ISYE-6740

Homework 4

1.

a) In the context of Support Vector Machines, the C parameter is what determines the allowable margin used to linearly separate our points for the defined classes (i.e. -1 and 1 based on the equality constraints. Generally, a value of 1 means that we're using a lower value of C and we want to create a larger margin to separate our training data points. However, this may cause us to incorrectly classify specific outlier observations that are contained within our margin.

When using a larger value of C, we see that the margin is much less. The tradeoff for this hyperparameter selection is that our smaller margin may perform better against our training data, but at the cost of overfitting and not reacting well to our test data. This can really depend on the data we're dealing with, and may results in the use of hyperparameter.

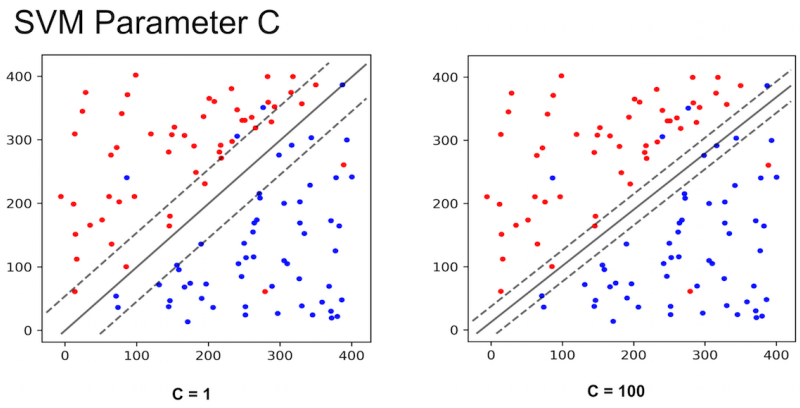

Figure 1: SVM C-Parameter Comparison

From an optimization standpoint, when we revert back to the lecture material, our problem is we're trying to maximize w and our scalar offset b based on the following:

$$\max_{w,b} = \frac{2c}{||w||} \tag{1}$$

When changing the value of c to a higher value, it will simply scale our separation vector w, and the scalar offset, which does not change the goodness of the classifier for different c values. Hence, as the professor mentioned we can set it to 1 for a cleaner problem.

b)

Based on dual representation for the Lagrangian multipliers, our function is set to:

$$L(w, b, \alpha) = \frac{1}{2} w^T w + \sum_{i=1}^{m} \alpha_i \left(1 - y^i \left(w^T x^i + b\right)\right) \tag{2}$$

To formalize the Lagrangian multiplier, we find the partial derivative with respect to w, b and alpha. But for our context of this question we just need to determine the partial derivative of the weighted vector w:

$$\frac{\partial L}{\partial w} = \frac{1}{2} 2w - \sum_{i=1}^{m} \alpha_i \left(y^i \left(x^i\right)\right) \tag{3}$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{m} \alpha_i \left(y^i \left(x^i\right)\right) \tag{4}$$

$$w = \sum_{i=1}^{m} \alpha_i \left(y^i \left(x^i\right)\right) \tag{5}$$

We've now proved the weighted sum of vector w can take the form expressed in the question.

c)
Our Lagrangian function is as follows:

$$L(w, b, \alpha) = f(w) + \sum_{i=1}^{n} \alpha_i \left(1 - y^i \left(w^T x^i + b\right)\right) \tag{6}$$
$$\alpha_i \geqq 0$$

Where the second term contains our function g(x).

The KKT Conditions State:

The KKT conditions

If there exists some saddle point of L, then the saddle point satisfies the following "Karush-Kuhn-Tucker" (KKT) conditions:

$$\frac{\partial L}{\partial w} = 0$$
$$\frac{\partial L}{\partial b} = 0$$
$$\frac{\partial L}{\partial \alpha} = 0$$
$$\frac{\partial L}{\partial \beta} = 0$$
$$g_i(w) \leq 0$$
$$h_i(w) = 0$$
$$\alpha_i \geq 0$$
$$\alpha_i g_i(w) = 0$$

