Instituto Politécnico Nacional

Centro de Investigación en cómputo

Chapter 2: Training versus testing.

Subject: Introducción a machine learning

Alumno:

- Carpintero Mendoza Marcos Mauricio

Profesor:
Menchaca Mendez Ricardo

14/10/19

Índice

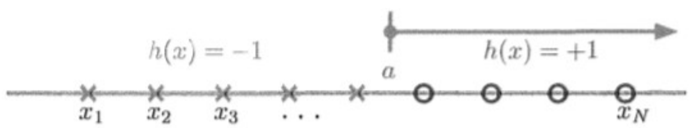
1. Exercises.

1.1. Theory of generalization.

1.1.1. Effective number of hypotheses.

Exercise 2.1 By inspection, find a break point k for each hypothesis set in *Example 2.2* (if there is one). Verify that $m_H(f) < 2^n$ using the formulas derived in the Example.
Solution:
We want to figure out the k such that the the hypothesis can not shatter the dichotomy, so we must do it iteratively. For all the hypothesis we should start with a $N = 2$, because with $N = 1$ is pretty trivial and obvious.

1. For linear rays, which $h(x) = sgn(x - a)$:

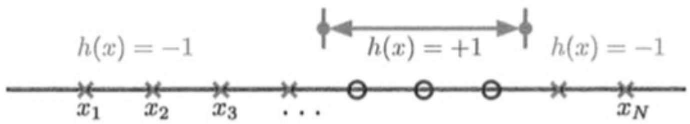


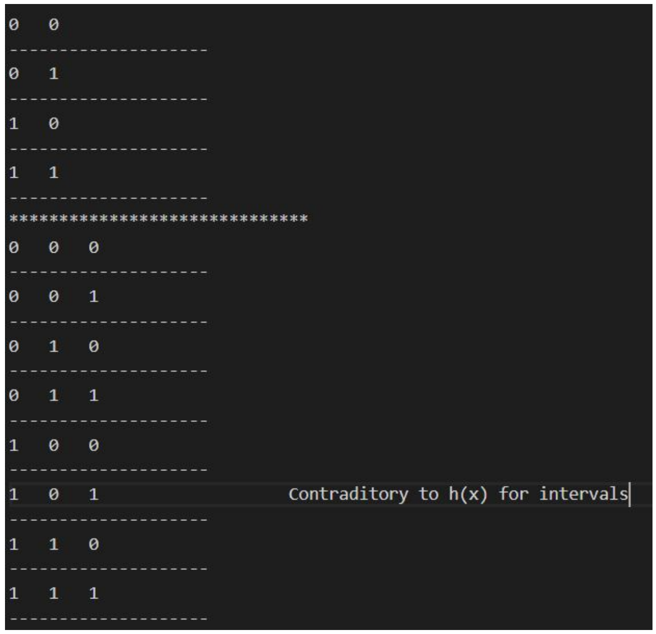

```
0   0
--------------------
0   1
--------------------
1   0                   Contraditory to h(x) for arrays
--------------------
1   1
```

Figura 1: We can observe that exists one case in which the hypothesis can not obtain that pattern.

For that we can say $k = 2$

2. For intervals as its name sounds $h(x) = +1$ if x is within the interval, and otherwise $h(x) = -1$.

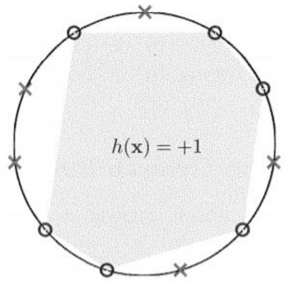


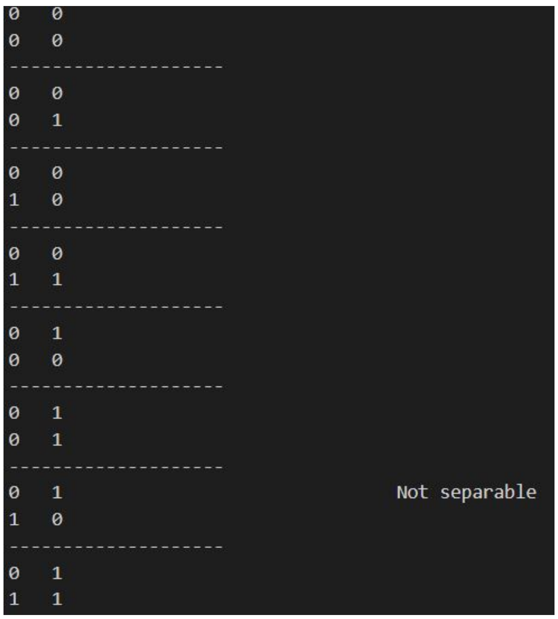


Here we must observe that for $N = 2$, which is a break point for the previous hypothesis, is not a break point here because the current hypothesis is more flexible. But with $N = 3$ an impossible case shows up and is the same analysis as before, because is contradictory to $h(x)$.

