CS155 Homework 2

Ty Limpasuvan

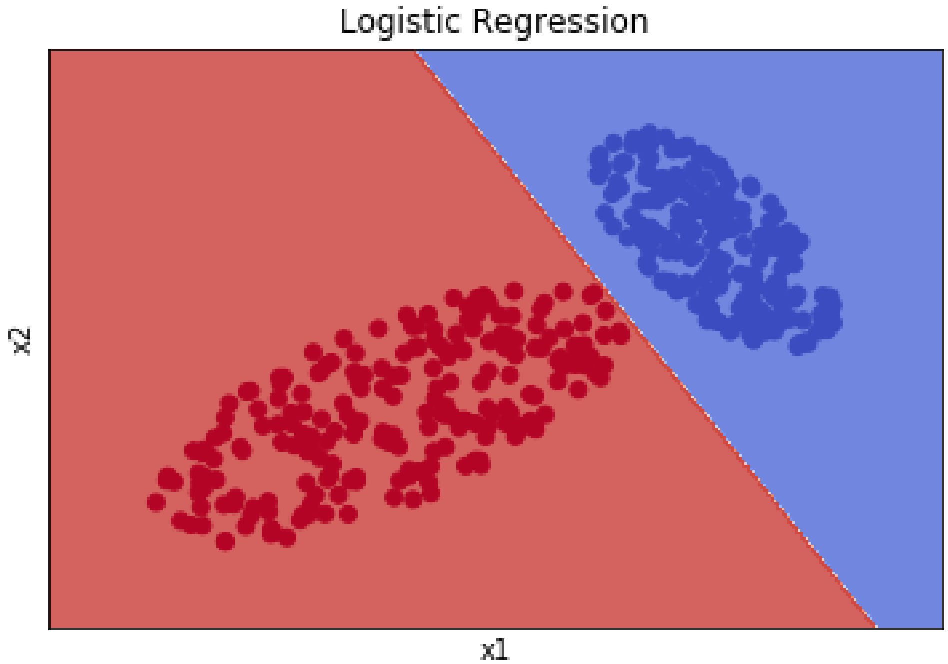
6 late hours used; 38 remaining

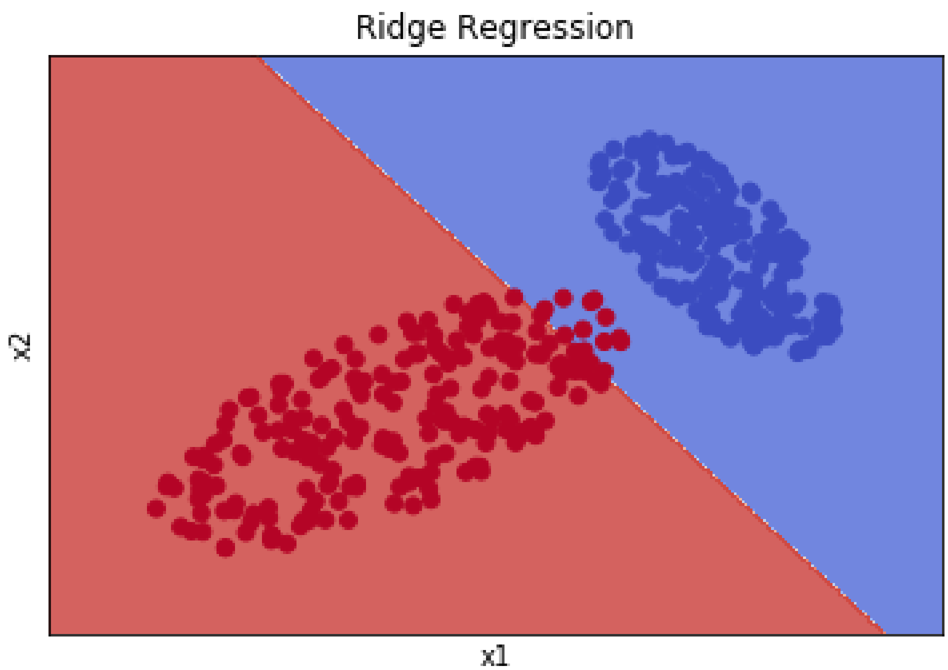
1 Comparing Different Loss Functions

1.1 A

Squared loss is a terrible choice of a loss function to train on for classification because outliers in the data can drastically affect the classifications based on squared loss.

