CS 5350/6350: Machine Learning Spring 2020

Homework 4
Britton Gaul
u0915408

Handed out: 19 Mar, 2020
Due date: 11:59pm, 4 Apr, 2020

1 Paper Problems [40 points + 10 bonus]

1.

(a)
If x_i breaks into the margin and is on then correct the side then $0 < \xi_i < 1$

(b)
If x_i is correctly classified then $\xi_i = 0$. If x_i is missclassified then $\xi_i > 1$. However it is still possible for it to stay on the outside of the margin, it would just be on the wrong side.

(c)
This term maximizes the margin and factors in how many points will break into the margin. This is to make sure not too many points are breaking into the margin when maximizing. If it is removed then a lot more points will be allowed to break into the margin, which can lead to missclassification.

2.

The Langrangian form of the primal is:
$min : w, b, \{\xi_i\}\ max : \{\alpha_i \geq 0, \beta_i \geq 0\}$
$L = \frac{1}{2} w^T w + C \sum_i \xi_i + \sum_i \beta_i(-\xi_i) + \sum_i \alpha_i(1 - \xi_i - y_i(w^T x_i + b))$
The Langrangian form of the dual is:
$min : w, b, \{\xi_i\}\ max : \{\alpha_i \geq 0, \beta_i \geq 0\}$
$L = \frac{1}{2} w^T w + C \sum_i \xi_i + \sum_i \beta_i(-\xi_i) + \sum_i \alpha_i(1 - \xi_i - y_i(w^T x_i + b))$

Because the function is differentiable the partial derivative can be set to equal 0 in order to get the minimum value.

$\frac{\partial L}{\partial w} = w - \frac{\partial(\sum_i \alpha_i y_i w^T x_i)}{\partial w} = 0$
which gives, $w = \sum_i \alpha_i y_i x_i$

$\frac{\partial L}{\partial b} = -\sum_i \alpha_i y_i = 0$
which gives, $\sum_i \alpha_i y_i = 0$

$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0$
which gives, $\alpha_i + \beta_i = C, \forall i$

Then by substituting in w and the two constraints into the Lagrangian form, the new objective function can be found

$L' = \frac{1}{2}(\sum_i \alpha_i y_i x_i)^T(\sum_i \alpha_i y_i x_i) - (\sum_i \alpha_i y_i)(\sum_i \alpha_i y_i x_i)^T(x_i) + \sum_i \alpha_i$
$L' = -\frac{1}{2}(\sum_i \alpha_i y_i x_i)^T(\sum_i \alpha_i y_i x_i) + \sum_i \alpha_i$
$L' = -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i$

Then the standard dual form is:
$max : \{\alpha_i \geq 0, \beta_i \geq 0\}\ L' = -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i$
with $\sum_i \alpha_i y_i = 0, \forall i$ and $\alpha_i + \beta_i = C, \forall i$
which is equivalent to:
$minL'' = \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_i \alpha_i$
with $\sum_i \alpha_i y_i = 0, \forall i \in N$ and $0 \leq \alpha_i \leq C, \forall i \in N$

3.

(a)
The parameter $\alpha_i = 0$ indicates the sample at x_i remains outside the margin.

(b)
$\beta_i \xi_i = 0, \forall i \in N$
$\alpha_i(y_i(w^T x_i + b) - 1 + \xi_i) = 0, \forall i \in N$
$\alpha_i + \beta_i = 0, \forall i \in N$

Using the equations above it can be seen that if $\alpha_i > 0, \xi_i = 0$, then x_i remains on the margin. Another way to check this is to look at the condition $\alpha_i > 0$ and $\beta_i > 0$ and if $\xi_i > 0$. If both of these conditions are true then it also means the support vectors are located within the margin.

4.
The prediction for linear SVM is calculated normall with the equation $sgn(\sum_i \alpha_i y_i x_i^T x)$ for the sample x. The kernel could be used to replace the dot product makeing the equation with the kernel, $sgn(\sum_i \alpha_i y_i K(x_i, x))$. Using the kernel also changes the dual form related to it to $min\{\alpha_i \in [0, C]\}, \sum_i \alpha_i y_i = 0 : \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_i \alpha_i$.

