CIS 419/519: Homework 1

Rutuja Moharil

February 23, 2019

Although the solutions are entirely my own, I consulted with the following people while working on this homework: {G.Sai,Gauri, Sagar,Siddharth,Roshan}

1 Gradient Descent

1. Taking a constant α_k mathematically means that the number by which the gradient vector is multiplied is constant across iterations . This can have following implications :
   (a) If the value of the constant α is large : We might overshoot and miss out the optimum .
   (b) If the value of constant α is small then we will take a long time to converge . It may also happen that we may get at a sub optimal point.
2. If we keep the learning as a function of iteration (k) , we are essentially making the learning rate adaptive.This adaptive choice of step size can help overcome problems like diverging or overshooting from the global optimum. So this would be a better technique for gradient descent.

2 Fitting SVM by hand

1. Let's find the mapped values first ::

Sr.No	x-coordinate	class	mapped value
1	0	-1	$[1,0,0]^T$
2	$\sqrt{2}$	1	$[1,2,2]^T$

   The vector parallel to w should be the vector between two mapped points.
   $[1,2,2]-[1,0,0]=[0,2,2]$ is the vector parallel to w.
2. The value of margin can be written as finding the euclidean distance between two points.

$$W_value = \sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}$$
$$W_value = \sqrt{(1-1)^2+(2-0)^2+(2-0)^2} = \sqrt{8}$$

3. Given that

$$margin = \frac{2}{\|w\|}$$
$$margin = \frac{2}{\sqrt{w_1^2+w_2^2+w_3^2}}$$

From the previous part we derived the margin value as W_value

$$\sqrt{8} = \frac{2}{\sqrt{w_1^2 + w_2^2 + w_3^2}}$$
$$w_1^2 + w_2^2 + w_3^2 = \frac{4}{8} = \frac{1}{2}$$

We also saw in the first part that the vector between the two mapped points is parallel to the w vector Assume k is some constant by which the magnitude of w vector differs from that of derived parallel vector

$$w = k[0, 2, 2]$$

Substituting $[k.0, k.2, k.2]$ in the equation above we get

$$k^2.0 + k \cdot 2^2 + k^2.2^2 = \frac{1}{2} 8k^2 = \frac{1}{2} \xrightarrow{k^2 = \frac{1}{16}}$$
$$k = \frac{1}{4}$$

Plugging this value for the w equation we get the following result

$$\text{Answer is } w = [0, \frac{1}{2}, \frac{1}{2}]$$

4. Solving for w_0 . Here we will use the value of w and the equations from the first two parts

Substituting the values in the constraint equation
First constraint equation
$$y = -1$$
$$-1 * ([w_1, w_2, w_3][1, 0, 0]^T + w_0) \geq 1$$
Simplifying we get
$$-w_1 - w_0 \geq 1 \text{ First element of the w matrix is 0}$$
$$0 - w_0 \geq 1$$
$$w_0 \leq -1$$
Second constraint equation
$$+1 * ([w_1, w_2, w_3][1, 2, 2]^T + w_0) \geq 1$$
Simplifying we get
$$w_1 + 2w_2 + 2w_3 + w_0 \geq 1 \text{ First element of the w matrix is 0}$$
$$0 + 2 * \frac{1}{2} + 2 * \frac{1}{2} + w_0 \geq 1$$
$$2 + w_0 \geq 1$$
$$w_0 \geq -1$$

So from the two constraint equations we see that $w_0 \geq -1$ and $w_0 \leq -1$. So the only possible solution is that $w_0 = -1$

5. So here we need to plug the values we found for w and w_0 in the equation given.

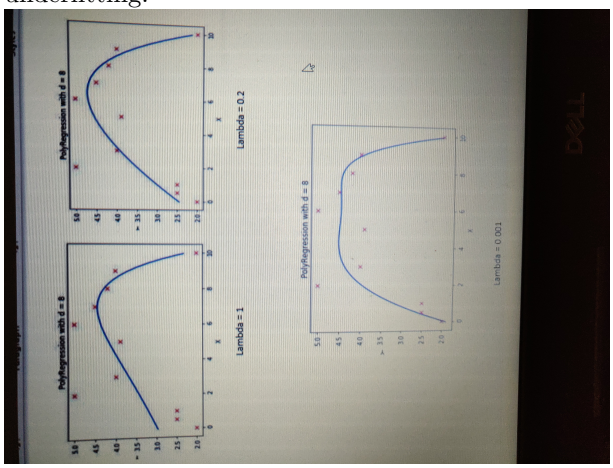
$$f(x) = w_0 + w^T \phi(x)$$
$$f(x) = -1 + [0, 0.5, 0.5]^T [1, \sqrt{2}, x^2]^T$$
$$f(x) = -1 + 0 + 0.5 * \sqrt{2}x + 0.5 * x^2$$
$$f(x) = -1 + 0.5(\sqrt{2}x + x^2)$$

