FOUNDATIONS OF STATISTICAL MACHINE LEARNING

CLASSICAL LEARNING MODEL

Domain Set $\longrightarrow$ Set of objects for labelling (X)

$\Rightarrow$ Depicted as feature vectors (input to the algorithm)

Label Set $\rightarrow$ Set of possible labels (Y)

Training Data $\longrightarrow$ $S \subseteq X \times Y$

Feature vectors + Labels.

Learner's Output $\longrightarrow$ Hypothesis function $h: X \rightarrow Y$

Prediction Algorithm $\rightarrow A$

Distribution $\longrightarrow$ Distrubution D exists on domain X

For labels $\rightarrow$ There exists correct labelling function (f)

Loss Function

$$L_{D,f}(h) = \Pr_{x \in D}\left[h(x) \neq f(x)\right]$$

EMPIRICAL RISK MINIMISATION

$f \to$ unknown $\quad D \to$ unknown

* Minimising error over training data

Learner has S (SAMPLE) according to D (DISTRIBUTION)

$S = (x_1, y_1), (x_2, y_2) \dots (x_m, y_m)$

where $x_i \sim D$

$y_i = f(x_i)$

$$L_S = \frac{|\{i \in [m] \mid h(x_i) \neq f(x_i)\}|}{m}$$

OVERFITTING :-

The distribution D may confuse a learner to learn a wrong rule. $\Rightarrow$ Because S is a part of D and does not represent D exhaustively

$\downarrow$

This rule will work fine on training data

ERM $\to$ Find h that minimizes L_S

* Finding h may not be an efficient way because all possible functions may be very large.

PROBLEM: Aiming for $L_S(h) = 0$ may lead to overfitting.

* USE ADDITIONAL INFO: $ERM_{\mathcal{H}}$ $\quad$ $\mathcal{H} \to$ Hypothesis Class (PRIOR KNOWLEDGE)

$$ERM_{\mathcal{H}}(S) = \underset{h \in \mathcal{H}}{\text{argmin}}\ L_S(h)$$

* A restrictive $\mathcal{H}$ may make learning the problem easy.

* If $\mathcal{H}$ is restricted, it can lead to large INDUCTIVE BIAS

Over which hypothesis classes $\mathcal{H}$ will $ERM_{\mathcal{H}}$ not lead to overfitting ??

$$h_S = \underset{h \in \mathcal{H}}{\text{argmin}}\ L_S(h)$$

$\downarrow$

Result of applying $ERM_{\mathcal{H}}$

* REALIZABILITY ASSUMPTION : $f \in \mathcal{H}$ $\leftarrow$ Non-intuitive

ACCURACY PARAMETER :

$L_{D,f}(h_S) \Rightarrow$ Probability of choosing an example $x \sim D$ such that $h_S(x) \neq f(x)$

* If $L_{D,f}(h_S) < \varepsilon$, then learner is approximately correct

$1 - \varepsilon \Rightarrow$ Accuracy Parameter

<u>THEOREM</u>: Let $\mathcal{H}$ be a finite hypothesis class of functions from domain X to range $\{0,1\}$. Then for every $f \in \mathcal{H}$ and every D over X,

$$\Pr_{S \sim D^m}\left[L_{D,f}\left(\underbrace{ERM_{\mathcal{H}}(S)}_{h_S}\right) > \epsilon\right] \leq \underbrace{|\mathcal{H}|}_{\text{Size of } \mathcal{H}} \cdot (1-\epsilon)^{m}$$

m: Sample size

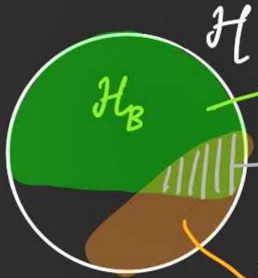
Probability goes down exponentially

★ Theorem holds for every m and every ϵ

★ As long as $e^{-\epsilon m}$ dominates $|\mathcal{H}|$ we can contain error in $L_{D,f}(h_S)$

★ Therefore, if $|\mathcal{H}|$ is bounded or finite, then we can bound $|\mathcal{H}| \cdot (1-\epsilon)^m$ by changing sample size m.

<u>PROOF</u>:



BAD FUNCTIONS $\{h \mid L_{D,f}(h) > \epsilon\}$

Unfavourable Conditions

SMALL ERROR FUNCTIONS $\{h \mid L_S(h) = 0\}$

GOAL: To bound $\Pr\left[\exists h \in \mathcal{H}_B \text{ such that } L_S(h) = 0\right]$

$$\Pr\left[\exists h \in \mathcal{H}_B \text{ and } L_S(h)=0\right] \leq |\mathcal{H}_B| \Pr_{S \sim D^m}\left[L_S(h) = 0\right]_{h \in \mathcal{H}_B}$$

Assumption about $\mathcal{H}_B \Rightarrow L_{D,f}(h) > \epsilon$

Thus $\Pr_{S \sim D^m}\left[L_S(h) = 0 \text{ for } h \in \mathcal{H}_B\right] \leq (1-\epsilon)^m$

