# Homework2 (late for 5 hours)

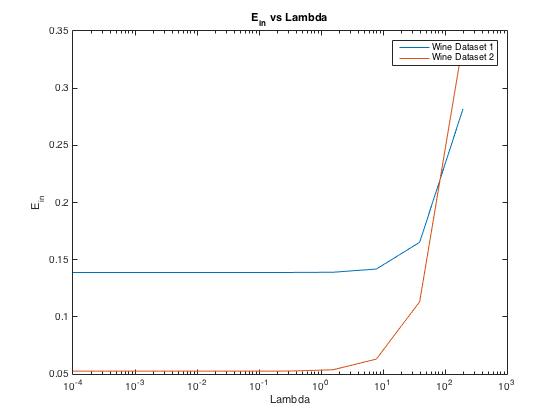
Leiya Ma

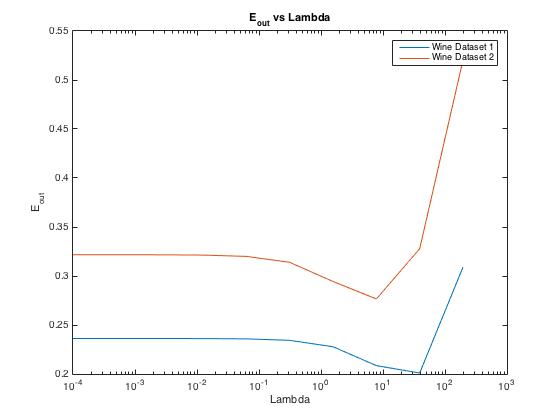
UID: 2018042

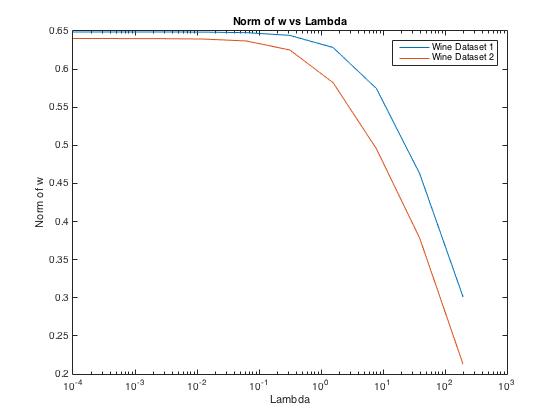
Email: lma2@caltech.edu

## 1. Effects of Regularization

* Question A: Adding the penalty term will not decrease the training(in-sample) error. This is because adding the penalty term is equivalent to constraining the model complexity which will not bring down the in-sample error. Instead, constraining the model complexity must either increase the in-sample error or keep it unchanged.
* Question B: regulation is rarely used because it involves norm of w, which is flat or discontinuous everywhere. Therefore, it is very difficult to optimize using gradient descent algorithm.
* Question C:







* Question D: For small , in wine\_training2 is slightly lower than in wine\_training1; But for big , wine\_training1 seems to give much lower . This means that more complex models are better fitting wine\_training2 and simpler models are worse fitting wine\_training2.

The reason for the former statement is that the data amount in wine\_training2 is much smaller than dataset1. Therefore, it is easier to more closely fit all the points with a more complex model in wine\_training2.

The reason for the later statement is that when more simple and generous model are used, the dataset with fewer samples may be affected more severely and thus does not fit well.

* Question E: From the beginning to the midway of the graph, we can see overfitting. We can see this because is increasing from its local minimum when is decreasing. This is overfitting because we are increasing the model complexity at the cost of accuracy when applied this model to a more generalized data.

Starting from the midway to the end of the graph, we can see underfitting. This is illustrated by the fact as we increase the , and increases dramatically. When increasing , which means we apply simpler models to the data that it cannot capture the data accurately enough.

* Question F: We can see that the norm of w decreases as increases while training with the data in wine\_training1.txt. This it because when increases, that means the propotion of the norm of w in the error function increases as well, and when we try to solve the optimization problem to minimize the error function, the norm of w gets more and more harmful. In summary, the larger is, the stricter constraint is displayed on the norm of w.
* Question G: if the model were trained with wine\_training2.txt, I would choose . Because it has the smallest .