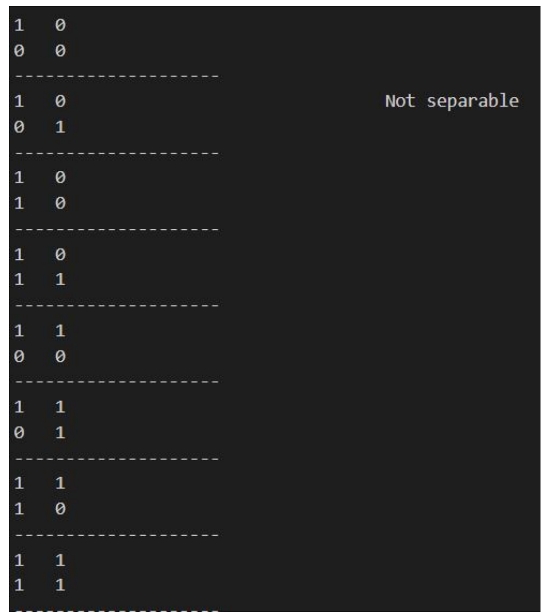
For that we can say $k = 3$

3. For covex sets, which $h(x) = +1$ if x is inside the convex and is negative if otherwise:





```
0   0
0   0
--------------------
0   0
0   1
--------------------
0   0
1   0
--------------------
0   0
1   1
--------------------
0   1
0   0
--------------------
0   1
0   1
--------------------
0   1                                   Not separable
1   0
--------------------
0   1
1   1
```



```
1   0
0   0
--------------------
1   0                                   Not separable
0   1
--------------------
1   0
1   0
--------------------
1   0
1   1
--------------------
1   1
0   0
--------------------
1   1
0   1
--------------------
1   1
1   0
--------------------
1   1
1   1
--------------------
```

Here there are all the patterns that can be generated with $N = 2$. With this hypothesis there are two cases in which we can not shatter them with a 2D perceptron.

For that we can say$k = 4$

Código: `breakpoint.py`

```
def printOne_D(pattern):
    listt = list(pattern)
    print(str.join('\t', listt))

def printTwo_D(pattern):
    print("%s\t%s" % (pattern[0], pattern[1]))
    print("%s\t%s" % (pattern[2], pattern[3]))


binaries = lambda x, n: format(x, 'b').zfill(n) # For 2D perceptron

if __name__ == '__main__':
    for i in range(16):
        bin_ = binaries(i, 4)
        printTwo_D(bin_)
        print("-"*20)

    print('*'*30)

    for j in range(2, 4, 1):
        for i in range(2**j):
            bin_ = binaries(i, j)
            printOne_D(bin_)
            print("-"*20)
        print('*'*30)
```

1.1.2. Bounding the Growth Function.

Exercise 2.2

1. Verify the bound of Theorem 2.4 in the three cases of Example 2.2:

   a) Positive rays: $\mathcal{H}$ consists of all hypotheses in one dimension of the form $h(x) = sign(x - a)$.

   b) Positive intervals: $\mathcal{H}$ consists of all hypotheses in one dimension that are positive within some interval and negative elsewhere.

   c) Convex sets: $\mathcal{H}$ consists of all hypotheses in two dimensions that are positive inside some convex set and negative elsewhere.

