 2^{-N[1+(1/2+\epsilon)\log_2(1/2+\epsilon)+(1/2-\epsilon)\log_2(1/2-\epsilon)]} \\
&\leq 2^{-\beta N}
\end{aligned}
$$

where, $\beta = 1 + (1/2+\epsilon)\log_2(1/2+\epsilon) + (1/2-\epsilon)\log_2(1/2-\epsilon)$.

Now, to check if the bound is monotonically decreasing in N, we need to take the derivative of β with respect to ϵ.

$$
\begin{aligned}
\beta &= 1 + (1/2+\epsilon)\log_2(1/2+\epsilon) + (1/2-\epsilon)\log_2(1/2-\epsilon) \\
\beta' &= \log_2(1/2+\epsilon) - \log_2(1/2-\epsilon)
\end{aligned}
$$

β' is positive for $0 < \epsilon < 1/2$, therefore, β is a monotonically increasing function. Hence, the bound is exponentially decreasing in N.

6 Exercise 5

Lemma 1 (Chernoff's method). Let X be a random variable. Then, for any $\varepsilon > 0$, we have

$$P(X > \varepsilon) \leq \inf_{s>0}\ e^{-s\varepsilon}\mathbb{E}[e^{Xs}] \text{ and } P(X < -\varepsilon) \leq \inf_{s>0}\ e^{-s\varepsilon}\mathbb{E}[e^{-Xs}].$$

Lemma 2 (Hoeffding's lemma). Suppose that $a \leq X \leq b$ and $\mu = \mathbb{E}[X]$. Then,

$$\mathbb{E}[e^{Xs}] \leq e^{s\mu}e^{\frac{s^2(b-a)^2}{8}}.$$

Hoeffding's inequality. Let $X_1, X_2, \ldots, X_N$ be i.i.d. observations such that $\mathbb{E}[X_n] = \mu$, $a \leq X_n \leq b$ and $\bar{X} = \frac{1}{N}\sum_{n=1}^{N} X_n$. Then, for any $\varepsilon > 0$,

$$P(|\bar{X} - \mu| > \varepsilon) \leq 2e^{-2N\varepsilon^2/(b-a)^2}.$$

Prove Hoeffding's inequality using the Chernoff's method and Hoeffding's Lemma, and without loss of generality, you can assume that $\mu = 0$.

Solution

Assume $\mu = 0$.

$$\begin{aligned} P(|\bar{X} - \mu| > \varepsilon) &= P(|\bar{X}| > \varepsilon) \\ &= P(\bar{X} > \varepsilon) + P(\bar{X} < -\varepsilon) \qquad \text{[As the two events are disjoint]} \end{aligned}$$

$$\begin{aligned} P(\bar{X} > \varepsilon) &= P(\frac{1}{N}\sum_{n=1}^{N} X_n > \varepsilon) \\ &= P(\sum_{n=1}^{N} X_n > N\varepsilon) \\ &\leq \inf_{s>0}\ e^{-sN\varepsilon}\mathbb{E}[e^{s\sum_{n=1}^{N} X_n}] \qquad \text{[Using Chernoff's method]} \\ &\leq \inf_{s>0}\ e^{-sN\varepsilon}\mathbb{E}[\prod_{n=1}^{N} e^{sX_n}] \\ &\leq \inf_{s>0}\ e^{-sN\varepsilon}\prod_{n=1}^{N}\mathbb{E}[e^{sX_n}] \\ &\leq \inf_{s>0}\ e^{-sN\varepsilon}\prod_{n=1}^{N} e^{s\mu}e^{\frac{s^2(b-a)^2}{8}} \qquad \text{[Using Hoeffding's lemma]} \\ &\leq \inf_{s>0}\ e^{-sN\varepsilon}\prod_{n=1}^{N} e^{\frac{s^2(b-a)^2}{8}} \qquad [\text{As } \mu = 0] \\ &\leq \inf_{s>0}\ e^{-sN\varepsilon}e^{N\frac{s^2(b-a)^2}{8}} \\ &\leq \inf_{s>0}\ e^{-sN\varepsilon+N\frac{s^2(b-a)^2}{8}} \end{aligned}$$

Let us minimize the following function:

$$f(s) = -sN\varepsilon + N\frac{s^2(b-a)^2}{8}.$$

A necessary condition for optimality is given by

$$f'(s) = -N\varepsilon + 2N\frac{s(b-a)^2}{8} = 0$$
$$\iff s^* = \frac{4\varepsilon}{(b-a)^2}$$

We can check that s^* is a minimum by verifying if the second derivative of $f(s)$ is positive at $s = s^*$.
By replacing s with s^* in the bound above, we obtain

$$P(\bar{X} > \varepsilon) \leq e^{-2N\varepsilon^2/(b-a)^2}.$$

Similarly, we have $P(\bar{X} < -\varepsilon) \leq e^{-2N\varepsilon^2/(b-a)^2}$.

