E0 270: MACHINE LEARNING (JAN-APRIL 2024)

PROBLEM SHEET #2
INDIAN INSTITUTE OF SCIENCE

1. Suppose we are given a dataset $D = \{x_n, y_n\}_{n=1}^N$. For solving the SVM problem on this dataset, suppose the four support vectors $x_{n_1}, x_{n_2}, x_{n_3}, x_{n_4}$ are given a priori. Is it possible to directly solve the SVM problem using these support vectors without needing to use the whole dataset? Justify.

   Solution:
   We have the optimization problem

$$\min_{w,b} \max_{\alpha \geq 0} \frac{||w||^2}{2} + \sum_{n=1}^{N} \alpha_n \left(1 - y_n \left(w^T x_n + b\right)\right).$$

   For a feasible solution, $1 - y_n \left(w^T x_n + b\right) \leq 0$, and so for x_n that are not support vectors, the expression is strictly negative and maximization w.r.t α means that the corresponding $\alpha_n = 0$. Therefore, the value of w given by

$$w = \sum_{n=1}^{N} \alpha_n y_n x_n.$$

   depends only on the support vectors since all the other terms become zero.

2. Let $Z(.)$ denote the vector of monomials upto degree d, i.e,

$$Z(x) = \begin{pmatrix} 1 \\ x \\ x^2 \\ \vdots \\ \vdots \\ x^d \end{pmatrix}.$$

   Let Q be a symmetric positive definite matrix. Is the following function a valid kernel? Justify.

$$k(x, y) = Z(x - y)^T Q Z(x - y).$$

   Solution:

The function is not a kernel since it is not symmetric. Let $d = 1$ and $Q = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, then

$$\begin{aligned} k(x,y) &= Z(x-y)^T Z(x-y) \\ &= \begin{pmatrix} 1 & x-y \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ x-y \end{pmatrix} \\ &= \begin{pmatrix} 1 & x-y \end{pmatrix} \begin{pmatrix} 2+x-y \\ 1+2x-2y \end{pmatrix} \\ &= 2 + 3(x-y) + (x-y)^2 + 1 + 2(x-y) + (x-y) + 2(x-y)^2 \\ &= 3 + 6(x-y) + 3(x-y)^2. \end{aligned}$$

3. Let $\mathcal{X} = [0, \frac{\pi}{2}]^2$. Is the following function defined on $\mathcal{X} \times \mathcal{X}$ a kernel? Justify.

$k(x,y) = \cos(x_1 - y_1)\cos(x_2 - y_2)$, $\forall x = (x_1, x_2), y = (y_1, y_2) \in \mathcal{X}$.

Solution:

For dataset $D = \{x^1, \ldots, x^N\}$, and $i \neq j$,

$$\begin{aligned} k(x^i, x^j) &= \cos(x_1^i - x_1^j)\cos(x_2^i - x_2^j) \\ &= \left(\cos x_1^i \cos x_1^j + \sin x_1^i \sin x_1^j\right)\left(\cos x_2^i \cos x_2^j + \sin x_2^i \sin x_2^j\right) \\ &= \begin{pmatrix} \cos x_1^i \\ \sin x_1^i \end{pmatrix}^T \begin{pmatrix} \cos x_1^j \\ \sin x_1^j \end{pmatrix} \quad \begin{pmatrix} \cos x_2^i \\ \sin x_2^i \end{pmatrix}^T \begin{pmatrix} \cos x_2^j \\ \sin x_2^j \end{pmatrix} \end{aligned}$$

Consider only the first dot product for now. Each (i,j)'th element of the kernel Gram matrix is due to the dot product corresponding to points x_i and x_j, and so can be written as

$$K_1 = \begin{pmatrix} \cos x_1^1 & \sin x_1^1 \\ \cos x_1^2 & \sin x_1^2 \\ \cos x_1^3 & \sin x_1^3 \\ \vdots & \vdots \\ \cos x_1^i & \sin x_1^i \\ \vdots & \vdots \\ \cos x_1^{N-1} & \sin x_1^{N-1} \\ \cos x_1^N & \sin x_1^N \end{pmatrix} \begin{pmatrix} \cos x_1^1 & \cos x_1^2 & \cos x_1^3 & : & \cos x_1^i & : & \cos x_1^{N-1} & \cos x_1^N \\ \sin x_1^1 & \sin x_1^2 & \sin x_1^3 & : & \sin x_1^i & : & \sin x_1^{N-1} & \sin x_1^N \end{pmatrix}$$

$$= M_1^T M_1.$$

Similarly,

$$K_2 = M_2^T M_2,$$

and the Gram matrix is $K = K_1 \odot K_2$ is positive semidefinite.

4. l_2 norm soft margin SVM:

   When introducing slack variables to soft margin SVM, instead of adding ξ_n's to the objective function, we can also add ξ_n^2, giving rise to the following optimization problem

$$\begin{aligned} \min_{w,b,\xi} \ & \frac{1}{2}||w||^2 + \frac{c}{2}\sum_{n=1}^{N}\xi_n^2 \\ \text{s.t.} \ & y_n\left(w^T x_n + b\right) \geq 1 - \xi_n, \ n \in [N] \end{aligned}$$

   (a) Compared to the standard soft margin SVM formulation, we have dropped the extra set of constraints $\xi_n \geq 0$. Show that these non-negativity constraints can be removed without affecting the optimal solution.
   (b) Write the Lagrangian for the above optimization problem.

   Solution:
   (a) Proof by contradiction: Suppose there is a feasible and optimal solution for which some $\xi_n < 0$. Then since the corresponding constraint is satisfied for $\xi_n < 0$, it will also be satisfied for $\xi_n = 0$.
   Therefore, the ξ_n value can be increased to zero without violating any constraints, decreasing the objective function due to the ξ_n^2 term, contradicting the assumption that the solution was optimal in the first place.
   (b)

$$\min_{w,b,\xi}\max_{\alpha\geq 0} L(w,b,\xi,\alpha_n) = \frac{1}{2}||w||^2 + \frac{c}{2}\sum_{n=1}^{N}\xi_n^2 + \sum_{n=1}^{N}\alpha_n\left[1 - \xi_n - y_n\left(w^T x_n + b\right)\right].$$

5. Given a function $y = f(x)$ with x and y taking scalar values, and loss function L. For values $x = \ln(10)$ and $\frac{dL}{dy} = 1$, answer the following questions:
   (a) What will be gradient of loss function with respect to x if f is sigmoid, $f(x) = \frac{1}{1+e^{-x}}$?
   (b) What will be gradient of loss function with respect to x if f is ReLU, $f(x) = max(0, x)$?
   (c) If you build a multi-layered (deep) neural network, what will be your preferred activation function between sigmoid and ReLU? Explain your answer based on the observation made from the previous questions.

   Solution:
   (a) $\frac{dL}{dx} = \frac{dL}{dy}\frac{dy}{dx} = f\left(\ln(10)\right)\left(1 - f\left(\ln(10)\right)\right) = \frac{10}{120}$.
   (b) Since $\ln(10) > 0$, $f(x) = x$ for $x = \ln(10)$, and $\frac{dL}{dx} = 1$.

(c) ReLU is preferred for multi-layered networks because when backpropagating, the gradient reduces in each layer due to the sigmoid resulting in a vanishing gradient in deep networks.