Thursday, 9/17/2020

EE 660

MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

Lecture 8

p. 1

Announcements

- Homework 2 (Week 3) is due tomorrow (Friday).
- Homework 3 will be posted.
- Homework late submission policy has been posted
  - In Content > Syllabus and Course Information

Today's Lecture

- 2 aspects of learning feasibility
- Noisy targets
- Generalization error (revisited)
- Toward an effective number of hypotheses
  - Dichotomies
  - Growth function
  - Shattering
  - Break points
  - VC dimension (if time)

Noisy Target Functions [AML 1.4.2]

Sometimes the target function (class labels or regression output) is noisy.

Instead of $y = y(\underline{x}) = f(\underline{x})$ = deterministic, we let:

$$\boxed{\begin{array}{l} f(\underline{x}) = \text{deterministic} \;; \quad y = y(\underline{x}) = \text{random} \sim p(y \mid \underline{x}) \\ \text{Data points come from } p(\underline{x}, y) = p(y \mid \underline{x})\, p(\underline{x}) \end{array}}$$

Ex: $y(\underline{x}) = f(\underline{x}) + \epsilon_{noise}$ (regression)

$y(\underline{x}) = \text{sgn}\{g(\underline{x}) + \epsilon_{noise}\}$ (classification) $\quad f(\underline{x}) = \text{sgn}\{g(\underline{x})\}$

$g(\underline{x})$ = discriminant function.

$\underline{x}_n$ is drawn from $p(\underline{x}_n)$; $y_n = y(\underline{x}_n)$ is drawn from $p(y \mid \underline{x} = \underline{x}_n)$

Then: $E_{\mathcal{D}}(h) = \frac{1}{N} \sum_{n=1}^{N} [\![h(\underline{x}_n) \neq y(\underline{x}_n)]\!]$

and $E_{out}(h) = P[h(\underline{x}) \neq y(\underline{x})]$

for classification (h, y are binary or integer)

We can reduce the noisy (probabilistic) case to the deterministic case:

$$\text{let } \quad p(y \mid \underline{x}) = [\![y = f(\underline{x})]\!]$$

Thus, the noisy target is the more general case.

For the rest of the AML material, we will use the notation of the deterministic case most of the time, with the understanding that it can be generalized to the noisy target case as described above.

Generalization Error (revisited)

From before:

$$M = |\mathcal{H}| = \#\text{ of hypotheses in } \mathcal{H}$$

(2) $P[\,|E_{in}(h) - E_{out}(h)| > \epsilon\,] \leq \underbrace{2Me^{-2\epsilon^2 N}}_{\delta}$, for any $\epsilon > 0$

↖ True for any h in $\mathcal{H}$ (including h_g)

Re-arrange:

$$P[\,|E_{out}(h) - E_{in}(h)| \leq \epsilon\,] \geq \overbrace{1 - 2Me^{-2\epsilon^2 N}}^{1-\delta}$$

δ is our tolerance. $\delta = 2Me^{-2\epsilon^2 N} \Rightarrow \epsilon = \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}} \triangleq \varepsilon_M$

Note: $|E_{out} - E_{in}| \leq \epsilon \Rightarrow$ [case $E_{out} \geq E_{in}$] (i) $E_{out} \leq E_{in} + \epsilon$

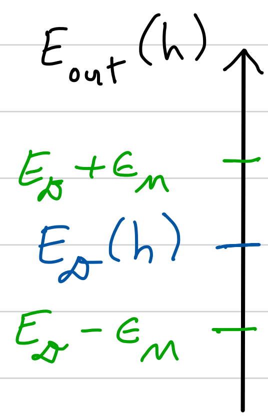
[case $E_{out} \leq E_{in}$] (ii) $E_{out} \geq E_{in} - \epsilon$

(2.1) $\therefore$ $\boxed{P\left[\, E_{out}(h_g) \leq E_{in}(h_g) + \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}\,\right] \geq 1 - \delta}$

(2.1)′ and $P\left[\, E_{out}(h) \geq E_{in}(h) - \underbrace{\sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}}_{\varepsilon_M}\,\right] \geq 1 - \delta$

(2.1) gives upper bound on $E_{out}(h_g)$ in terms of δ.

Note N, M dependence of ϵ_M



"error bar"
$E_{out}(h)$ is in this range with probability $(1-\delta)$.

(2.1)' gives lower bound on $E_{out}(h)$, for any $h \in \mathcal{H}$.

