Learning Theory-Infinite Hypothesis sets

DINA ELREEDY

Generalization bound (M hypotheses)

- Hoeffdeing's Inequality of the selected g hypothesis out of M hypotheses using N training points.

$$\boxed{\Pr[|Ein(g) - Eout(g)| > \varepsilon] \le 2Me^{-2\varepsilon^2 N}}$$

Then, with probability at least 1-δ:

$$Eout(g) \le Ein(g) + \varepsilon$$

$$\boxed{Eout(g) \le Ein(g) + \sqrt{\frac{1}{2N}\ln(\frac{2M}{\delta})}}$$

where $\delta = 2Me^{-2\varepsilon^2 N}$

Generalization bound

- What if we have an infinite set of hypotheses?
- Example: the perceptron learning algorithm.
- Can we derive a generalization bound?

Review: Union bound of M hypotheses

Remember that M comes from applying a union bound.

Let the bad events B_i be:

$\Pr[B_i] = \mathrm{P}[|Ein(h_i) - Eout(h_i)\,| > \varepsilon]$

The union bound for M hypotheses:

$\mathrm{P}[B_1 \; or \; B_2 or \; \ldots or B_M] \leq \mathrm{P}[B_1] + \mathrm{P}[B_2] + \ldots \mathrm{P}[B_M] \leq \sum_{i=1}^{M} \Pr[B_i]$

M terms

Since P(|Ein(h_i)-Eout(h_i)|>ε)≤$2e^{-2\varepsilon^2 N}$

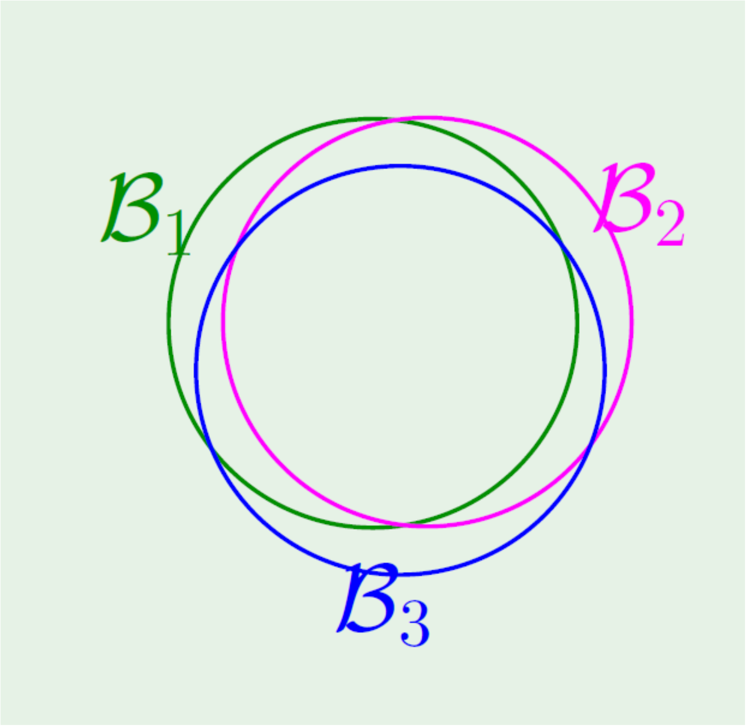
Then, P(|Ein(g)-Eout(g)|>ε)≤$2Me^{-2\varepsilon^2 N}$

