COM S 573: Machine Learning
Homework #2

Abdurahman Ali Mohammed

1. The toy dataset is given as $\{([0,0],-1),([2,2],-1),([2,0]),+1)\}$. Before we setup the dual problem, let's add the bias to the data that we are given as follows:

$$\{([1,0,0],-1),([1,2,2],-1),([1,2,0]),+1)\}$$

The next thing to do is to put the lagrange equation as:

$$L(w,\alpha) = \frac{1}{2}||w||^2 + \sum_{i=1}^{N} \alpha_i(1 - y_i(w^T x_i + w_0))$$

Expanding the equation we will get:

$$L(w,\alpha) = \frac{1}{2}(w_0^2 + w_1^2 + w_2^2) + \alpha_1(1 + w_0) + \alpha_2(1 + w_0 + 2w_1 + 2w_2) + \alpha_3(1 - w_0 - 2w_1)$$

And now, we will differentiate the equation with respect to w_0, w_1 and w_2.

$$\frac{\partial L(w,\alpha)}{\partial w_0} = w_0 + \alpha_1 + \alpha_2 - \alpha_3 = 0$$

$$w_0 = -\alpha_1 - \alpha_2 + \alpha_3$$

$$\frac{\partial L(w,\alpha)}{\partial w_1} = w_1 + 2\alpha_2 - 2\alpha_3 = 0$$

$$w_1 = -2\alpha_2 + 2\alpha_3$$

$$\frac{\partial L(w,\alpha)}{\partial w_2} = w_2 + 2\alpha_2 = 0$$

$$w_2 = -2\alpha_2$$

$$\frac{\partial L(w,\alpha)}{\partial \alpha_1} = 1 + w_0 = 0$$

And we replace w_0 with its value.

$$1 - \alpha_1 - \alpha_2 + \alpha_3 = 0$$

$$\alpha_1 = 1 + \alpha_3 - \alpha_2$$

$$\frac{\partial L(w,\alpha)}{\partial \alpha_2} = (1 + w_0 + 2w_1 + 2w_2) = 0$$

$$1 + \alpha_3 - \alpha_2 - \alpha_1 + 2(2\alpha_3 - 2\alpha_2) + 2(-2\alpha_2) = 0$$

Substitute the value of α_1

$$1 + \alpha_3 - \alpha_2 - 1 - \alpha_3 + \alpha_2 + 2(2\alpha_3 - 2\alpha_2) + 2(2\alpha_2) = 0$$

$$\alpha_2 = \frac{\alpha_3}{2} = 0$$

$$\frac{\partial L(w,\alpha)}{\partial \alpha_3} = (1 - w_0 - 2w_1) = 0$$

Substitute the value of w_0

$$1 + \alpha_3 - \alpha_2 - \alpha_1 = 0$$

$$\alpha_1 = 1 + \alpha_3 - \alpha_2$$

$$1 - \alpha_3 - \alpha_2 - \alpha_1 - 2(2\alpha_3 - 2\alpha_2) = 0$$

Substitute the value of α_2

$$1 - \alpha_3 - \frac{\alpha_3}{2} - 1 - \alpha_3 + \frac{\alpha_3}{2} - 2(2\alpha_3 - 2\frac{\alpha_3}{2}) = 0$$

$$1 - 2\alpha_3 + \alpha_3 = 0$$

$$1 - \alpha_3 = 0$$

$$\alpha_3 = 1$$

Now, Substituting α_3 in the other equations we will get the values of α_1 and α_2 as follows.

$$\alpha_1 = \frac{3}{2}\ ,\ \alpha_2 = \frac{1}{2} \text{ and } \alpha_3 = 1$$

2. For this question we will consider the toy dataset $\{([0,0],+1),([1,0],+1),([0,1]),-1)\}$. Adding the bias the dataset will look like $\{([1,0,0],+1),([1,1,1],+1),([1,0,1]),-1)\}$. Let's first write the langrangian equation and expand it.

$$L(w,\alpha) = \frac{1}{2}||w||^2 + \sum_{i=1}^{N} \alpha_i(1 - y_i(w^T x_i + w_0))$$

$$L(w,\alpha) = \frac{1}{2}(w_0^2 + w_1^2 + w_2^2) + \alpha_1(1 - w_0) + \alpha_2(1 - w_0 - w_1) + \alpha_3(1 + w_0 + w_2)$$

Now, let's differentiate it with respect to $w_0, w_1, w_2, \alpha_1, \alpha_2, \alpha_3$.

