Machine Learning Homework 1

Yuxiao Qi

September 21, 2018

1 Problem 1

For this problem, I tested d from 1 to 40. Following Figure 1 to Figure 6 is the plot for different d.

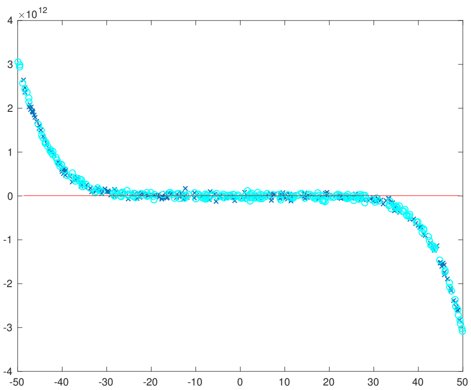

Figure 1: d=1

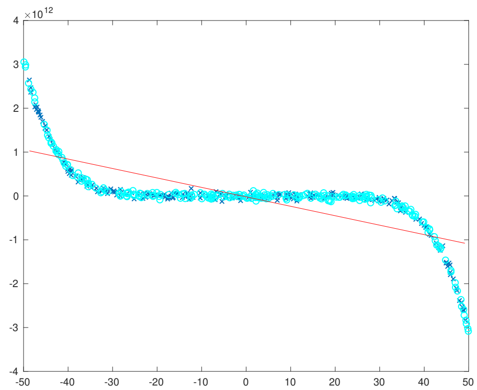

Figure 2: d=2

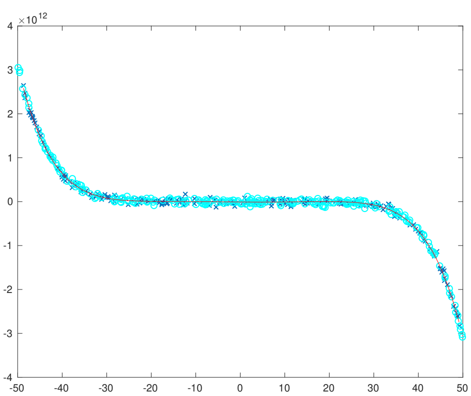

Figure 3: d=8

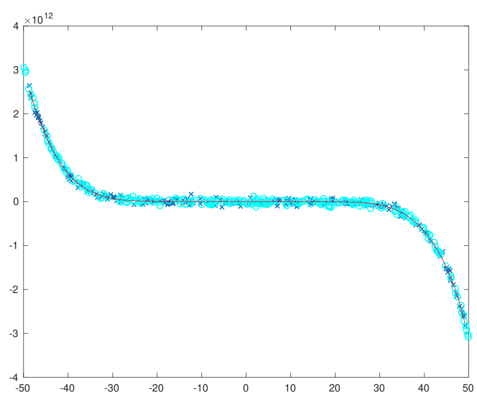

Figure 4: d=15

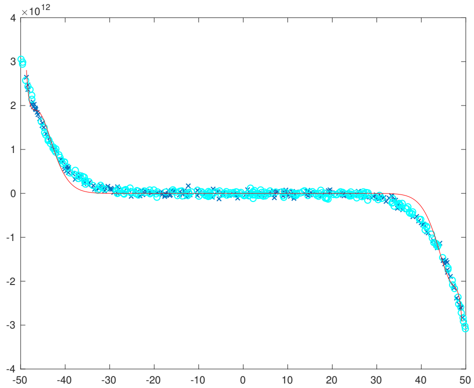

Figure 5: d=30

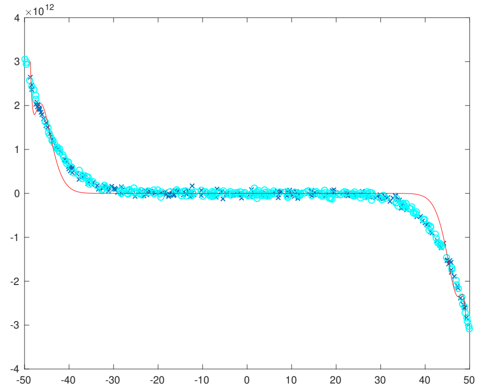

Figure 6: d=40

Following is the plot for Cross-Validation of test error. We can see from this plot, the test error is minimized when $d = 8$. When $d > 8$, juding fromt the Test Error, it became overfit data. We can also see that from Figure 4 to Figure 6