If we had chosen a different hypothesis h_g' in $\mathcal{H}$, such that $E_{\mathcal{D}}(h_g') > E_{\mathcal{D}}(h_g)$,

then
$$\underbrace{E_{\mathcal{D}}(h_g') - \epsilon}_{\text{lower bound on } E_{out}(h_g')} > \underbrace{E_{\mathcal{D}}(h_g) - \epsilon}_{\text{lower bound on } E_{out}(h_g)}$$

Two Aspects of Learning Feasibility [AML 1.3.3]

From last time:

$$\begin{cases} P\left[\,|E_{\mathcal{D}}(h_g) - E_{out}(h_g)| > \epsilon\,\right] \le 2Me^{-2\epsilon^2 N}, \text{ for any } \epsilon > 0 \\ M = |\mathcal{H}| = \#\text{ of hypotheses in } \mathcal{H} \end{cases}$$

$\boxed{\text{Feasibility of learning: Can we get } E_{out}(h_g) \text{ small enough?}}$

$\boxed{\text{Can we get } |E_{\mathcal{D}}(h_g) - E_{out}(h_g)| \text{ or } \epsilon_M \text{ small enough?}}$

$\boxed{\text{Can we make } E_{\mathcal{D}}(h_g) \text{ small enough?}}$

↑ Generalization error

↑ Measured error of h_g on $\mathcal{D}$

Comment: if we use $\epsilon = \varepsilon_M$ for generalization error with a linear perceptron classifier: $(\underline{w} = \underline{w}^{(t)})$

$$\mathcal{H} = \{ h_{\underline{w}}(\underline{x}) = \operatorname{sgn}(\underline{w}^T \underline{x}) \mid \underline{w} \in \mathbb{R}^{D+1} \}$$

then $M = |\mathcal{H}| = \#\text{hypotheses} = ? = \infty$

$$\Rightarrow \text{generalization error} = \varepsilon_M = \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}} = \infty$$

$\Rightarrow$ We need a better measure than $M = |\mathcal{H}|$ for complexity of $\mathcal{H}$.

Toward an Effective Number of Hypotheses [AML 2.1.1]

For 2-class (binary) classification problems
$\Rightarrow f(\underline{x}) \in \{-1, +1\}$ or $\{0, +1\}$

Consider how each $h_i(\underline{x}) \in \mathcal{H}$ behaves on N data points $\underline{x}_n$, $n = 1, 2, 3, \cdots, N$, drawn from $\mathcal{X}$.

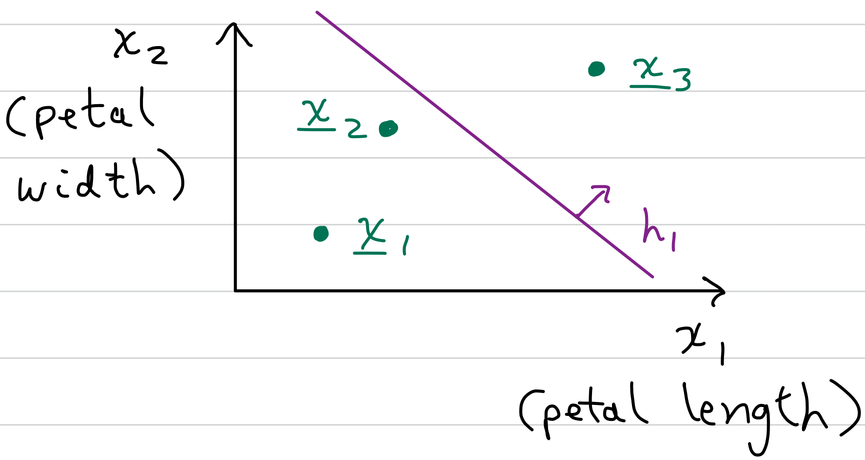
Def: The set of dichotomies generated by $\mathcal{H}$ on $\{\underline{x}_n\}_{n=1}^{N}$ is