Therefore,

$$P(|\bar{X}| > \varepsilon) \leq 2e^{-2N\varepsilon^2/(b-a)^2}$$

7 Exercise 6

Which of the following are possible growth functions $m_{\mathcal{H}}(N)$ for some hypothesis set:

$$1+N; 1+N+\frac{N(N-1)}{2}; 2^N; 2^{\lfloor\sqrt{N}\rfloor}; 2^{\lfloor N/2\rfloor}; 1+N+\frac{N(N-1)(N-2)}{6}.$$

Solution

There are two cases for the growth function:

- $d_{\text{VC}} = \infty$ and $m_{\mathcal{H}}(N) = 2^N$ for all N.
- d_{VC} is finite and $m_{\mathcal{H}}(N) \leq N^{d_{\text{VC}}} + 1$

Now let's see the growth functions:

(i) $m_{\mathcal{H}}(N) = 1 + N$
Here, $d_{\text{VC}} = 1$ as $m_{\mathcal{H}}(2) = 3 < 2^2$. Therefore, $m_{\mathcal{H}}(N)$ must be bounded by $N^1 + 1$ which is true. Hence, $m_{\mathcal{H}}(N) = 1 + N$ is a possible growth function.

(ii) $m_{\mathcal{H}}(N) = 1 + N + \frac{N(N-1)}{2}$
Here, $d_{\text{VC}} = 2$ as $m_{\mathcal{H}}(3) = 7 < 2^3$. Therefore, $m_{\mathcal{H}}(N)$ must be bounded by $N^2 + 1$ for all N, which is true here. Hence, $m_{\mathcal{H}}(N) = 1 + N + \frac{N(N-1)}{2}$ is a possible growth function.

(iii) $m_{\mathcal{H}}(N) = 2^N$
Here, $d_{\text{VC}} = \infty$. Hence, $m_{\mathcal{H}}(N) = 2^N$ is a possible growth function.

(iv) $m_{\mathcal{H}}(N) = 2^{\lfloor\sqrt{N}\rfloor}$
Here, $d_{\text{VC}} = 1$ as $m_{\mathcal{H}}(2) = 2 < 2^2$. Therefore, $m_{\mathcal{H}}(N)$ must be bounded by $N^1 + 1$ for all N. Consider an example $N = 25$, $m_{\mathcal{H}}(25) = 32 \geq 25 + 1$. So, the bound $N^1 + 1$ for all N is not true here. Hence, $m_{\mathcal{H}}(N) = 2^{\lfloor\sqrt{N}\rfloor}$ is **not** a possible growth function.

(v) $m_{\mathcal{H}}(N) = 2^{\lfloor N/2\rfloor}$
Here, $d_{\text{VC}} = 0$ as $m_{\mathcal{H}}(1) = 1 < 2^1$. Therefore, $m_{\mathcal{H}}(N)$ must be bounded by $N^0 + 1 = 2$ for all N. Consider an example $N = 4$, $m_{\mathcal{H}}(4) = 4 \geq 2$. So, the bound $N^0 + 1$ for all N is not true here. Hence, $m_{\mathcal{H}}(N) = 2^{\lfloor N/2\rfloor}$ is **not** a possible growth function.

(vi) $m_{\mathcal{H}}(N) = 1 + N + \frac{N(N-1)(N-2)}{6}$
Here, $d_{\text{VC}} = 1$ as $m_{\mathcal{H}}(2) = 3 < 2^2$. Therefore, $m_{\mathcal{H}}(N)$ must be bounded by $N^1 + 1$ for all N. Consider an example $N = 3$, $m_{\mathcal{H}}(3) = 5 \geq 3^1 + 1$. So, the bound $N^0 + 1$ for all N is not true here. Hence, $m_{\mathcal{H}}(N) = 1 + N + \frac{N(N-1)(N-2)}{6}$ is **not** a possible growth function.