$$\frac{\partial L(w,\alpha)}{\partial w_0} = w_0 - \alpha_1 - \alpha_2 + \alpha_3 = 0$$
$$w_0 = \alpha_1 + \alpha_2 - \alpha_3$$

$$\frac{\partial L(w,\alpha)}{\partial w_1} = w_1 - \alpha_2 = 0$$
$$w_1 = \alpha_2$$

$$\frac{\partial L(w,\alpha)}{\partial w_2} = w_2 + \alpha_3 = 0$$
$$w_2 = -\alpha_3$$

$$\frac{\partial L(w,\alpha)}{\partial \alpha_1} = 1 - w_0 = 0$$

And we will substitute w_0 with the value we got above.

$$1 - \alpha_1 - \alpha_2 + \alpha_3 = 0$$

$$\alpha_1 = 1 - \alpha_2 + \alpha_3 = 0$$

$$\frac{\partial L(w,\alpha)}{\partial \alpha_2} = (1 - w_0 - w_1) = 0$$
$$1 - \alpha_1 - \alpha_2 + \alpha_3 = 0$$
$$1 - 1 - \alpha_3 + \alpha_2 + \alpha_3 - 2\alpha_2 = 0$$
$$\alpha_2 = 0$$

$$\frac{\partial L(w,\alpha)}{\partial \alpha_3} = (1 + w_0 + w_2) = 0$$
$$1 + \alpha_1 + \alpha_2 - 2\alpha_3 = 0$$
$$2 - \alpha_3 = 0$$
$$\alpha_3 = 2$$

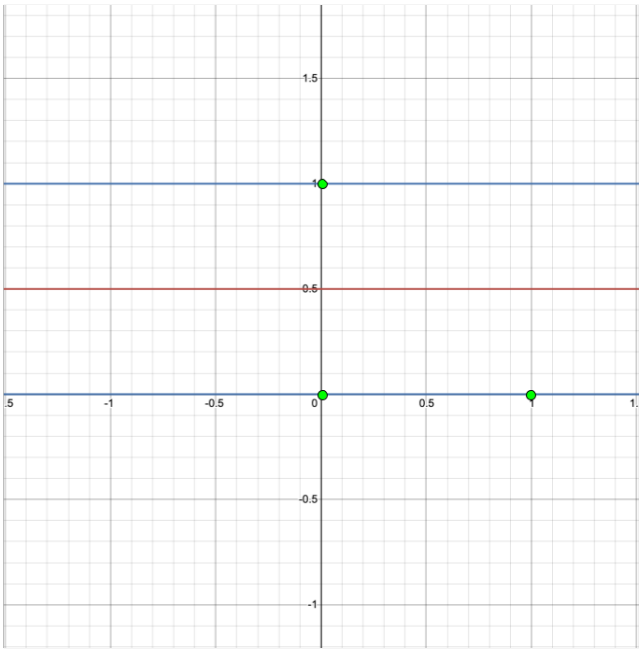
And after solving for α_1, we will get

$$\alpha_1 = 3 \ , \ \alpha_2 = 0 \text{ and } \alpha_3 = 2$$

Now we will put it the α values to find the ws.

$$w_0 = 1 \ , \ w_1 = 0 \text{ and } w_2 = -2$$

. Hence, the decision boundary will have value of $x_2 = \frac{-w_0}{w_2} = 0.5$ and the margins $x_2 = 0$ and $x_2 = 1$ with $\alpha_2 = 0$.



As we can see in the above plot, the data point $([1, 0]), +1$ which had $\alpha_2 = 0$ is lying on the margin. This tells us that it is possible for a data point to be on the margin even though its α is 0 since the condition $y_i(w^T x_i) = 1$ is satisfied.

3. We know that the optimization problem for the soft-margin SVM equation will look as follows:

$$\min_{w} ||w||^2 + C\sum_{i=1}^{N} \xi_i$$

subject to $y_i(w^T x_i) \geq 1 - \xi_i$ and $\xi_i \geq 0, i = 1, 2, \ldots, N$

To solve this optimization problem we will use Langrange equation

$$\mathcal{L}(w, \alpha, \mu) = \min_{w,\xi} ||w||^2 + \sum_{n=1}^{N} \alpha_n g_n(w) + \sum_{m=1}^{N} \mu_n h_n(w)$$

Since we are dealing with the soft margin support vector machine, Lagrangian multipliers α and μ will be introduced.

$$\mathcal{L} = (\frac{1}{2})||w||^2 + C\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i(1 - \xi_i - y_i(w^T x_i)) - \sum_{i=1}^{N} \mu_i \xi_i$$

where $||w|| = w^T w$

And the dual problem for the hard margin SVM looks like

$$\mathcal{L} = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{N} \alpha_i$$

such that $0 \leq \alpha_i$ Let's differentiate $\mathcal{L}$ over ξ_i

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 + C - \alpha_i - \mu_i = 0$$
$$\mu_i = C - \alpha_i$$