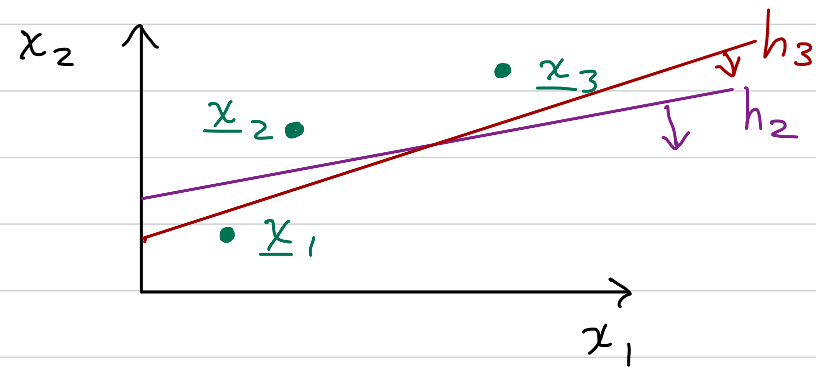
$$\mathcal{H}(\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_N) = \{ \underbrace{(h_i(\underline{x}_1), h_i(\underline{x}_2), \cdots, h_i(\underline{x}_N))}_{N\text{-tuple}} \mid h_i \in \mathcal{H} \}.$$

(Duplicate N-tuples count as the same member of $\{\cdot\}$.)

Ex: Linear perceptron in 2D

Let $N=3$, $h=+1 \Rightarrow$ setosa, $h=-1 \Rightarrow$ virginica





h_1: $(h_1(\underline{x}_1), h_1(\underline{x}_2), h_1(\underline{x}_3))$
$= (-1, -1, +1)$

h_2: $(1, -1, -1)$
h_3: $(1, -1, -1)$

Let $|\mathcal{H}(\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_N)| = \underline{\text{Cardinality of } \mathcal{H}(\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_N)}$
= # of dichotomies (or # of unique N-tuples) that $\mathcal{H}$ can realize on the points $\{\underline{x}_n\}_{n=1}^N$.

Given N points, what is the maximum # of dichotomies? 2^N

Growth Function

Def: Growth function for a given $\mathcal{H}$ is:

$$m_{\mathcal{H}}(N) \triangleq \max_{\underline{x}_1, \underline{x}_2, \underline{x}_3, \cdots, \underline{x}_N \in \mathcal{X}} \left| \mathcal{H}(\underline{x}_1, \underline{x}_2, \underline{x}_3, \cdots, \underline{x}_N) \right|$$

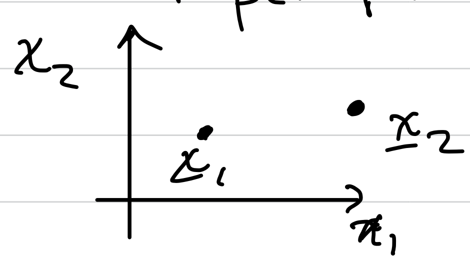
i.e., find the set of N points that maximizes the # of dichotomies that $\mathcal{H}$ can realize. $m_{\mathcal{H}}(N)$ = this # of dichotomies.

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

[Example: in text and in discussion 4.]

Def: If $\mathcal{H}$ can realize all possible dichotomies of a set of N points $\underline{x}_1, \underline{x}_2, \cdots \underline{x}_N$, then $\mathcal{H}$ can shatter $\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_N$.

Ex: Can a linear perceptron in 2D shatter $N=2$ data points shown below?



→ Yes.

Break Points

Def: 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

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

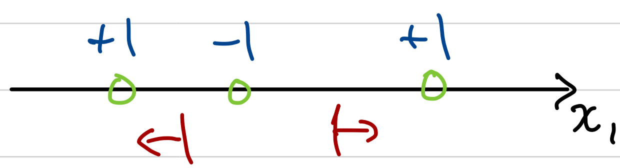
Ex: For $\mathcal{H}_L^{(1D)} = \{\text{linear perceptron in 1D}\}$,

$N=2$:

Can be shattered by $\mathcal{H}_L^{(1D)}$

$$\Rightarrow \quad m_{\mathcal{H}_L^{(1D)}}(N=2) = 2^2 = 4$$

$N=3$:



$\leftarrow$ $(+1, -1, +1)$ can't be realized by $\mathcal{H}_L^{(1D)}$.

True for all sets of $N=3$ distinct data points

$\Rightarrow$ $k=3$ is a break point for $\mathcal{H}_L^{(1D)}$

Any $k \geq 3$ is a break point for $\mathcal{H}_L^{(1D)}$

VC Dimension [AML 2.1.3]

(Vapnik–Chervonenkis) (binary classification problems)

VC dimension (or VCdim) is a measure of the "flexibility" or "complexity" of the hypothesis set (or decision boundary).

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 $d_{vc}(\mathcal{H}) = \infty$.