8 Exercise 7

Compute the maximum number of dichotomies, $m_{\mathcal{H}}(N)$, for these learning models, and consequently compute d_{VC}, the VC dimension.

(a) Positive or negative ray: $\mathcal{H}$ contains the functions which are $+1$ on $[a, \infty)$ (for some a) together with those that are $+1$ on $(-\infty, a]$ (for some a).

(b) Positive or negative interval: $\mathcal{H}$ contains the functions which are $+1$ on $[a, b]$ and -1 elsewhere or -1 on an interval $[a, b]$ (for some a) together and $+1$ elsewhere.

(c) Two concentric spheres in $\mathbb{R}^d$: $\mathcal{H}$ contains the functions which are $+1$ for $a \leq \sqrt{x_1^2 + x_2^2 + \cdots + x_d^2} \leq b$

Solution

(a) **Positive or negative ray:** The growth function for positive rays is $N + 1$. Now, for negative rays we get $N - 1$ numbers of new dichotomies (the opposite of the ones from positive rays - two dichotomies where all points are +1 and where all are -1. Hence,

$$
\begin{aligned}
m_{\mathcal{H}}(N) &= N + 1 + N - 1 = 2N \\
d_{\text{VC}} &= 2 \qquad [\text{As } m_{\mathcal{H}}(3) = 6 < 2^3]
\end{aligned}
$$

(b) **Positive or negative interval:** The growth function for positive interval is $\binom{N+1}{2} + 1$ as we can place interval ends in two of $N + 1$ spots. Similarly, growth function for negative interval is $\binom{N+1}{2} + 1$. Now, we need to subtract the number of dichotomies covered by both the positive and negative interval. The overlap is $2N$ where all the positive and negative points are grouped separately (Consider for $N = 3$, 6 dichotomies are covered by both). Hence,

$$
\begin{aligned}
m_{\mathcal{H}}(N) &= \binom{N+1}{2} + 1 + \binom{N+1}{2} + 1 - 2N = N^2 - N + 2 \\
d_{\text{VC}} &= 3 \qquad [\text{As } m_{\mathcal{H}}(4) = 14 < 2^4]
\end{aligned}
$$

Now, we need to add the number of dichotomies for negative interval. For negative interval, we can place interval ends in two of $N - 1$ spots as we need to subtract two spots where all points are +1 and all are -1. Therefore, for negative interval, the number of dichotomies is $\binom{N-1}{2}$. Hence,

$$
\begin{aligned}
m_{\mathcal{H}}(N) &= \binom{N+1}{2} + 1 + \binom{N-1}{2} = N^2 - N + 2 \\
d_{\text{VC}} &= 3 \qquad [\text{As } m_{\mathcal{H}}(4) = 14 < 2^4]
\end{aligned}
$$

(c) **Two concentric spheres in $\mathbb{R}^d$:** We can map two concentric sphere from $\mathbb{R}^d$ to $[0, \infty)$ by using the function below:

$$
f : (x_1, x_2, \ldots, x_d) \mapsto r = \sqrt{x_1^2 + x_2^2 + \cdots + x_d^2}
$$

Now, the problem of two concentric circles in $\mathbb{R}^d$ is equivalent to the problem of positive interval. Hence,

$$
\begin{aligned}
m_{\mathcal{H}}(N) &= \binom{N+1}{2} = \frac{N^2}{2} + \frac{N}{2} + 1 \\
d_{\text{VC}} &= 2 \qquad [\text{As } m_{\mathcal{H}}(3) = 7 < 2^3]
\end{aligned}
$$

9 Exercise 8

Show that $B(N,k) = \sum_{i=0}^{k-1} \binom{N}{i}$ by showing the other direction to

$$B(N,k) \geq \sum_{i=0}^{k-1} \binom{N}{i}.$$

To do so, construct a specific set of $\sum_{i=0}^{k-1} \binom{N}{i}$ dichotomies that does not shatter any subset of k variables. [**Hint:** Try limiting the number of -1's in each dichotomy.]