Union bound of M hypotheses

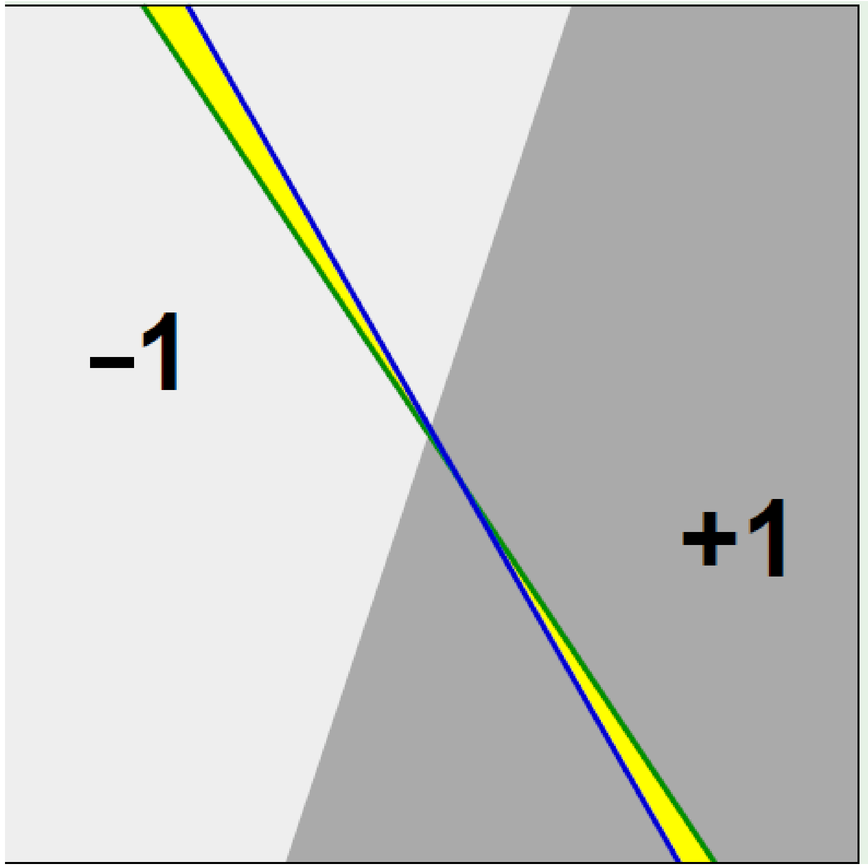
- The good news is that the B events are overlapping!!
- Since many hypotheses are similar, so their corresponding

B events $\mathrm{P}|Ein(h_i) - Eout(h_i)\ | > \varepsilon]$ are overlapping.

- Thus, the derived union bound was loose.
- We shall derive a tighter bound for infinite hypotheses.



Example- linear classifiers



Dichotomies

- For a binary classification problem:

A hypothesis h: X -> {+1,-1}.

- Apply a hypothesis h to a finite sample of input points not the whole input space.
- Assume we have N sample points $\{x_1, x_2, \ldots x_N\}$
- Apply h $\in$H to the N points, we get a dichotomy which is an N-tuple of $(h(x_1), h(x_2), \ldots h(x_N))$.

Dichotomies (cont.)

- For a hypothesis set H, the dichotomies generated by H are defined as:

$$H(x_1, x_{2,\ \dots}\ x_N) = \{(h(x_1), h(x_2), \dots h(x_N)) | h \in H\}$$

- A hypothesis h: X -> {+1,-1}.
- A dichotomy: $\{x_1, x_2, \dots x_N\}$->{+1,-1}
- Number of hypotheses |H| can be infinite.
- Maximum number of dichotomies $|H(x_1, x_{2,}\ \dots x_N)|$ is 2^N.
- The greater the number of dichotomies $|H(x_1, x_{2,}\ \dots x_N)|$ is the more diverse and powerful the hypothesis set H.

Growth Function

- The growth function $m_H(N)$ for a hypothesis set H is defined as:
  - The largest number of dichotomies that can be generated by H on any N points.

$$m_H(N) = \max_{x_1, x_2, \ldots, x_N \in X} |H(x_1, x_2, \ldots x_N)|$$

bt2olak akbur 3dd mn el t2semat elly t2dr t3mlha, dichatres -> t2sema.

usually we don't consider special cases.

- Since the maximum number of dichotomies $|H(x_1, x_2, \ldots x_N)|$ is 2^N, then:

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

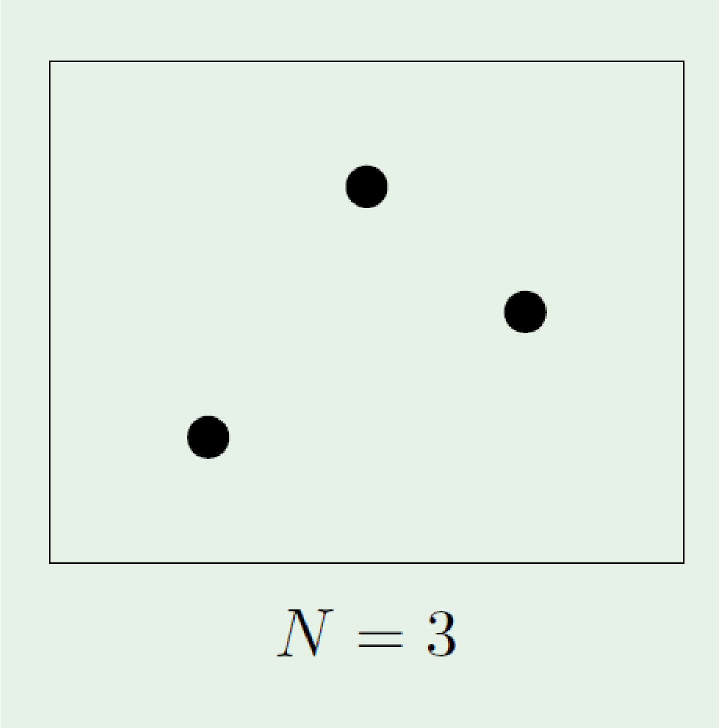
Growth Function (cont.)