Finally $\Pr\left[\exists h \in \mathcal{H}_B \text{ and } L_S(h)=0\right] \leq |\mathcal{H}_B| (1-\epsilon)^m$

$$\leq |\mathcal{H}| (1-\epsilon)^m$$

PROBABILISTICALLY APPROXIMATELY CORRECT

A class $\mathcal{H}$ is said to be PAC learnable, if there is a function $m_{\mathcal{H}} : (0,1) \times (0,1) \rightarrow \mathbb{N}$ and learning algorithm A such that for every $\varepsilon, \delta \in (0,1)$, for every Distribution D and for every $f \in \mathcal{H}$

$$\Pr_{S \sim D^m} \left[L_{D,f}(A(S)) \leq \varepsilon \right] > 1-\delta$$

for any $m \geq m_H(\varepsilon, \delta)$

$\varepsilon \rightarrow$ Accuracy Parameter

$1-\delta \rightarrow$ Confidence Parameter

What Algorithm to use? $\Rightarrow ERM_h$

Since $|\mathcal{H}|(1-\varepsilon)^m \leqslant |\mathcal{H}| e^{-\varepsilon m} \leq \delta$

$$m_H(\varepsilon, \delta) = \frac{\ln(|\mathcal{H}|) + \ln(1/\delta)}{\varepsilon}$$

STRENGTH: Valid for any distribution

WEAKNESS: Assumption that $f \in \mathcal{H}$

Used when we assumed $L_S(h) = 0$ if $h = ERM_{\mathcal{H}}(S)$

How to remove this assumption??

* Agnostic PAC learning

Removing realisability assumption

$D \rightarrow$ Distribution over $X \times Y$

$D_x \rightarrow$ Distribution over unlabelled domain points $\nearrow$ Marginal

$D((x,y)|x) \longrightarrow$ conditional probability of labels for each domain point.

TRUE ERROR / RISK : The likelihood of h making an error when labelled points are drawn as per D

$$L_D(h) = \Pr_{(x,y)\sim D}[h(x) \neq y]$$

EMPIRICAL RISK :

$$L_S(h) = \frac{|\{i \in [m] \mid h(x_i) \neq y_i\}|}{m}$$

BAYES OPTIMAL PREDICTOR

$$f_D(x) = \begin{cases} 1 & \text{if } \Pr[y=1|x] \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

Prove that: For any classifier g, $L_D(f_D) \leq L_D(g)$

PROBLEM : Learner does not know D

AGNOSTIC PAC LEARNING

A class $\mathcal{H}$ is said to be agnostic PAC learnable, if there is a function $m_{\mathcal{H}} : (0,1) \times (0,1) \rightarrow \mathbb{N}$ and a learning algorithm A such that for every distribution D on $X \times Y$,

$$\Pr_{S \sim D^m} \left[L_D(A(S)) \leq \min_{h \in \mathcal{H}} \{ L_D(h) \} + \varepsilon \right] > 1 - \delta$$

for any $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$

→ Cannot guarantee 0 error
This is best possible function

STRENGTH: Realizability assumption not necessary

★ If $H_1 \subseteq H_2$, H_1 is easier to learn.

WEAKNESS: The error is relative

★ with respect to $\min_{h \in \mathcal{H}} \{ L_D(h) \}$

OTHER LEARNING TASKS:

① Multiclass Prediction
② Real valued Prediction (Regression)

→ Different notion of loss

General Setup for Learning:

① Binary Label Prediction → $\text{loss} = \begin{cases} 1 & h(x) \neq y \\ 0 & h(x) = y \end{cases}$

② Multiclass Prediction → $\text{loss} = \begin{cases} 1 & h(x) \neq y \\ 0 & h(x) = y \end{cases}$

③ Regression → $\ell(h, (x,t)) = (h(x) - t)^2$

④ Classification → $\ell((c_1 \ldots c_k), z) = \min_{1 \leq i \leq k} \| c_i - z \|^2$

LEARNING PARADIGM : UNIFORM CONVERGENCE

DEFINITION : A sample $S = ((x_1, y_1), (x_2, y_2) \cdots (x_n, y_n))$ is ϵ-representative of a class $\mathcal{H}$ with respect to the distribution D if $\forall h \in H \ |L_S(h) - L_D(h)| \leq \epsilon$

where $\quad L_S(h) = \frac{1}{m} \sum_{(x_i, y_i) \in S} \ell(h, (x_i, y_i))$

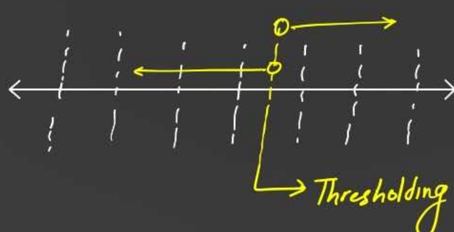
★ Note that if S is indeed representative of H with respect to D then ERM_H is a good learning strategy