Solution

Let's assume that we have N points and $m_{\mathcal{H}}(N) = 2^N$. Now, we focus on the dichotomies that contains $(k-1)$ -1's. These dichotomies are:

- Number of dichotomies that doesn't contain -1 : $\binom{N}{0} = 1$.
- Number of dichotomies that contain one -1 : $\binom{N}{1} = N$.
- Number of dichotomies that contain two -1's : $\binom{N}{2}$.
- Number of dichotomies that contain three -1's : $\binom{N}{3}$.
- ...
- Number of dichotomies that contain $(k-1)$ -1's : $\binom{N}{k-1}$.

In total there are $\sum_{i=0}^{k-1} \binom{N}{i}$ such dichotomies. Moreover, these dichotomies do not shatter any subset of k variables and the set does not have any dichotomy which contains k (-1)s. Hence,

$$B(N,k) \geq \sum_{i=0}^{k-1} \binom{N}{i}$$

Now, Sauer's lemma is:

$$B(N,k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

Therefore, we can conclude that,

$$B(N,k) = \sum_{i=0}^{k-1} \binom{N}{i}$$

10 Exercise 9

Prove by induction that $\sum_{i=0}^{D}\binom{N}{i} \le N^D + 1$, hence

$$m_{\mathcal{H}}(N) \le N^{d_{\mathrm{VC}}} + 1$$

Solution

To prove the inequality, let's look at the following cases:

- For $D = 0$,

$$1 = \binom{N}{0} \le N^0 + 1$$

- Consider the inequality is true for D ($D \ge 1$),

$$\sum_{i=0}^{D}\binom{N}{i} \le N^D + 1$$

- Now, we need to prove that it is true for $D+1$,

$$\begin{aligned}\sum_{i=0}^{D+1}\binom{N}{i} &= \sum_{i=0}^{D}\binom{N}{i} + \binom{N}{D+1}\\ &\le N^D + 1 + \binom{N}{D+1}\\ &\le N^D + 1 + \frac{N!}{(D+1)!(N-D-1)!}\end{aligned}$$

Now, we want to prove that $\frac{N!}{(N-D-1)!} \le N^{D+1}$,

$$\frac{N!}{(N-D-1)!} = \prod_{i=0}^{D}(N-i) \le N^{D+1}$$

Therefore, we can write,

$$\begin{aligned}\sum_{i=0}^{D+1}\binom{N}{i} &\le N^D + 1 + \frac{N^{D+1}}{(D+1)!}\\ &\le N^D + 1 + \frac{N^{D+1}}{2} \qquad [\text{As } D \ge 1, \text{ we have } (D+1)! \ge 2 \iff \frac{1}{(D+1)!} \le \frac{1}{2}]\end{aligned}$$

Moreover, we have assumed $N \ge D+1$ (otherwise, $\binom{N}{D+1} = 0$). So, $N \ge 2$ and consequently

$$\frac{1}{N} \le \frac{1}{2} \iff \frac{N^D}{N^{D+1}} \le \frac{1}{2} \iff N^D \le \frac{N^{D+1}}{2}$$

Hence we can write,

$$\begin{aligned}\sum_{i=0}^{D+1}\binom{N}{i} &\le N^D + 1 + \frac{N^{D+1}}{2}\\ &\le \frac{N^{D+1}}{2} + 1 + \frac{N^{D+1}}{2}\end{aligned}$$

$$\leq N^{D+1} + 1$$

So, we have proved $\sum_{i=0}^{D} \binom{N}{i} \leq N^D + 1$. Now,

$$\begin{aligned} m_{\mathcal{H}}(N) &\leq \sum_{i=0}^{d_{\text{VC}}} \binom{N}{i} \\ &\leq N^{d_{\text{VC}}} + 1 \end{aligned}$$

11 Exercise 10

1. Let $\mathcal{H} = \{h_1, h_2, \ldots, h_M\}$ with some finite M. Prove that $d_{\text{VC}}(\mathcal{H}) \leq \log_2 M$.
2. For hypothesis sets $\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_K$ with finite VC dimensions $d_{\text{VC}}(\mathcal{H}_k)$, derive and prove the highest upper and lower bound that you can get on $d_{\text{VC}}(\cap_{k=1}^K \mathcal{H}_k)$.
3. For hypothesis sets $\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_K$ with finite VC dimensions $d_{\text{VC}}(\mathcal{H}_k)$, derive and prove the highest upper and lower bound that you can get on $d_{\text{VC}}(\cup_{k=1}^K \mathcal{H}_k)$.