- Given H, if H can generate all possible dichotomies on data points $(x_1, x_{2, \dots} x_N)$ such that $|H(x_1, x_{2, \dots} x_N)| = 2^N$ then H **shatters** $(x_1, x_{2, \dots} x_N)$.

Growth function-Perceptron

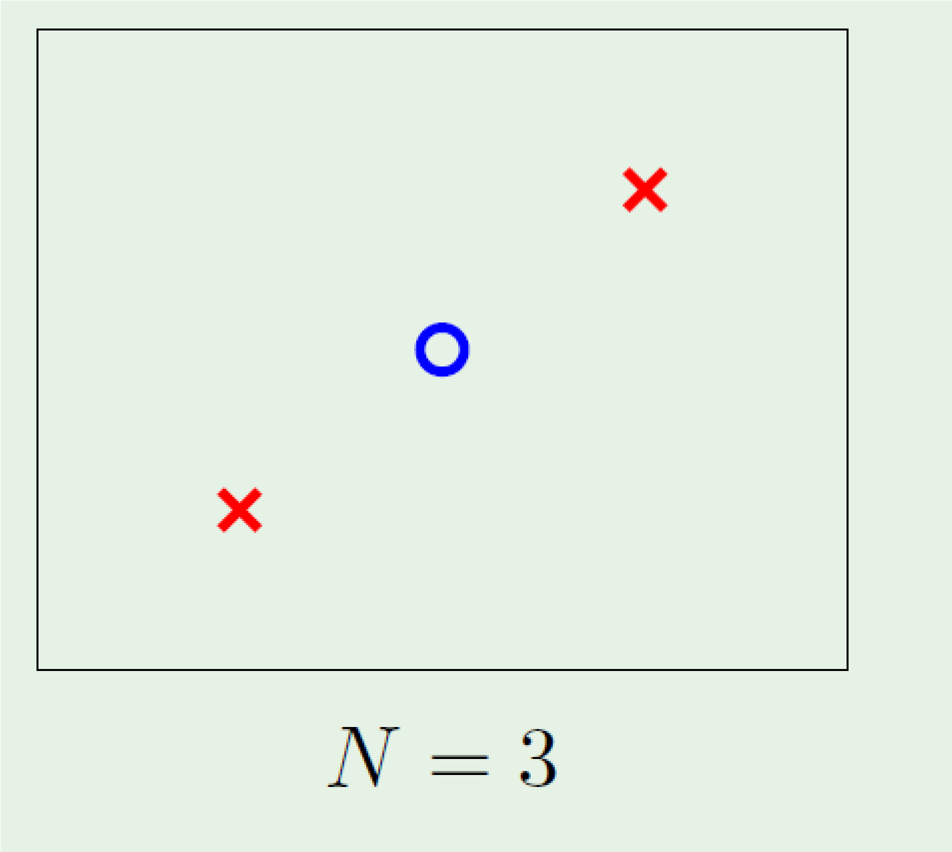
N=3

$m_{H(3)} = 8$



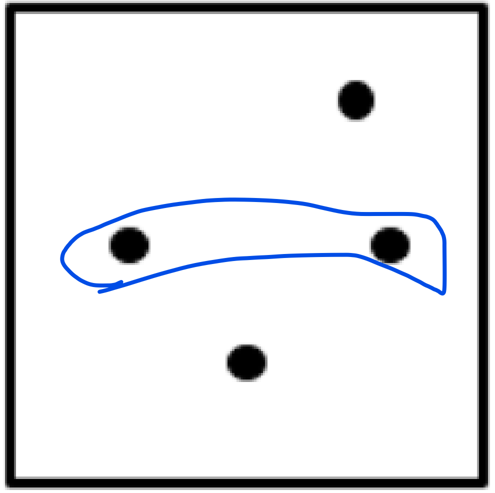