Solution:

1. Using the Theorem 2.4:

$$\text{If } m_h(k) < 2^k \quad \Longrightarrow \quad m_h(N) \leqslant \sum_{i=0}^{k-1} \binom{N}{i}$$

   a) Positive rays, for k = 2.

$$m_h(N) \leqslant \sum_{i=0}^{2-1} \binom{N}{i}$$
$$= \binom{N}{0} + \binom{N}{1} = 1 + \frac{(N)(N-1)!}{(N-1)!}$$
$$= 1 + N$$

   Result the same as the previous growth function.

   b) Intervals, for k = 3.

$$m_h(N) \leqslant \sum_{i=0}^{3-1} \binom{N}{i}$$
$$= \binom{N}{0} + \binom{N}{1} + \binom{N}{2} = 1 + \frac{(N)(N-1)!}{(N-1)!} + \frac{(N)(N-1)(N-2)!}{(N-2)!}$$
$$= 1 + N + \frac{N(N-1)}{2!} = 1 + N + \frac{N^2 - N}{2} = 1 + \frac{N}{2} + \frac{N^2}{2}$$

   Result the same as the previous growth function.

   c) Convex sets, for k = 4.

$$m_h(N) \leqslant \sum_{i=0}^{4-1} \binom{N}{i}$$
$$= \sum_{i=0}^{2} \binom{N}{i} + \binom{N}{3} = 1 + \frac{N}{2} + \frac{N^2}{2} + \frac{N!}{(N-3)!3!}$$
$$= 1 + \frac{N^3}{6} + \frac{5N}{6}$$
$$2^N \leqslant 1 + \frac{N^3}{6} + \frac{5N}{6}$$

1.1.3. The VC Dimension.

Exercise 2.4

Consider the input space $X = \{1\} \times R^d$ (including the constant coordinate $x_0 = 1$). Show that the dimension of the perceptron (with $d+1$ parameters, counting w_0) is exactly $d+1$ by showing that it is at least $d+1$ and at most $d+1$, as follows:

a) To show that $d_{\text{vc}\geqslant d+1}$, find $d+1$ points in X that the perceptron can shatter.

b) To show that $d_{\text{vc}\leqslant d+1}$, show that no set of $d+2$ points in X can be shattered by perceptron.

Solution:

a) Let $X \in \mathcal{M}_{\text{d+1 x d+1}}$, so is a square matrix which rows are x vector with $d+1$ elements.

$$X_{d+1,d+1} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_{d+1}^T \end{pmatrix}$$

We know that $det(X) \neq 0$, so X is invertible. Now let $Y \in \mathcal{M}_{\text{d+1 x 1}}$

$$Y_{d+1,1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{pmatrix}$$

Finally, we know that perceptron works as follow:

$$\text{sgn}(Xw) = Y \quad \Longleftarrow \quad (Xw) = Y \quad \Longleftrightarrow \quad w = X^{-1}Y$$

$$\text{We get that X can be shattered} \quad \Longrightarrow \quad d_{\text{vc}} \geqslant d+1$$

b) Let $X \in \mathcal{M}_{\text{d+2 x d+1}}$, so X is matrix which has more rows than columns.

$$X_{d+2,d+1} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_{d+2}^T \end{pmatrix}$$

It means that in the last row exist a linear dependency:

$$x_{d+2} = \sum_{i=0}^{d+1} \alpha_i x_i \quad \exists! \alpha_i \neq 0$$

That results in a dichotomy which can not be shattered, so $d+2$ is a breakpoint.

$$\Longrightarrow \quad d_{\text{vc}} \leqslant d+1$$

1.2. Interpreting the Generalization Bound

Exercise 2.5

Suppose we have a simple learning model whose growth function is $m_h(N) = N + 1$, hence $d_{\text{vc}} = 1$. Use the VC bound to estimate the probability that E_{out} will be within 0.1 of E_{in} given 100 training examples.

Solution:

We have to use the VC bound that is:

$$E_{\text{out}(g)} \leqslant E_{\text{in}}(g) + \sqrt{\frac{8}{N} ln(\frac{4m_H(2N)}{\delta})}$$

$$E_{\text{out}(g)} - E_{\text{in}}(g) \leqslant \sqrt{\frac{8}{100} ln(\frac{4m_H(200)}{.1})}$$

$$E_{\text{out}(g)} - E_{\text{in}}(g) \leqslant \sqrt{\frac{2}{25} ln(\frac{804}{.1})}$$

$$E_{\text{out}(g)} - E_{\text{in}}(g) \leqslant 0.84816$$

2. Problems

2.1) In Equation (2.1), set $\delta = 0.03$ and let

$$\epsilon(M, N, \delta) = \sqrt{\frac{1}{2N} ln(\frac{2M}{\delta})}$$

a) For M = 1, how many examples do we need to make $\epsilon \leqslant 0.05$?
b) For M = 100, how many examples do we need to make $\epsilon \leqslant 0.05$?
c) For M = 10,000, how many examples do we need to make $\epsilon \leqslant 0.05$?

