
Problem Set 4 for Machine Learning 15 Fall

Jingyuan Liu
AndrewId: jingyual
jingyual@andrew.cmu.edu

1 VC dimension

1.1 Show the VC dimension of linear classifier

To prove that the linear classifier h with x in R^n has the VC dimension of $n + 1$, we need to prove that $VCdim(h_n) \geq n + 1$, and then prove that $VCdim(h_n) \leq n + 1$.

(a). Prove $VCdim(h_n) \geq n + 1$

First of all, we could use the Mathematical Induction to prove that $VCdim(H) \geq n + 1$:

For $n = 1$, it is easy to get $VCdim(h_1) \geq 2$

For $n = 2$, we could also get $VCdim(h_2) \geq 3$, which could be proved using following figures:

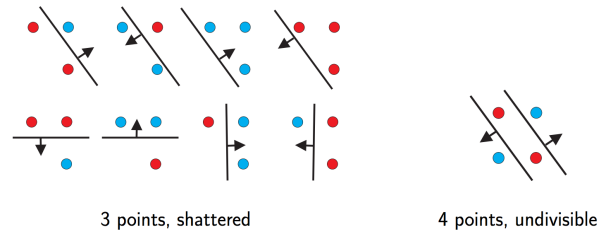


Figure 1: The representation of $VCdim(h_2)$

For $n = i$, we assume that the $VCdim(h_i) \geq i + 1$. We could form a matrix of $(i+1) \times (i+1)$, with the rank of $i+1$, since the VCdim is $i+1$.

For $n = i+1$, we could form a matrix of $(i+2) \times (i+2)$, since the variable x have one more independent dimension. Therefore, $VCdim(h_{i+1}) \geq VCdim(h_i) + 1 \geq i + 2$.

Therefore, we have $VCdim(h_n) \geq n + 1$. We could use a more straight forward detailed example to show that. Say we have n dimensions of the data, then we could construct a dataset:

1th: $(1, 0, 0, \dots, 0)$
2th: $(0, 1, 0, \dots, 0)$
 i th: $(0, 0, \dots, 1, \dots, 0)$
 n th: $(0, 0, 0, \dots, n)$
 $n + 1$ th: $(0, 0, \dots, 0)$

For these dataset, we could see that for the i th data point, the i th dimension is 1, otherwise 0. And for the $n+1$ th data, all dimension are 0. In this case, we could see that the linear classifier could shatter this dataset.

(b). Prove $VCdim(h_n) \leq n + 1$

Then we should prove that $VCdim(h_n) \leq n + 1$:

Suppose we could shatter $n+2$ points, then using the convex combinations of points in C , we could separate points into $S1$ and $S2$, that:

$$conv(s1) \cap conv(s2) \neq \emptyset \quad (1)$$

However, if $VCdim(h_n) = n + 2$, then we could split points to half space contains $s1$ and the complement of the half space contains $s2$. This implies that the both half space contains the convex hull of $s1$ and the complement of the half space the contains the convex hull of $s2$. Thus, we get:

$$conv(s1) \cap conv(s2) = \emptyset \quad (2)$$

These two observations are contradictory to each other. So we know that for the $n+2$ points here, we could not use the linear classifier to shatter all of them.

In conclusion, we prove that $VCdim(h_n) \geq n + 1$ and $VCdim(h_n) \leq n + 1$ for the linear classifier, then we have $VCdim(h_n) = n + 1$

1.2 Show the VC dimension of axis-aligned boxes

Similarly, for this case, to prove that the axis-aligned boxes classifier h with x in R^n has the VC dimension of $2n$, we need to prove that $VCdim(h_n) \geq 2n$, and then prove that $VCdim(h_n) \leq 2n$.

(a). Prove $VCdim(h_n) \geq 2n$

First of all, we need to prove that for any case, $VCdim(h_n) \geq 2n$, which means if the x has n dimensions, we can use the axis-aligned boxes classifiers to shatter $2n$ points.

Suppose we have n dimensions, then we can map all the x to a dimension i . In the dimension i , we can have x_{max}^i and x_{min}^i . Therefore, in this dimension i , we could always shatter at least 2 points via a split between the x_{max}^i and x_{min}^i .