Growth function-Perceptron

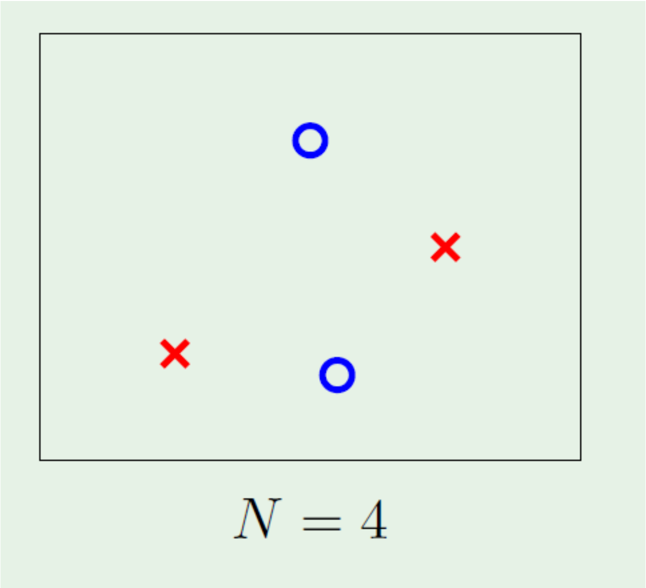
N=3 (co-linear points)



Growth function-Perceptron

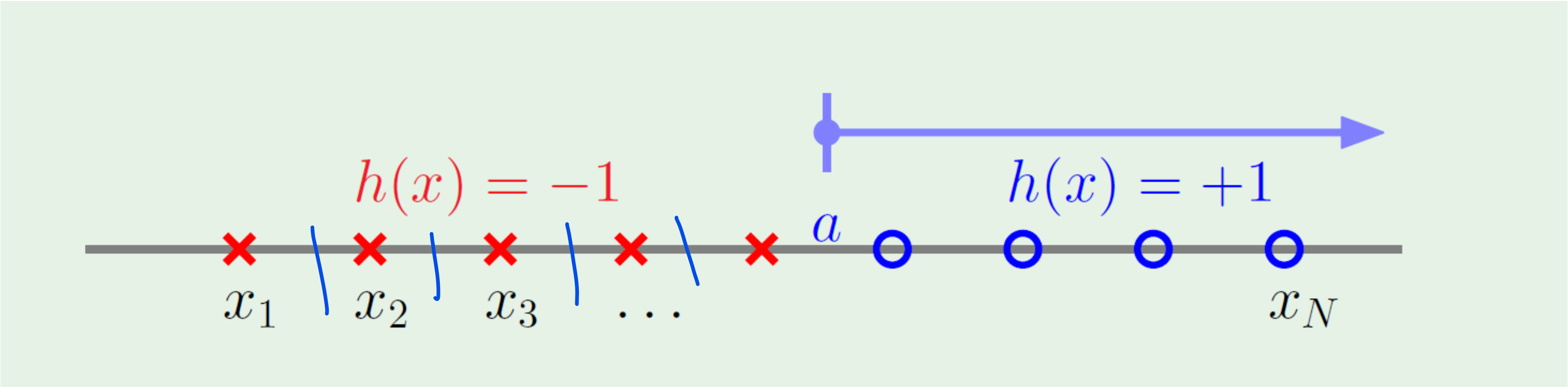
$m_{H}(4) = 14$ not 2^4 -> two cases are not possible





Growth function- Positive rays

- H consists of all hypotheses h of the form: h(x)=sign(x-a)