Solution:
First we have to move N:

$$N \geqslant \frac{1}{2N\epsilon^2} ln(\frac{2M}{\delta})$$

a) $N \geqslant \frac{1}{2 \cdot 0.05^2} ln(\frac{2}{0.03}) = 840$
b) $N \geqslant \frac{1}{2 \cdot 0.05^2} ln(\frac{200}{0.03}) = 1761$
c) $N \geqslant \frac{1}{2 \cdot 0.05^2} ln(\frac{20000}{0.03}) = 2682$

2.2) Show that for the learning model of positive rectangles (aligned horizontally or vertically), $m_H(4) = 24$ and $m_H(5) < 25$. Hence, give a bound for $m_H(N)$.
Solution:
Using the theorem:

$$\text{If } m_H(k) < 2^k \text{ for some value } k \quad \Longrightarrow \quad m_H(N) \leqslant \sum_{i=0}^{k-1} \binom{N}{i} \quad \forall N$$

$$k = d_{\text{vc}} + 1$$

$$k = 5 \quad \Longrightarrow d_{\text{vc}} = 4$$

Finally:

$$m_H(N) \leqslant N^4 + 1$$

2,3) Compute the maximum number of dichotomies, $m_H(N)$, for these learning models, and consequently compute d_{vc} , the VC dimension.

a) Positive or negative ray: H contains the functions which are +1 on [a, ∞) (for some a) together with those that are +1 on (-∞, a] (for some a).
b) Positive or negative interval: H contains the functions which are + 1 on an interval [a, b] and -1 elsewhere or -1 on an interval [a, b] and +1 elsewhere.
c) Two concentric spheres in $\mathbb{R}^d$: H contains the functions which are +1 for $a \leqslant \sqrt{x_1^2, ..., x_d^2} \leqslant b$

Solution:

a) We already know the growth function for positive rays $m_H(N) = N + 1$.
Let us count the dichotomies in $(-\infty, a]$ (negative rays) which it results in $N + 1$, now let us count the dichotomies in $[a, \infty)$ (positive rays) it and results the same number of dichotomies, but we have counted 2 times the same dichotomy, then:

$$m_H(N) = 2N$$

In order to obtain the d_{vc} we must find out the maximum value for N such that $2N \leqslant N^2$, and we get $d_{\text{vc}} = 2$.

b) The growth function for positive intervals is: $1 + \frac{N}{2} + \frac{N^2}{2}$.
Now let us count the negative intervals which results that for each N there are some dichotomies identical in positive intervals, the only differents are the ones in the middle of the interval taken one for one, it results: $N - 2$ for negative intervals. Finally:

$$m_H(N) = N\frac{3}{2} + \frac{N^2}{2} - 1$$

As in a) we must find out the maximum value for N such that:

$$N\frac{3}{2} + \frac{N^2}{2} - 1 \leqslant N^2$$
$$\frac{9}{2} + \frac{9}{2} - 1 \leqslant 3^2$$
$$8 \leqslant 8 \quad \Longrightarrow \quad d_{\text{vc}} = 3$$

c) Firstly, we hate to make a non-linear transformation in order to have an interval such as $[0, +\infty)$.

$$\Phi(x_1, ..., x_d) \longrightarrow r = \sqrt{x_1^2 + ... + x_d^2}$$

Now it is similar to the positive intervals, then $m_H(N) = 1 + \frac{N}{2} + \frac{N^2}{2}$ because no exists negative intervals here.

$$N\frac{1}{2} + \frac{N^2}{2} + 1 \leqslant N^2$$
$$\frac{2}{2} + \frac{4}{2} + 1 \leqslant 2^2$$
$$4 \leqslant 4 \quad \Longrightarrow \quad d_{\text{vc}} = 2$$

2.4) Show that $B(N, K) = \sum_{i=0}^{k-1} \binom{N}{i}$ by showing the other direction to Lemma 2.3, namely that:

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

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.