Solution

(a) Let, $d_{\text{VC}} = d$, then, $m_H(d) = 2^d$ (by definition). Now,

$$\begin{aligned}
m_{\mathcal{H}}(d) &= \max_{x_1, x_2, \ldots, x_d} |\mathcal{H}(x_1, x_2, \ldots, x_d)| \\
&= \max_{x_1, x_2, \ldots, x_d} |\{(h(x_1), h(x_2), \ldots, h(x_d)) : h \in \mathcal{H}\}| \\
&= \max_{x_1, x_2, \ldots, x_d} |\{(h(x_1), h(x_2), \ldots, h(x_d)) : h \in \{h_1, h_2, \ldots, h_M\}\}| \\
&\leq |\mathcal{H}| = M
\end{aligned}$$

Therefore, we can write,

$$\begin{aligned}
&2^d \leq M \\
\iff &d \leq \log_2(M)
\end{aligned}$$

(b) At worst, we have $\cap_{k=1}^K \mathcal{H}_k = \{h\}$. Here, trivially VC dimension is 0 as $m_{\mathcal{H}}(N) = 1$ for all N. So, we can write $d_{\text{VC}}(\cap_{k=1}^K \mathcal{H}_k) \geq 0$.
Now, we will prove that,

$$d_{\text{VC}}(\cap_{k=1}^K \mathcal{H}_k) \leq \min_{1 \leq k \leq K} d_{\text{VC}}(\mathcal{H}_k)$$

To prove that let's assume,

$$d_{\text{VC}}(\cap_{k=1}^K \mathcal{H}_k) > \min_{1 \leq k \leq K} d_{\text{VC}}(\mathcal{H}_k) = d$$

It means that $\cap_{k=1}^K \mathcal{H}_k$ can shatter $d+1$ points, let $x_1, \ldots, x_{d+1}$ be those points. Now, we may write,

$$\begin{aligned}
\{-1, +1\}^{d+1} &= \cap_{k=1}^K \mathcal{H}_k(x_1, \ldots, x_{d+1}) \\
&= \{(h(x_1), \cdots, h(x_{d+1})) : h \in \cap_{k=1}^K \mathcal{H}_k\} \\
&\subseteq \{(h(x_1), \cdots, h(x_{d+1})) : h \in \mathcal{H}_k\} \quad \text{for all } k = 1, \cdots, K
\end{aligned}$$

If we compute the cardinality of these sets, we obtain,

$$\begin{aligned}
&2^{d+1} \leq |\{(h(x_1), \cdots, h(x_{d+1})) : h \in \mathcal{H}_k\}| \leq 2^{d+1} \quad \text{for all } k = 1, \cdots, K \\
&\Rightarrow |\{(h(x_1), \cdots, h(x_{d+1})) : h \in \mathcal{H}_k\}| = 2^{d+1} \quad \text{for all } k = 1, \cdots, K
\end{aligned}$$

Therefore, any $\mathcal{H}_k$ can shatter $d+1$ points.
Now, let $\min_{1 \leq k \leq K} d_{\text{VC}}(\mathcal{H}_k) = d_{\text{VC}}(\mathcal{H}_{k0})$. Then we have,

$$d = d_{\text{VC}}(\mathcal{H}_{k0}) \geq d + 1$$

which is not possible. Hence,

$$0 \leq d_{\text{VC}}(\cap_{k=1}^K \mathcal{H}_k) \leq \min_{1 \leq k \leq K} d_{\text{VC}}(\mathcal{H}_k)$$

(c) Let, $d_{\mathrm{VC}}(H_k) = d_k$ for all $k = 1, \cdots, K$. This implies $\mathcal{H}_k$ shatters d_k points $x_1, \cdots, x_{dk}$,

$$\begin{aligned}\{-1,+1\}^{d_k} &= \{(h(x_1), \cdots, h(x_{d_k})) : h \in \mathcal{H}_k\} \\ &\subset \{(h(x_1), \cdots, h(x_{d_k})) : h \in \cup_{k=1}^{K} \mathcal{H}_k\}\end{aligned}$$