1.2 B





With low regularization, the logistic model appeared to do a better job with classifying the points. In the ridge graph, some of the red points were classified on the blue side of the line.

1.3 C

Hinge loss gradient is given by -yx.
Log loss gradient is given by -yx(1 + e$^{yw^T x}$)$^{-1}$.
$Point1 : Hinge : (-1, -0.5, -3)$
$Log : (-0.3775, -0.1888, -1.1326)$
$Point2 : Hinge : 0$
$Log : (-0.7311, -1.4621, 1.4621)$
$Point3 : Hinge : 0$
$Log : (0.0474, -0.1423, 0.0474)$

1.4 D

The hinge loss gradient can converge to 0; its value will be zero whenever the value of 1 - yw$^T xisnegative.Theloglossgradientw$
$yxis0.Inalinearlyseparabledataset, movingthetrainingpointsthemselveswithoutadjustingtheboundarycanreducethetrainin$

1.5 E

Adding the lambda penalty term can prevent the classification from acting upon L-hinge directly.

2 Effects of Regularization

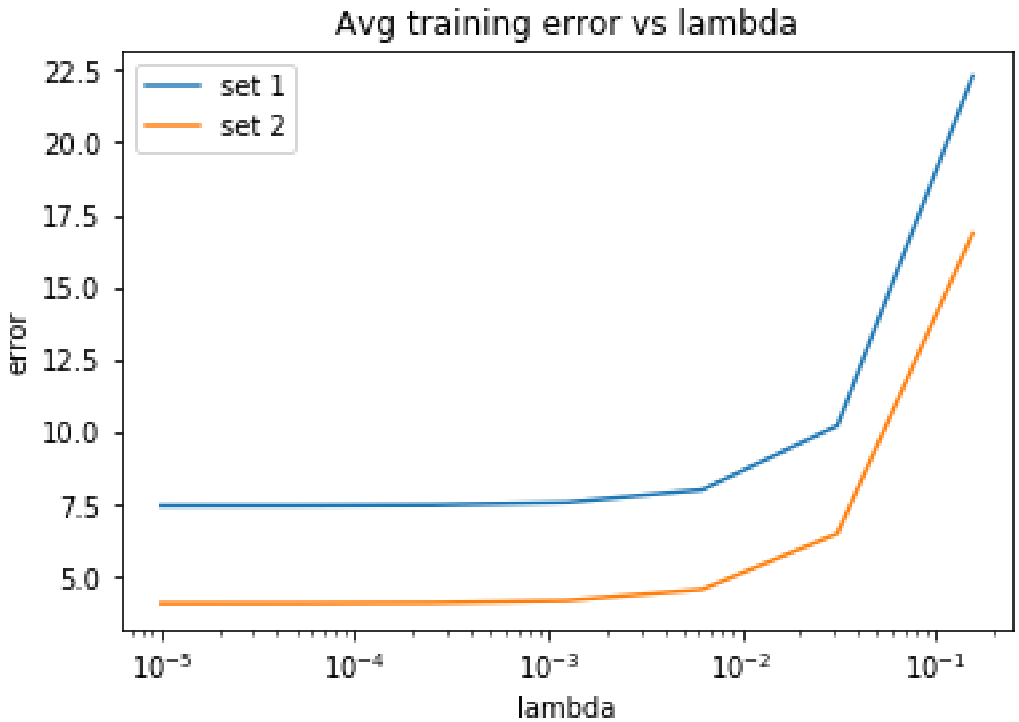
2.1 A

Adding the penalty term cannot decrease the training error; normally the loss term is minimized, but adding the penalty value will prevent the in-sample error from decreasing. It will not always decrease the out of sample error. The model does have a chance to be improved, but reducing the overfitting has the potential to increase the out of sample error.