Highlighted in red our last condition in red shows the following:

$$\alpha_i \left(1 - y^i \left(w^T x^i + b\right)\right) = 0 \tag{7}$$

From this we can produce inequalities to show that for different values of theta, we will produce an observation either within the margin or not:

As stated on the following slide:

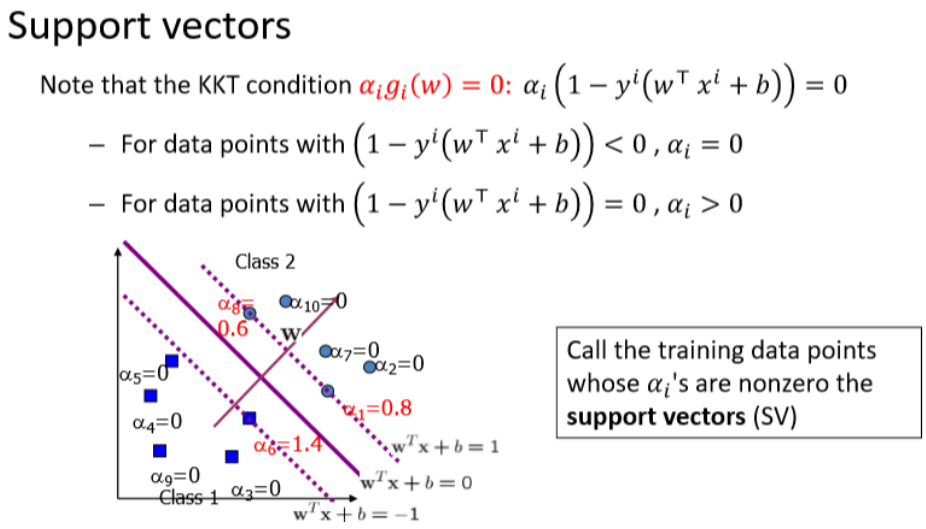

Figure 2: Support Vector Inequality Conditions

When alpha = 0, for $\left(1 - y^i\left(w^T x^i + b\right)\right) < 0$

When alpha > 0, for $\left(1 - y^i\left(w^T x^i + b\right)\right) = 0$

Therefore we can conclude that values of theta would have to be greater than 0 to be able to within the margin, and are the only values that can contribute to the sum for the Lagrangian function.

2

a)

When plotting lines for our ranges of the variable h we can see the following:

$0 < h < 1$ (Line 1): You can see from Line 1 that the intersection at of the line at (1,1) show the margin in which a higher value then 1 would produce a misclassified point.

$h > 4$ (Line 2): You can see from Line 2 that the intersection at of the line at (4,1) show the margin in which a smaller value then 4 would produce a misclassified point.

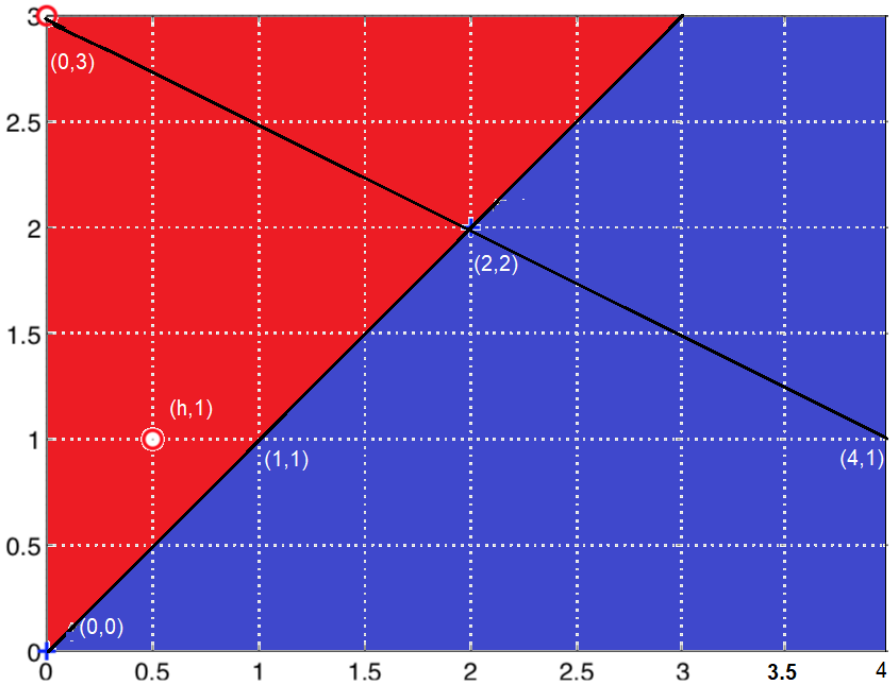

Figure 3:

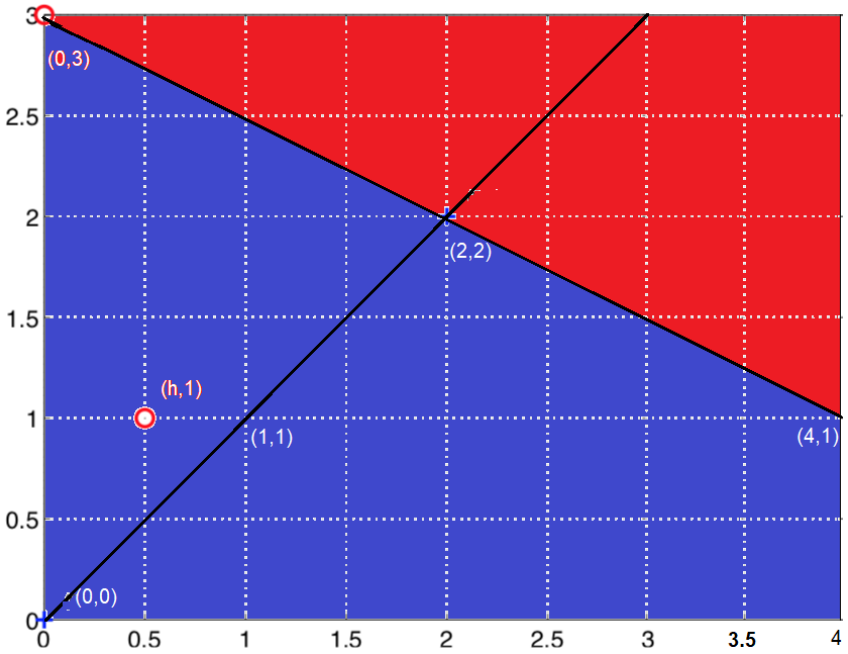

Figure 4: SVM Range of h $>$ 4

b)
We can see from the last result, that when h $>$ 4, we can create a function of h to form the linear separation. Creating vectors from the points where h $>$ 4, we want to ensure that the result is non-zero.

$$f(h) = \begin{bmatrix} 2 \\ 2 \end{bmatrix} - \frac{1}{2}\left(\begin{bmatrix} h \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 3 \end{bmatrix}\right) \tag{8}$$

$$\text{When } h = 4 \; f(h) = \begin{bmatrix} 2 \\ 2 \end{bmatrix} - \frac{1}{2}\left(\begin{bmatrix} 4 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 3 \end{bmatrix}\right) \tag{9}$$

$$f(h) = \begin{bmatrix} 2 \\ 2 \end{bmatrix} - \frac{1}{2}\left(\begin{bmatrix} 4 \\ 4 \end{bmatrix}\right) \tag{10}$$

$$f(h) = \begin{bmatrix} 2 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \end{bmatrix}) \tag{11}$$

$$f(h) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{12}$$

Therefore, we've shown that with the function that any value of h $>$ 4 constituting the max margin decision boundary can change as a function of h. Since any result $>$ 4 will produce a non-zero result capturing the distance from point (2,2), which on a decision boundary line represented by Line 2 in the figure.

