Tianyong Zhang
CS M146
HW 2
404-743-024

Problem 1:

a) AND: $y = \theta^T x + b$, $\theta^T = (w_0, w_1, b)$

w_0	w_1	result
0	0	-1
0	1	-1
1	0	-1
1	1	1

Therefore:
$$\begin{cases} -w_0 - w_1 + b < 0 \\ -w_0 + w_1 + b < 0 \\ w_0 - w_1 + b < 0 \\ w_0 + w_1 + b > 0 \end{cases}$$

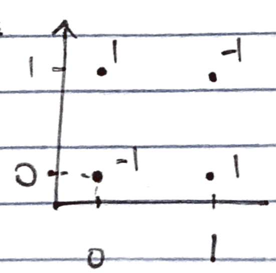
one solution is: $w_0 = 1$, $w_1 = 1$, $b = -1$
another solution is: $w_0 = 1$, $w_1 = 2$, $b = -2$
$\therefore$ It is not unique.

b). XOR: $y = \theta^T x + b$, $\theta^T = (w_0, w_1, b)$

w_0	w_1	result
0	0	-1
0	1	1
1	0	1
1	1	-1

$$\begin{cases} -w_0 - w_1 + b < 0 & ① \\ -w_0 + w_1 + b > 0 & ② \\ w_0 - w_1 + b > 0 & ③ \\ w_0 + w_1 + b < 0 & ④ \end{cases}$$

There is no perceptron exsist, since:
① + ④ $= 2b < 0$, ② + ③ $= 2b > 0$, which is contradict, also, by graph:



There is no perceptron exsit to separate -1 and 1.

Problem 2:

$$J(\theta) = -\sum_{n=1}^{N}\left[y_n \log h_\theta(x_n) + (1-y_n)\log(1-h_\theta(x_n))\right]$$

$h_\theta(x_n) = \sigma(\theta^T x)$, by differentiate:

$$\frac{\partial J}{\partial \theta_j} = -\sum_{n=1}^{N} y_n x_{n,j}(1-\sigma(\theta^T x_n)) - x_{n,j}(1-y_n)(\sigma(\theta^T x_n))$$

$$= -\sum_{n=1}^{N} y_n x_{n,j} - \cancel{y_n x_{n,j}\sigma(\theta^T x_n)} - x_{n,j}\sigma(\theta^T x_n) + \cancel{y_n x_{n,j}\sigma(\theta^T x_n)}$$

$$= -\sum_{n=1}^{N} y_n x_{n,j} - x_{n,j}\sigma(\theta^T x_n)$$

$$= -\sum_{n=1}^{N} x_{n,j}(y_n - \sigma(\theta^T x_n))$$

$$= \sum_{n=1}^{N} x_{n,j}(\sigma(\theta^T x_n) - y_n)$$

Transfer back to h_θ:

$$= \sum_{n=1}^{N} x_{n,j}(h_\theta(x_n) - y_n)$$

Problem 3:

a). $$\frac{\partial J}{\partial \theta_0} = \sum_{n=1}^{N} 2w_n(\theta_0 + \theta_1 x_{n,1} - y_n)$$

$$\frac{\partial J}{\partial \theta_1} = \sum_{n=1}^{N} 2w_n(\theta_0 + \theta_1 x_{n,1} - y_n)\cdot x_{n,1}$$

b). $\sum 2w_n(\theta_0+\theta_1 x_{n,1} - y_n) = 0$ - - - ①

$\sum 2w_n(\theta_0+\theta_1 x_{n,1} - y_n)\cdot x_{n,1} = 0$ - - - ②

From ①:

$\sum w_n\theta_0 = \sum(w_n y_n - w_n\theta_1 x_{n,1})$ - - - ③

From ②:

$\sum w_n\theta_0 x_{n,1} = \sum(w_n y_n x_{n,1} - w_n\theta_1 x_{n,1}^2)$ - - - ④

From ③:

$$\theta_0 = \frac{\sum(w_n y_n - w_n\theta_1 x_{n,1})}{\sum w_n} \text{ - - - ⑤}$$

