Foundations of Data Science & Machine Learning

Summary — Week 04

Devansh Singh Rathore
111701011

B.Tech. in Computer Science & Engineering
Indian Institute of Technology Palakkad

March 22, 2021

Abstract

We study about Support Vector Machine (SVM) and Maximum - Margin Separating Hyperplane (MMSHP) mathematically. Later, we look into generalisation of rule 'h' for future unknown points.

1 Generalisation (Cont'd.)

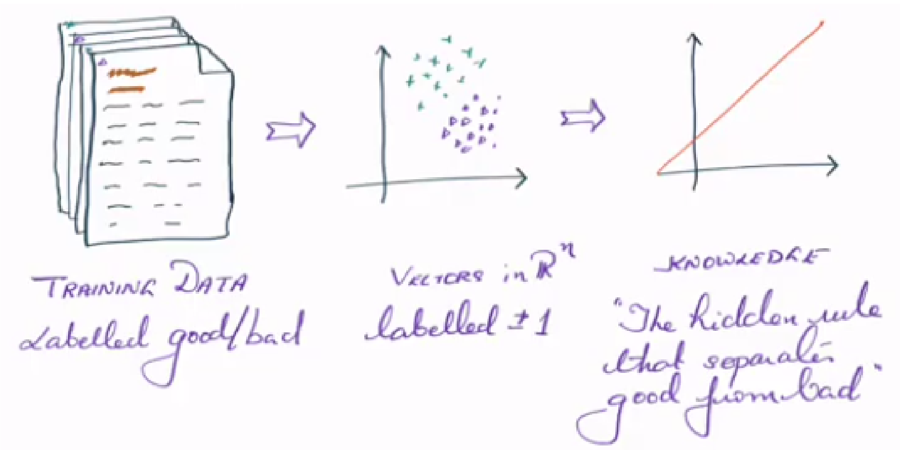

Fig 1.0 Hope of ML

→ The "In Sample Error" i.e. $E_{in}(g)$ is also known as **Empirical error**.
→ The "Out of Sample Error" i.e. $E_{out}(g)$ is also known as **True error**.
→ Perceptron, SVM are deterministic algorithms.
→ Guarantee: If $2(1-\epsilon)^n < \delta$, then every separating line (through origin) of S has out sample error at most ϵ with probability $1-\delta$.

$$P[\cup_{g \in H, E_{in}(g)=0}[E_{out}(g) > \epsilon]] < \delta$$

i.e. even the 'worst' line that separates S will generalise to X.
→ The randomness and thus the probability is associated with choosing the S.
→ Solving 'n' i.e. number of training samples, in terms of ϵ and δ:

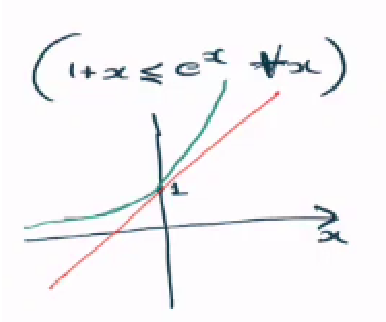

Fig 1.1 $e^x \geq (1+x)$

we want:

$$2(1-\epsilon)^n < \delta$$
$$2e^{-\epsilon n} < \delta$$
$$e^{\epsilon n} > 2/\delta$$
$$n > (1/\epsilon)ln(2/\delta)$$

Theorm 1.0: let $X \subseteq \mathbb{R}^2$ and $f : X \to \{+1, -1\}$ be a set of points which are separable by a line through origin. For any $\delta, \epsilon > 0$, if we pick a set S of size $n > (1/\epsilon)ln(2/\delta)$ points from X independently and uniformly at random, then, with probability at least $1 - \delta$, every line through origin that separates S also separates at least $1 - \epsilon$ fraction of points in X.

Observation:

- $|X|$ can be ¡ n. (Sampling without replacement)
- $|X|$ can be ∞.
- X can be all of $\mathbb{R}^2$. But then how do we choose S? because uniform sampling is impossible.

2 PAC (Probably Approximately Correct) Model

$\to$ D is the probability distribution on $\mathbb{R}^2$. S is n independent D-samples from $\mathbb{R}^2$ i.e. $S\tilde{\ }D^n$. And E_{out} is estimated assuming real data also obeys D.

$$E_{out}(g) = E_D[g \neq f] = Pr_{x\tilde{\ }D}[g(x) \neq f(x)]$$
$$= \sum_{x \in X, f(x) \neq g(x)} P_D(x) \quad \text{(if D is discrete and } P_D \text{ is its p.m.f.)}$$
$$= \int_{x' \in \mathbb{R}^2} f_D(x) 1_{f \neq g}(x') dx'$$

$\to$ Hope: if x is sampled from D and $E_{in}(g) = 0$, then:

$$Pr_{x\tilde{\ }D}[g(x) \neq f(x)] (\text{w.p.} > 1 - \delta)$$

$\to$ δ is referred to as a **confidence** parameter, while ϵ is referred to as a **accuracy** parameter.

