MUTHUKUMARASWAMY, SHWEKA
5042889.

ASSIGNMENT-2

1. (a) $K_1, K_2, K_3, K_4, \ldots, K_m$ are all valid kernels.

$w_j \geq 0$.

To Prove :- $\sum_{j=1}^{M} w_j K_j$ is also a valid kernel.

To prove that a kernel is valid we need to prove the following two properties,

(1) Symmentric (2) Positive semi definite.

(1) SYMMENTRIC:-

$$K(x,y) = \sum_{j=1}^{M} w_j K_j(x,y).$$

(ie) $K(x,y) = w_1 K_1(x,y) + w_2 K_2(x,y) + w_3 K_3(x,y) + \cdots + w_m K_m(x,y)$. — (1)

$K(y,x) = w_1 K_1(y,x) + w_2 K_2(y,x) + w_3 K_3(y,x) + \cdots + w_m K_m(y,x)$

$\because K_1, K_2, K_3, \ldots K_m$ are all valid kernels,

$K_1, K_2, K_3 \ldots K_m$ are symmentric

$\therefore K_1(y,x) = K_1(x,y)$; similarly for other K kernels.

$\therefore K(y,x) = w_1 K_1(x,y) + w_2 K_2(x,y) + w_3 K_3(x,y) + \cdots + w_m K_m(x,y)$ — (2)

$\Rightarrow K(y,x) = K(x,y)$ (comparing equation (1) and (2))

(2) POSITIVE SEMI DEFINITE :-

To prove this property we have to prove that,

$$K = u^T G u$$

where u is any random matrix, G is the Gram matrix.

Let us multiply K by $u^T u$ as follows,

$$u^T K u = u^T \left(\sum_{j=1}^{M} w_j G_j \right) u$$

$$= u^T (w_1 G_1) u + u^T (w_2 G_2) u + \cdots + u^T (w_m G_m) u$$

$$= w_1 (u^T G_1 u) + w_2 (u^T G_2 u) + \cdots + w_m (u^T G_m u)$$

$\because$ ($w_1, w_2, \ldots w_m$ are all scalars).

$\because$ $K_1, K_2, K_3, \ldots K_m$ are all valid kernels.

$$u^T G_1 u ; u^T G_2 u , u^T G_3 u \ldots \geq 0 \quad \text{—— (3)}$$

Also it is given that $w_1, w_2 \ldots w_m \geq 0$ —— (4)

Multiplication of 3 and 4 ≥ 0

$\therefore$ The sum of all such term ≥ 0.

$$\therefore \sum_{j=1}^{M} w_j u^T G_j u \geq 0$$

$$\Rightarrow u^T G u \geq 0$$

$\therefore$ K is positive semi definite

$\therefore$ The Kernel k is valid.

ⓑ To prove Hadamard Product is a kernel.

So that $K(x_i, x_j) = k_1(x_i, x_j) k_2(x_i, x_j)$

Kernels k_1 and k_2 are valid (given)

$\Rightarrow$ k_1 and k_2 are both positive semi definite and symmetric

Any matrix that is positive semidefinite can be assumed as a covariance matrix.

$\therefore$ Let the kernels ~~of~~ 1 and 2 be a covariance matrix.

(ii) Let the distribution $(u_1, u_2, u_3, \ldots u_n)$ have covariance as G_1 and we can assume without any loss of generality that the mean is zero and similarly let $(v_1, v_2, v_3, v_4, \ldots v_n)$ be represented with mean 0 and covariance G_2

Now consider

$$(w_1, w_2, w_3 \ldots, w_n) = (u_1 v_1, u_2 v_2, u_3 v_3 \ldots u_n v_n)$$

Then the covariance,

$$G = E[(\omega-\mu_\omega)(\omega-\mu_\omega)^T]$$
$$= E[\omega\omega^T]$$
$$= E[(uv)(uv)^T]$$
$$= E[uvv^Tu^T] = E[(uu^T)\cdot(vv^T)] \quad \text{--- (1)}$$

$\because$ The distribution considered is independent

(1) can be written as,

$$= E[uu^T]\cdot E[vv^T].$$

$G = G_1 \cdot G_2$ (from the assumption)

$\Rightarrow$ Product of the two covariance matrices is also a positive semi definite matrix.

