Lecture 9 **EE 660** Sep 18, 2018

Announcements

- Homework 4 was due today.
- Homework 5 has been posted.
- Homework 3 solutions have been posted.
- My office hours tomorrow will be changed to 12:00 -1:00 PM.

Today's Lecture

- Need a better measure of complexity of H
- Towards an effective number of hypotheses
  - Dichotomies, growth function, shattering, break points
  - VC dimension

FROM LAST LECTURE:

$$(2.1) \quad P\left[E_{out}(h) \le E_{in}(h) + \underbrace{\sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}}_{\varepsilon_M}\right] > 1-\delta$$

$\uparrow$ or $\mathcal{D}$

TODAY:

TRY USING M TO EVALUATE COMPLEXITY OF $\mathcal{H}_1$ AND $\mathcal{H}_2$

LET $\varepsilon_M = \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}$ = GENERALIZATION ERROR BOUND BASED ON M.

$$E_{\mathcal{D}}(h) \xrightarrow{\pm\varepsilon_M} E_{out}(h)$$

Ex: 2-class problem (setosa vs. virginica), $D = 2$ features.

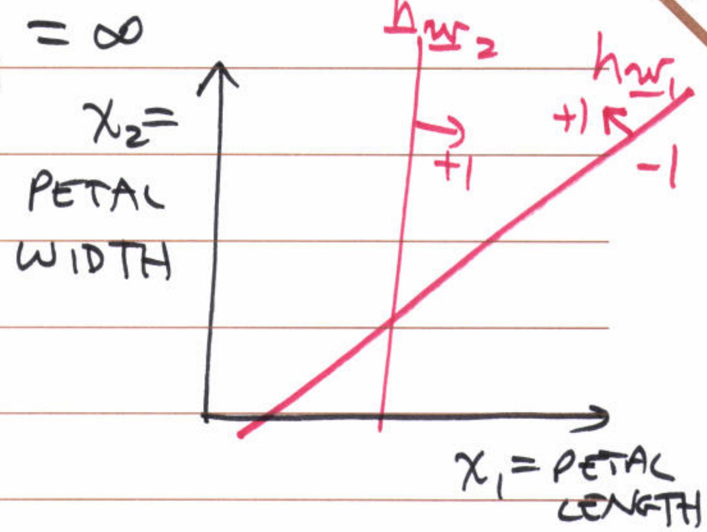
CASE 1: LINEAR PERCEPTRON

$$g_1(\underline{x}) = w_0 + w_1 x_1 + w_2 x_2$$

$$\mathcal{H}_1 = \{ h_{\underline{w}}(\underline{x}) = \text{sgn}[w_0 + w_1 x_1 + w_2 x_2] \mid w_j \in \mathbb{R},\ j = 1, 2, 3 \}$$

\# HYPOTHESES $M_1 = \infty$

$\mathcal{E}_{M_1} = \infty$



CASE 2: QUADRATIC PERCEPTR.

$$g_2(\underline{x}) = \underline{w}^T \underline{\phi}(\underline{x})$$

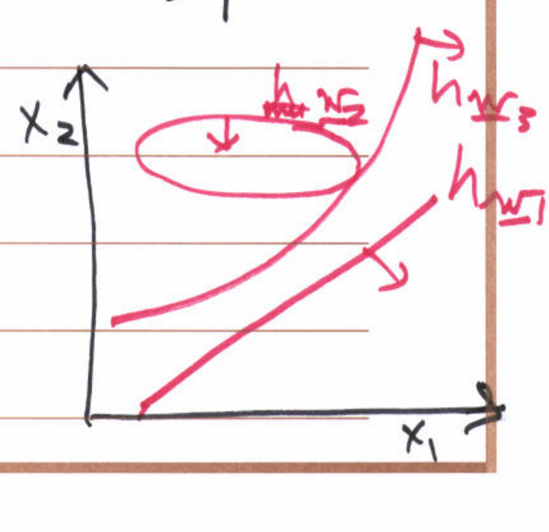
$$\underline{\phi}(\underline{x}) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix}$$

$$\cancel{\mathcal{H}_2 = \{ h_{\underline{w}}(\underline{x}) \mid = \underline{w}^T \underline{\phi}(\underline{x})}$$

$$\mathcal{H}_2 = \{ h_{\underline{w}}(\underline{x}) = \text{sgn}\left[\underline{w}^T \underline{\phi}(\underline{x})\right] \mid \underline{w} \in R^6 \}$$

