Homework2 (late for 5 hours)

Leiya Ma

UID: 2018042

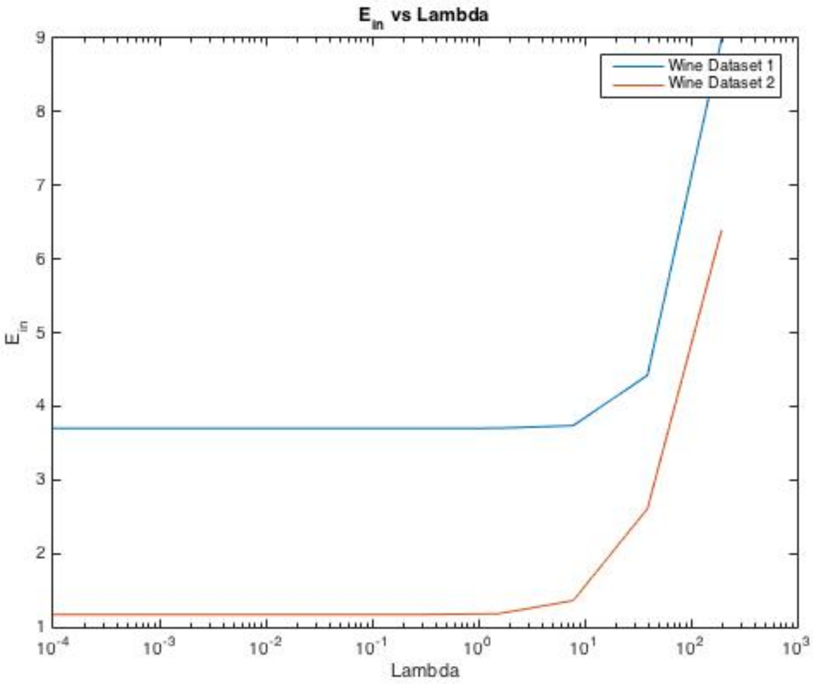
Email: lma2@caltech.edu

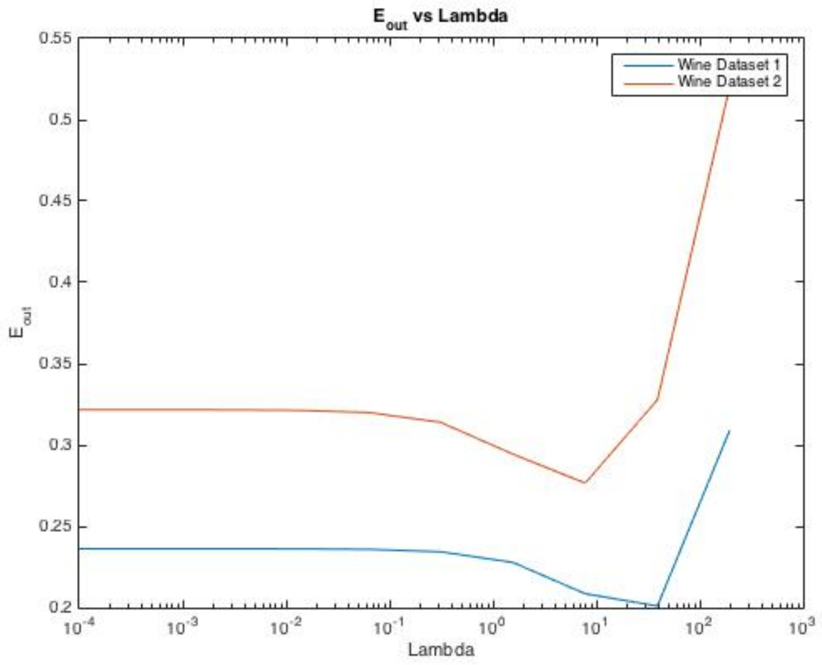
1. Effects of Regularization

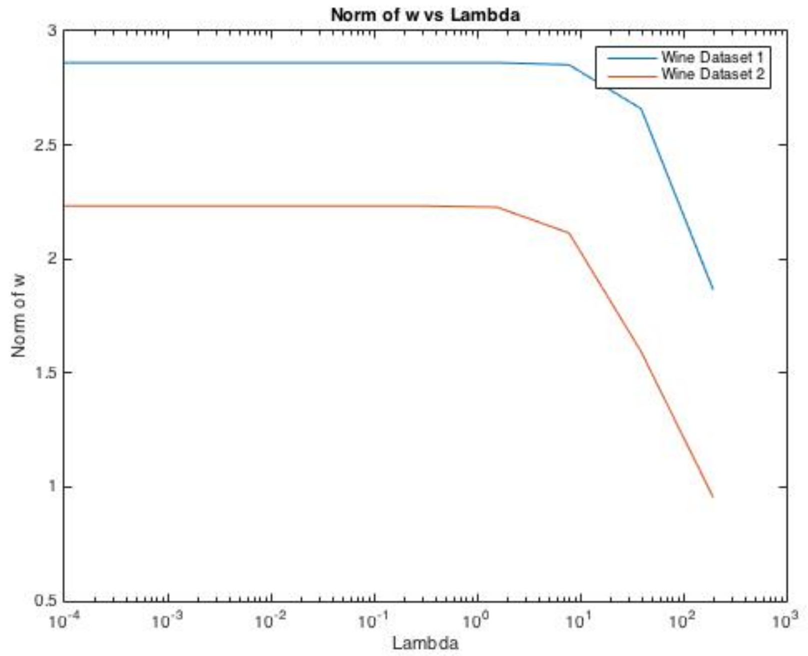
- Question A: Adding the penalty term will not decrease the training(in-sample) error. This is because adding the penalty term is equivalent to constraining the model complexity which will not bring down the in-sample error. Instead, constraining the model complexity must either increase the in-sample error or keep it unchanged.

- Question B: ℓ_0 regulation is rarely used because it involves ℓ_0 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 λ, E_{in} in wine_training2 is slightly lower than E_{in} in wine_training1; But for big λ, wine_training1 seems to give much lower E_{in}. 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 $\mathrm{E}_{\mathrm{out}}$ is increasing from its local minimum when λ 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 λ, $\mathrm{E}_{\mathrm{out}}$ and E_{in} increases dramatically. When increasing λ, 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 λ increases while training with the data in wine_training1.txt. This it because when λ 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 λ 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 $\lambda = 39.0625$. Because it has the smallest $\mathrm{E}_{\mathrm{out}}$.