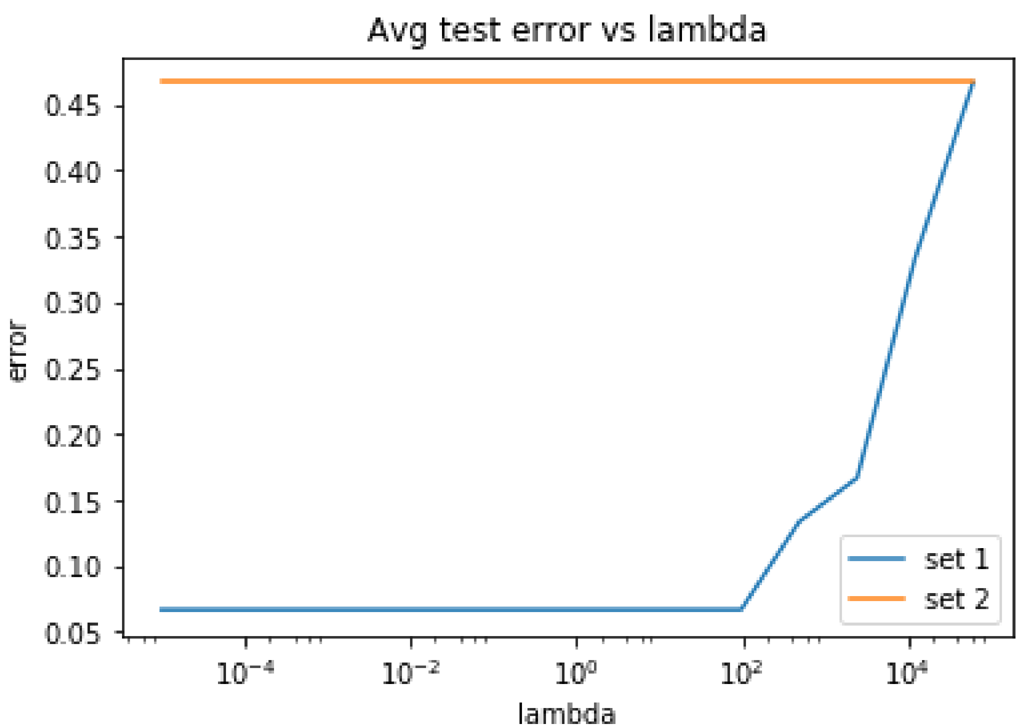
2.2 B

l0 regularization is sparse, but it is not continuous. This property makes it less desirable than l1 in most cases.

