Foundations of Data Science & Machine Learning

Summary — Week 03
Devansh Singh Rathore
111701011

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

March 14, 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 Support Vector Machines (SVM)

→ Is one separating hyperplane better than other?

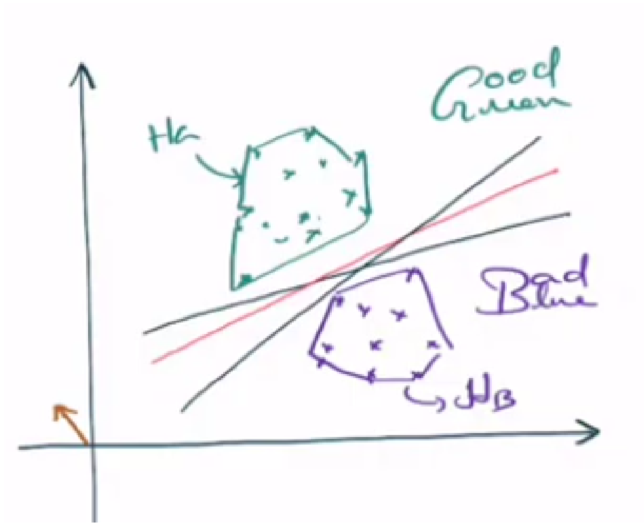

Fig 1.0 Separating hyperplanes

→ In general, we try to find the hyperplane which is at the maximum distance from all the points i.e. $x \in G \cup B$.

→ If a hyperplane equations depend on a and b where $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ and $||a|| = 1$.

Then we have to $Maximise_{a,b}(Min_x| \langle x, a\rangle - b|)$ among all the hyperplanes.

1.1 Maximum - Margin Separating HyperPlane (MMSHP)

→ We also need to ensure that the hyperplane should be separating Good and Bad points i.e. lies between the H_G and H_B.

Definition: For two linearly separable sets G, B $\in \mathbb{R}^n$, the **Maximum - Margin Separating HyperPlane (MMSHP)** is defined as $a \in \mathbb{R}^n$, $||a|| = 1$ and $b \in \mathbb{R}$ which maximises:

$$min_{x \in G \cup B}\,(\langle x, a\rangle - b)f(x)$$

$$where\; f(x) = \left\{ \begin{smallmatrix} 1, x \in G \\ -1, x \in B \end{smallmatrix} \right.$$

→ Perceptron algorithm finds one of the separating hyperplanes which need not be the MMSHP. While on the other hand, SVM finds MMSHP.

→ Intutively, MMSHP is unique and separates the Good and Bad points in the most efficient manner. We will discuss this in later part of summary.

Input: G, B, where $X := G \cup B$ and $\delta : X \to \{1, -1\}$
Objective: $Maximise_{(a,b)}\ Min_{x \in X}(\langle x, a\rangle - b)f(x)$
Constraint: $||a|| = 1$

→ Slight modification:
→ Idea is to collect the (a,b) with **margin** ($= Min_{x \in X}(\langle x, a\rangle - b)f(x)) \geq 1$ and pick the tuple with minimum $||a||$.

$$\forall x \in G,\ margin\ = \langle x, a\rangle - b \geq 1$$

$$Dist(x, l) = \langle x, a/||a||\rangle - b/||a|| \geq 1/||a||$$

→ So our new **Objective**: find line l which gives minimum $||a||^2$, to maximise Dist(x, l).
Constraints:

$$\forall x \in G,\ \langle x, a\rangle - b \geq 1$$

$$\forall x \in B,\ \langle x, a\rangle - b \leq -1$$

→ This is solved using **"Quadratic Programming"**.

1.2 SVM with Embedding

→ $\phi : \mathbb{R}^n \to V$
Minimise: $||a||^2$ among $(a \in V)$
Subject to: $(\langle a, \phi(x)\rangle - b)f(x) \geq 1$

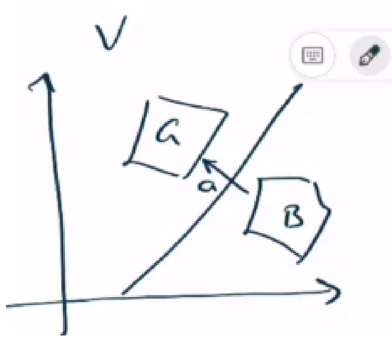

Fig 1.1 Embedding plot

$\mathbb{R}^n : V$

$x_1, x_2, ...x_n : \phi(x_1), \phi(x_2), ...\phi(x_n)$
$a \in \mathbb{R}^n$, $b \in \mathbb{R} : a \in V$, $b \in \mathbb{R}$
$K(x,y) = \langle \phi(x), \phi(y) \rangle$

