Assignment 1 - Regularization and Sparsity

Nirmal Roy, 4724429

February 27, 2018

Question 1

Loss Function vs r_+

Under the given setting the Loss Function becomes:-

$$L = \frac{1}{2}\{(1+r_+)^2 + (1-r_+)^2\} + \lambda|1-r_+| \tag{1}$$

Fig 1 shows the plot of Loss function for various values of r_+ and λ

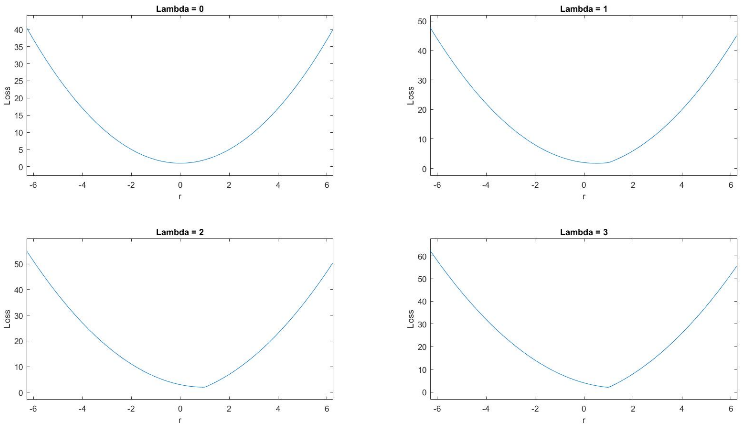

Figure 1: Plots of loss function versus r_+ for $\lambda = \{0, 1, 2, 3\}$. As can be seen for non zero values of lambda there's a discontinuity(sharp bent) at $r_+ = 1$. For $\lambda = 0$ it's a parabola and for the remaining the curves are nothing but the merging of two parabolas from the two cases $r_+ < 1$ and $r_+ > 1$. With increasing value of λ the bend becomes more prominent and the curve also becomes steeper

Minimizers

In this section we find the minimizers and the minimum value for each case of the loss function as mentioned above.

$\lambda = 0$

Taking derivative of the Loss Function $L = \frac{1}{2}\{(1+r_+)^2 + (1-r_+)^2\}$ with respect to r_+ and equating it to 0 we get $r_+ = 0$. Hence, at $r_+ = 0$ the L attains the minimum value which is 1 (putting $r_+ = 0$ in L).

For non-zero value of λ we have to consider two cases, $1 - r_+ > 0$ or $r_+ < 1$ and $1 - r_+ < 0$ or $r_+ > 1$. For the first case, loss function is $L_1 = \frac{1}{2}\{(1+r_+)^2 + (1-r_+)^2\} + \lambda(1-r_+)$ and for the second case it is $L_2 = \frac{1}{2}\{(1+r_+)^2 + (1-r_+)^2\} - \lambda(1-r_+)$. We will take derivative of L_1 and L_2 and find the minimizer for both the cases respectively.

$\lambda = 1$

Taking derivative of L_1 w.r.t. r_+ and equating it to zero we get $r_+ = 1/2$. Hence, the graph to the left of $r_+ = 1$ in the second image of fig 1 reaches it's minimum at $r_+ = 1/2$. Similarly, Taking derivative of L_2 w.r.t. r_+ and equating it to zero we get $r_+ = -1/2$. But this doesn't satisfy the second constraint $1 - r_+ < 0$. Hence, the derivative of the curve to the right of $r_+ = 1$ doesn't reach zero. Hence the minimizer of this loss function is $r_+ = 0.5$ and the minimum value is 7/4.

$\lambda = 2$

We first differentiate L_1 with respect to r_+ and equate it to zero to obtain $r_+ = 1$. But this violates the case $1 - r_+ > 0$ and hence for in the bottom left image of fig 1 we don't have an r_+ for which the derivative becomes zero. Similarly we differentiate L_2 and do the respective computations to obtain $r_+ = -1$ which again violates the case $1 - r_+ < 0$. Also the derivative of the loss function is undefined at $r_+ = 1$. Hence, for $\lambda = 2$ we don't have a point where the derivative of the loss function becomes zero. Thinking analytically, we can see that both the curves on each side of $r_+ = 1$ are convex and hence if we climb up from their respective point of minima we will reach $r_+ = 1$ where the two curves meet and form the 'bend'. Thus, that meeting point, $r_+ = 1$ will be the point where the loss function is minimum. For this case of $\lambda = 2$ the minimum value of L is 2. This analysis holds true for all values of $\lambda \geq 2$.

