CS155 Set 2

Timothy Liu

January 18, 2018

1 Problem 1

1.1 Problem A

Square loss is a poor choice because a point that is correctly classified but has a predicted value far from the target will still register as a large loss. Square loss is very sensitive to extreme outliers.

1.2 Problem B

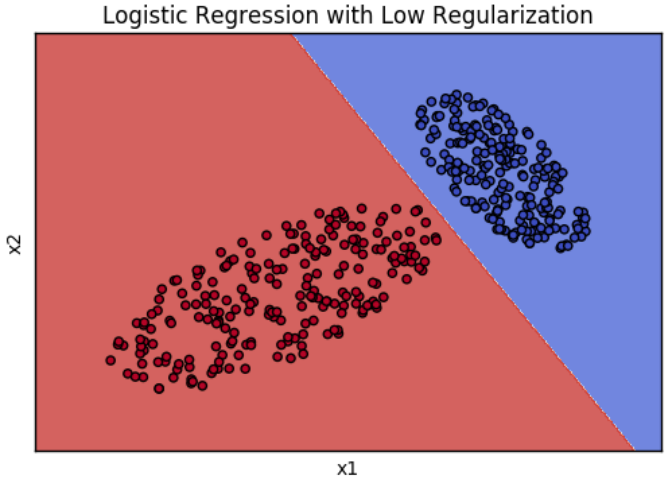

Figure 1: Logistic regression with low regularization; C = 1000

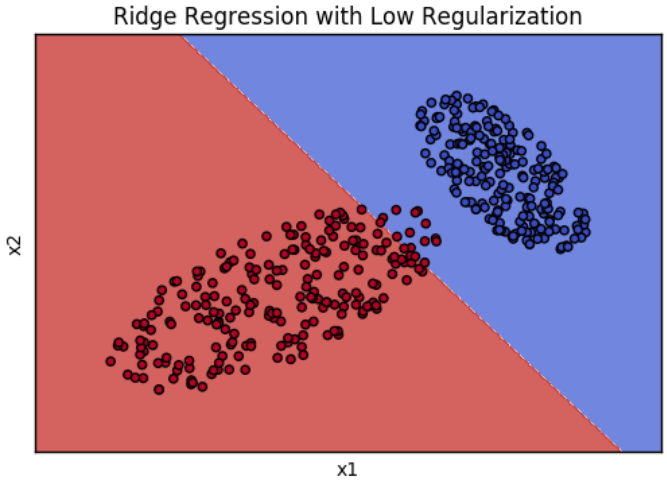

Figure 2: Ridge regression with low regularization; alpha = 0.001

The ridge regression model misclassifies some points and is closer to the red points. The ridge regression classification line appears to be drawn closer to the center of mass of the red points, resulting in the misclassification.

1.3 Problem C

Gradient of hinge loss:

$$\nabla L_{hinge} = \nabla max(0, 1 - y\mathbf{w}^T\mathbf{x})$$

if $y\mathbf{w}^T\mathbf{x} > 1$:

$$\nabla L_{hinge} = 0$$

if $y\mathbf{w}^T\mathbf{x} < 1$:

$$\nabla L_{hinge} = \nabla 1 - y\mathbf{w}^T\mathbf{x}$$
$$= -y\mathbf{x}$$

Gradient of log loss:

$$\nabla L_{log} = \nabla ln(1 + e^{-y\mathbf{w}^T\mathbf{x}})$$
$$= \nabla(1 + e^{-y\mathbf{w}^T\mathbf{x}})\frac{1}{1 + e^{-y\mathbf{w}^T\mathbf{x}}}$$
$$= \frac{-ye^{-y\mathbf{w}^T}}{1 + e^{-y\mathbf{w}^T\mathbf{w}}}\mathbf{x}$$

Point	Classifier	Hinge Grad	Log Grad
[1, 0.5, 3]	1	[-1. -0.5 -3.]	[-0.37754067 -0.18877033 -1.13262201]
[1, 2, -2]	1	[0, 0, 0]	[-0.11920292 -0.23840584 0.23840584]
[1, -3, 1]	-1	[0, 0, 0]	[0.04742587 -0.14227762 0.04742587]

Figure 3: Gradient of the weight vector for the hinge and log loss weight functions.