Considering we have n dimensions, and the mapping of each dimension of the data is independent to the mapping of other dimension of the data. Therefore, we could at least shatter $2*n$ points via the axis-aligned boxes. Therefore, we have $VCdim(h_n) \geq 2n$.

(b). Prove $VCdim(h_n) \leq 2n$

Then we need to prove that $VCdim(h_n) \leq 2n$. Suppose we have $2n + 1$ points, we could always find the $2n$ "boundries", which is the min value and max value for each dimension, and form an "area". Then the $2n+1$ th points will be guaranteed to appear within the selected "area".

For any $2n+1$ points, we could always transfer to the above mentioned scenario. In this scenario, we could not classify the $2n+1$ th point, because it is contained within the "area", and no rules are guaranteed to rightly classify it.

Therefore, by mapping the data to each dimension and form an "area", we could prove that $VCdim(h_n) \leq 2n$.

In conclusion, we prove that $VCdim(h_n) \geq 2n$ and $VCdim(h_n) \leq 2n$ for the axis-aligned boxes, then we have $VCdim(h_n) = 2n$

2 AdaBoost

2.1 Justify the Update Rule

Based on the definition, if we change the distribution, the previous learned hypothesis h_t would be a “random guess” to the new distribution of data, therefore:

$$\epsilon_t = \text{err}_{D_t}(h_t) = \Pr_{x \sim D_t}(y \neq h_t(x)) = \frac{1}{2} \quad (3)$$

The error rate is $\frac{1}{2}$ for a random guess with margin $\gamma_t = 0$. Then we would have, for all i :

$$\alpha_t = \frac{1}{2} \log\left(\frac{1 - \epsilon_t}{\epsilon_t}\right) = 0 \quad (4)$$

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} e^{-\alpha_t y_i h_t(x_i)} \quad (5)$$

$$= \frac{D_t(i)}{Z_t} \quad (6)$$

Therefore, we could see that the $D_{t+1} = D_t$, so:

$$\text{err}_{D_{t+1}}(h_t) = \Pr_{x \sim D_t}(y \neq h_t(x)) \quad (7)$$

$$= \sum_{y_i \neq h_t(x_i)} D_{t+1}(i) \quad (8)$$

$$= \sum_{y_i \neq h_t(x_i)} D_t(i) \quad (9)$$

$$= \text{err}_{D_t} = \frac{1}{2} \quad (10)$$

2.2 Show the $D_{t+1}(i)$

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} e^{-\alpha_t y_i h_t(x_i)} \quad (11)$$

$$= \frac{e^{-y_i}}{Z_t} e^{\alpha_t h_t(x_i)} D_t(i) \quad (12)$$

$$= \frac{e^{-y_i}}{Z_t} e^{\alpha_t h_t(x_i)} \cdot \frac{e^{-y_i}}{Z_{t-1}} e^{\alpha_{t-1} h_{t-1}(x_i)} D_{t-2}(i) \quad (13)$$

$$= \frac{e^{-y_i}}{Z_t} e^{\alpha_t h_t(x_i)} \dots \frac{1}{m} e^{\alpha_1 h_1(x_i)} \quad (14)$$

$$= (m \prod_t Z_t)^{-1} e^{-y_i \sum_t \alpha_t h_t(x)} \quad (15)$$

2.3 Show the error upperbound

$$err_S(H) = Pr(y \neq H(x)) = \frac{1}{m} \sum_{H(i) \neq y_i} D_S(i) \quad (16)$$