Substitute ⑤ into ④:

$$\frac{\sum(w_n y_n - w_n\theta_1 x_{n,1})}{\sum w_n}\cdot\sum w_n x_{n,1} = \sum(w_n y_n x_{n,1} - w_n\theta_1 x_{n,1}^2)$$

$$\Downarrow$$

$$\sum(w_n y_n)\sum(w_n x_{n,1}) - \sum(w_n\theta_1 x_{n,1})\sum(w_n x_{n,1}) = \sum w_n\cdot\sum(w_n y_n x_{n,1} - w_n\theta_1 x_{n,1}$$

$$\Downarrow$$

$$\sum w_n\sum w_n\theta_1 x_{n,1}^2 - \sum(w_n\theta_1 x_{n,1})\sum w_n x_{n,1} = \sum w_n\sum w_n y_n x_{n,1} - \sum w_n y_n\sum w_n x_{n,1}$$

$$\Downarrow$$

$$\theta_1\left(\sum w_n\sum w_n x_{n,1}^2 - \left(\sum w_n x_{n,1}\right)^2\right) = \sum w_n\sum w_n y_n x_{n,1} - \sum w_n y_n\sum w_n x_{n,1}$$

$$\boxed{\theta_1 = \frac{\sum_{n=1}^{N} w_n \sum_{n=1}^{N} w_n y_n x_{n,1} - \sum_{n=1}^{N} w_n y_n \cdot \sum_{n=1}^{N} w_n x_{n,1}}{\sum_{n=1}^{N} w_n \cdot \sum_{n=1}^{N} w_n x_{n,1}^2 - \left(\sum_{n=1}^{N} w_n x_{n,1}\right)^2}}$$

$$\boxed{\theta_0 = \frac{\sum_{n=1}^{N}(w_n y_n - w_n\theta_1 x_{n,1})}{\sum_{n=1}^{N} w_n}}$$ substitute θ_1 with

Problem 4:

a). Since data set D is linear separable, exist a hyperplane $\vec{v}^T\vec{x} + p$ such that:

$$\min_{(\vec{x},y)\in D,\, y=1} (\vec{v}^T\vec{x}+p) \geq 0 > \max_{(\vec{x},y)\in D,\, y=-1} (\vec{v}^T\vec{x}+p)$$

Therefore, let $\vec{x}_i$ be the positive sample which is closet to $\vec{v}^T\vec{x}+p$, let $\vec{x}_j$ be the negative sample which is closet to $\vec{v}^T\vec{x}+p$.

let $P_{pos} = \vec{v}^T\vec{x}_i + p$

$P_{neg} = \vec{v}^T\vec{x}_j + p$

From definition of linear separable, $P_{pos} \geq 0 > P_{neg}$, therefore, exist $\eta \geq 0$ s.t $P_{pos} - \eta \geq P_{neg} - \eta$.

For some value of η, we have $\vec{v}^T\vec{x}+p-\eta = 0$ separate D for both $\vec{x}_i$ and $\vec{x}_j$, and the distance from $\vec{x}_i$ and $\vec{x}_j$ to $\vec{v}^T\vec{x}+p-\eta = 0$ is the same. So we have:

$$\frac{|\vec{v}^T\vec{x}_i + p - \eta|}{\|\vec{v}\|} = \frac{|\vec{v}^T\vec{x}_j + p - \eta|}{\|\vec{v}\|}$$

$$P_{pos} - \eta = -(P_{neg} - \eta)$$

$$\eta = \frac{P_{pos} - P_{neg}}{2}$$

With $\min_{(\vec{x},y)\in D,\, y=1} (\vec{v}^T\vec{x}+p-\eta) = \frac{P_{pos} - P_{neg}}{2}$

$$\max_{(\vec{x},y)\in D,\, y=-1} (\vec{v}^T\vec{x}+p-\eta) = -\frac{P_{pos} - P_{neg}}{2}$$

Therefor, $y(\vec{v}^T\vec{x}+p-\eta) \geq \frac{P_{pos} - P_{neg}}{2}$ for all $(\vec{x},y)\in D$

Since $P_{pos} > P_{neg}$, $\eta = \frac{P_{pos} + P_{neg}}{2}$, it becomes:

$$y(\vec{w}^T\vec{x}+\theta) \geq 1-\delta, \ \forall (\vec{x},y)\in D$$

where $\vec{w} = \frac{\vec{v}}{\eta}$, $\theta = \frac{p-\eta}{\eta}$, and δ with optimal solution $\delta = 0$.