Solution:

Firstly, let us construct a set of dichotomies for N points, then it possibly will contains 2^N dichotomies, but we just are looking for those with at most k-1 called (-1). Thus we obtain:

$$\text{Dichotomies that have just j - (-1), are} \binom{N}{j} \qquad j = 0, 1, 2, ..., k-1$$

$$= \sum_{i=0}^{k-1} \binom{N}{i}$$

But in this set not exists the dichotomy which can shatter k elements, for that we must include the k-dichotomy. This results in:

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

$$B(N, K) \leqslant \sum_{i=0}^{k-1} \binom{N}{i} \wedge B(N, K) \geqslant \sum_{i=0}^{k-1} \binom{N}{i} \Longrightarrow B(N, K) = \sum_{i=0}^{k-1} \binom{N}{i}$$

2.5) Prove by induction $\sum_{i=0}^{D} \binom{N}{i} \leqslant N^D + 1$ hence $m_H(N) \leqslant N^D + 1$.
Solution:

a) First prove for D = 1

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

b) Our induction hypothesis is D = k, we assume is correct here.

$$\sum_{i=0}^{k} \binom{N}{i} \leqslant N^k + 1$$

c) Now let us demonstrate for D = k + 1

- We start with the demonstration, and we replace for the base case.

$$\sum_{i=0}^{k+1} \binom{N}{i} = \sum_{i=0}^{k} \binom{N}{i} + \binom{N}{k+1} \leqslant N^k + 1 + \binom{N}{k+1}$$

- We must prove that:

$$\frac{N!}{(N-k-1)!} \leqslant N^{k+1}$$
$$\frac{1}{N^{k+1}} \cdot \frac{N!}{(N-k-1)!} \leqslant 1$$
$$\frac{1}{N^{k+1}} \cdot \prod_{i=0}^{k} (N-i) \leqslant 1$$

- Now we can rewrite as follow:

$$\sum_{i=0}^{k+1} \binom{N}{i} \leqslant N^k + 1 + \frac{N!}{(N-k-1)!(k+1)!}$$
$$\leqslant N^k + 1 + \frac{N^{k+1}}{(k+1)!}$$

- We know that $k \in \mathcal{R}$, so (k+1)! $\geqslant 2$:

$$\sum_{i=0}^{k+1} \binom{N}{i} \leqslant N^k + 1 + \frac{N^{k+1}}{2}$$

In addtion:

$$N \geqslant k+1 \rightarrow N \geqslant 2$$

$$\frac{1}{N} < \frac{1}{2} \iff \frac{N^k}{N^{k+1}} < \frac{1}{2} \iff N^k < \frac{N^{k+1}}{2}$$

Finally:

$$\sum_{i=0}^{k+1} \binom{N}{i} \leqslant N^k + 1 + \frac{N^{k+1}}{2}$$

$$\leqslant \frac{N^{k+1}}{2} + 1 + \frac{N^{k+1}}{2} = N^{k+1} + 1$$

2.10) Show that $m_h(2N) \leqslant m_h(N)^2$, and hence obtain a generalization bound which only involves $m_h(N)$.
Solution:
By definition:

$$m_H(N) = max_{x_1,\ldots,x_N} |\{h(x_1), \ldots, h(x_N) : h \in H\}| \leqslant 2^N$$

$$\implies \quad m_H(N) \leqslant 2^N$$

$$m_H(2N) = m_H(N+N) \leqslant 2^{N+N}$$

$$m_H(2N) \leqslant 2^N \cdot 2^N$$

$$m_H(2N) \leqslant m_H(N) \cdot m_H(N)$$

$$m_H(2N) \leqslant m_H(N)^2$$

Now, let us introduce this result in the generalization bound:

$$E_{\text{out}(g)} \leqslant E_{\text{in}}(g) + \sqrt{\frac{8}{N} ln(\frac{4m_H(2N)}{\delta})} \leqslant \sqrt{\frac{8}{N} ln(\frac{4m_H(N)^2}{\delta})}$$

2.13) *a*) Let $H = \{h_1, h_2, ..., h_M\}$ with some finite M. Prove that $d_{\text{vc}}(H) \leqslant log_2(M)$.

b) For any sets $H_1, H_2, ..., H_k$ with finite VC dimensions $d_{\text{vc}}(H_k)$ derive and prove the tightest upper and lower bound that you can get on $d_{\text{vc}} = (\cap_{k=1}^{K} H_k)$

c) For any sets $H_1, H_2, ..., H_k$ with finite VC dimensions $d_{\text{vc}}(H_k)$ derive and prove the tightest upper and lower bound that you can get on $d_{\text{vc}} = (\cup_{k=1}^{K} H_k)$