Key Obs.: 'a' returned by SVM (a for the MMSHP) is linear combination of $\phi(dataVector)$ i.e. $a = \sum_{i=1}^{N} \alpha_i \phi(x_i)$.

$$\begin{aligned}\langle a, \phi(x) \rangle &= \left\langle \sum_{i=1}^{N} \alpha_i \phi(x_i), \phi(x) \right\rangle \\ &= \sum_{i=1}^{N} \alpha_i \langle \phi(x_i), \phi(x) \rangle \\ &= \sum_{i=1}^{N} \alpha_i K(x_i, x_j)\end{aligned}$$

$$\begin{aligned}||a||^2 &= \langle a, a \rangle \\ &= \left\langle \sum_{i=1}^{N} \alpha_i \phi(x_i), \sum_{j=1}^{N} \alpha_j \phi(x_j) \right\rangle \\ &= \sum_{i=1}^{N} \alpha_i \left\langle \phi(x_i), \sum_{j=1}^{N} \alpha_j \phi(x_j) \right\rangle \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \langle \phi(x_i), \phi(x_j) \rangle \\ &= \sum_{i,j=1}^{N} \alpha_i \alpha_j K(x_i, x_j)\end{aligned}$$

1.3 SVM with Kernel

Given $X \subseteq \mathbb{R}^n$, $f : X \to \{-1, +1\}$, and a kernel $K : (\mathbb{R}^n \times \mathbb{R}^n) \to \mathbb{R}$.
Minimise $\sum_{i,j=1}^{N} \alpha_i \alpha_j K(x_i, x_j)$, where N is the number of total points.
Subject to: $(\sum_i^N \alpha_i K(x_i, x) - b) f(x) \geq 1$, $\forall x \in X$

Arguments for the **Key Obs.**:

1. Obvious if $\{\phi(x) : x \in X\}$ span V (a $\in$ V)
2. Since the normal to the MMSHP is a difference between two points in the two hulls.
3. a is a linear combination of those embedded points which are nearest to the MMSHP.

$\rightarrow$ These vectors, whose embeddings are meant to build and define MMSHP are called "**Support Vectors**" of MMSHP. Hence the name, SVM.

2 Intro. to Generalisation

$\rightarrow$ We are given the sets of Good points (+1) and the Bad points (-1). By training on the existing data points, we try to find out a rule i.e. 'h'. Thus derived 'h' can be further used to classify and categorize future points as Good or otherwise Bad.

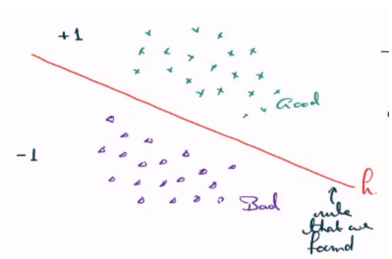

Fig 2.0 General Classification task

$\rightarrow$ Hopefully, 'h' will separate unknown future good and bad points too.

→ Important Questions:

- Will the perfect separator for the test data separate real data reasonably well?
- Will your separator generalise?
- Did you algorithm gain any knowledge from the data?
- When and how well can we learn?
- What are the most lenient conditions under which generalisation happen?
- Can we have quantitative bounds on real world errors under these conditions?

2.1 Supervised Learning Scenario

→ Let's say we have 10000 data points in X, and a function $f : X \to \{+1, -1\}$ in real world.

Now we are given just 100 points from X and is called **Training Sample Set or S**. We also are given the class of each data point i.e. f(x) for those x ∈ S.

If we derive a separating hyperplane (a,b) which separates good and bad points in S i.e. $\forall x \in S$, $g(x) = sign(\langle x, a\rangle - b)$.

We have to check whether the function g works equally well on points in X i.e. can we generalise g on X i.e. g(x) = f(x).

