Com S 573 Spring Semester 2021
Machine Learning

Homework 2

Instructor: Hongyang Gao Submitted by: Ibne Farabi Shihab

Question number 1

Given that,

$L(w, \alpha) = \frac{1}{2}W^TW + \sum_{i=1}^{N} \alpha_i(1 - y_i(W^Tx_i))$

$x, y = ([0, 0], -1), ([2, 2], -1), ([2, 0], +1)$

Adding values to the equation,

$L(w, \alpha) = \frac{1}{2}(w_1^2 + w_2^2) + \alpha(1 - (-1)(0 \times w_1 + 0 \times w_2) + \alpha(1 - (-1)(2 \times w_1 +$

$2 \times w_2) + \alpha(1 - (+1)(2 \times w_1 + 0 \times w_2)$

$L(w, \alpha) = \frac{1}{2}(w_1^2 + w_2^2) + \alpha(3 + 1 \times 0 + 2w_1 + 2w_2 - 2w_1)$

$L(w, \alpha) = \frac{1}{2}(w_1^2 + w_2^2) + \alpha(3 + 2w_2)$

$\frac{\partial L(w,\alpha)}{\partial w_2} = w_2 + 2\alpha = 0$

$\frac{\partial L(w,\alpha)}{\partial w_1} = w_1 = 0$

Using the value of w_1 and w_2 we get,

$w_2 = -2\alpha$

$\frac{\partial L(w,\alpha)}{\partial \alpha} = 3 + 2w_2 = 0$

$0 = 3 - 4\alpha$

$\alpha = \frac{3}{4}$

Question 2

Let us consider three points -

$(x, y) = ([0, 0], +1), ([1, 0], +1), ([0, 1], -1)$

$L(w, \alpha) = \frac{1}{2}(w_1^2 + w_2^2) + \alpha(1 - (-1)(0 \times w_1 + 0 \times w_2) + \alpha(1 - (-1)(1 \times w_1 +$

$0 \times w_2) + \alpha(1 - (+1)(0 \times w_1 + 1 \times w_2)$

$L(w, \alpha) = \frac{1}{2}(w_1^2 + w_2^2) + \alpha(1 - 0) + \alpha(1 - w_1) + \alpha(1 + w_2)$

$L(w, \alpha) = \frac{1}{2}(w_1^2 + w_2^2) + \alpha(3 - w_1 + w_2)$

$\frac{\partial L(w,\alpha)}{\partial w_1} = w_1 - \alpha = 0$

$w_1 = \alpha$

$\frac{\partial L(w,\alpha)}{\partial w_2} = w_2 + \alpha = 0$

$w_2 = -\alpha$

$w = [w_1, w_2] = [\alpha, -\alpha]$

For the case of

$y_i(W^T x_i) = 1$ ..... (1) where-

$y = +1$; $x = [x_1, x_2]$ and $W = \begin{bmatrix} \alpha \\ -\alpha \end{bmatrix}$

For equation 1 to be true, $w^T x$ must be equal to 1 only, if $w^T x = 1$

$W = \begin{bmatrix} \alpha \\ -\alpha \end{bmatrix} [x_1, x_2] = 1$

$\alpha x_1 - \alpha x_2 = 1$

$\alpha x_1 = \alpha x_2 + 1$

$x_1 = x_2 + \frac{1}{\alpha}$

Now the co ordinates will be valid if, $\alpha > 0$

If $\alpha <= 0$

the co ordinate will be undefined.

Question 3

The equation of soft margin is as follow: $L(w, \alpha) = \frac{1}{2} W^T W + C \sum_{i=1}^{N} \epsilon_i + \sum_{i=1}^{N} \alpha_i (1 - y_i (W^T x_i)) - \sum_{i=1}^{N} \epsilon_i \mu_i$

Taking derivatives with respect to w, α and ϵ we get,

w^*= w- $\sum_{i=1}^{N} \alpha_i y_i x_i$

$\alpha^* = 1 - \sum_{i=1}^{N} \epsilon_i - \sum_{i=1}^{N} w^T y_i x_i$

$\epsilon^* = C - \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \mu_i$ .... eq(1)

The equation of hard margin is as follow:

$L(w, \alpha) = \frac{1}{2} W^T W + \sum_{i=1}^{N} \alpha_i (1 - y_i (W^T x_i))$

Taking derivatives with respect to w, α and ϵ we get,

w^*= w- $\sum_{i=1}^{N} \alpha_i y_i x_i$

$\alpha^* = 1 - \sum_{i=1}^{N} w^T y_i x_i$

Now, for equation(1), setting $\epsilon^* = 0$, we get, $n * \alpha + n\mu_i = C$. If we get rid of $n\mu_i$ part, we see that C is bounding α(that is the max α can go) and it is $0 \leq \alpha_i \leq C$ (We can ignore n as an constant).

This constraint indicate that we cannot include too much weight on any point (at most C). In the hard margin case, we saw, via complementary slackness, that $\alpha_i > 0$ only when the corresponding example is on a margin

Then looking at the derivatives of soft and hard margin, it is noticeable (max value of α's is 1 in hard margin while for soft margin it is C) that both derivatives are same if we ignore the impact of C. Hence,we can proved the claim in the question

Question 4

Given function

$K(x_i, x_j) = -x_i^T x_j$ .... (eq 3)

Let us consider $x_i = [x_1, x_2]$ and $x_j = [y_1, y_2]$

In that case the Kernel matrix according to (1) will be -

$$(-1)\begin{bmatrix} x_1y_1 & x_1y_2 \\ x_2y_1 & x_2y_2 \end{bmatrix}$$
.

In order to be Kernel function a matrix has to be - Symmetric and Positive semi-definite

In order to be positive semi definite all its determinant must be equal to or greater than zero. But here the whole matrix is negative. Assuming all output values from kernel is ¿0, then the determinant become negative.

In order for a matrix to be symmetric, the matrix should be equal to its transpose. Now, if we can show that they are not symmetric, then we can say it is not a right kernel function.

To satisfy that condition x_i and x_j has to be equal.

Kernel	Accuracy	Support vector
Linear	95.5189	543
Polynomial	95.9906	163
RBF	96.2264	211

Table 1: Accuracy and no of Support vectors for different Kernels

Value of C	Accuracy	Support vector
0.01	88.6792	1112
0.1	95.9906	424
1	95.9906	163
2	95.9906	131
3	96.2264	114
5	96.2264	102

Table 2: Accuracy and no of Support vectors for different values of C

For this reason this is not an appropriate Kernel Matrix.

Question 5

If we look at the table 2, for the value of C=5 model works the best. Though it has the same accuracy as C=3 valued model, it has used less number of support vector. Hence, in terms of efficiency and accuracy it is the best one. This model works best, because, it got more flexibility as we soften the margin a lot which helps the support vectors to classify properly with less number of support vectors with compare to the lowest value of C one.

However, if we look at the tab1, there is no single best model in terms of both accuracy and support vectors. Most likely the data is very high dimensional and that is why RBF works well then others but used a slightly more support vectors. It is most likley because of using the RBF kernel the algorithm needed for support vectors to classify.

One interesting think to notice that, instead of using a complex kernel, using a linear kernel with a value of C-=5, performs the same as RBF kernel.