3.
a)
We can see from logistic regression that the probability distribution

$$P(y = 1|, x, w) = \frac{1}{1 + \exp\left(-w^T x\right)}$$

This exp function can be seen as a sigmoid function where similar to the activation function used in neural networks:

$$\alpha(u) = \frac{1}{1 + \exp(-u)}$$

Therefore, without a hidden layer, our models are essentially doing the exact same thing, we're taking a set of outputs and passing them through a sigmoid function to get a binary output of either 0 or 1.
b)

$$l(w, \alpha, \beta) = \sum_{i=1}^{m} \left(y^i - \alpha\left(w^T z^i\right)\right)^2 \tag{13}$$

So we know the partial derivative of w is given by the partial derivative of u^i times the partial derivative of w with respect to u^i

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{m} \frac{\partial}{\partial u^i}\left(y^i - \alpha\left(w^T z^i\right)\right)^2 \frac{\partial u^i}{\partial w} \tag{14}$$

Using chain rule, we bring the 2 down from the exponent and multiply by -1 due to the sigmoid function when treating it as a constant, then we sub in the value of the sigmoid function which is

$\alpha(x) = \frac{1}{1+e^{-x}}$ we get the following:

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{m} 2(y^i - \alpha\left(w^T z^i\right)(-1) \cdot \frac{\partial}{\partial u^i}\left(\frac{1}{1+e^{-u}}\right) \cdot \frac{\partial u^i}{\partial w} \tag{15}$$

Now to solve for the partial derivative of $\frac{\partial}{\partial u^i}$, so we know that derivative in this case will be the derivative of a fraction $\frac{d}{dx}\left(\frac{1}{x}\right) = -x^{-1-1} = -\frac{1}{x^2}$, where $x = 1+e^{-x}$. Then we know that the $\frac{d}{dx}(x)$ for a euler is simply $= e^{-x}$. We also know that $\frac{\partial u^i}{\partial w}$ is simply $= z^i$ since our w^T will simply to 1. Therefore we will have:

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{m} 2\left(y^i - \alpha\left(w^T z^i\right)(-1) \cdot \frac{1}{\left(1+e^{-u^i}\right)^2}. e^{-u^i}(-1)\right) \cdot z^i \tag{16}$$

To multiply the following $\frac{1}{\left(1+e^{-u^i}\right)^2}. e^{-u^i}(-1)$ we can simply factor out of a full term from $\frac{1}{\left(1+e^{-u^i}\right)^2}$ and then we get the following:

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{m} 2\left(y^i - \alpha\left(w^T z^i\right)(-1) \cdot \frac{1}{\left(1+e^{-u^i}\right)}\left(\frac{e^{-u^i}}{\left(1+e^{-u^i}\right)}\right)(-1)\right) \cdot z^i \tag{17}$$

Now we can do some further simplification, and breakout the factorization a little more to the following:

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{m} 2\left(y^i - \alpha\left(w^T z^i\right)(-1) \cdot \frac{1}{\left(1+e^{-u^i}\right)}\left(1 - \frac{1}{\left(1+e^{-u^i}\right)}\right)(-1)\right) \cdot z^i \tag{18}$$

Now when we convert the fraction back to $\alpha(u)$ we get the following:

$$\frac{\partial L}{\partial w} = -\sum_{i=1}^{m} 2\Big(y^i - \alpha\big(u^i\big)\Big)\alpha\big(u^i\big)\Big(1-\alpha\big(u^i\big)\Big).\ z^i \tag{19}$$

For our second part of the question, we need to find the gradient of l(w,α,β), now with respect to the coefficients α and β. Which will be the result of the following:

$$\frac{\partial l}{\partial \alpha} = \frac{\partial l}{\partial u^i}\frac{\partial u^i}{\partial z_1^i}\frac{\partial z_1^i}{\partial \alpha} \tag{20}$$

We know the derivative of our sigmoid function $\frac{\partial l}{\partial u^i} = \alpha^T x^i$, so we plug this into our x term and the partial derivative of our coefficient α with respect to z_1^i,

$$\frac{\partial z_1^i}{\partial \alpha} = \frac{\partial}{\partial \alpha}\frac{1}{1+\ e^{-\alpha^T x^i}} \tag{21}$$

Now we will use chain rule. We know from our previous calculations when we factored out the fraction where we took the derivative of our fraction times the derivative of the euler, and we also know the derivative of $-\alpha^T x^i$ in the exponent term is simply $-x^i$ when holding alpha as a constant. Therefore, this will result in the following:

$$\frac{\partial z_1^i}{\partial \alpha} = \frac{(-1)e^{-\alpha^T x^i}}{\big(1+e^{-\alpha^T x^i}\big)^2}\big(-x^i\big) \tag{22}$$