1.4 Problem D

The gradients from hinge loss are zero if the point is correctly classified or if the weight vector multiplied by x multiplied by the classification is at least 1. The log loss gradients are never zero, unlike the hinge loss functions. The log loss gradients will converge to zero when the product $yw^{\mathbf{T}}x$ is very large (near infinity). The training error can be reduced by maximizing the margin.

1.5 Problem E

Minimizing only L_{hinge} will result in a correct decision boundary; there may be multiple decision boundaries that have equivalent loss functions using hinge loss that have differing margins. Minimizing the term $\lambda||w||^2$ places a penalty on having a small margin and will result in a minimum margin, not just a correct decision boundary.

2 Problem 2

2.1 Problem A

Adding the penalty term will not decrease the in sample error. Without regularization, the model attempts to minimize the in sample error, but regularization is attempting to minimize a different quantity.

Adding a penalty term will decrease the out of sample errors in cases where there is overfitting in the training model. If the penalty term is excessively large and the out of sample set matches closely to the in sample set, then regularization will result in a higher out of sample error.

2.2 Problem B

The L_0 norm is rarely used because maximizing the L_0 norm is an NP-hard problem, making it computationally very challenging and inefficient. The L_0 norm is also not a continuous function.

2.3 Problem C

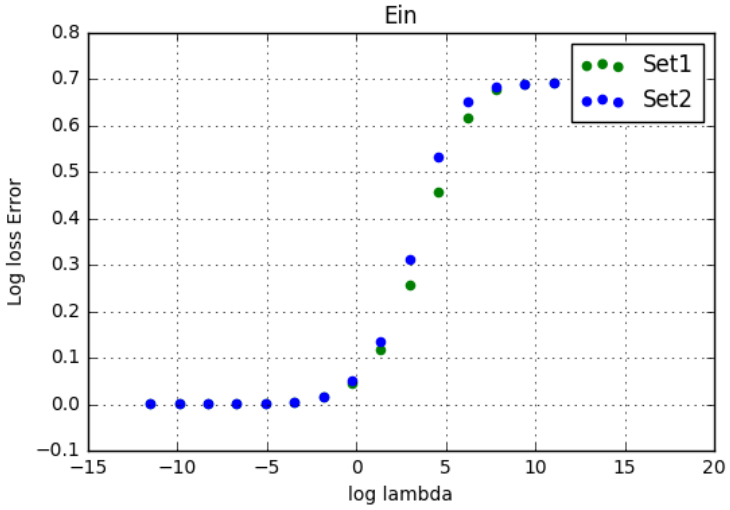

Figure 4: In sample error for varying lambda.

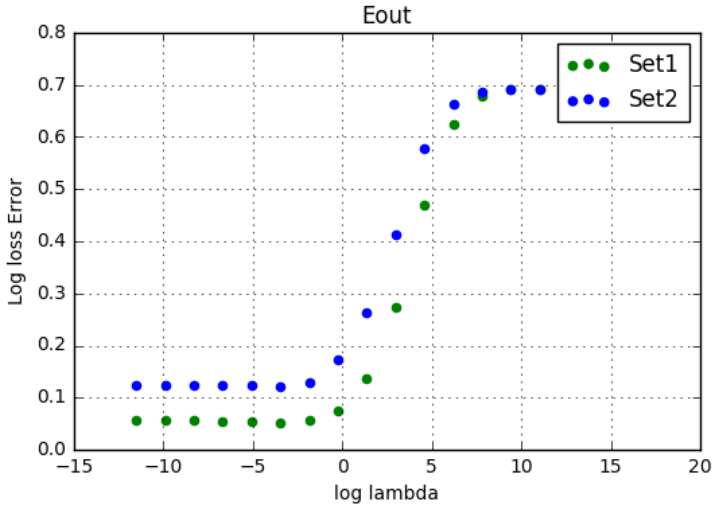

Figure 5: Out of sample error for varying lambda.