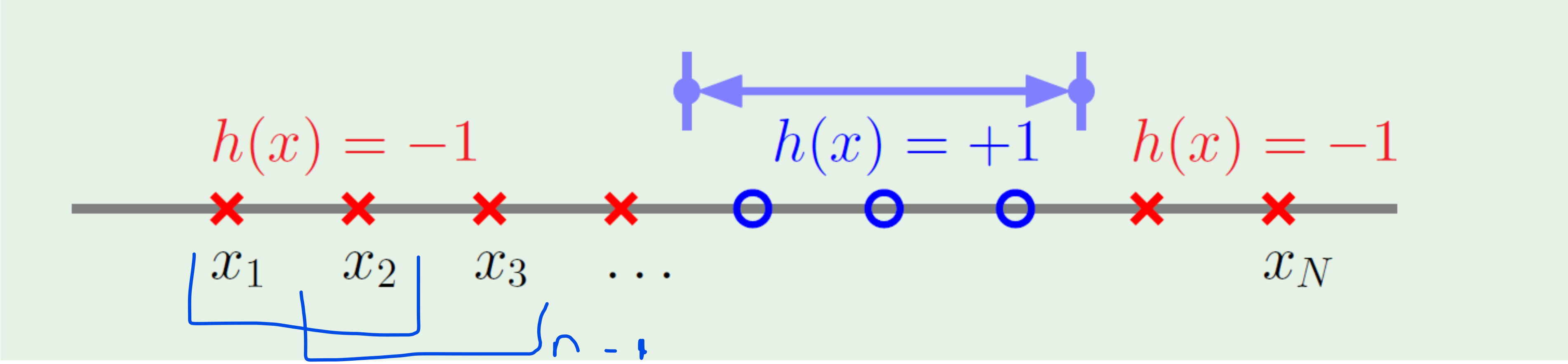


- $m_{H(_N)} = \text{N} + 1$

number of separator in between = n -1 , w wa7ed 3la el ymen khales, w wa7ed 3la el shmal 5ales.

Growth function- Positive intervals

- H consists of all hypotheses in one dimension that returns +1 within some interval and -1 otherwise.
- Each hypothesis is specified by the two end values of the interval (a,b).



- $m_{H(N)} = \binom{N+1}{2} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$

I was too close!! :D

Convex sets

A set is convex if the line between any two points in the set entirely lies within the set.

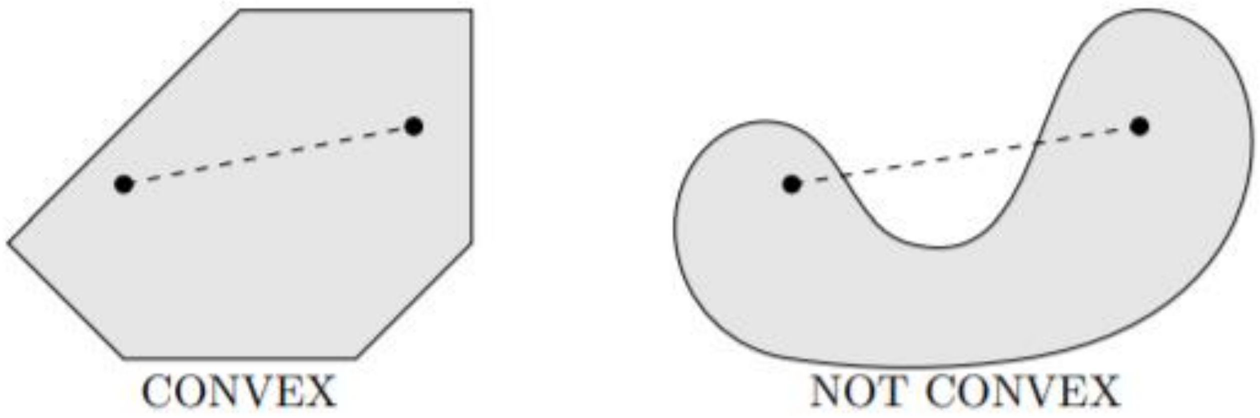
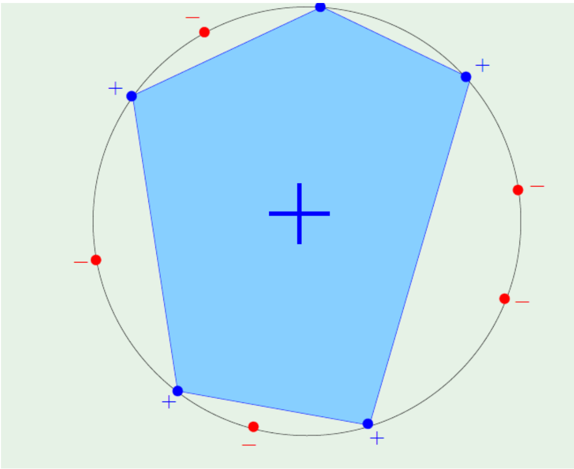


Image source:
https://faculty.math.illinois.edu/~mlavrov/docs/484-spring-2019/ch2lec1.pdf

Growth function- Convex sets

- H consists of all the hypotheses in two dimensions that are positive inside a convex set and negative elsewhere.
- Choose N points on the circumference of a circle.
- Connect positive points with a polygon.

el circle de akbur convex shape, fa bnftrdo akbur shape, w b3den ay convex shape 3ndy, ana a2dr akwno gowa el cirlcle de, fa ayan kan shaklo a2dr a7oto gowa el circle.
w el convex shape bnftrdo eno el class el positive w el ba2y el 3la el circle -ve, fa keda el possible cases 3ndy hyb2o 2^N



- $m_{H}(N) = 2^{N}$
- H shatters these N points.