When $H(i) \neq y_i$, we have $y_i f(x_i) < 0$, and $exp(-y_i f(x_i)) > 0$, therefore:

$$err_S(H) < \frac{1}{m} \sum_i exp(-y_i f(x_i)) \quad (17)$$

For the right term, according to the definition of Z and W , we have:

$$w_{t,i} exp(-\alpha_t y_i h_t(x_i)) = Z_m w_{t+1,i} \quad (18)$$

$$\frac{1}{m} \sum_i exp(-y_i f(x_i)) = \frac{1}{m} \sum_i exp(-\sum_t \alpha_t y_i h_t(x_i)) \quad (19)$$

$$= \sum_{1,i} w_{1,i} \prod_t exp(-\alpha_t y_i h_t(x_i)) \quad (20)$$

$$= Z_1 \sum_{2,i} w_{2,i} \prod_{t=2}^T exp(-\alpha_t y_i h_t(x_i)) \quad (21)$$

$$= Z_1 Z_2 \sum_{3,i} w_{3,i} \prod_{t=3}^T exp(-\alpha_t y_i h_t(x_i)) \quad (22)$$

$$= Z_1 Z_2 \cdots \sum_i w_{T,i} exp(\alpha_T y_i H(x_i)) \quad (23)$$

$$= \prod_t Z_t \quad (24)$$

Therefore, we have:

$$err_S(H) < \frac{1}{m} \sum_i exp(-y_i f(x_i)) = \prod_t Z_t \quad (25)$$

2.4 Show the $\prod_t Z_t$ upperbound

$$Z_t = \sum_i exp(-\alpha_t y_i h_t(x_i)) \quad (26)$$

$$= \sum_{h_t(x_i)=y_i} w_i e^{-\alpha_t} + \sum_{h_t(x_i) \neq y_i} w_i e^{\alpha_t} \quad (27)$$

$$= (1 - \epsilon_t) e^{-\alpha_t} + \epsilon_t e^{\alpha_t} \quad (28)$$

$$= 2\sqrt{\epsilon_t(1 - \epsilon_t)} = \sqrt{1 - 4\gamma_t^2} \quad (29)$$

Therefore, we could use Taylor expansion series to get $\sqrt{1 - 4\gamma_t^2} \leq exp(-2\gamma_t^2)$, so we have:

$$\prod_t Z_t \leq e^{-2 \sum_t \gamma_t^2} \quad (30)$$

2.5 Express big-O notation of T

From above questions, we could get:

$$\epsilon \leq e^{-2 \sum_t \gamma_t^2} \quad (31)$$

With the smallest margin of γ , we could transfer to

$$\epsilon \leq e^{-2T\gamma^2} \quad (32)$$

$$T = O\left(\frac{-\log(\epsilon)}{\gamma^2}\right) \quad (33)$$

2.6 implementation

For each h, we choose the classifier with smallest error in each iteration. The result is:

t	ϵ_t	α_t	$D_t(1)$	$D_t(2)$	$D_t(3)$	$D_t(4)$	$D_t(5)$	$D_t(6)$	$D_t(7)$	$D_t(8)$	$D_t(9)$	$err_S(H)$
1	0.222	0.626	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.222
2	0.143	0.896	0.071	0.071	0.071	0.071	0.071	0.071	0.071	0.25	0.25	0.222
3	0.125	0.973	0.042	0.042	0.042	0.042	0.25	0.25	0.042	0.146	0.146	0

Table 1: AdaBoost results

The classification rule: first, if $x > 2.5$, pos; second, if $x > 3.5$, pos; third, if $x < 4.5$, pos.

3 Gaussian Mixture Model

3.1 Show the expectation

$$E(x) = \int p(x)dx = \sum_k p(x_k)x_k \quad (34)$$

$$= \sum_k \pi_k E_k(x) \quad (35)$$

We know that for each k, the implicit distribution of x is a Gaussian distribution, and $E_k(x) = \mu_k$, μ_k is the mean of the kth gaussian distribution, so we have:

$$E(x) = \sum_k \pi_k \mu_k \quad (36)$$

3.2 Show the covariance

$$cov(x) = E(xx^T) - E(x)E(x)^T \quad (37)$$

$$= \sum_k \pi_k E_k(xx^T) - E(x)E(x)^T \quad (38)$$

$$= \sum_k \pi_k (\Sigma_k + \mu_k \mu_k^T) - E(X)E(x)^T \quad (39)$$

4 K-Means

5 Collaboration

I dicussed with Zheng Chen with problem 1.2 on how to calculate the parameters, and he also informed me about 2.2.1(b).