5.
$N = 3$ so $C = \frac{1}{3}$
For the first step (x_1, y_1):
$\nabla J = [w_0, 0] - 3C y_1 x_1$
$\nabla J = [0, 0, 0, 0]^T - 3C[0.5, -1, 0.3, 1]^T$
$\nabla J = [-0.5, 1, -0.3, -1]^T$

$w^1 = w^0 - 0.01[-0.5, 1, -0.3, -1]^T$
$w^1 = [0.005, -0.01, 0.003, 0.01]^T$

For the second step (x_2, y_2):
$\nabla J = [0.005, -0.01, 0.003, 0] - CNy_2x_2$
$\nabla J = [-0.995, -2.01, -1.997, 1]^T$
$w^2 = w^1 - 0.005[-0.995, -2.01, -1.997, 1]^T$
$w^2 = [-0.990, -1.9999, -1.9870, 0.9950]^T$

For the third step (x_3, y_3):
$\nabla J = [w^2T, 0]^T$
$\nabla J = [-0.990, -1.9999, -1.9870, 0]^T$
$w^3 = w^2 - 0.0025[-0.990, -1.9999, -1.9870, 0]^T$
$w^3 = [-0.9875, -1.9949, -1.9820, 0.9950]^T$

2 Practice [60 points + 10 bonus]

1. GitHub link: https://github.com/BritGaul/CS5350

2. (a)
   The chosen values are: $\gamma_0 = 2.3$ and $d = 1$
   Training and Testing Error for each C Data:

C	Training Error	Testing Error	Weight
$\frac{100}{873}$	0.039	0.048	[-1.2668, -0.6846, -0.7392, -0.2697, 0.0]
$\frac{500}{873}$	0.044	0.052	[-1.7982, -0.9413, -1.0412, -0.4045, 0.0]
$\frac{700}{873}$	0.038	0.048	[-1.8919, -0.9910, -1.1521, -0.4060, 0.0]

   (b)
   The chosen values are: $\gamma_0 = 2.3$ and $T = 100$
   Training and Testing Error for each C Data:

C	Training Error	Testing Error	Weight
$\frac{100}{873}$	0.039	0.046	[-1.3206, -0.7079, -0.7585, -0.2930, 0.0]
$\frac{500}{873}$	0.039	0.046	[-1.8717, -1.1230, -1.2014, -0.5189, 0.0132]
$\frac{700}{873}$	0.084	0.096	[-2.0623, -1.0124, -1.4387, -0.5426, 0.0]

(c)

For the first rate schedule from part a, both the training errors and testing errors remained some what simiar for each of the values of C. For the second rate schedule from part b, the training and testing errors were alos similar. However, for the last C value tested they both increased dramatically. The weights for both scheduled rates seemed to increase as the C value increased, but the weights for the schedule rate in part b were slightly larger then the ones in part a.

3.

(a)

Training and Testing Error for each C Data:

C	Training Error	Testing Error	Weight
$\frac{100}{873}$	0.071	0.078	[-9.4292e-01, -6.5149e-01, -7.3372e-01, -4.1021e-02, 2.3959e-10]
$\frac{500}{873}$	0.056	0.066	[-1.5639e+00, -1.0140e+00, -1.1806e+00, -1.5651e-01, 4.4810e-09]
$\frac{700}{873}$	0.056	0.062	[-2.0425e+00, -1.2806e+00, -1.5135e+00, -2.4902e-01, -3.3224e-09]

The testing and training errors from the dual svm are higher then compared to the primal errors in question2. The weights are pretty similar although they are a bit higher in the dual svm then the primal. Also the b values are much higher in the dual then the primal. This is most likely because the dual svm is less accurate then the primal.

(b)

Training and Testing Error for each C Data:

γ	C	Training Error	Testing Error
0.1	$\frac{100}{873}$	0.0	0.002
0.5	$\frac{100}{873}$	0.0	0.002
1	$\frac{100}{873}$	0.0	0.002
5	$\frac{100}{873}$	0.008	0.006
100	$\frac{100}{873}$	0.315	0.28
0.1	$\frac{500}{873}$	0.0	0.002
0.5	$\frac{500}{873}$	0.0	0.002
1	$\frac{500}{873}$	0.0	0.002
5	$\frac{500}{873}$	0.003	0.006
100	$\frac{500}{873}$	0.298	0.266
0.1	$\frac{700}{873}$	0.0	0.002
0.5	$\frac{700}{873}$	0.0	0.002
1	$\frac{700}{873}$	0.0	0.002
5	$\frac{700}{873}$	0.003	0.006
100	$\frac{700}{873}$	0.298	0.266

The best combination is having gamma as 0.1. This was true for all the different values of C. This means that the data becomes linearly seperable with these values in place. Compared with the linear SVM, this version results in much lower training and teating errors.

(c)

Number of Support Vectors:

γ	$\frac{100}{873}$	$\frac{500}{873}$	$\frac{700}{873}$
0.1	869	871	872
0.5	858	751	790
1	821	613	781
5	722	527	611
100	459	444	450

Number of overlapping Support Vectors when $C = \frac{500}{873}$:

γ	Number of Overlaps
0.1	750
0.5	592
1	487
5	228
100	0

The number of support vectors decreases as gamma increases for all the different values of C. Also for each gamma values as C increases the number of support vectors for each C value increases.