3 Support vectors

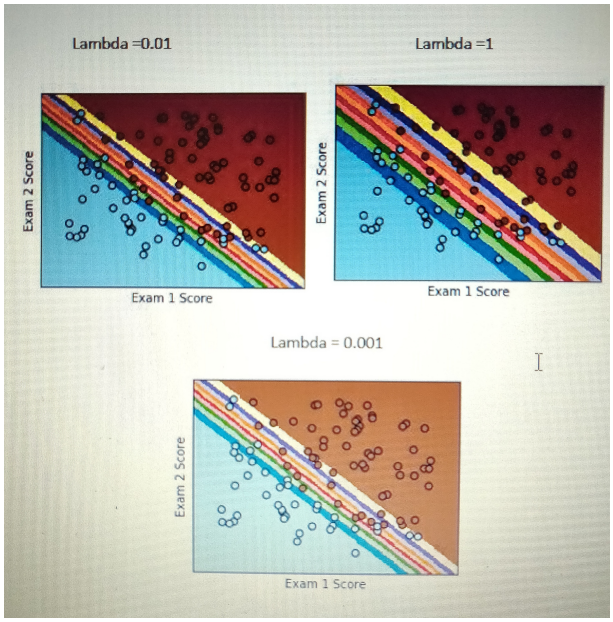
Here our support vectors are represented by the equation constraints of a maximization optimization problem. If we remove a support vector we are essentially dropping some constraints in a constrained maximization problem. So dropping a constraint (just puts lesser restriction in a way) we get an optimal value which would be at least as good the previous one.
So, for the weaker constraints, the old optimal solution is still available and there may be additions solutions that are even better.So in terms of SVM when removing some support vector the margin can stay the same or increase depending on the geometry.

4 Programming Part Analysis

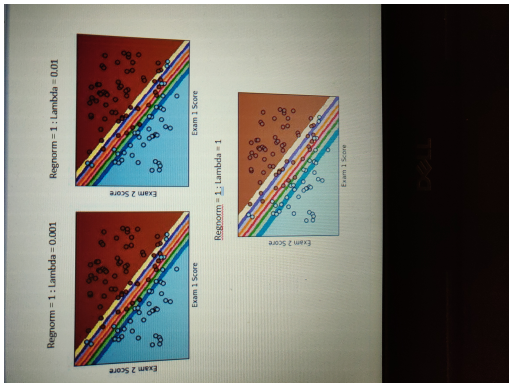
1. Polyreg Analysis
   In Polyreg we were asked to implement the closed form solution . Now in the graph generated we can see that the curve is passing through many points which is an indication of over-fitting. So basically this phenomenon means that the curve has almost neatly and perfectly fit the current data set such that given a new data set our polyreg prediction won't be right.
   Increase of the regularization or lambda value, we find a reduction in overfitting and this is observed from the graph shown in this . So the catch here is that we need to avoid over-fitting but at the same time not underfit also . Now if λ is increased by a lot this would cause high variance and hence underfitting.



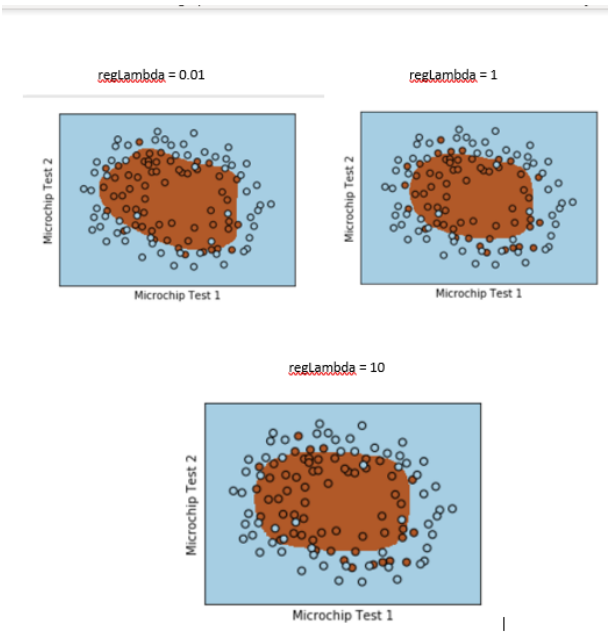
2. Logistic regression L1 Vs L2 regularization analysis in Logistic regression
   So for the L1 and L2 regularization I observed the following curves : Logreg Linear performance for L1 Vs L2



3. For the L1 normalization we have the following graphs



4. Non linear features for logreg



(47).png

5. L1 regularization is also known as Lasso and L2 is called as Ridge regression. In the L2 norm adds the squared magnitude of the coefficient. We can also think this is as the penalty term being .The penalty term is a square one . Now Lambda if very large adds too much weight and hence leads to underfitting .
   For L1 regularization we can see that the L1 technique shrinks the less important feature's coefficient to zero. This works well for feature selection.
   Now let's talk about the case of this non-linear features now HW. Here if we increase Lambda it so happens that a huge increase of λ causes L1 to classify only one variable. For Smaller values of λ they are kind of same. That's why there's a need for the rainbow plot . Because only through those plots were we able to see L1 and L2 differences. L1 has sparse solution and is robust to outliers. However L2 is not robust to outliers because the error term is squared and it blows up .