Solution:

a) Firstly, by definition:

$$d_{\text{vc}}(H) = 2^N \quad \Longrightarrow \quad \text{N is the largest value such that } m_h(N) = 2^N$$

$$\Longrightarrow \quad 2^N = m_H(N) = max_{x_1,...,x_N} |\{h(x_1), ..., h(x_N) : h \in H\}|$$

$$\leqslant |H| = M$$

$$d_{\text{vc}}(H) = 2^N \quad \Longrightarrow \quad N \leqslant log_2(M)$$

b) If we have the minimum set $H = \{h\}$ we got that the $d_{\text{vc}}(H) = 0$, because $m_H(N) = 1$.

$$\text{The lower bound} = 0 \leqslant \cap_{k=1}^{K} H_k$$

For the upper bound, as is an intersection we are interested in:

$$min_{1,...,k} d_{\text{vc}}(H_k)$$

In order to demonstrate that, we suppose:

$$\cap_{k=1}^{K} H_k > min_{1,...,k} d_{\text{vc}}(H_k) = c$$

It means that the LHS can shatter c+1 points:

$$\cap_{k=1}^{K} H_k(x_1, ..., x_{c+1}) \subset \{h(x_1), ..., h(x_{c+1}) : h \in H\} = H_k(x_1, ..., x_{c+1})$$

It results:

$$2^{c+1} \leqslant |\{h(x_1), ..., h(x_{c+1}) : h \in H\}| \leqslant 2^{c+1}$$

$$\Longrightarrow \quad |\{h(x_1), ..., h(x_{c+1}) : h \in H\}| = 2^{c+1} \quad \text{for k = 1, 2, ..., K}$$

Then, any H_k can shatter c+1 points, let set $min_{1 \leqslant k \leqslant K} d_{\text{vc}}(H_k) = d_{\text{vc}}(H_{k_0})$

$$c = d_{\text{vc}}(H_{k_0}) \geqslant c + 1$$

$$0 \leqslant \cap_{k=1}^{K} H_k \leqslant min_{1 \leqslant k \leqslant K} d_{\text{vc}}(H_k)$$

c) Let $d_{\text{vc}}(H_{k_k}) = d_k$ for k = 1, ..., K. It means that H_{k_k} can shatter any d_k points, so:

$$\{+1, -2\}^{d_k} = \{h(x_1), ..., h(x_{c+1}) : h \in H\} \subset \{h(x_1), ..., h(x_{c+1}) : h \in \cup_{k=1}^{K} H_k\}$$

$$2^{d_k} \leqslant |\{h(x_1), ..., h(x_{c+1}) : h \in \cup_{k=1}^{K} H_k\}| \leqslant 2^{d_k}$$

$$2^{d_k} \leqslant |\{h(x_1), ..., h(x_{c+1}) : h \in \cup_{k=1}^{K} H_k\}|$$

We obtain:

$$max_{1 \leqslant k \leqslant K} d_k \leqslant d_{\text{vc}}(\cup_{k=1}^{K} H_k)$$

Secondly, we as it is a union we want to find out the sum for $k = 1, 2, ..., K$.

$$d_{\text{vc}}(\cup_{k=1}^{K} H_k) \leqslant K - 1 + \sum_{i=1}^{K} d_{\text{vc}}(H_k)$$

As is an union let us prove for k = 2:

$$\begin{aligned} d_{\text{vc}}(\cup_{k=1}^{2} H_k) &\leqslant m_{H_1}(N) + m_{H_2}(N) \\ &\leqslant \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=0}^{d_2} \binom{N}{i} \\ &\leqslant \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=0}^{d_2} \binom{N}{N-1} \\ &\leqslant \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=N-d_2}^{N} \binom{N}{i} \\ < \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=N-d_2}^{N} \binom{N}{i} + \sum_{i=d_1+1}^{N-d_2-1} \binom{N}{i} &= \sum_{i=1}^{N} \binom{N}{i} = 2^N \end{aligned}$$

If prove for K - 1, we will prove for K. We have that:

$$\begin{aligned} d_{\text{vc}}(\cup_{k=1}^{K} H_k) &= d_{\text{vc}}(\cup_{k=1}^{K-1} H_k \cup H_k) \\ &= 1 + d_{\text{vc}}(\cup_{k=1}^{K-1} H_k) + d_{\text{vc}}(H_k) \\ &= 1 + (K-2) + \sum_{k=1}^{K-1} d_{\text{vc}}(H_k) + d_{\text{vc}}(H_k) \\ &= K - 1 + \sum_{k=1}^{K} d_{\text{vc}}(H_k) \end{aligned}$$