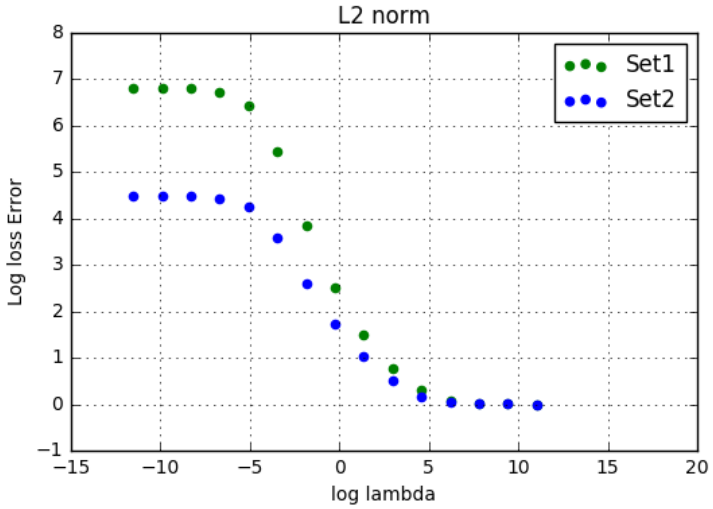

Figure 6: L2 norm for varying lambda.

2.4 Problem D

The training and errors for the two sets are almost identical. This is expected, since they are training off the same technique. The test errors for the larger dataset are consistently smaller. The larger training set is better able to generalize to the out of sample set, since the larger training set has a lower variance.

2.5 Problem E

The out of sample error for both test sets decreases slightly with larger lambda up to the sixth value of lambda for training set 1. Smaller values of lambda experience overfitting as the out of sample error rises even as the in sample error falls. Above this value of lambda, there is underfitting and the out of sample error rises significantly from the 8th to the 12th value of lambda.

2.6 Problem F

As λ increases, L2 norm for training set 1 falls from about 48 to close to zero. As lambda increase, L2 norm first falls slowly before rapidly plummeting from the 4th to 9th lambda.

2.7 Problem G

For training set 2, $\lambda = 0.00125$ has the lowest test set error. This is the best choice of λ if the data is trained on training set 2.

3 Problem 3

3.1 Problem A

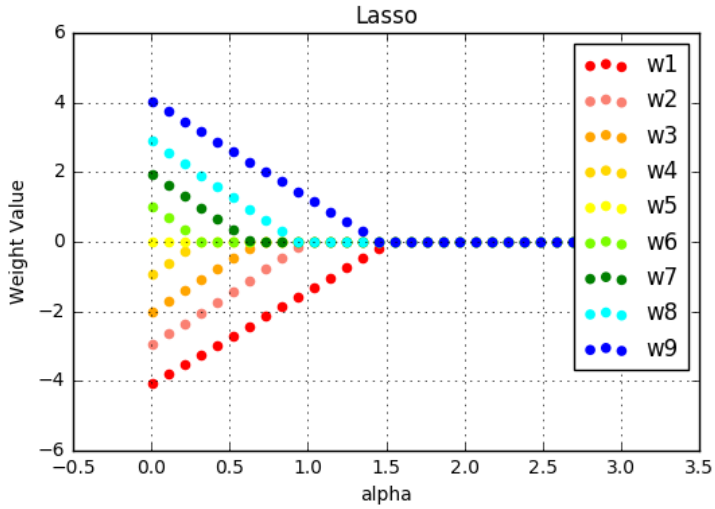

Figure 7: Value of weights as a function of alpha for lasso regularization.

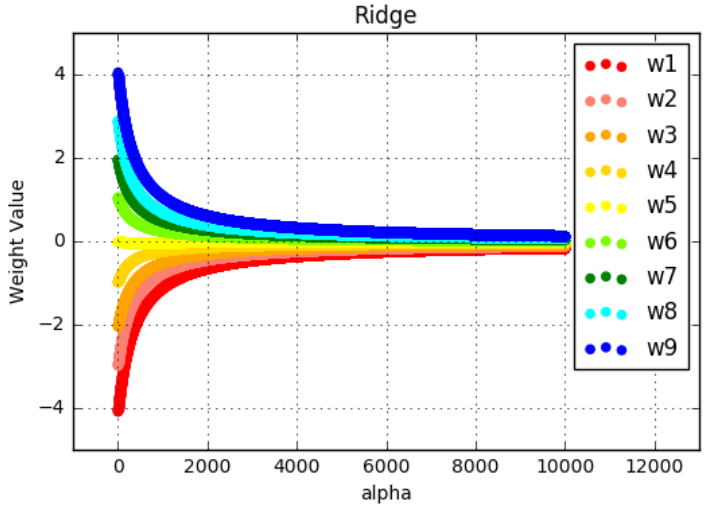

Figure 8: Value of weights as a function of alpha for ridge regularization.