CLAIM : If S is ϵ-representative of $\mathcal{H}$ with respect to D then for any ERM_H function h_s

$$L_D(h_s) \leq \min_{h \in H} \left(L_D(h)\right) + 2\epsilon$$

PROOF: $L_D(h_s) \leq L_S(h_s) + \epsilon \leq \min_{h \in H} \left[L_S(h) + \epsilon\right] \leq \min_{h \in H} \left[L_D(h)\right] + 2\epsilon$

★ Practical approach to infinite classes $\longrightarrow$ Discretization



EXAMPLE $\longrightarrow$ $H_\alpha^{thr} = \left\{ h_r : h_r(x) = \begin{cases} 0 & x \leq r \\ 1 & x > r \end{cases}, \ r \in \left\{0, \frac{1}{\alpha}, \frac{2}{\alpha} \cdots 1\right\} \right\}$

★ In theory, H_α^{thr} may not be a good approximation of H^{thr}.

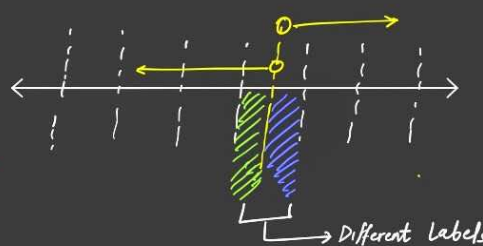
$$\min_{h \in H^{thr}} L_D(h) \ll \min_{h \in H_\alpha^{thr}} L_D(h)$$

EXAMPLE

$D(\{(\frac{3}{2}\alpha, 0)\}) = 0.5$

$D(\{(\frac{7}{4}\alpha, 1)\}) = 0.5$

$\hookrightarrow$ Put everything on 2 point



NO FREE LUNCH THEOREM :

Let $\mathcal{X}$ be a domain of size n.

Let H_n^{all} be the set of all possible labellings.

$H_n^{all} = \{h : \mathcal{X} \to [0,1]\} \qquad |H_n^{all}| = 2^n \qquad$ NFL proves that $m_{H_n^{all}}\left(\frac{1}{8}, \frac{1}{7}\right) \geq \frac{n}{2}$

Therefore, the class of labelling functions over an infinite domain is not PAC-learnable.

TRADE OFF :

$$\epsilon_{app} = \min_{h \in H} L_D(h) \qquad \epsilon_{est} = L_D(h_S) - \epsilon_{app}$$

→ By Free Lunch Theorem this is large.

THEOREM :

Let A be a learning algorithm for the task of binary classication over the domain set X. Let $m < |X|/2$. Then $\exists$ D over $X \times \{0,1\}$ such that

$\exists\, f : X \to \{0,1\}$ such that $L_D(f) = 0$

$$\boxed{\Pr_{S \sim D^m} \{ L_D(A(S)) \geq 1/8 \} \geq 1/7}$$

MOTIVATION FOR VC DIMENSION :

→ Comes from proof of no-free-lunch theorem.

★ NOTION OF SHATTERING A SET

A set of points $S \subseteq X$ is said to be shattered by $\mathcal{H}$ for every possible labellings of points in S as 0-1, there is a function $h \in \mathcal{H}$ that realises this labelling.

EXAMPLE 1 :

$\mathcal{H} = \{[0,b] \mid b \in \mathbb{R}^{\geq 0}\}$

★ $\mathcal{H}$ can shatter any set of size 1. EXCEPTION: $x=0$ case.

★ $\mathcal{H}$ cannot shatter any sets of size 2.

EXAMPLE 2 :

$\mathcal{H} = \{[a,b] \mid a,b \in \mathbb{R} \text{ and } a<b\}$

★ $\mathcal{H}$ can shatter any set of size 2

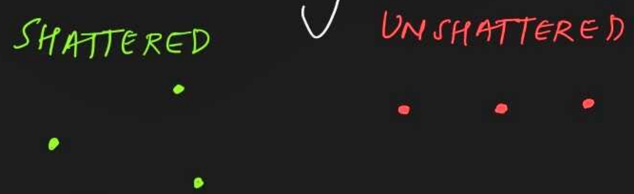
★ $\mathcal{H}$ cannot shatter any set of size 3

EXAMPLE 3 :

$\mathcal{H}$ is a set of linear functions

★ $\mathcal{H}$ can shatter any set of size 2

★★ $\mathcal{H}$ can shatter SOME sets of size 3



VAPNIK - CHERVONENKIS DIMENSION :

Let $\mathcal{H}$ be a hypothesis class over functions from X to $\{0,1\}$.
The VC - Dimension of $\mathcal{H}$ is the size of the largest finite subset of X such that it is shattered by $\mathcal{H}$.

① If $\exists$ a subset of size d such that it can be shattered by $\mathcal{H}$, then VC dimension of $\mathcal{H}$ is at least d.

② If no subset of size d is shattered by $\mathcal{H}$, then VC dimension of $\mathcal{H}$ is less than d.