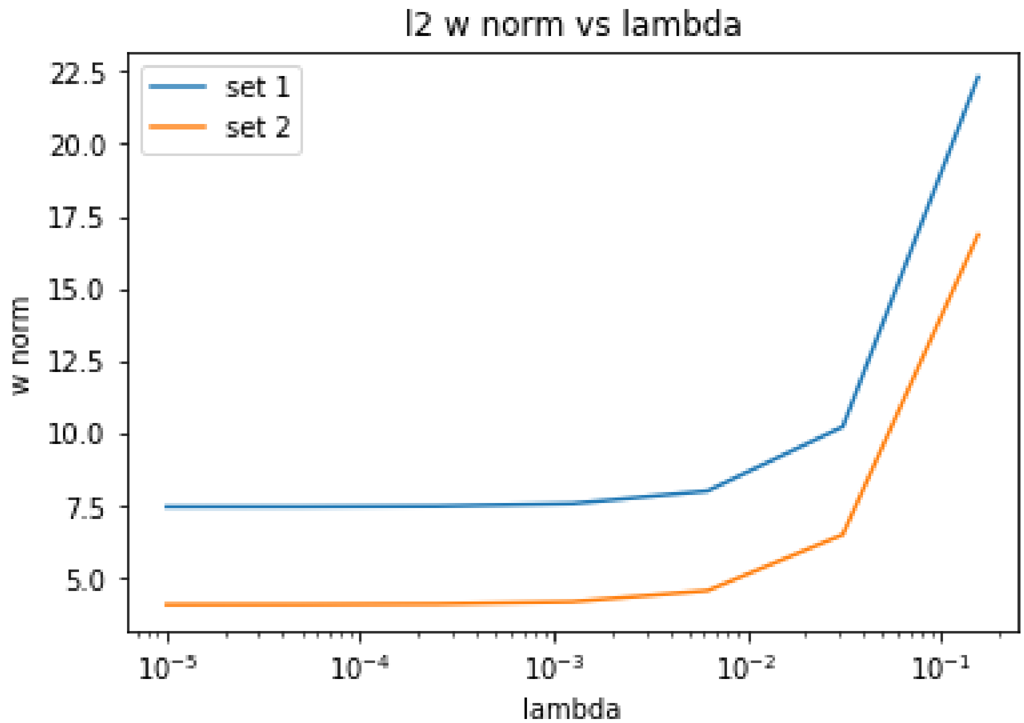
2.3 C



Avg training error vs lambda
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
error
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
lambda
set 1
set 2



Avg test error vs lambda
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
error
10^{-4}
10^{-2}
10^{0}
10^{2}
10^{4}
lambda
set 1
set 2



2.4 D

Set 2 generally had higher amounts of testing error; fitting from fewer point resulted in a lower accuracy. The training error was lower for Set 2; having fewer points allowed the model to better match the points.

2.5 E

The errors generally increased with the lambda values. Overfitting appears to occur with lambda values at the scale of 100 and higher; the out of sample error increases drastically.

2.6 F

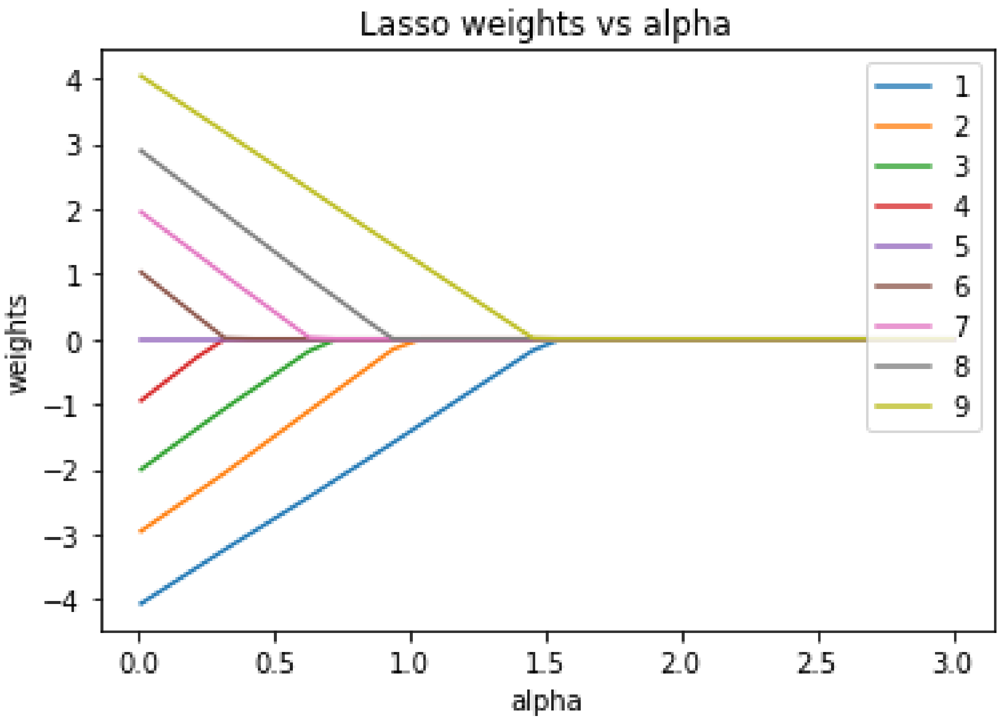
The l2 norm for weights also generally increased with lambda.