→ Since few errors are acceptable,

$$"In\,Sample\,Error" \text{ or } E_{in}(g) = |\{x \in S : f(x) \neq g(x)\}|/|S|$$
$$"Out\,of\,Sample\,Error" \text{ or } E_{out}(g) = |\{x \in X : f(x) \neq g(x)\}|/|X|$$

→ Is generalisation possible if:

- **"Negation of Realisation Assumption" i.e. f on X is far from linearly separable?** - NO
- **"Sampling Bias" i.e. The 100 points of S is chosen by a malicious "teacher"?** - NO, because we might not even get proper set of points required for training. eg. selecting all 100 points to be good in S.
- **S is chosen by a good "teacher"?** - YES, but its difficult to happen. eg. we can choose points in S from convex hulls of G and B.
- **S is chosen uniformly at random from X i.e. Indifferent teacher?** - YES (to be discussed next.)
- **S is much smaller?** - NO, because size of S matters.

→ **Assumptions for $\mathbb{R}^2$ case:**

1. $X \subseteq \mathbb{R}^2 (10000\,points)$, $f : X \to \{-1, 1\}$
2. (X,f) is linearly separable by a line through origin, called **"Realisation Assumption"**.
3. S is obtained by picking 100 points from X independently and uniformly at random(i.e. with replacement).

→ General Observation: X is separable ⇒ S is separable.
→ We run PLA to find a line defined by a normal 'a' $\in \mathbb{R}^2$. This line classifies S correctly.

$$g(x) = sign(\langle x, a \rangle),\ g : \mathbb{R}^2 \rightarrow \mathbb{R}$$
$$E_{out}(g) = |\{x \in X : f(x) \neq g(x)\}|/|X| << \epsilon << 1$$

→ **Claim:** $E_{out}(g) \leq \epsilon$ with high probability.

$$P(E_{out}(g) \leq \epsilon) \geq 1 - \delta \ \ (where\, \delta << 1)$$

→ **Analysis:** $P(E_{out}(g) > \epsilon)$
So $E_{out}(g) > \epsilon$ under two types of cases:

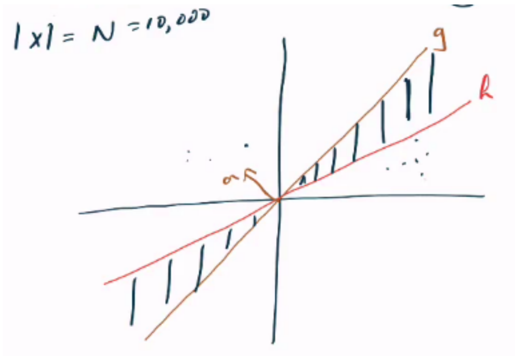

Fig 2.1 Case(a).

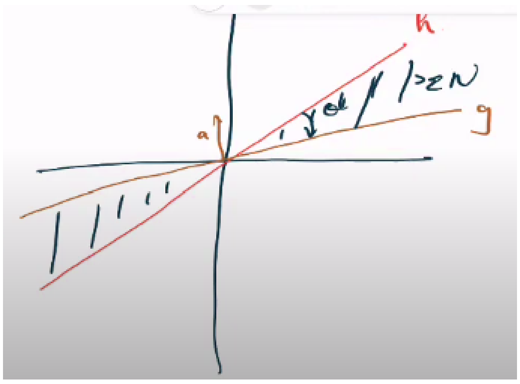

Fig 2.2 Case(b).

→ The case (a) could have occurred probably because the points lying in shaded region are not there in S. And using prob., there are $\epsilon.N$ points of X in the shaded region.
→ For case (a),

$$P(PLA\, outputs\, g) = the\, prob.\, that\, no\, point\, from\, the\, shaded\, region\, is\, chosen$$
$$= (1 - \epsilon)^{100}$$

→ Consider theta (θ) to be the smallest angle so that the shaded region has $> \epsilon.N$ points.

$$P(PLA\, outputs\, a\, line\, which\, is\, at\, an\, angle \geq \theta) \leq (1 - \epsilon)^{100}$$

→ From case (a) and (b), $E_{out}(g) > \epsilon$ only if PLA outputs a separating line which has $\geq \theta$ counterclockwise rotated or otherwise $\geq \theta'$ clockwise rotated from 'h'.
P$(E_{out}(g) > \epsilon) \leq (1 - \epsilon)^{100} + (1 - \epsilon)^{100}$
$\leq 2(1 - \epsilon)^{100}$
$\leq 2e^{-100\epsilon} \leq \delta$