CS717: Class 2

Introduction to SVM:

Data D

$\langle x_1, y_1 \rangle$ . . .

$\langle x_2, y_2 \rangle$.

$\langle x_n, y_n \rangle$ . .

$\longrightarrow$ build a model M s.t

Assume that $y(x)$ is ideal

$x \longrightarrow M(x)$

$M(x)$ is as accurate as possible w.r.t $(y(x))$

n pairs of <input, expected o/p>

Eg: of x_i →
- Document (text)
- molecule
- student
- Course

x_i: Complex objects

- spam/not spam
- carcenogenic/not
- career choice
- rating

y_i: labels/numbers

We would need some signals on each x to characterize the x & build a model M.

Documents $\longrightarrow$ words or sequences of words or sets of words or word occuring in a specific section

One part of Relational learning is about either implicitly capturing or enumerating these signals

OR

explicitly identifying the "good" signals

Note: These signals have different levels of expressivity.

Eg: a → an individual word

ab → sequence

{a, b} → set (ignore order of occurrence)

a [3] b → subsequence (between set & seqnce)

gap of upto 3

in_title (a) → predicate logic (first order logic)

in_section (title, a)

more general instance of pred. logic

Each of These signals will be denoted by $\phi_j(x_i)$

$x_i \rightarrow [\phi_1(x_i), \phi_2(x_i) \ldots \phi_m(x_i)]$

i.e m "signals" OR features/attrbs

Design of $\phi_j(x_i)$ can happen either

By domain expert

Induced using Relational Learning [Learn composite ϕ's from simple ϕ's]

sets | sequences | subsequences etc of words

word

$\phi_j(x_i)$ could be
- discrete valued $\in \{v_1 \ldots v_k\}$
- binary $\in \{0, 1\}$
- integer valued (# of C–H bonds)
- Real valued (say size of molecule)

Say you want to represent a "sequence" feature ("a", "aa", "ab", "ac" . .)

① $\phi_{\text{"aa"}}(x_i) \in \{0, 1\}$

② $\phi_{\text{string}}(x_i) \in \{\text{"a"}, \text{"aa"}, \ldots \text{"ac"}\}$

(prob: Dealing with multiple sequences occurring)

③ $\phi_{\text{string}}(x_i) \in 2^{\{\text{"a"}, \text{"aa"} \ldots \text{"ac"}\}}$

(problem: sparsity)

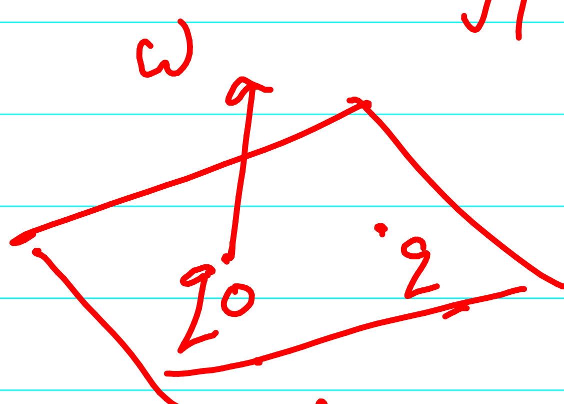
$\hookrightarrow$ 4 $\phi_{\text{"}a'a\text{"}}(x_i) \in \mathbb{R}$

Svm idea:

1) We will assume $\phi_j(x_i) \in \mathbb{R}$.

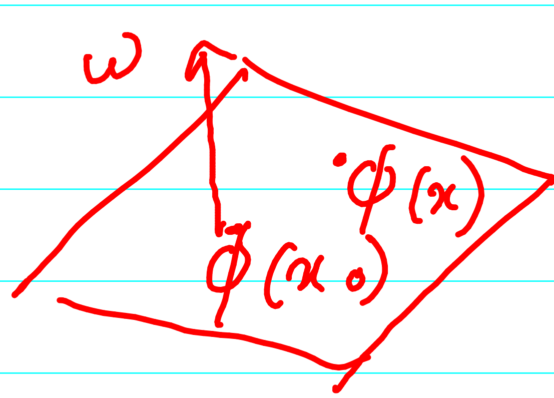
2) We will assume that model $\mathcal{M}$ is a hyperplane