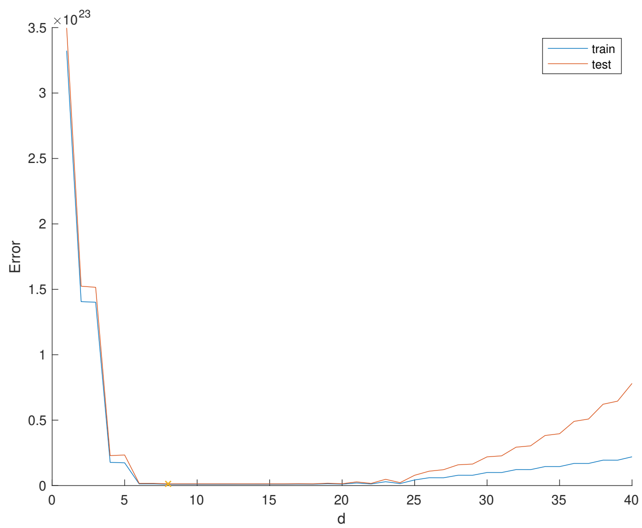

Figure 7: Relation between Test Error's and d

2 Problem 2

From the l_2 loss function, by solving gradient=0, we can get:

$$\begin{aligned}
\nabla R_{\text{reg}}(\theta) &= 0 \\
\nabla_{\boldsymbol{\theta}}\Big(\frac{1}{2N}||\mathbf{y}-\mathbf{x}\boldsymbol{\theta}||^2+\frac{\lambda}{2N}||\boldsymbol{\theta}||^2\Big) &= 0 \\
\frac{1}{2N}\nabla_{\boldsymbol{\theta}}\Big((\mathbf{y}-\mathbf{x}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{y}-\mathbf{x}\boldsymbol{\theta})+\lambda\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}\Big) &= 0 \\
\frac{1}{2N}\nabla_{\boldsymbol{\theta}}(\mathbf{y}^{\mathrm{T}}\mathbf{y}-2\mathbf{y}^{\mathrm{T}}\mathbf{x}\boldsymbol{\theta}+\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}^{\mathrm{T}}\mathbf{x}\boldsymbol{\theta}+\lambda\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}) &= 0 \\
\frac{1}{2N}(-2\mathbf{y}^{\mathrm{T}}\mathbf{x}+2\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}^{\mathrm{T}}\mathbf{x}+2\lambda\boldsymbol{\theta}^{\mathrm{T}}) &= 0 \\
\mathbf{x}^{\mathrm{T}}\mathbf{x}\boldsymbol{\theta}+\lambda\boldsymbol{\theta} &= \mathbf{x}^{\mathrm{T}}\mathbf{y} \\
\boldsymbol{\theta}^{*} &= (\mathbf{x}^{\mathrm{T}}\mathbf{x}+\lambda\boldsymbol{I})^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{y}
\end{aligned}$$

Seeing from the following plot, as λ increased, training error increased. In the mean while, test error decreased drastically, and when λ is around 792, we got the minimized test error.

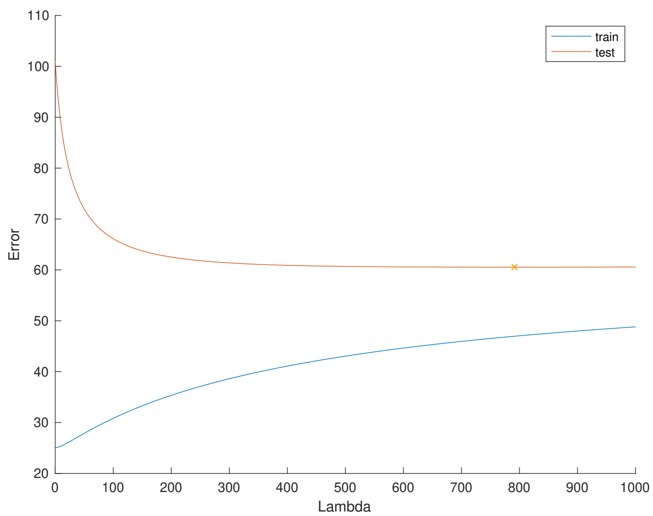

Figure 8: Relation between Test Error's and d

3 Problem 3

*Proof.*1 Prove $g(z) = 1 - g(z)$ when $g(z) = \dfrac{1}{1+\exp(-z)}$.

$$\begin{aligned}
g(-z) &= \frac{1}{1+\exp(z)} && (1)\\
1-g(z) &= 1-\frac{1}{1+\exp(-z)}\\
&= \frac{1+\exp(-z)-1}{1+\exp(-z)}\\
&= \frac{\exp(-z)}{1+\exp(-z)}\\
&= \frac{1}{\frac{1}{\exp(-z)}+1}\\
&= \frac{1}{1+\exp(z)} && (2)
\end{aligned}$$

Thus, from (1) and (2), we proved $g(z) = 1 - g(z)$.