\# HYPOTH $M_2 = \infty$

$\mathcal{E}_{M_2} = \infty.$



$\Rightarrow$ WE NEED A BETTER MEASURE OF COMPLEXITY OF $\mathcal{H}$

TOWARD AN EFFECTIVE # OF HYPOTHESES

BINARY-VALUED $f(\underline{x})$ (2 CLASS PROBLEMS).

CONSIDER HOW EACH $h_i(\underline{x}) \in \mathcal{H}$ BEHAVES ON N DATA POINTS $\underline{x}_n$, $n=1, \cdots, N$, DRAWN FROM $\mathcal{X}$.

DEF: THE SET OF DICHOTOMIES GENERATED BY $\mathcal{H}$ ON $\{\underline{x}_n\}_{n=1}^{N}$:

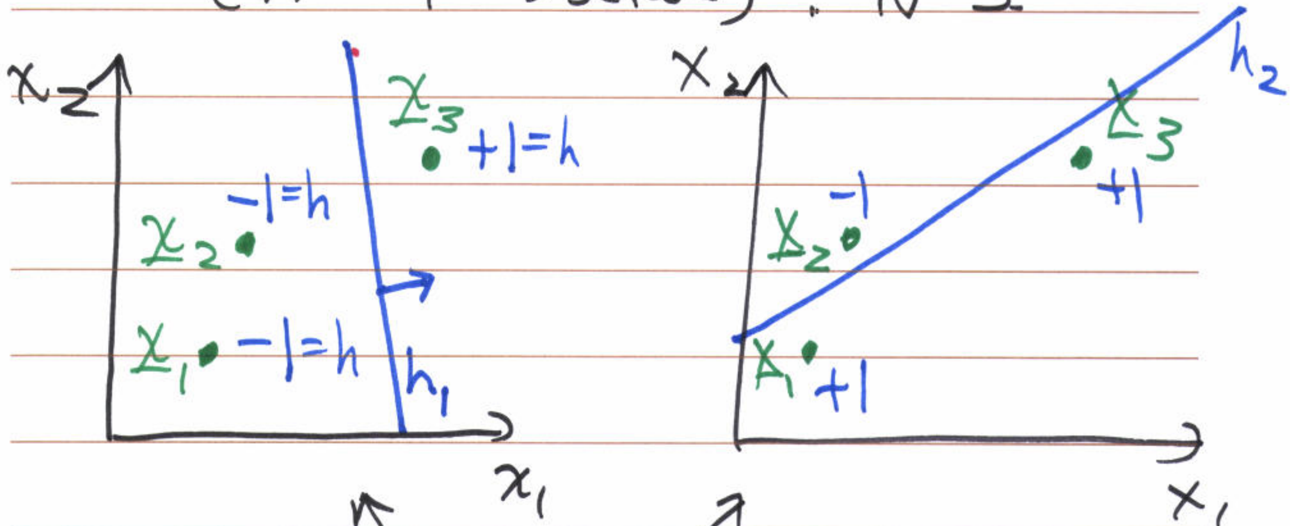
$$\mathcal{H}(\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_N)$$

$$= \left\{ \left(h(\underline{x}_1), h(\underline{x}_2), \cdots, h(\underline{x}_N)\right) \middle| h \in \mathcal{H} \right\}.$$

(DUPICATE N-TUPLES COUNT AS SAME MEMBER OF $\{\cdot\}$).

Ex: LINEAR PERCEPTR IN 2D:

($h = +1 \Rightarrow$ setosa), $N=3$



2 dichotomies

$|\mathcal{H}(\underline{x}_1, \cdots, \underline{x}_N)|$ = CARDINALITY OF $\mathcal{H}(\underline{x}_1, \cdots, \underline{x}_N)$
= # OF DICHOTOMIES.