2. Lasso($\boldsymbol{\ell}_1$) vs. Ridge($\boldsymbol{\ell}_2$) Regularization

- Question A:

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

$$\mathrm{p}(w|\mathcal{D}) = \frac{\mathrm{p}(\mathcal{D}|w)\mathrm{p}(w|\lambda)}{\mathrm{p}(\mathcal{D})}$$

we first have

(here the $argmax$ and $argmin$ 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.)

$$\hat{\mathrm{w}} = argmax\ p(w|\mathcal{D})$$

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

$$\hat{\mathrm{w}} = argmax\ \mathrm{p}(\mathcal{D}|w)\mathrm{p}(w|\lambda)$$

$$\hat{w} = argmax\ \mathrm{p}(\mathcal{D}|w) \prod_{j=1}^{\mathcal{D}} \frac{\lambda}{2} e^{-\lambda|w_j|}$$

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

$$\hat{w} = argmax\ \log\ (\mathrm{p}(\mathcal{D}|w) \prod_{j=1}^{\mathcal{D}} \frac{\lambda}{2} e^{-\lambda|w_j|})$$

$$\hat{w} = argmin - \log\ (\mathrm{p}(\mathcal{D}|w) \prod_{j=1}^{\mathcal{D}} \frac{\lambda}{2} e^{-\lambda|w_j|})$$

$$\hat{w} = argmin - (\mathrm{logp}(\mathcal{D}|w) + \sum_{\mathrm{j}=1}^{\mathcal{D}} \log \frac{\lambda}{2} e^{-\lambda|w_j|})$$

$$\hat{w} = argmin - (\mathrm{logp}(\mathcal{D}|w) + \sum_{\mathrm{j}=1}^{\mathcal{D}} \log \frac{\lambda}{2} + \sum_{j=1}^{\mathcal{D}} loge^{-\lambda|w_j|})$$