Next, we will use the new value of μ_i that we got in the Soft-margin Lagrange equation.

$$\begin{aligned}
\mathcal{L} &= (\frac{1}{2})||w||^2 + C\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i(1 - \xi_i - y_i(w^T x_i)) - \sum_{i=1}^{N}(C - \alpha_i)\xi_i \\
&= (\frac{1}{2})||w||^2 + C\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i \xi_i - \sum_{i=1}^{N} \alpha_i y_i(w^T x_i) - C\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i \xi_i \\
&= \frac{1}{2}(w^T w) + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i y_i(w^T x_i) \\
&= \frac{1}{2}\sum_{i=2}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_j x_i^T + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \\
&= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_j x_i^T + \sum_{i=1}^{N} \alpha_i
\end{aligned}$$

We now have derived the equation for the hard-margin SVM's dual problem by substituting the value of μ_i with $C - \alpha_i$. This indicates that the soft-margin SVM is almost identical to hard-margin SVM except αs are bounded by the tradeoff parameter C.

4. Given the function $K(x_i, x_j) = -x_i^T x_j$, we want to check if it is a valid kernel function or not. We know that there are two requirements for a function to be a. kernel function. Its corresponding kernel matrix is

   (a) Symmetric

   (b) Positive semi-definite.

   So first we will check if the given function is satisfies the first condition. To check that we will interchange x_i and x_j. This will result the function to be $-x_j^T x_i$ which will still have the same value. Hence, we can say that the function is symmetric and satisfies our first condition.

   Now, let's check the second condition. To check if it is semi-definite, we will check if the following equation is true or not.

$$\sum_{i=1}^{M}\sum_{j=1}^{M} u_i^T u_j(-x_i^T x_j) \geq 0$$
$$= -\sum_{i=1}^{M} u_i^T x_i^T \sum_{j=1}^{M} u_j x_j \geq 0$$
$$= -X^T X \geq 0$$

   The above equation fails to satisfy being ≥ 0 if the value of X is negative. This proves that the given function $K(x_i, x_j) = -x_i^T x_j$ fails to satisfy the second requirement of being a kernel function. Therefore, we can say that it is not a valid kernel function.

5. We need to show that the maximum number of support vectors in the naive support vector machine for linearly separable data. We know that all the support vectors which are on the margin line have the characteristics $y_i(w^T x_i + b) = 1$. These are the points that determine the decision line. Hence, they will have an α value ≥ 0. We know that the weights are determined only by points on the boundary. Since removing any data point that is not a support vector won't affect the decision boundary. We will write the matrix format as follows $A = \begin{bmatrix} y_1 & y_1 x_1^T \\ y_2 & y_2 x_2^T \\ . & \\ . & \\ . & \\ y_n & y_n x_n^T \end{bmatrix}$ and let's considering the augmented matrix $A^* = [A1]$. Since the data is linearly separable, we must have at least one optimal solution. We know that the rank of matrix A is $d+1$ since only $d+1$ rows are independent and these $d+1$ rows are capable of determining the weights. Therefore, $rank(A) \leq d+1$. This tells us that the independent rows are the support vectors and are the ones that decide the decision boundary of our SVM model.

6. (a)
```
------------Accuracy with different C values--------------
Cost= 0.01 : Score= 0.894
Support vectors:  1080
Cost= 0.1 : Score= 0.962
Support vectors:  414
Cost= 1 : Score= 0.960
Support vectors:  162
Cost= 2 : Score= 0.960
Support vectors:  129
Cost= 3 : Score= 0.962
Support vectors:  117
Cost= 5 : Score= 0.962
Support vectors:  104
```

(b)
```
------------Accuracy with different Kernels--------------
Kernel= linear : Score= 0.960
Support vectors: 162
Kernel= rbf : Score= 0.962
Support vectors: 90
Kernel= poly : Score= 0.958
Support vectors: 75
```

(c) For the first part, We experimented with different cost values ranging from 0.01 to 5. As we can see from the results, the larger the value of c gets, the accuracy of the predictions increased from 89.3% to 96.2%. In the mean time, the number of support vectors decreased. This happened because larger values of C will narrow the margin decreasing the number of support vectors.

For the second method, we experimented with three different kernels, namely, linear, polynomial and radial basis function. Among the three kernels, RBF exhibited the best performance with 96.2%. It had 90 support vectors in total. Looking at the linear kernel, It had the most number of support vectors but still has a slightly better accuracy than the polynomial classifier. The reason behind the relatively lower accuracy of the polynomial kernel is due to the fact that it tries to map the data to a higher dimension which leads to more features being used.