9/12/2018

ASSIGNMENT - 07

Soft Margin SVM and Kernel

SUBMITTED BY: VINDHYA SINGH ENROLLMENT NO: 03693296
and
WASIQ RUMANEY ENROLLMENT NO: 03694978

Solution 1: Assuming that we have a linearly separable dataset D, on which soft-margin SVM is fitted, we can NOT guarantee that all training samples in D will be assigned correct label by the fitted model.

From slide 55, lecture 6 & 7, we know that,

$$f_o(w, b, \xi) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \quad - (1)$$

where,

$f_o(w, b, \xi)$:= cost function

ξ_i := slack variable

and $C > 0$ determines how heavily a violation is punished.

If there exists some points that are very close to the decision boundary, then misclassifying these data points would lead to a the margin being significantly increased. From (1), the variation would depend on the cost factor C.

Solution 2 [From slide 59, lec 6 & 7,]

C value : How much one wants to avoid misclassifying every training example.

If

$C > 0$, the hyperplane margin will be smaller if the optimisation results in a hyperplane that correctly classifies all the training data points.

Now, if $C < 0$ OR $C = 0$,

Case I : $C < 0$: The optimiser will look for larger margin hyperplanes at the cost of misclassifying certain data points.

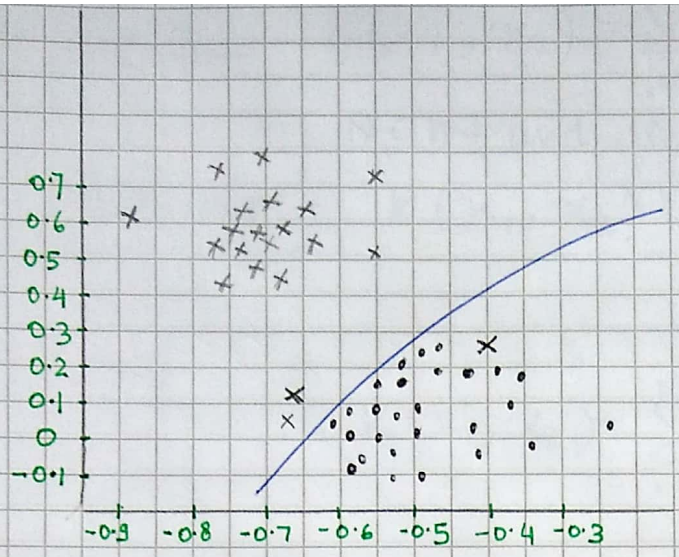
Case II : $C = 0$: So, for any misclassification, there is no penalty imposed. Thus, all constraints will be trivially satisfied by

$$\xi_i = 1 - y_i (w^T x_i + b)$$

$$\{ i : y_i (w^T x_i + b) = 0 \}.$$

Thus, we need to ensure that $C > 0$ in the slack variable formulation of soft-margin SVM.

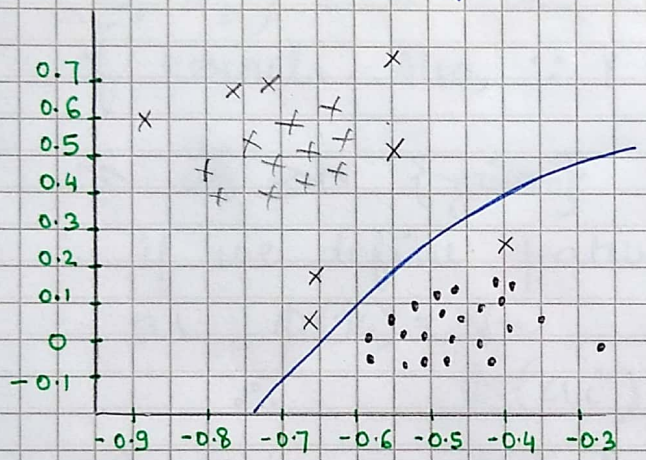
Problem 3



(a) $C = 10^{10}$

Fig 1 : Sketch of decision boundary of an SVM with a Quadratic Kernel.

High value of C means more penalty for misclassified samples. Thus, decision boundary lies closer to the data points. Case of Hard Margin SVM.



(b) $C = 10^{-10}$

Fig 2: Sketch of decision boundary of an SVM with a Quadratic Kernel.

Low value of C in this case. Thus, lower penalty for the misclassified samples implying larger margin. Case of Soft margin SVM.

Sol 4

Given : $K(x_1, x_2) = \sum_{i=1}^{N} a_i (x_1^T x_2)^i + a_0$

$N \in \mathbb{N}$ and $a_i \geq 0$ with $i \in [0, N]$

To Prove : $K(x_1, x_2)$ is a valid kernel.

Proof :

$$K(x_1, x_2) = \sum_{i=1}^{N} a_i (x_1^T x_2)^i + a_0 \quad \{\text{given}\}$$

① $x_1^T x_2$ is a valid kernel because it is a scalar product of input vectors.