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

$$\hat{w} = argmin - (\mathrm{logp}(\mathcal{D}|w) + \sum_{j=1}^{\mathcal{D}} loge^{-\lambda|w_j|})$$

$$\hat{w} = argmin - (\mathrm{logp}(\mathcal{D}|w) + \sum_{j=1}^{\mathcal{D}} -\lambda|w_j|)$$

$$\hat{w} = argmin - \mathrm{logp}(\mathcal{D}|w) +\ \lambda\|w\|_1)$$

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

$$\mathrm{p}(w|\mathcal{D}) = \frac{\mathrm{p}(\mathcal{D}|w)\mathrm{p}(w|\lambda)}{\mathrm{p}(\mathcal{D})}$$

we first have

$$\hat{w} = argmax\ p(w|\mathcal{D})$$

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

$$\hat{w} = argmax\ \mathrm{p}(\mathcal{D}|w)\mathrm{p}(w|\lambda)$$

$$\hat{w} = argmax\ \mathrm{p}(\mathcal{D}|w) \prod_{j=1}^{\mathcal{D}} \sqrt{\frac{\lambda}{\pi}} e^{-\lambda {w_j}^2}$$

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

$$\widehat{\mathrm{w}} = argmax \log\,(\mathrm{p}(\mathcal{D}|w)\prod_{j=1}^{\mathcal{D}}\sqrt{\frac{\lambda}{\pi}}e^{-\lambda {w_j}^2})$$

$$\widehat{\mathrm{w}} = argmin - \log\,(\mathrm{p}(\mathcal{D}|w)\prod_{j=1}^{\mathcal{D}}\sqrt{\frac{\lambda}{\pi}}e^{-\lambda {w_j}^2})$$

$$\widehat{\mathrm{w}} = argmin - (\mathrm{logp}(\mathcal{D}|w) + log\prod_{j=1}^{\mathcal{D}}\sqrt{\frac{\lambda}{\pi}}e^{-\lambda {w_j}^2})$$

$$\widehat{\mathrm{w}} = argmin - (\mathrm{logp}(\mathcal{D}|w) + \sum_{\mathrm{j}=1}^{\mathcal{D}}\log\sqrt{\frac{\lambda}{\pi}} + \sum_{j=1}^{\mathcal{D}} loge^{-\lambda {w_j}^2})$$

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

$$\widehat{\mathrm{w}} = argmin - (\mathrm{logp}(\mathcal{D}|w) + \sum_{j=1}^{\mathcal{D}} loge^{-\lambda {w_j}^2})$$

$$\widehat{\mathrm{w}} = argmin - \mathrm{logp}(\mathcal{D}|w) + \ \lambda\|w\|_2^2$$

III. we have

$$\mathrm{y}{\sim}\mathcal{N}(\mathrm{Xw}, \sigma^2\mathrm{I}), \mathcal{D}\text{ contains X and y}$$

we start with

$$\widehat{\mathrm{w}} = argmax\ p(\mathcal{D}|w)$$

The multivariate normal distribution has a density function:

$$\frac{1}{\sqrt{(2\pi)^2|\sigma^2 I|}}e^{-\frac{1}{2}(y-Xw)^T\left(\sigma^2 I\right)^{-1}(y-Xw)}$$

This density function describes the relative likelihood for generating the dataset $\mathcal{D}$ given a certain weight vector w, which is obviously equivalent to $p(\mathcal{D}|w)$.

$$\widehat{\mathrm{w}} = argmax\ \frac{1}{\sqrt{(2\pi)^2|\sigma^2 I|}}e^{-\frac{1}{2}(y-Xw)^T\left(\sigma^2 I\right)^{-1}(y-Xw)}$$

Exclude the irrelevant item,

$$\widehat{\mathrm{w}} = argmax\ e^{-\frac{1}{2}(y-Xw)^T\left(\sigma^2 I\right)^{-1}(y-Xw)}$$