*Proof.*2 Given $y = g(z) = \dfrac{1}{1+\exp(-z)}$, prove $g^{-1}(y) = \ln(\dfrac{y}{1-y})$.

$$\begin{aligned}
\ln(\frac{y}{1-y}) &= \ln(\frac{\frac{1}{1+\exp(-z)}}{1-\frac{1}{1+\exp(-z)}})\\
&= \ln(\frac{1}{1+\exp(-z)-1})\\
&= \ln(1)-\ln(-z)\\
&= z && (3)
\end{aligned}$$

Thus, we proved $g^{-1}(y) = \ln(\dfrac{y}{1-y})$.

4 Problem 4

Given classification function:

$$f(\mathbf{x};\boldsymbol{\theta}) = \big(1+\exp(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x})\big)^{-1}$$

Empirical risk with logistic loss:

$$R_{emp}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(y_i-1)\log\big(1-f(\mathbf{x};\boldsymbol{\theta})\big) - y_i\log\big(f(\mathbf{x};\boldsymbol{\theta})\big).$$

$$\begin{aligned}
\nabla_\theta R &= \nabla_\theta\Big(\frac{1}{N}\sum_{i=1}^{N}(y_i-1)\log\big(1-f(\mathrm{x}_i;\theta)\big) - y_i\log\big(f(\mathrm{x}_i;\theta)\big)\Big)\\
&= \nabla_\theta\Big(\frac{1}{N}\sum_{i=1}^{N}(y_i-1)\log\big(\frac{\exp(\theta^{\mathrm{T}}\mathrm{x}_i)}{1+\exp(\theta^{\mathrm{T}}\mathrm{x}_i)}\big) - y_i\log\big(\frac{1}{1+\exp(\theta^{\mathrm{T}}\mathrm{x}_i)}\big)\big)\Big)\\
&= \frac{1}{N}\sum_{i=1}^{N}(y_i-1)\frac{d}{d_\theta}\Big(\big(\log\big(\exp(\theta^{\mathrm{T}}\mathrm{x}_i)\big) - \log\big(1+\exp(\theta^{\mathrm{T}}\mathrm{x}_i)\big)\Big) - \frac{d}{d_\theta}y_i\Big(-\log\big(1+\exp(\theta^{\mathrm{T}}\mathrm{x}_i)\big)\Big)\\
&= \frac{1}{N}\sum_{i=1}^{N}(y_i-1)\Big(-\mathrm{x}_i + \frac{\mathrm{x}_i\exp(\theta^{\mathrm{T}}\mathrm{x}_i)}{1+\exp(\theta^{\mathrm{T}}\mathrm{x}_i)}\Big) - y_i\Big(\frac{\mathrm{x}_i\exp(\theta^{\mathrm{T}}\mathrm{x}_i)}{1+\exp(\theta^{\mathrm{T}}\mathrm{x}_i)}\Big)
\end{aligned}$$

We consider the following combination of ϵ and η:

Table 1: The value of ϵ and η

	ϵ	η
1	0.005	0.1
2	0.002	0.5
3	0.001	1
4	0.001	2

Scenario 1

In Scenario 1, decision boundary is $\theta = [0.5868, 2.3928, -0.2696]$, the plot of decision boundary, and binary classification error and the empirical risk is as below:

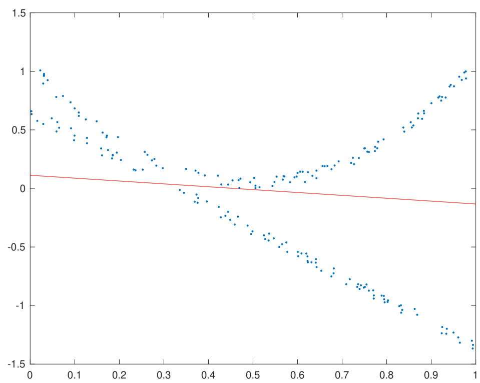

Figure 9: Result of decision boundary

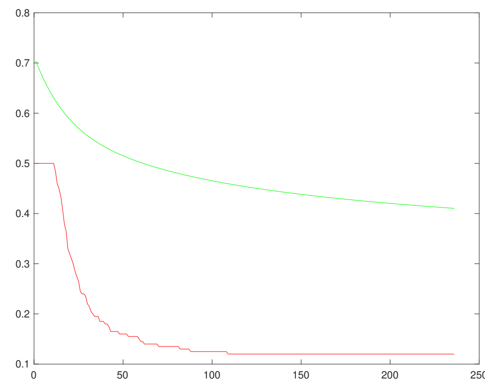

Figure 10: Classification error and the empirical risk

Scenario 2

In Scenario 2, decision boundary is $\theta = [20.8247, 12.9278, -8.8047]$, the plot of decision boundary, and binary classification error and the empirical risk is as below:

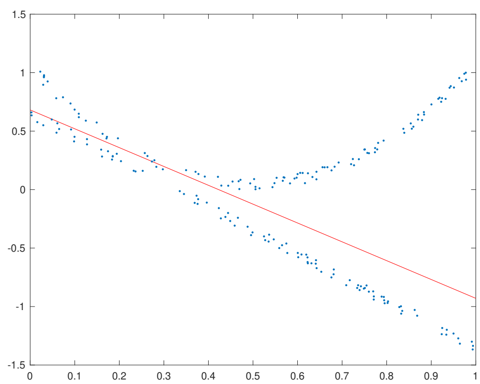

Figure 11: Result of decision boundary

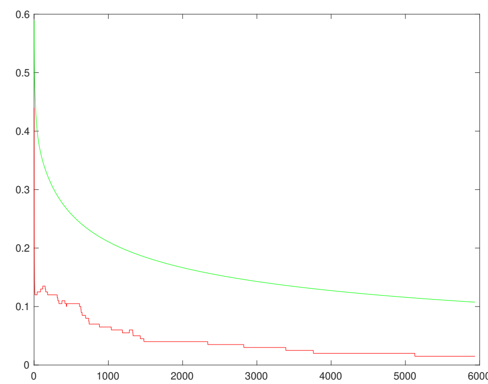

Figure 12: Classification error and the empirical risk

Scenario 3

In Scenario 3, decision boundary is $\theta = [48.8850, 26.8970, -19.2476]$, the plot of decision boundary, and binary classification error and the empirical risk is as below:

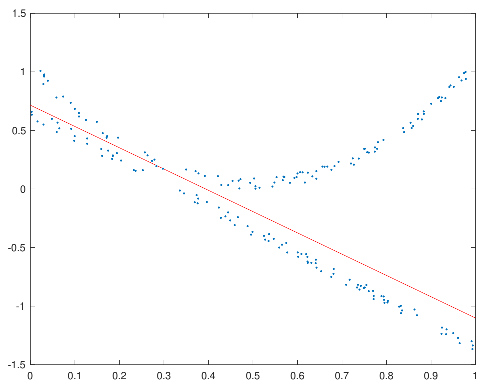

Figure 13: Result of decision boundary

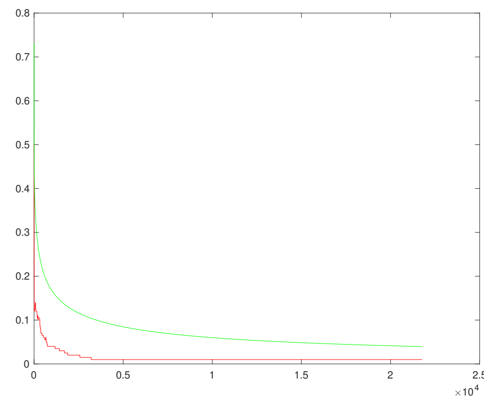

Figure 14: Classification error and the empirical risk

Scenario 4

In Scenario 4, decision boundary is $\theta = [68.6235, 36.5859, -26.4126]$, the plot of decision boundary, and binary classification error and the empirical risk is as below:

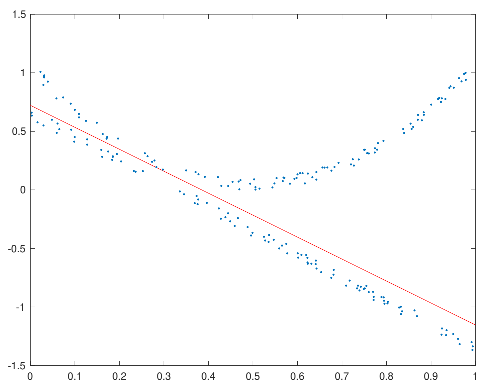

Figure 15: Result of decision boundary

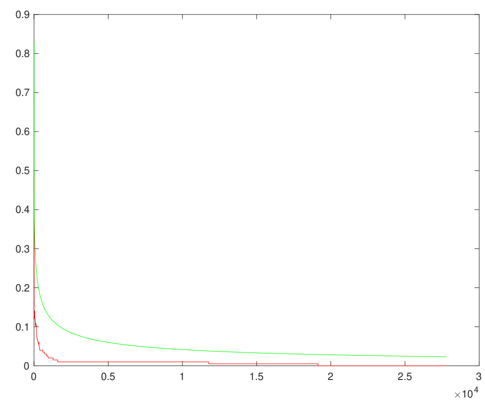

Figure 16: Classification error and the empirical risk