As regularization increases, the number of model weights equal to zero increases for Lasso regularization. At $\alpha = 0$, exactly one model weight is zero. When α exceeds 1.5, all of the model weights are zero.

Ridge regression has none of the weights going to zero. The weights tend towards zero, but they don't converge to zero.

3.2 Problem B

3.2.1 Problem i

$$E_{in} = ||\mathbf{y} - \mathbf{x}w||^2 + \lambda||w||_1$$
$$\nabla E_{in} = -2\mathbf{x}^T(\mathbf{y} - \mathbf{x}w) + \lambda\frac{w}{||w||_1} = 0$$
$$\mathbf{x}^T\mathbf{y} - \mathbf{x}^T\mathbf{x}w - \frac{1}{2}\lambda\frac{w}{||w||_1} = 0$$

For $w > 0$:

$$\mathbf{x}^T\mathbf{y} = \mathbf{x}^T\mathbf{x}w + \frac{1}{2}\lambda$$
$$\mathbf{x}^T\mathbf{x}w = \mathbf{x}^T\mathbf{y} - \frac{1}{2}\lambda$$
$$w = (\mathbf{x}^T\mathbf{x})^{-1}(\mathbf{x}^T\mathbf{y} - \frac{1}{2}\lambda)$$

For $w < 0$:

$$\mathbf{x}^T\mathbf{y} = \mathbf{x}^T\mathbf{x}w - \frac{1}{2}\lambda$$
$$\mathbf{x}^T\mathbf{x}w = \mathbf{x}^T\mathbf{y} + \frac{1}{2}\lambda$$
$$w = (\mathbf{x}^T\mathbf{x})^{-1}(\mathbf{x}^T\mathbf{y} + \frac{1}{2}\lambda)$$

For $w = 0$ we employ subgradients:

$$\nabla E_{in} = -2\mathbf{x}^T(\mathbf{y} - \mathbf{x}w) + [-1, 1]\lambda = 0$$
$$\mathbf{x}^T\mathbf{y} = \mathbf{x}^T\mathbf{x}w + [-\frac{1}{2}, \frac{1}{2}]\lambda$$
$$\mathbf{x}^T\mathbf{x}w = \mathbf{x}^T\mathbf{y} - [-\frac{1}{2}, \frac{1}{2}]\lambda$$
$$\mathbf{x}^T\mathbf{x}w = \mathbf{x}^T\mathbf{y} + [-\frac{1}{2}, \frac{1}{2}]\lambda$$
$$w = \mathbf{x}^T\mathbf{x}^{-1}(\mathbf{x}^T\mathbf{y} + [-\frac{1}{2}, \frac{1}{2}]\lambda)$$

3.2.2 Problem ii

The weight vector is zero when

$$(\mathbf{x}^T\mathbf{y} + [-\frac{1}{2}, \frac{1}{2}]\lambda) = 0$$

The smallest value that this occurs at is $\lambda = \pm 2\mathbf{x}^T\mathbf{y}$. Since we are only considering positive λ, the $\pm$ is determined by whichever value will make λ positive.

3.2.3 Problem iii

$$E_{in} = ||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 + \lambda||\mathbf{w}||_2^2$$
$$E_{in} = ||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 + \lambda\mathbf{w}^T\mathbf{w}$$
$$\nabla E_{in} = ||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 + \lambda\mathbf{w}^T\mathbf{w} = 0$$
$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w} = 0$$
$$\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda I\mathbf{w} = 0$$
$$\mathbf{X}^T\mathbf{y} - (\mathbf{X}^T\mathbf{X} + \lambda I)\mathbf{w} = 0$$
$$(\mathbf{X}^T\mathbf{X} + \lambda I)\mathbf{w} = \mathbf{X}^T\mathbf{y}$$
$$\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

3.2.4 Problem iv

No, there is no finite value for λ that will give a weight value of zero. This is demonstrated in problem 3a. In general, the matrix $\mathbf{X}^T\mathbf{X}$ is a symmetric matrix, and when the regularization term is added ($\mathbf{X}^T\mathbf{X} + \lambda I$) is also a symmetric matrix. As λ grows larger, there is no guarantee that the matrix $(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}$ will have any zero terms. If this matrix has no zero terms, then the product $\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$ is not guaranteed to lead to any zeros, meaning that there is no finite λ that will necessarily lead to $w_i = 0$. λ must be infinity for the weights to go to zero.