Breakpoint

el la7za elly b3dha, mbn2drsh n2sm el data points le el classes bkol el ashkal el momkena.
zy el case bta3t el 4 points fl linear separator keda.
fa keda el 4 hya el break point lel linear separator.

- If no data set of size k can be shattered by H, then k is said to be a breakpoint for H.
- If k is a breakpoint then:

$$m_{H(k)} < 2^k$$

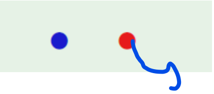
Breakpoint Examples

- For 2D perceptron:

  k=4

- For the 1D positive rays:

k=2

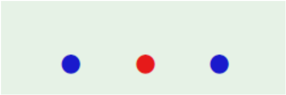


m2drsh akhly el awal -ve wl odam +ve

-ve

- For the 1D positive interval:

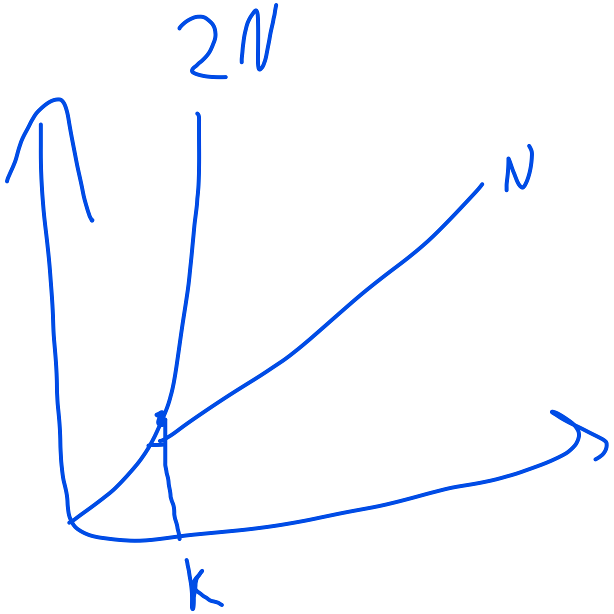
k=3



mn2drsh nkhly el gowah -ve w el bara +ve.

- For the convex sets:

  k=∞



el VC hya akbur no2ta t2dr 3ndha te3ml shuttering.

shuttering --> enk tgeb 2^N.

The VC Dimension

- The VC dimension of a hypothesis set H denoted by $d_{vc}(H)$ is the largest value of N for which $m_{H(N)} = 2^N$.

in case of linear separator -> dvc(H) = 3

- $d_{vc(H)} = k - 1$

- If there is no break point for the hypothesis set H, $m_{H(N)} = 2^N \;\; \forall N$, then $d_{vc}(H)$=∞.

VC dimension-examples

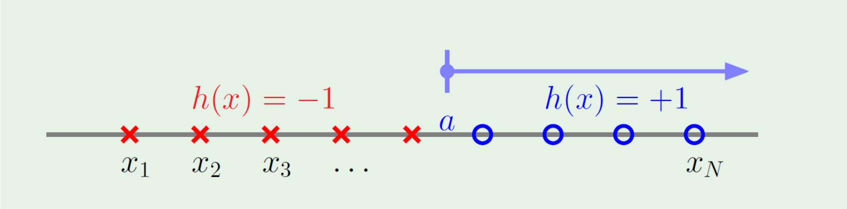
- The VC dimension for 2D perceptron is $d_{vc(H)}$ =3.

number of degree of freedom elly a2dr a8yr fehom 34an aghyr el hypothesis bt3ty.

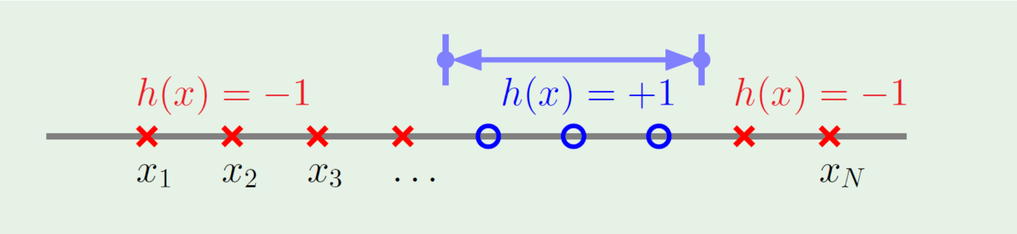
- The VC dimension for d-dimension perceptron is $d_{vc(H)}$ =d+1.