$\therefore$ Hadamard product of 2 kernels is positive semi definite.

$\therefore$ Hadamard Product of 2 valid kernel is a valid kernel.

(c) $k(x,x') = (xx'+1)^{2015}$

<u>Symmentric:-</u>

$$K(x,x') = (xx'+1)^{2015}$$
$$K(x',x) = (x'x+1)^{2015}$$

By property of multiplication,

$$K(x',x) = (xx'+1)^{2015} = K(x,x')$$

$\therefore$ symmentric

A Kernel is a valid kernel if it can be expressed in the form $\phi(x)^T\cdot\phi(x')$

Consider the given kernel,

$(xx'+1)^{2015}$

Using Binomial expansion formula,

$$(a+x)^n = \sum_{k=0}^{n} \frac{n!}{(n-k)!\,k!} a^{n-k}.x^k.$$

$$\Rightarrow (xx'+1)^{2015} = \sum_{k=0}^{n} \frac{n!}{(n-k)!\,k!} (xx')^{n-k}.(1)^k.$$

$\frac{n!}{(n-k)!\,k!}$ is a constant. Let us represent it as $C=[c_1, c_2, \ldots c_n]$

$$\Rightarrow (xx'+1)^{2015} = [C] \sum_{k=0}^{n} (xx')^{n-k}. \quad (1^k = 1 \ \forall k=0,\ldots n)$$

This can be written as,

$$(xx'+1)^{2015} = [c] \sum_{k=0}^{n} (x)^{n-k} (x')^{n-k} \quad \text{—— (1)}$$

Consider vector notation

$$\phi(x) = [1, x, x^2, x^3, x^4 \ldots x^n]$$

$$\phi^T(x_1) = \begin{bmatrix} 1 \\ x_1 \\ x_1^2 \\ \vdots \\ x_1^n \end{bmatrix}$$ (considering $x_1 = x'$ for the ease of calculation)

$$\phi(x)\,\phi^T(x_1) = 1 + xx_1 + x^2x_1^2 + \cdots x_n x_n^2$$

$$= \sum_{k=0}^{n} x^k . x_1^k .$$

This is same as 1 formula (reverse order) (summation part)

Bring c inside

Assume $\phi(x) = [\sqrt{c_0}, \sqrt{c_1}\,x, \sqrt{c_2}\,x^2 \ldots]$ —— (3)

and $\phi(x_1) = [\sqrt{c_0}, \sqrt{c_1}\,x_1, \sqrt{c_2}\,x_1^2 \ldots]$ —— (4)

The representations (3) and (4) will give the required formula.

$\therefore (xx'+1)^{2015} = \phi(x) \cdot \phi^T(x')$

$\therefore$ It is a valid kernel.

(d) $k(x,x') = \exp\left(\frac{-(x-x')^2}{2}\right)$

$$= \exp\left(-\frac{(x^2+x'^2-2xx')}{2}\right)$$

$$= \exp\left(\frac{-x^2}{2}\right) \cdot \exp\left(\frac{-x'^2}{2}\right) \cdot \exp(xx')$$

$$= \underbrace{g(x) \cdot g(x')}_{K_1} \cdot \underbrace{\exp(K_1(x,x'))}_{K_2}.$$

consider K_2,

$$\exp(x) = \lim_{A\to\infty}\left(1+x+\frac{x^2}{2!}+\cdots\frac{x^A}{A!}\right)$$

$\therefore$ $\exp(x)$ can be expressed as a sum and product of Polynomial and it has all positive co-efficients.

$\therefore$ From Proofs 1(a) and 1(b)

$\exp(K_1(x,x'))$ is also a valid kernel.

$K_1 = g(x) \cdot g(x')$

By definition of kernels

$K_2 = \langle \phi(x), \phi(x') \rangle$ (dot Product notation)

$K(x,x') = \langle g(x)\phi(x), g(x')\phi(x') \rangle$

let $\phi'(x) = g(x)\phi(x)$ and

$\phi'(x') = g(x')\phi(x')$

$K(x,x') = \langle \phi'(x), \phi'(x') \rangle$ which is a valid kernel

Also, $K(x,x') = \underbrace{g(x)\cdot g(x')}_{K_1(x)} \cdot K_2(x)$

Here the kernel $K_1(x)$ can be expressed as the outer product of two vectors and hence it is symmentric and positive semi definite.