Then similar to before, we factor out the fraction squared term to get the following:

$$\frac{\partial z_1^i}{\partial \alpha} = \frac{-e^{-\alpha^T x^i}}{1+\ e^{-\alpha^T x^i}}\cdot\frac{1}{1+\ e^{-\alpha^T x^i}}\big(-x^i\big) \tag{23}$$

When we plug this back into our original equation for z_1^i we get the following:

$$= \big(1-z_1^i\big)z_1^i x^i \tag{24}$$

Now the final step is plugging this back into the derivative of alpha equation along with the derivative of our weight coefficient w and we get:

$$\frac{\partial l}{\partial \alpha} = -\sum_{i=1}^{m} 2\Big(y^i - \alpha\big(u^i\big)\Big)\alpha\big(u^i\big)\Big(1-\alpha\big(u^i\big)\Big).\big(1-z_1^i\big)z_1^i x^i\ .\ z_1^i \tag{25}$$

The same would hold true for the beta coefficient, just the subscript would change the z term.

4. Comparing SVM and Neural Networks

Part 1) Divorce classification/prediction

a-c)

Provided is the performance / model accuracy for each of the 3 classifiers. This chart shows the iterations with the number of features used and the dataset used, in our case the divorce

and digits dataset.

Model_Name	Num_Features	Param_Name	Param_Value	Dataset	
neural_net	2	hidden_layer_sizes	(5, 2)	divorce	0.905882
	All	hidden_layer_sizes	(5, 2)	digits	0.510176
				divorce	0.463235
svm	2	C	1	divorce	0.897059
			10	divorce	0.905882
			100	divorce	0.897059
			5	divorce	0.897059
	All	C	1	digits	0.507412
				divorce	0.982353
			10	digits	0.515075
				divorce	0.969118
			100	digits	0.518719
				divorce	0.969118
			5	digits	0.508668
				divorce	0.977941

Figure 5: Model Accuracy Grouped By Dataset, Parameters and Number of Features

Based on our results we can determine that the training scores after 1 iteration is higher then the scores used for 2 features. However, this is clearly due to the model having only chosen the first 2 features in the dataset, which means we're excluding other features that have high predictive power and account for significant variance in the target variable.

b)
Decision boundary plots for 2-features used in the divorce dataset:

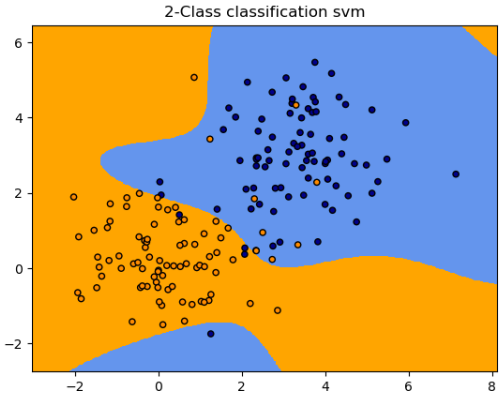

Figure 6: SVM Decision Boundary

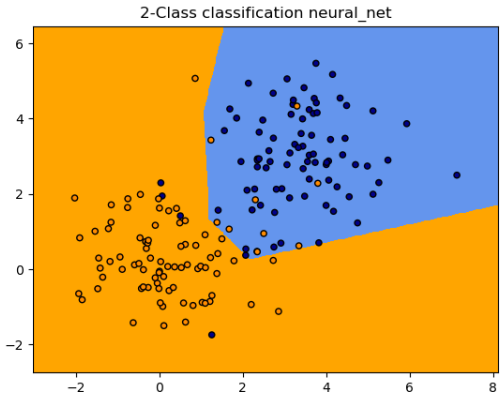

Figure 7: Neural Network Decision Boundary

We can see with the decision boundaries that the results are non-linearly separable, which means our classifiers will be able to equally account for a majority of the variance and have strong accuracy when predicting against our target variable. The results are very good, and seem to be quite aligned with our results from homework 3. However, this is due to the volume of data we're dealing with and overfitting since we're using a train test split of 80, 20.