VC dimension-examples (cont.)

- The VC dimension for positive rays $d_{vc(H)}$ =1.



- The VC dimension for positive intervals $d_{vc(H)} = 2$.



- VC dimension can be interpreted as the effective number of parameters (degrees of freedom).

Theorem

- If H has a break point, then $m_{H(N)}$ is bounded by a polynomial in N.

k is the breakpoint

$$\boxed{m_{H(N)} \leq \sum_{i=0}^{k-1} \binom{N}{i}}$$

Nc0 + Nc1 + Nc2 .... NcK-1

$$\boxed{m_{H(N)} \leq \sum_{i=0}^{d_{vc}(H)} \binom{N}{i}}$$

VC-Generalization bound for Infinite hypothesis set

The VC inequality using the growth function instead of M:

$$\boxed{\Pr[|Ein(g) - Eout(g)| > \varepsilon] \leq 4m_H(2N)e^{-\frac{1}{8}\varepsilon^2 N}}$$

de nfs el equation bta3t el mo7adra el fatet
bs el fr2 en bdl l M el hya total number of hypothesis
7tena el m el hya el gross function

VC Generalization bound

- The VC inequality:

$$\Pr[|Ein(g) - Eout(g)| > \varepsilon] \le 4m_{H(2N)}e^{-\frac{1}{8}\varepsilon^2 N}$$

- Thus, with probability at least 1-δ where $\delta = 4m_{H(2N)}e^{-\frac{1}{8}\varepsilon^2 N}$:

$$Eout(g) \le Ein(g) + \varepsilon$$

- Accordingly, the VC generalization bound:

$$\boxed{Eout(g) \le Ein(g) + \sqrt{\frac{8}{N}\ln\left(\frac{4m_{H(2N)}}{\delta}\right)}}$$

- If VC dimension is finite, then the growth function $m_{H(2N)}$ is polynomial, and the generalization error converges to zero as N increases.

lw 3ndy bound, el term el howa el ln, hyb2a byzed brkm a2al mn el N, fa sa3tha lama N -> inf, el limit hyb2a 0 -> fa el error hyb2a b 0 t2rebn

lakn lw mfesh bound, hytl3 3ndk error 16 log 2 + rkm +ve, fa keda 3ndk el bound 3omro ma hy2l mahma zwdt el data set.

Sample Complexity

- The sample complexity means the number of training examples N needed to achieve a certain
- generalization performance.

- The generalization performance is characterized by two parameters:

ε: Error tolerance defines the allowed generalization error

δ: Defines how often the error tolerance ε is violated.

Sample Complexity (cont.)

- To get a generalization error $\leq \varepsilon$:

$$\sqrt{\frac{8}{N}\ln(\frac{4m_{H}(2N)}{\delta})} \leq \varepsilon$$

- Then, the sample complexity N to achieve that generalization error would be:

$$\boxed{N \geq \frac{8}{\varepsilon^{2}} \ln\left(\frac{4m_{H}(2N)}{\delta}\right)}$$

grb t7lha 3la wr2a lw m3ndksh breakpoint -> mh(2N) = 2^2N

which if function of N, solve it using numerical iterative methods.

dayman bnb2a 3uzen el dv mykonsh b infinity.

As a rule of thumb $N \geq 10\, d_{vc(H)}$

3auzen nrkz 3la el model, mynf34 ngeb msln el positive ray model w nst5dmo fe problem zy determine whether the number is even or odd.
kda el Ein hyb2a kber awy mhma kbrna el data set bt3ty.

Model Complexity

- With probability at least *1- δ:*

$$Eout(g) \leq Ein(g) + \sqrt{\frac{8}{N}\ln(\frac{4m_{H}(2N)}{\delta})}$$

- With fixed N (number of training samples), the term $\sqrt{\frac{8}{N}\ln(\frac{4m_{H}(2N)}{\delta})}$ can be regarded as model complexity:

$$\Omega(N,H,\delta)=\sqrt{\frac{8}{N}\ln(\frac{4m_{H}(2N)}{\delta})}$$

Thus, with probability at least *1- δ:*

$$\boxed{Eout(g) \leq Ein(g) + \Omega(N,H,\delta)}$$

Model complexity vs. generalization error Trade-off

Increases as dvc increases (complex models)

leh da by7sl, ka intuation msh ka equation -> eb2a 2ul l hany.