And hence both $K_1(x)$ and $K_2(x)$ are valid kernels

$\therefore K(x,x')$ is also a valid kernel.

Citation :- cs.berkeley.edu

② Summary Of Methodused.

Objective Function:-

Maximize the following objective function.

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{i=1}^{l} y_i y_j K(\vec{x_i}, \vec{x_j}) \alpha_i \alpha_j$$

METHOD:-

The maximisation is done according to SMO (Sequential Minimal Optimisation method).

(1) The method chooses two α's at any given iteration and will optimize the equation for those 2 values and will Proceed thus till all the values are optimised.

* One α chosen is chosen in order till we finish traversing through the loop once. Then we take values that are not 'α' or '0'.

* The second alpha is considered based on the following heuristic.

* Find all non-zero, non-c alpha and Pick the One with highest value / lowest value of α based on $|E_2-E_1|$

If that step does not work then loop through α's and try will all indices.

If that does not work loop through all Possible values.

Now the quadratic programming problem we have to solve has only two variables.

To calculate new α_1 and α_2 = (subproblem solution)

Error is calculated as = SVM output at that point (say i_1) $- y(i_1)$

$$\alpha_2^{new'} = \alpha_2^{old} - \frac{y_2(E_2 - E_1)}{\eta} \quad \text{——} \ (1)$$

where η is the second derivative and it is given as,

$$\eta = 2K(\vec{x_1}, \vec{x_2}) - K(\vec{x_1}, \vec{x_1}) - K(\vec{x_2}, \vec{x_2})$$

For us to get maximum value second derivative must be less than zero.

So check if the value of η is lesser than zero

If it is lesser than formula (1) holds or we have to set it to L or H depending on the condition,

$$\alpha_2^{new} = \begin{cases} H & \text{if } \alpha_2^{new'} \geq H \\ L & \text{if } \alpha_2^{new'} \leq L \end{cases}$$

where L is given as,

$$L = \max(0, \alpha_2^{old} - \alpha_1^{old}) \quad H = \min(C, C + \alpha_2^{old} - \alpha_1^{old}) \quad y_1 \neq y_2$$

$$L = \max(0, \alpha_1^{old} + \alpha_2^{old} - C) \quad H = \min(C, \alpha_1^{old} + \alpha_2^{old}) \quad y_1 = y_2.$$

Then update α_1 using the formula,

$$\alpha_1^{new} = \alpha_1^{old} + s(\alpha_2^{old} - \alpha_2^{new})$$

This Procedure is continued till convergence is reached
It is also proved in the paper that the sub-problem cannot be made more smaller (ie) you need atleast 2 equations to get an optimal solution.

(3) 2 class classification with labels set as ± 1 can be solved by the algorithm.

The aim/objective function is given as,

$$\min_{w} \frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_{(x,y) \in S} \ell(w; (x,y))$$

$$\text{where } \ell(w,(x,y)) = \max\{0, 1 - y\langle w, x\rangle\}$$

where w represents the weight matrix,
x represents the training data and
y represents the label

WORKING:-

The algorithm considers only a part of the data-set available and it finds gradient descent for that and takes decision accordingly.

A set of data is chosen at random and we check if it satisfies the condition,

$A_t^+ \rightarrow y\langle w, x\rangle < 1.$ ← sets that suffer nonzeroloss.

Then to update w we use the formula,

$$w_{t+\frac{1}{2}} = (1 - \eta_t \lambda) w_t + \frac{\eta_t}{K} \sum_{(x,y) \in A_t^+} y x.$$

$$w_{t+1} = \min\left\{1, \frac{1/\sqrt{\lambda}}{\| w_{t+1/2} \|}\right\} w_{t+\frac{1}{2}}$$

This is done till the norm difference (ie) $\|w_{t+1} - w_t\| < \epsilon$.