GIVEN N PTS. — HOW MANY DICHOTOMIES ARE POSSIBLE? $\underline{\underline{2^N}}$.

DEF: GROWTH FUNCTION FOR $\mathcal{H}$ IS:

$$m_{\mathcal{H}}(N) = \max_{\underline{x}_1, \cdots, \underline{x}_N \in \mathcal{X}} |\mathcal{H}(\underline{x}_1, \cdots, \underline{x}_N)|$$

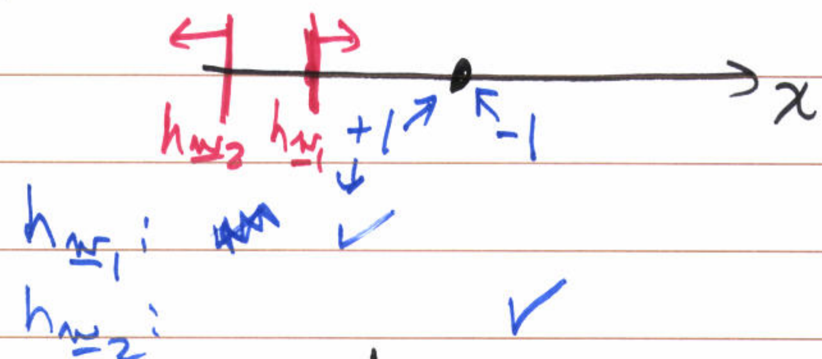
I.E., FIND THE SET OF N POINTS THAT MAXIMIZES THE # DICHOTOMIES THAT $\mathcal{H}$ CAN REALIZE. $m_{\mathcal{H}}(N)$ = THIS # DICHOTOMIES.

$$m_{\mathcal{H}}(N) \leq 2^N \text{ ALWAYS.}$$

<u>Ex</u>: LET $\mathcal{H}_L$ BE THE SET OF ALL 2-CLASS LINEAR CLASSIFIERS (E.G., PERCEPTRONS) IN 1D (1 FEATURE), SO THAT:

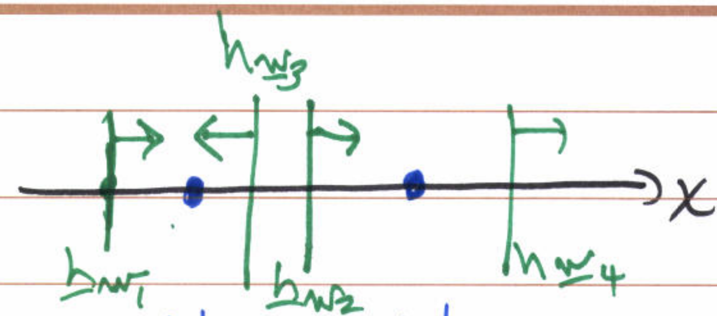
$$\mathcal{H}_L = \{ h_{\underline{w}}(\underline{x}) = \underline{w}^T \underline{x} \mid \underline{w} \in \mathbb{R}^2 \} \quad \text{(augm.)}$$

LET $N = 1$:



$$m_{\mathcal{H}}(1) = 2 = 2^1.$$

$N=2$:



$h_{\underline{w}_1}$: $+1$ $+1$ ✓

$h_{\underline{w}_2}$: -1 $+1$ ✓

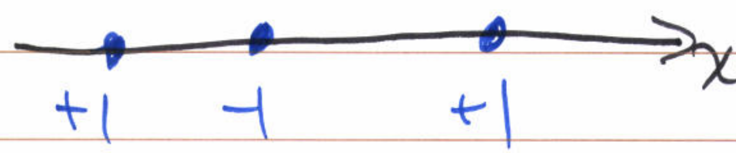
$h_{\underline{w}_3}$: $+1$ -1 ✓

$h_{\underline{w}_4}$: -1 -1 ✓

$\Rightarrow m_{\mathcal{H}}(2) = 4. = 2^2.$

Is $m_{\mathcal{H}}(3) = 2^3 = 8$?

No, BECAUSE:



CAN'T BE REALIZED BY $\mathcal{H}_L$.