$\lambda = 3$

This scenario is exactly similar to the previous one. Differentiating L_1 we get $r_+ = 1.5$ which violates the case $1 - r_+ > 0$ and differentiating L_2 we get $r_+ = -1.5$ which violates the case $1 - r_+ < 0$. Hence, even for this curve there's no point where the differentiation is zero and the minimizer is at $r_+ = 1$ giving a minimum value of $L = 2$.

Question 2

A nearest mean classifier will have high variance when the number of data points per class is really low. This is because if there is say one point per class then the mean will change with new data and as a result the classifier will also change. The regularizer in the loss function equation tries to reduce the variance of the classifier by putting a constraint on the difference between the representor points. Without the regularizer the representor points are nothing but the means. Hence, the regularizer tries to minimize the distance between the representor points and effectively reduce variance of the classifier even in settings with low data points.

If the value of λ gets higher and higher the importance of the constraint part dominates the error part in the loss function. With larger λ the effective allowed distance between the representor points becomes lower. When λ becomes large enough(close to infinity) the error part of the loss function becomes inconsequential to the optimizer. Hence, in that scenario the only way to minimise the loss function is to make the difference between the representor as much close to zero as possible, and in extreme values of λ they will coincide.

Question 3

Contour Lines

In general cases, the contour lines for the Loss function need to be computed for the two cases of $r_- > r_+$ and $r_- < r_+$. Hence, we will get an ellipse (or circle when the number of data points in both the classes are equal) for each of the two cases. Thus the contour lines will be an aggregation of the ellipses(or circles). On one side of the line $r_- = r_+$, the contour will be the ellipse(or circle) corresponding to $r_- > r_+$ and on the other side it will be the ellipse(or circle) corresponding to $r_- < r_+$. The distance between the center of the ellipses(or circles) is controlled by the value of λ. Larger λ corresponds to larger distance(on the opposite sides of $r_- = r_+$ lines) , effectively narrowing the ellipse, and vice versa. Finally, we can say that the contour lines will look like boat/leaf shaped depending on the radius(dependent on the value of L) and separation between the two figures.

Optimal Solution for large λ

If we have a large enough λ as mentioned in Question 2, the regularizer completely dominates the loss function and only way to minimize the function is to lower down the difference between r_- and r_+ and when $\lambda \to \infty$, $r_- = r_+ = r$(say). In this scenario the Loss function given the data points become

$$L = \frac{1}{2}\{(1+r)^2 + (1-r)^2 + (3-r)^2 + (1+r)^2\} \tag{2}$$

To find the optimal r, we have to take the derivative of L with respect to r and equate it to 0. We get $4r - 2 = 0$ and hence $r = 1/2$. The exact solution is given by $(r_-, r_+) = (1/2, 1/2)$

Question 4

Search Strategy

The loss function in question is discontinuous and hence not differentiable when $r_- = r_+$. Hence, in order to find the optimal solution we cannot perform a normal gradient descent search. Thus we resort to a variant of gradient descent known as sub gradient descent[1] search. The sub gradient of $|x|$ is given by $sgn(x)$. Thus the gradient of our loss function with respect to r_- and r_+ are:

$$\delta(r_-) = \left(-2 * \sum_{i=1}^{N_-} \frac{1}{N_-}(x_i - r_-) \right) + \lambda sgn(r_- - r_+) \tag{3}$$

$$\delta(r_+) = \left(-2 * \sum_{i=1}^{N_+} \frac{1}{N_+}(x_i - r_+) \right) - \lambda sgn(r_- - r_+) \tag{4}$$

Hence, our update step becomes:

$$r_-(k+1) = r_-(k) - \beta * \delta(r_-(k)) \tag{5}$$

$$r_+(k+1) = r_+(k) - \beta * \delta(r_+(k)) \tag{6}$$

where β is the step size of the sub gradient descent algorithm. We kept β to be inversely proportional to the norm of gradient δ. This means that the step size becomes small when the gradient is large and it becomes larger when the gradient is small. We initialize r_- and r_+ from a random point in the respective classes.

Stopping criterion

Although fig 2 shows a nice convergence of our optimisation algorithm when we trained on the entire dataset but this might not always be the case. Since the subgradient method is not a descent method, it is common to keep track of the best point found so far, i.e., the one with smallest function value. When the minimum value does not change for a considerable amount of time we stop iterating.