5 Comparison of Algorithms

1. Here we have been presented with the 3 medical data sets . The data sets have been described in short below :

Data set Name	Instances,Attributes	Label
Retinopathy	1151 instances,20 attributes	Class label =1 or 0 means signs of DR or no signs of DR respectively
Breast Cancer	569 instances,32 attributes	Class label =M or B means signs of cancer malignant or benign respectively
Diabetes	768 instances,9 attributes	Class label =tested positive or negative means affected by diabetes or not respectively

   (a) The first data set is the Retinopathy data set . The data set is multivariate.There are no values missing in the data set The features consists of results of quality assessment,pre-screening,MA detection ,Euclidean distance of the center of macula and center of optic disc,Center of optic disc ,result of AM/FM classification.

(b) The second data set is the Breast cancer data set. It is again a multivariate data setIt consists of features like radius ,mean size of tumour,contours of the tumour (including it's concavity and severity of concavity),fractal dimension and standard error for standard deviation of grayscale values .

(c) The third data set for the PIMA indians Diabetes data set . The features that were used in it were no. of pregnancies the patient has had, Glucose level,Blood pressure ,, Skin thickness,Insulin,BMIT, diabetes pedigree function,age.

Here is a tabulated form of the performances of Adagrad,Logistic regression and SVM on the three data sets

Data set Name	Algorithm	α	λ/C	Norm	Accuracy
Diabetes	Logistic regression	0.01	0.01	L2	0.7467532
	Logistic regression	0.01	0.001	L2	0.7792207
	Logistic regression	0.0001	0.01	L2	0.764935
	Logistic regression	0.0001	0.01	L1	0.793506
	Logistic regression	0.01	0.001	L1	0.780519
	Adagrad	0.01	0.01	L2	0.7415584
	Adagrad	0.0001	0.01	L2	0.62727
	Adagrad	0.01	0.001	L2	0.768831
	Adagrad	0.01	0.0001	L1	0.7701298
	Adagrad	0.001	0.0001	L1	0.5792207
	SVM	N/a	C=0.01	Linear	0.7857142
	SVM	N/a	C=1	Linear	0.779220

Data set Name	Algorithm	α	λ/ C	Norm	Accuracy
Breast Cancer	Logistic regression	0.01	0.001	L2	0.963157
	Logistic regression	0.001	0.01	L2	0.0.97368421
	Logistic regression	0.001	0.01	L1	0.980701
	Logistic regression	0.01	0.001	L1	0.950877
	Adagrad	0.001	0.01	L2	0.968421
	Adagrad	0.01	0.001	L2	0.0.97017
	Adagrad	0.01	0.001	L1	0.957894
	Adagrad	0.001	0.01	L1	0.9543859
	SVM	N/a	C=0.01	Linear	0.956140
	SVM	N/a	C=1	Linear	0.9754385

Data set Name	Algorithm	α	λ/ C	Norm	Accuracy
Retinopathy	Logistic regression	0.01	0.01	L2	0.67304
	Logistic regression	0.001	0.001	L2	0.74174
	Logistic regression	0.01	0.01	L1	0.67304
	Logistic regression	0.01	0.01	L1	0.67304
	Adagrad	0.01	0.01	L2	0.51043
	Adagrad	0.001	0.001	L2	0.62608
	Adagrad	0.001	0.001	L1	0.60782
	SVM	N/a	C=0.01	Linear	0.7534
	SVM	N/a	C=1	Linear	0.72365

The observations was tabulated and they are show above I have tried to keep the $\lambda\alpha$ values in a permutation of 0.001 and 0.01. So on running all previous function I saw that most common values for good accuracies

were on 0.01 and 0.001 permutation and combinations . Now with these values the convergence is fast
My expectation was that adagrad on sgd would work better and give good values ,however Logistic regression outshined.
In our initial expectation we expected Logistic Adagrad to perform better than the Logistic Algorithm . We can explain the performance of the Logistic Regression with L2 norm to be higher since in Logisitc Adagrad, we have been running 5000 iterations .In each iteration of ours , we divide our learning rate with the history of sum of squares of our gradient . Running the adagrad over 10000 iterations makes gradient sum (saved in the history) which is in the denominator ,becomes exponentially larger overpowers the alpha or learning rate value that was supposed to help perfom the adagrad better. So because the denominator was overpowering maybe that is one of the reasons why our Adagrad didnt't "outshine" the other algorithms as expected to.

6 Learning Curves

We know that the learning curves are a way of For the values of λ and α as 0.02 and 0.02 respectively I found the following curves.
For 1000 iterations and using cross validation we get