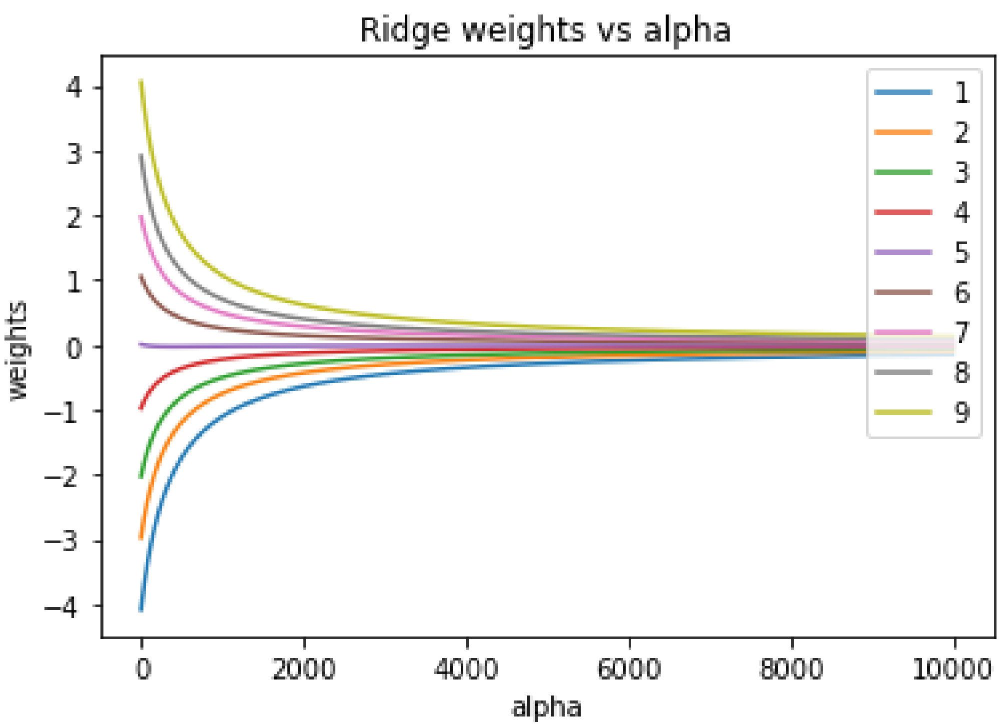
2.6.1 G

I would use a lambda value on the scale of 10^{-2}; it appears to be the point after which the training error and the norm weights increase drastically.

3 Lasso vs Ridge Regularization

3.1 A





With Lasso, the number of models with 0 weight reach zero after at most an alpha of 1.5 and stay zero for every alpha beyond 1.5. For Ridge, the number of models with exactly 0 weight decreases and approaches 0, but still doesn't reach 0 even with an alpha of 10000.

3.2 B i

$argmin|y - Xw|^2 + \lambda|w|$
$\frac{d}{dw}(|y - xw|^2 + \lambda|w|)$
$(y - xw)^T(y - xw) + \lambda|w|$
$(y^T - wx^T)(y - xw + \lambda|w|)$
$y^Ty - 2wx^Ty + w^2x^Tx$
and there are three cases for $\lambda|w|$:
-$\lambda|w| < 0$
$\lambda * C \in [-1, 1]|w| = 0$ $\lambda|w| > 0$
Let us define these 3 cases as a single variable z and carry on
$-2x^Ty + 2wx^Tx + z$
$z - 2x^Ty = -2wx^Tx$
$(x^Ty - \frac{1}{2}z)\frac{1}{x^tx} = w$
Simplifying gives:
$xy < \frac{-\lambda}{2}$
$xy > \frac{\lambda}{2}$
$xy \in [\frac{-\lambda}{2}, \frac{\lambda}{2}]$

3.3 B ii

$\frac{-\lambda}{2}$ is the smallest value for w = 0. The value xy must be between $\frac{-\lambda}{2}$ and $\frac{\lambda}{2}$ and $\frac{-\lambda}{2}$ is the smallest.