$\mathcal{H} = \{ q \mid (q - q_0)^T w = 0 \}$



More specifically, for svm

$\mathcal{H} = \{ \phi(x) \mid (\phi(x) - \phi(x_0))^T w = 0 \}$



Given a new point x

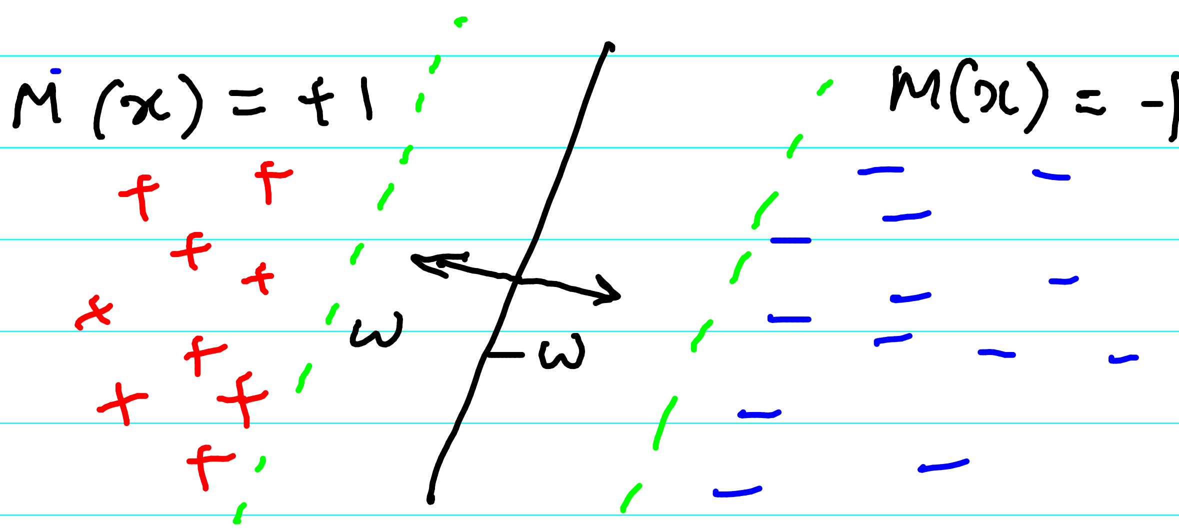
SVM classification model says: if $(\phi(x) - \phi(x_0))^T \omega$

≥ 0 → $M(x) = +1$

< 0 → $M(x) = -1$

In other words:

$\phi(x)^T \omega - \phi^T(x_0)\omega$ is our decision function (model decision fn)



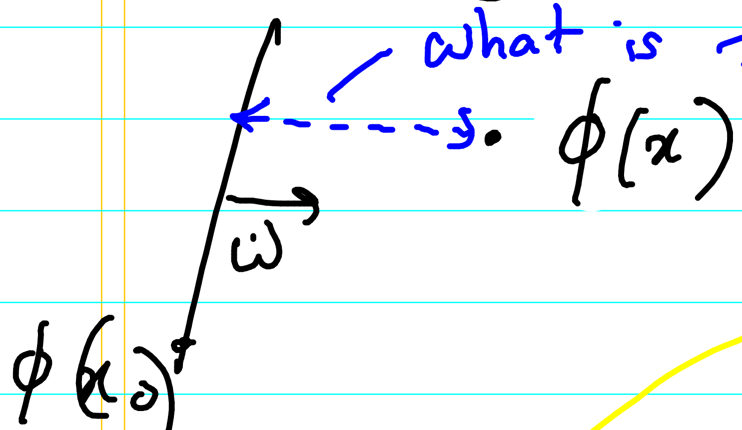
SVM: Learn the green margin s.t. no +/- ve pt enters the margin Advtg

"Advtg" → Good generalization from one data to another

→ Statistical learning theory

Basic formulation

Maximize the distance of the "closest" training pt from the separating surface



what is this distance?

Ans: $\dfrac{(\phi(x) - \phi(x_0))^T w}{\|w\|} y$

H/W.

$y = \pm 1$ is the class of x

Multiplication of signed distance by y gives unsigned distance

We will substitute $\phi^T(x_0).w = -w_0$

SVM decision function $M(x)$ or $f(x)$ is $\phi^T(x_0) w + w_0$

& SVM optimization objective

$$\max_{w, w_0} \min_{x \in D} y \frac{(\phi^T(x) w + w_0)}{\|w\|} \qquad (*)$$

Q: If ω, ω_0 is solution to (*),

$\lambda\omega, \lambda\omega_0$ is also a solution to (*)

Solution 1: $||w|| = 1$ (constrain space of w)

Solution 2: Put a constraint on (ω, ω_0) s.t inner optimization problem is resolved.

$$\min_{x \in D} \frac{y(\phi^T(x)\omega + \omega_0)}{||w||} = \frac{1}{||w||}$$

i.e $\quad y(\phi^T(x)\omega + \omega_0) \geq 1$

SVM objective (hard margin)

$$\max_{\omega, \omega_0} \frac{1}{||w||}$$

s.t $\forall x \in D \quad y(\phi^T(x)\,\omega + \omega_0) \geq 1$

[H/W: Understand intuitively that Karush Kuhn Tucker Optimality conditions from the notes & link(s) below:

http://www.cse.iitb.ac.in/~cs717/notes/classNotes/BasicsOfConvexOptimization.pdf

All references to sections, theorems will be with respect to these notes. It will be denoted for instance as CoOpt,4.1.1

For some basic definitions of optimization, see sections *CoOpt,4.1.1* and *CoOpt,4.1.2*. To find the maximum minimum values of an objective function, see section *CoOpt,4.1.3* from Theorem 38 to Theorem 56. A set of systematic procedures to find optimal value is listed in Procedures 1, 2, 3 and 4 (*CoOpt, pages 10, 16, 17*).

The notion of *directional derivative* (D), can be obtianed from *CoOpt, Section 4.1.4* and that of level curves from *CoOpt, Figure 4.12* and the associated explanation. Briefly, for a function $f(\bar{x})$, defined on domain D, is given by

$$\text{Level Curve } (C) = \{\bar{x} \mid \bar{x} \in D, f(\bar{x}) = c\}$$

The level curves applet can be found at `http://www.slu.edu/classes/maymk/banchoff/LevelCurve.html`.

CoOpt, Theorem 62 and Corollary 63 list the necessary and sufficient conditions for local maxima (resp. minima) as follows:

For definitions of relative boundary, relative interior and an example concerning the same, refer to *CoOpt, pg 250, 251.*

Read *CoOpt, Section 4.4.1* on Lagrange Multipliers for a detailed exposition

Refer to *CoOpt, Section 4.4.4* for a detailed explanation and proofs. of KKT conditions