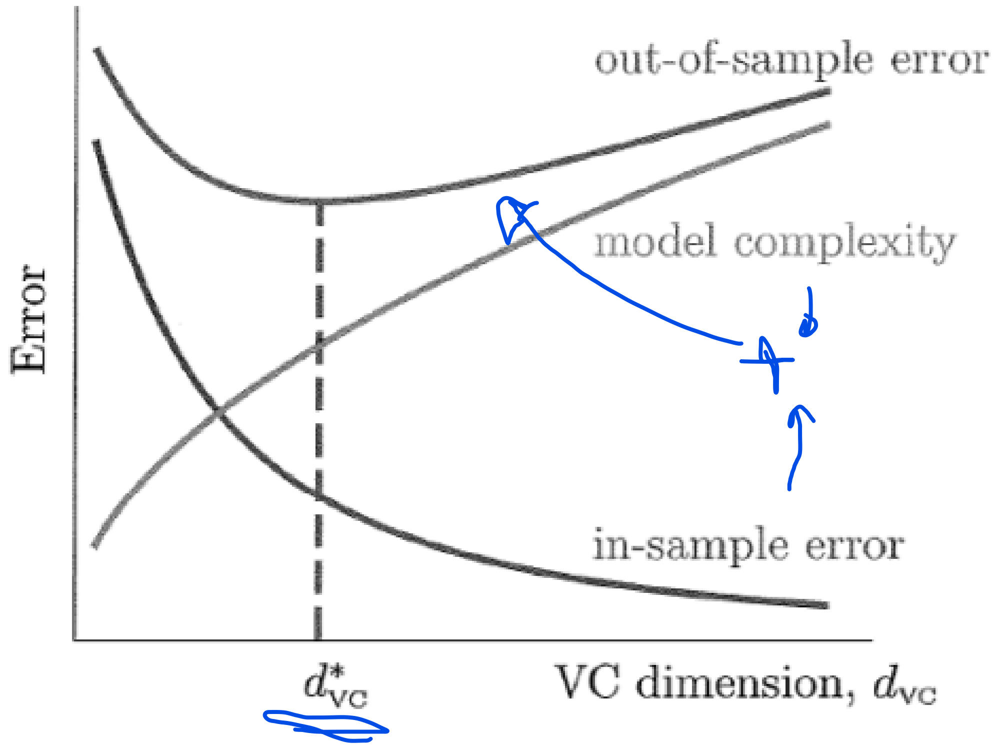
$$Eout(g) \leq Ein(g) + \sqrt{\frac{8}{N}\ln\left(\frac{4m_{H(2N)}}{\delta}\right)}$$

Decreases as dvc increases (complex models)

- The more complex H, the higher $m_{H(2N)}$ and the generalization error would be larger (worse).
- However, the more complex H, the less the in-sample error.

Generalization Error

- A very simple model with low d_{vc}, will not fit the training data well and will have a high in-sample error.
- A more complex learning model with higher d_{vc} would fit the training data better, resulting in a lower in-sample error, but the generalization will be worse.
- Some intermediate d^*_{vc} represents a trade-off between the two errors.



sweet point

Ein vs. Eout

- Ein is the error on the training data. (in-sample error)
- Eout is the error over the entire input space X. (out of sample error)

- To estimate Eout, we should use new test points that are never used for training.

Estimating Eout in practice

- The VC bound is loose, we need a more accurate estimate of Eout for real-world applications.

Evaluate a sample estimate for Eout as follows:

- Use a fresh new test set of size K not involved in the training process.
- Test the final hypothesis g on the test set and report Etest.

- However, what about the generalization error between Etest and Eout?

We can apply Hoeffding's Inequality as g is not affected by test data.

$$\boxed{P(|Etest(g) - Eout(g)| > \varepsilon) \leq 2e^{-2\varepsilon^2 K}}$$

where K is the test set size.

el hadaf mn el slide hena eno y2olak enk b3d ma tkhls training, yb2a 3ndk test set btmrn el model 3leh, w 34an tet2kd hwa hytsrf ezay fl wak3, lazm ykon 3ndk test set w t7sb el error bta3ha 3aml ezay.

w lw el accuracy mkntsh kwysa, hat test set gdeeda, w b3d ma t3ml t3delat 3la el model w tmrno, eb2a e3ml test behom, mt3mlsh test b nfs el set.

Summary

- Dichotomies
- Growth function
- Breakpoint
- VC dimension
- VC generalization bound for infinite hypotheses
- Sample complexity
- Model Complexity and trade-off between in-sample and generalization error
- Estimate Eout in practice