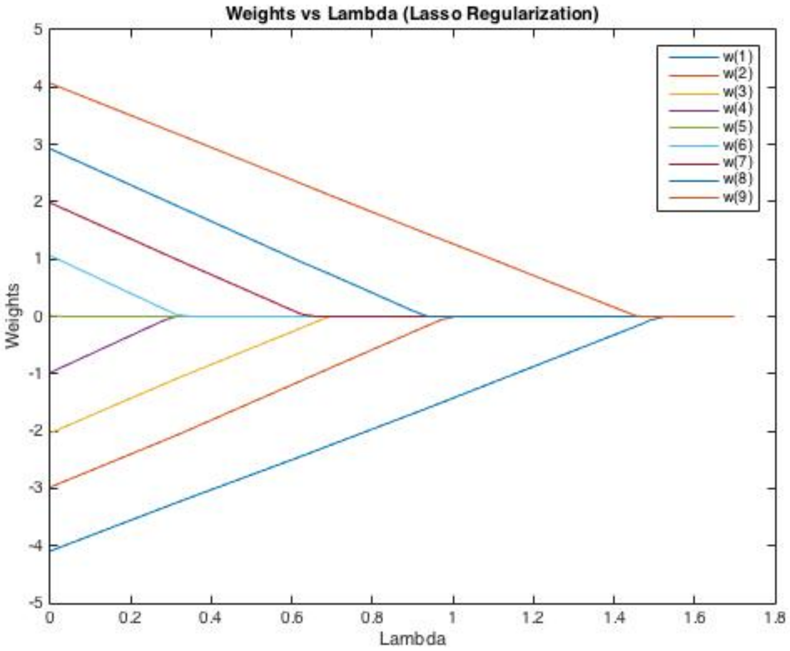
$$\widehat{\mathrm{w}} = argmax \log\,(e^{-\frac{1}{2}(y-Xw)^T\left(\sigma^2 I\right)^{-1}(y-Xw)})$$

$$\widehat{\mathrm{w}} = argmax\ (-\frac{1}{2}(y - Xw)^T(\sigma^2 I)^{-1}(y - Xw))$$

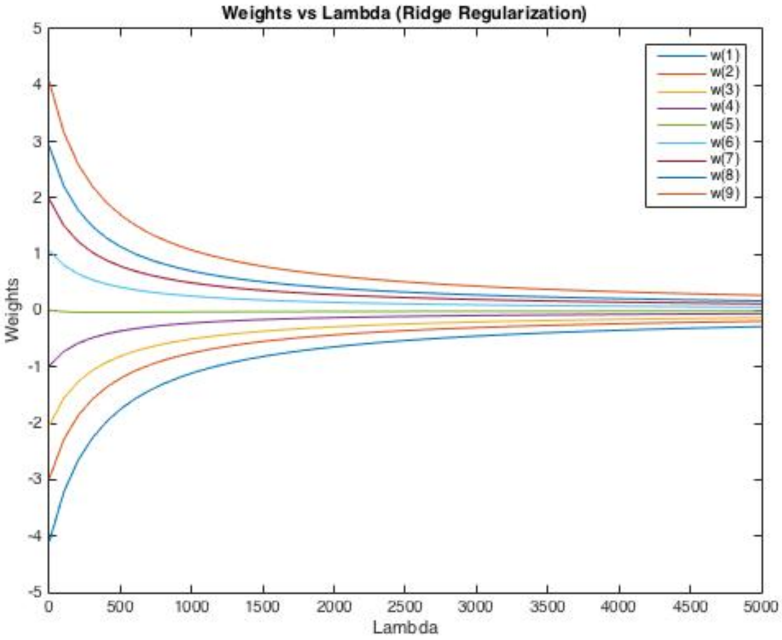
$$\hat{w} = argmin\ \frac{1}{2}(y - Xw)^T(y - Xw)$$