② $(x_1^T x_2)^i$ is a valid kernel as, it is a product of kernels. Also, $\because i \in [0, N]$ wher $N \in \mathbb{N}$ {given}

③ $\because a_0 \geq 0$ {given} a_0 is a constant

$\therefore$ if we define feature map $\phi(a_i) = \phi(x_1) \cdot \phi(x_2)$

as $\phi(x) = \sqrt{a}$

$\therefore \quad \phi(a_i) = (\sqrt{a}) \cdot (\sqrt{a}) = (\sqrt{a})^2 = a$

$\Rightarrow a_0$ is a valid kernel.

④ $\because$ Sum of recursive kernels is a kernel

$\therefore \sum_{i=1}^{N} a_i (x_1^T x_2)^i$ is a valid kernel.

⑤ $\because$ Sum of 2 valid kernels is a valid kernel

$\therefore$ From ①, ②, ③, ④ and ⑤,

$$K(x_1, x_2) = \sum_{i=1}^{N} a_i (x_1^T x_2)^i + a_0$$

is a valid kernel.

Hence, Proved

Sol 5

Given : $K(x_1, x_2) = \frac{1}{1 - x_1 x_2}$

with $x_1, x_2 \in (0, 1)$

To find : Feature Transformation $\phi(x)$ corresponding to kernel $K(x_1, x_2)$

Solution : Considering an infinite-dimensional space, i.e using the concept of Hilbert Space,

$$K(x_1, x_2) = \phi(x_1)^T \cdot \phi(x_2)$$
$$= \phi(x_1) \cdot \phi(x_2) \quad \text{---------- (1)}$$

{ $\langle v, w \rangle = v^T w$ Traditional finite vector space inner product }

Defining $\phi(x) = [1, x, x^2, x^3, \ldots]$
i.e infinite dimensional

From (1), and using value of $\phi(x)$ defined above,

$$K(x_1, x_2) = \phi(x_1) \cdot \phi(x_2)$$
$$= [1, x_1, x_1^2, x_1^3, \ldots] \cdot [1, x_2, x_2^2, x_2^3, \ldots]$$
$$= 1 + x_1 x_2 + x_1^2 x_2^2 + x_1^3 x_2^3 + \ldots$$
$$= 1 + x_1 x_2 + (x_1 x_2)^2 + (x_1 x_2)^3 + \ldots$$
$$= \frac{1}{1 - x_1 x_2}$$

$\because$ given that $x_1, x_2 \in (0, 1)$

$\therefore \quad x_1 \cdot x_2 < 1$

A Geometric Series converges to $\frac{1}{1 - x_1 x_2}$ where geometric series is a sum of an infinite.

Sol 6:

a) The algorithm matches every element of the given two character strings x and y of length m and n respectively and returns the number of characters (s) common in both strings.

b) The kernel $k : S \times S \to \mathbb{R}$ (where S = set of strings over a finite alphabet v) on a pair of strings x and y is a valid kernel.

Let,

Σ = alphabet of size v

Σ^n = set of strings of length n.

for a given index i such that $i = (1 \le i_1 < i_2 \cdots < i_r \le |s|)$

where we define $s(i) = s(i_1)s(i_2)\cdots s(i_r)$

and s' be the number of common characters b/w 2 strings. (x ~~and~~ and y in this case)

We define a parameter λ such that $0 \le \lambda \le 1$ and it defines $\left[\phi_n(s)\right]$ a map with $|\Sigma^n|$ components

$$\therefore \quad \left[\phi_n(s)\right]_u = \sum_{i=u} \lambda^{s'}$$

The kernel gives the number of elements common in x and y as a scalar product i.e $x^T y$ which is a valid inner product. Thus, a valid kernel too.

Sol 7.

We know that,

$$K_G(x_1, x_2) = \exp\left(-\frac{|x_1 - x_2|^2}{2\sigma^2}\right)$$

if $\sigma \to 0$

$$K_G(x_1, x_2) = \begin{cases} 1 & , \text{ if } x_1 = x_2 \\ 0 & , \text{ if } x_1 \neq x_2 \end{cases} \quad \text{---(i)}$$

Basically, $K_G(x_1, x_2)$ becomes identity matrix I.

In order to correctly classify data points, the following should hold, i.e,

$$y_i(w^T \phi(x_i) + b) > 0 \quad \text{---(ii)}$$

i.e $$y_i\left(\sum_j y_j \alpha_j \phi(x_j)^T \phi(x_i) + b\right) > 0$$

{ substituting $w^T = \sum_j y_j \alpha_j \phi(x_j)^T$ in (ii) }.

$$= y_i\left(\sum_j y_j \alpha_j K(x_i, x_j) + b\right) > 0 \quad \{ \because K(x_1, x_2) = \phi(x_1)^T \phi(x_2)$$

Thus, from (i), the above becomes.

$$y_i^2 \alpha_i + y_i b > 0$$

If in the above inequality, if $\alpha_i > 0$ and $b = 0$ for all i, then every vector is a support vector as $(\alpha > 0)$.

Hence, we can say that if the variance chosen is small enough then all finite sets of points can be linearly separated using the Gaussian Kernel.