## 2. Lasso() vs. Ridge() Regularization

* Question A:

I. Using the given priors and Bayes rule, we have

we first have

(here the and all means to find the w which maximum or minimum the function after them, I don’t know how to write the w below the argmax and argmin in work equations so first state it here.)

then we can replace it using the Bayes rule, and ignore the denominator because it has nothing to do with w

we can replace the right side with log function because it is monotonically increasing,

we can ignore the second item as it has nothing to do with w,

II. Using the given priors and Bayes rule, we have

we first have

then we can replace it using the Bayes rule, and ignore the denominator because it has nothing to do with w

we can replace the right side with log function because it is monotonically increasing,

we can ignore the second item as it has nothing to do with w,

III. we have

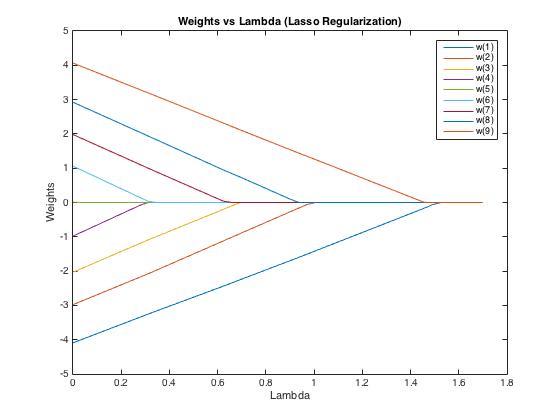
we start with

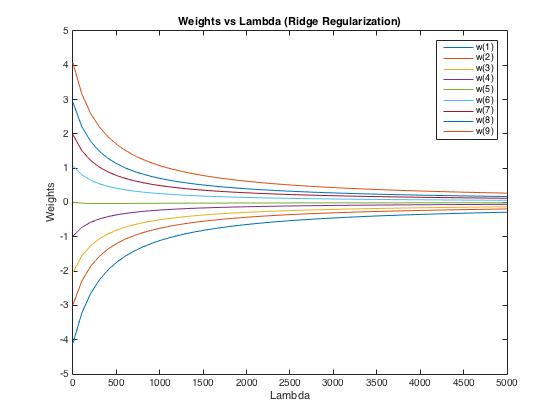
The multivariate normal distribution has a density function:

This density function describes the relative likelihood for generating the dataset given a certain weight vector w, which is obviously equivalent to .

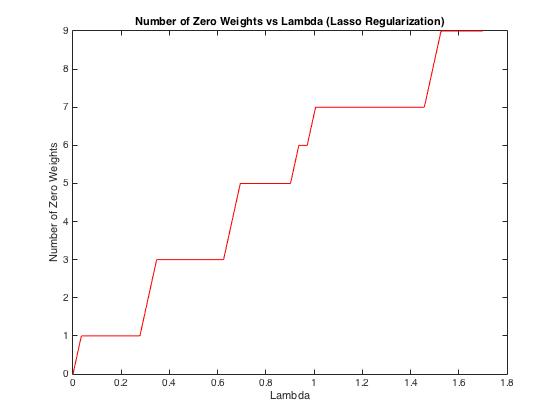
Exclude the irrelevant item,

* Question B: I.



II.

III. As regularization parameter with Lasso Regression increases, all the estimated weights go to zero (before ). As the regularization parameter with Ridge Regression increases, none of the estimated weights go to zero. The plot below shows this trend.



* Question C:

I. For , we want to find the w where its corresponding equals 0.

II. we want to explore whether there exist a value for and find out the smallest such value. We already got this from I.

therefore, we want

now we have two situation that when w>0,

when w<0, ; but notice that when w<0, , therefore, our two situation can convert to a uniform one,

which is the smallest value.

III. For , we want to find the w where its corresponding equals 0.

therefore, we have

IV. when . If we want to find a value of such that w=0,

we have

from this equation, we cannot find out the such that w=0 because the inverse of some matrix cannot equal 0. Therefore, there is no such .