$$\hat{w} = argmin\ \|(y - Xw)\|_2^2$$

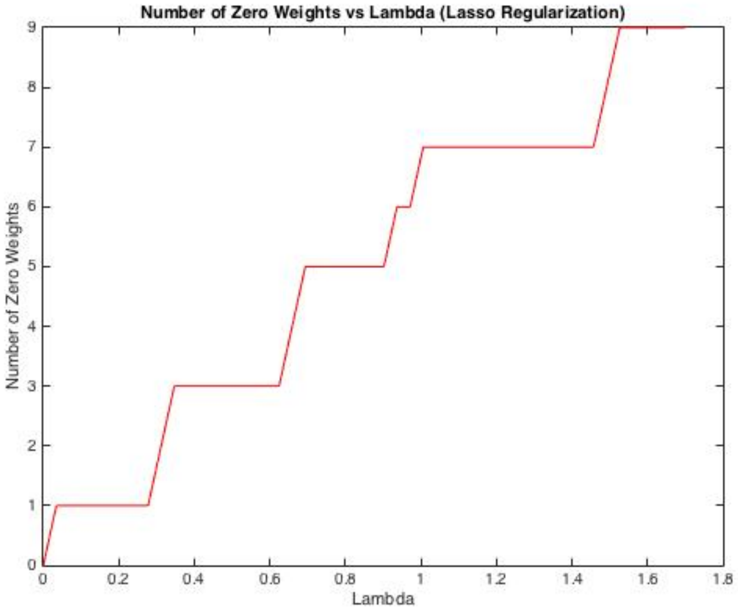
- Question B: I.



II.



III. As regularization parameter with Lasso Regression increases, all the estimated weights go to zero (before $\lambda = 1.8$). 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 $\operatorname{argmin} \|y - w^T x\|^2 + \lambda \|w\|_1$, we want to find the w where its corresponding $\nabla f(x)$ equals 0.

$$f(x) = \|y - w^T x\|^2 + \lambda \|w\|_1$$

$$\nabla f(x) = -2\mathrm{X}^{\mathrm{T}}(y - Xw) \pm \lambda I$$

$$\nabla f(x) = -2\mathrm{X}^{\mathrm{T}}\mathrm{y} + 2\mathrm{X}^{\mathrm{T}}\mathrm{Xw} + \frac{\lambda w}{|w|} = 0$$

$$\mathrm{w} = (2\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}(2X^T y \mp \lambda I)$$

II. we want to explore whether there exist a value for $\mathrm{w}_1 = 0$ and find out the smallest such value. We already got this from I.

$$\mathrm{w} = (2\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}(2X^T y \mp \lambda I)$$

therefore, we want

$$\mathrm{w} = (2X^T y \mp \lambda I) = 0$$

now we have two situation that when w>0, $0 = 2\mathrm{X}^{\mathrm{T}} y - \lambda \mathrm{I} \Longrightarrow \lambda = 2X^T y$
when w<0, $0 = 2\mathrm{X}^{\mathrm{T}} y + \lambda I \Longrightarrow \lambda = -2X^T y$; but notice that when w<0, $\mathrm{X}^{\mathrm{T}} y < 0$,
therefore, our two situation can convert to a uniform one,

$$\lambda = \|2X^T y\|$$

which is the smallest value.

III. For $\mathrm{argmin}\|y - w^T x\|^2 + \lambda\|w\|_2^2$, we want to find the w where its corresponding $\nabla f(x)$ equals 0.

$$f(x) = \|y - w^T x\|^2 + \lambda\|w\|_2^2$$
$$\nabla f(x) = 2X^T(Xw - y) + 2\lambda w = 0$$
$$2X^T Xw - 2X^T y + 2\lambda w = 0$$

therefore, we have

$$\mathrm{w} = (\mathrm{X^T X} + \lambda I)^{-1} X^T y$$

IV. when $\lambda = 0, \mathrm{w_i} \neq 0$. If we want to find a value of λ such that w=0,

$$\mathrm{w} = (\mathrm{X^T X} + \lambda I)^{-1} X^T y$$

we have

$$(\mathrm{X^T X} + \lambda I)^{-1} X^T y = 0$$

$$(\mathrm{X^T X} + \lambda I)^{-1} = 0$$

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