$\to$ We want guarantee for every correct learning algo. (i.e. every h $\in$ H that has $E_S(h) = 0$).

→ Condition for H to be PAC-learnable:

$$P_{S\text{\textasciitilde}D^n}[\forall h \in H \ (E_S(h) = 0) \Rightarrow (E_D(h) \leq \epsilon)] \geq 1 - \delta$$

Examples for H:

- **Ex1.** H = lines in $\mathbb{R}^2$ passing through origin.
- **Ex2.** Any finite H.
- **Ex3.** Any finite X.
- **Ex4.** H = lines in $\mathbb{R}^2$.
- **Ex5.** H = hyperplanes in $\mathbb{R}^n$, where n is finite

→ **Ex2. Any finite H.** Setup:

1. X can be any domain.
2. D be any probability distribution over X.
3. H be any finite class of $\{+1, -1\}$ functions over X.
4. Realisability Assumption: $f : X \to \{+1, -1\}$ be the true labelling and f ∈ H.
5. $\epsilon, \delta \in (0, 1), (Accuracy, Confidence)$

Goal: Prove that $\exists n$ (which may depend in ϵ, δ and H) such that,

$$P_{S\text{\textasciitilde}D^n}[\forall h \in H \ (E_S(h) = 0) \Rightarrow (E_D(h) \leq \epsilon)] \geq 1 - \delta$$

$$i.e. \ P_{S\text{\textasciitilde}D^n}[\exists h \in H \ (E_S(h) = 0) \wedge (E_D(h) > \epsilon)] < \delta$$

Proof: S $\text{\textasciitilde}D^n$
Let h be a fixed hypothesis s.t. $E_D(h) > \epsilon$

$$P_{S\text{\textasciitilde}D^n}[(E_S(h) = 0) \wedge (E_D(h) > \epsilon)] < \delta$$

$$\text{Region of Disagreement} = h \triangle f = \{x \in X : h(x) \neq f(x) = (Good_f \neq Good_h)\}$$

$$E_D(h) > \epsilon \text{ is same as } P_D(h \triangle f) > \epsilon$$

If $E_S(h) = 0 \Rightarrow$ S contain no point from $h \triangle f = (S \cap (h \triangle f) = \phi)$
Let $H_\epsilon \subseteq H$ be all those hypothesis s.t. $E_D > \epsilon$

$$H_\epsilon = \{h \in H : E_D(h) > \epsilon\}$$

Let $h \in H_\epsilon$,

$$\begin{aligned} P_{S\text{\textasciitilde}D^n}[(E_S(h) = 0)] &\leq P_{S\text{\textasciitilde}D^n}[S \cap (h \triangle f) = \phi] \\ &\leq (1 - \epsilon)^n \\ P_{S\text{\textasciitilde}D^n}[\exists h \in H_\epsilon : (E_S(h) = 0)] &\leq |H_\epsilon|(1 - \epsilon)^n \\ &\leq |H|(1 - \epsilon)^n \\ &\leq |H|e^{-\epsilon n} \end{aligned}$$

Now we want:

$$|H|e^{-\epsilon n} < \delta$$

which simplifies to:

$$n > (1/\epsilon)(ln|H| + ln(1/\delta))$$

→ **Ex3. Any finite X.**

Obs.: No. of "distinct" hypothesis possible is $\leq 2^{|X|}$

if $|X| = n$, then 2^N boolean functions

If X is finite, then H is finite. Which implies that we can use the proof from Ex2. But you consider two separation to be same if the executing good set is the same.

$h1 = h2$ if $h1^{-1} = h2^{(} - 1)$

But, this is almost useless:

$$n > (1/\epsilon)(ln(2^{|X|}) + ln(1/\delta)) = (1/\epsilon)(|X| + ln(1/\delta))$$

Better estimation method:

H = set of lines in $\mathbb{R}^2$

$X \subseteq \mathbb{R}^2, |X| = N$

To find: # distinct hypothesis in H?

distinct good sets among N points in $\mathbb{R}^2$ that can be carved out by straight lines

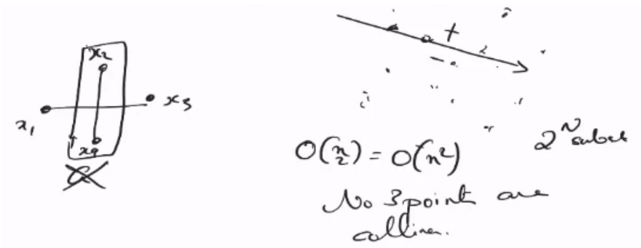

Fig 2.0 Proof by Combinatorics

(Exercise:) Ans: $\binom{n+1}{2} + 1$, using induction on n.

$$\binom{n+1}{2} + 1 \leq n^2$$

So now,

$$n > (1/\epsilon)(2ln(N) + ln(1/\delta))$$