Now if we compute the cardinality of these sets, we obtain

$$\begin{aligned}&2^{d_k} \leq |\{(h(x_1), \cdots, h(x_{d_k})) : h \in \cup_{k=1}^{K} \mathcal{H}_k\}| \leq 2^{d_k} \\ &\Rightarrow |\{(h(x_1), \cdots, h(x_{d_k})) : h \in \cup_{k=1}^{K} \mathcal{H}_k\}| = 2^{d_k} \quad \text{for all } k = 1, \cdots, K\end{aligned}$$

More simply we can write,

$$\begin{aligned}&m_{\cup_{k=1}^{K} \mathcal{H}_k}(d_k) = 2^{d_k} \quad \forall k \\ &\Rightarrow d_{\mathrm{VC}}(\cup_{k=1}^{K} \mathcal{H}_k) \geq d_k \quad \forall k \\ &\Rightarrow d_{\mathrm{VC}}(\cup_{k=1}^{K} \mathcal{H}_k) \geq \max_{1 \leq k \leq K} d_k = \max_{1 \leq k \leq K} d_{\mathrm{VC}}(\mathcal{H}_k)\end{aligned}$$

Now, consider $K = 2$ and $d_{\mathrm{VC}}(\mathcal{H}_1) = d_1$ and $d_{\mathrm{VC}}(\mathcal{H}_2) = d_2$. The number of dichotomies generated by $\mathcal{H}_1 \cup \mathcal{H}_2$ is at most the sum of the dichotomies generated by $\mathcal{H}_1$ and by $\mathcal{H}_2$. Therefore,

$$\begin{aligned}m_{\mathcal{H}_1 \cup \mathcal{H}_2} &\leq m_{\mathcal{H}_1} + m_{\mathcal{H}_2} \\ &\leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=0}^{d_2} \binom{N}{i} \\ &\leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=0}^{d_2} \binom{N}{N-i} \\ &\leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=N-d_2}^{d_2} \binom{N}{i} \\ &< \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=d_1+1}^{N-d_2-1} \binom{N}{i} + \sum_{i=N-d_2}^{d_2} \binom{N}{i} = \sum_{i=0}^{N} \binom{N}{i} = 2^N\end{aligned}$$

$\forall N$ such that $d_1 + 1 \leq N - d_2 - 1 \iff N \geq d_1 + d_2 + 1$. So, we can deduce that

$$d_{\mathrm{VC}}(\mathcal{H}_1 \cup \mathcal{H}_2) \leq d_1 + d_2 + 1$$

Now, we will prove by induction,

$$d_{\mathrm{VC}}(\cup_{k=1}^{K} \mathcal{H}_k) \leq K - 1 + \sum_{k=1}^{K} d_{\mathrm{VC}}(\mathcal{H}_k)$$

- For $K = 2$,

$$d_{\mathrm{VC}}(\mathcal{H}_1 \cup \mathcal{H}_2) \leq 1 + \sum_{k=1}^{2} d_{\mathrm{VC}}(\mathcal{H}_k) \quad \text{[Already proven]}$$

- Consider it is true for $K - 1$,

$$d_{\mathrm{VC}}(\cup_{k=1}^{K-1} \mathcal{H}_k) \leq K - 2 + \sum_{k=1}^{K-1} d_{\mathrm{VC}}(\mathcal{H}_k)$$

- For K,

$$\begin{aligned}
d_{\mathrm{VC}}(\cup_{k=1}^{K}\mathcal{H}_k) &= d_{\mathrm{VC}}((\cup_{k=1}^{K-1}\mathcal{H}_k) \cup \mathcal{H}_K) \\
&\leq 1 + d_{\mathrm{VC}}(\cup_{k=1}^{K-1}\mathcal{H}_k) + d_{\mathrm{VC}}(\mathcal{H}_K) \\
&\leq 1 + K - 2 + \sum_{k=1}^{K-1} d_{\mathrm{VC}}(\mathcal{H}_k) + d_{\mathrm{VC}}(\mathcal{H}_K) \\
&\leq K - 1 + \sum_{k=1}^{K} d_{\mathrm{VC}}(\mathcal{H}_k)
\end{aligned}$$

Finally, we obtain,

$$\max_{1\leq k\leq K} d_{\mathrm{VC}}(\mathcal{H}_k) \leq d_{\mathrm{VC}}(\cup_{k=1}^{K}\mathcal{H}_k) \leq K - 1 + \sum_{k=1}^{K} d_{\mathrm{VC}}(\mathcal{H}_k)$$