<u>DEF</u>: IF $\mathcal{H}$ CAN REALIZE ALL POSSIBLE DICHOTOMIES ON A SET OF POINTS $\underline{x}_1, \cdots, \underline{x}_N$, THEN $\mathcal{H}$ CAN <u>SHATTER</u> $\underline{x}_1, \cdots, \underline{x}_N$.

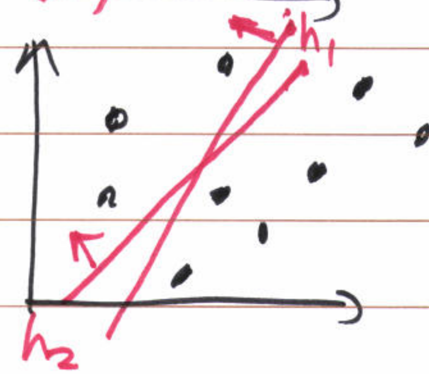
e.g., $\mathcal{H}_L$ SHATTERS THE $N=2$ POINTS PLOTTED ABOVE.

NOTE: WITH M,

$M = |\mathcal{H}|$

WITH

$$m_{\mathcal{H}}(N) = \underbrace{\max}_{x_1, \cdots, x_N \in \mathcal{X}} |\mathcal{H}(x_1, \cdots, x_N)|$$



BREAK POINT k

IF THERE IS NO SET OF k (DISTINCT) POINTS THAT CAN BE SHATTERED BY $\mathcal{H}$, THEN k IS A BREAK POINT FOR $\mathcal{H}$,

AND $m_{\mathcal{H}}(k) < 2^k$.

$\mathcal{H}_L$ BREAK POINTS ARE: ANY $k \geq 3$.

THEOREM 2.4:

IF k IS A BREAK POINT FOR $\mathcal{H}$, THEN:

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

NOTE: $\binom{N}{i} = \dfrac{N!}{i!(N-i)!} = O(N^{N-(N-i)})$

$= O(N^{i}).$

POLYNOMIAL IN N, OF DEGREE $k-1$.

VC DIMENSION

(VAPNIK–CHERVONENKIS)

VC DIM. IS A MEASURE OF THE "FLEXIBILITY" OR "COMPLEXITY" OF THE HYPOTHESIS SET.

DEF: THE VC DIMENSION OF $\mathcal{H}$, $d_{vc}(\mathcal{H})$, IS THE LARGEST VALUE OF N FOR WHICH $m_{\mathcal{H}}(N) = 2^N$. IF $m_{\mathcal{H}}(N) = 2^N \ \forall N$, THEN $d_{vc}(\mathcal{H}) = \infty$.

NOTES:

1. GIVEN $d_{VC}(\mathcal{H})$, THEN $k = d_{VC} + 1$ (OR ANY $k \geq d_{VC} + 1$) IS A BREAK POINT OF $\mathcal{H}$.
2. $d_{VC}(\mathcal{H})$ = MAX. N THAT $\mathcal{H}$ CAN SHATTER.
3. FOR OUR $\mathcal{H}_L$ (LINEAR PERC. IN 1D), $d_{VC}(\mathcal{H}_L) = 2$. (NOTE: FOR $\mathcal{H}_L$, $M = \infty$).

EFFECTIVE # HYPOTHESES AND VC GENERALIZATION BOUND

FROM Th'm 2.4,

$m_{\mathcal{H}}(N) \leq$ polyn. IN N OF DEGREE d_{VC}.

ONE CAN SHOW:

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

Th'm 2.5: VC GENERALIZATION BOUND

FOR ANY TOLERANCE $\delta > 0$,

$$E_{out}(h_g) \le E_{in}(h_g) + \underbrace{\sqrt{\frac{8}{N} \ln \frac{4 m_{\mathcal{H}}(2N)}{\delta}}}_{\varepsilon_{eff}}$$

WITH PROBABILITY $\ge 1-\delta$.

PROOF: IN AML APPENDIX (N.R.F.)

WILL EXPRESS ε_{eff} USING d_{vc}, NEXT TIME.