[1] http://stanford.edu/class/ee364b/lectures.html

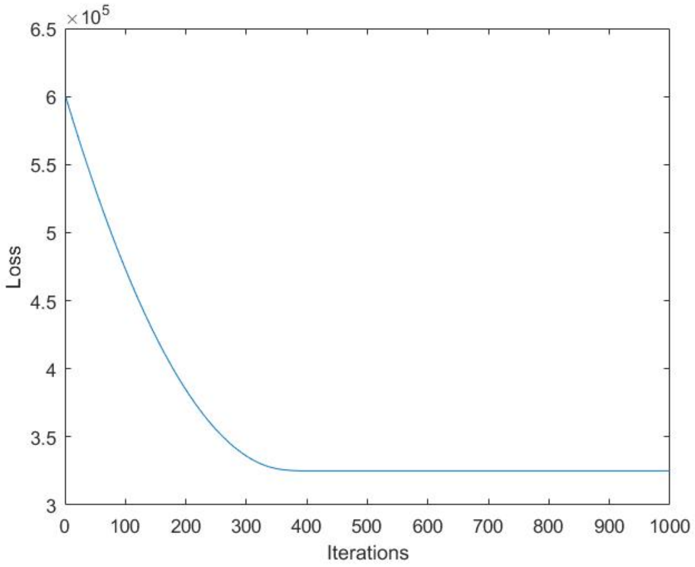

Figure 2: Behavior of the loss function for 1000 iterations of the sub gradient descent algorithm and $\lambda = 0.1$.

Representor Images

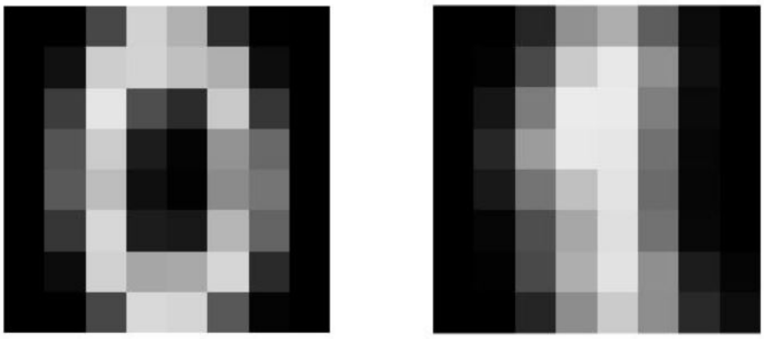

Figure 3: Representor images for $\lambda = 0$. These are images represent mean of each class of the dataset, since for $\lambda = 0$ the algorithm is nothing but gradient descent and it converges at the mean. The two images can be clearly distinguished from each other.

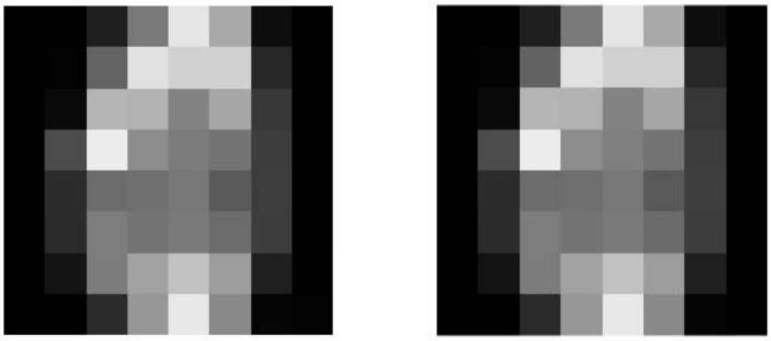

Figure 4: Representor images for $\lambda = 1000$. On further increase of λ the representor images do not change. And as expected they are almost identical meaning they have converged at the same point.

Regularization Curve

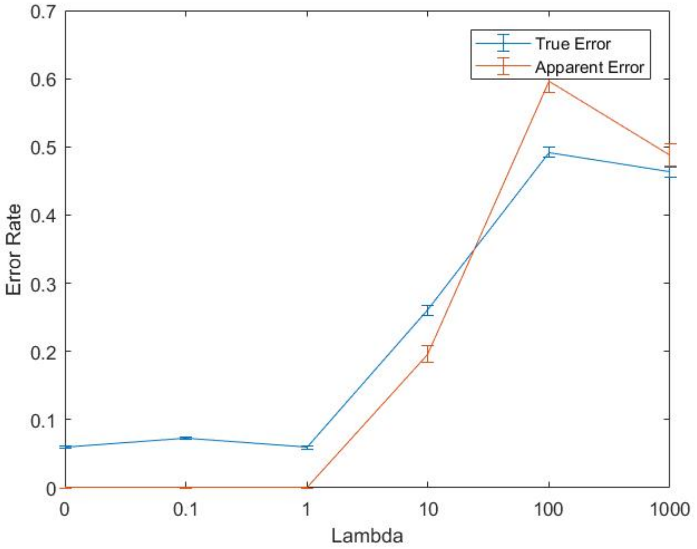

Figure 5: Regularization Curve with the estimates of true error and apparent error against various values of λ. The vertical bars represent the standard error of the mean. The errors have been averaged over 250 runs picking up random training sample from the dataset. The minimum true error of 5.39% occurs at $\lambda = 1$