b). if there is optimal $\delta = 0$, then:

$$y_i(\vec{w}^T \vec{x}_i + \theta) \geq 1 \ , \ \forall (\vec{x}_i, y_i) \in D$$

Therefore:

$$y(\vec{w}^T \vec{x} + \theta) \geq 1 \geq 0 \ , \ \forall (\vec{x}, y) \in D, \ y = 1$$

$$y(\vec{w}^T \vec{x} + \theta) \leq -1 < 0 \ , \ \forall (\vec{x}, y) \in D, \ y = -1$$

which satisfy the condition of linear separable.

c) With $\delta > 0$, if $1 - \delta > 0$, we can easily know the data set is linear separable by using same method as b). If $\delta \geq 1$, we cannot make sure it is linear separable. If the minimal $\delta \geq 1$, the data set is not linear separable.

d). The optimal solution is $\vec{w} = 0$, $\theta = 0$, $\delta = 0$. The issue with this formula is that it is not a hyperplane with this optimal solution.

e). (Seems the question should be $\vec{x}_1^T = [1\ 1 \dots 1]$, $\vec{x}_2^T = [-1\ -1 \dots -1]$ Since $\vec{x}_i$ is n-dimensional vector)

The data set is separable since there are only two samples. Hence the optimal $\delta = 0$, and $\vec{w}$, θ follow the constraints:

$$w_1 + w_2 + \dots + w_n + \theta \geq 1$$

$$-(-w_1 - w_2 - \dots - w_n + \theta) \geq 1$$

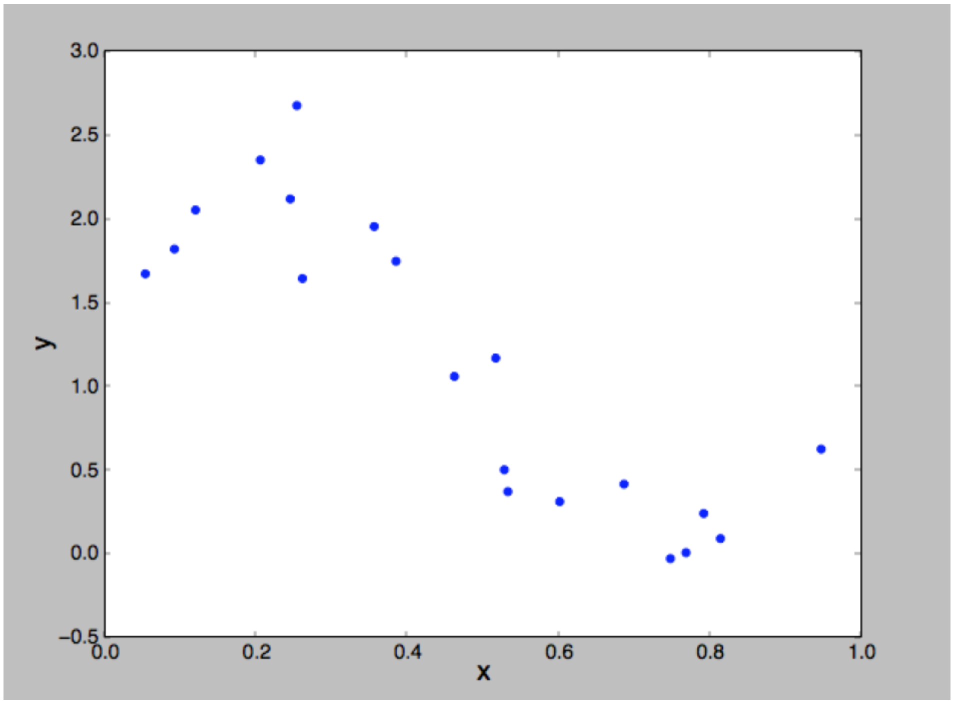
Therefore $w_1 + w_2 + \dots + w_n \geq 1 + |\theta|$

Thus, optimal solution would be $(\vec{w}, \theta, \delta)$ with $\delta = 0$, $w_1 + w_2 + \dots + w_n \geq 1 + |\theta|$.

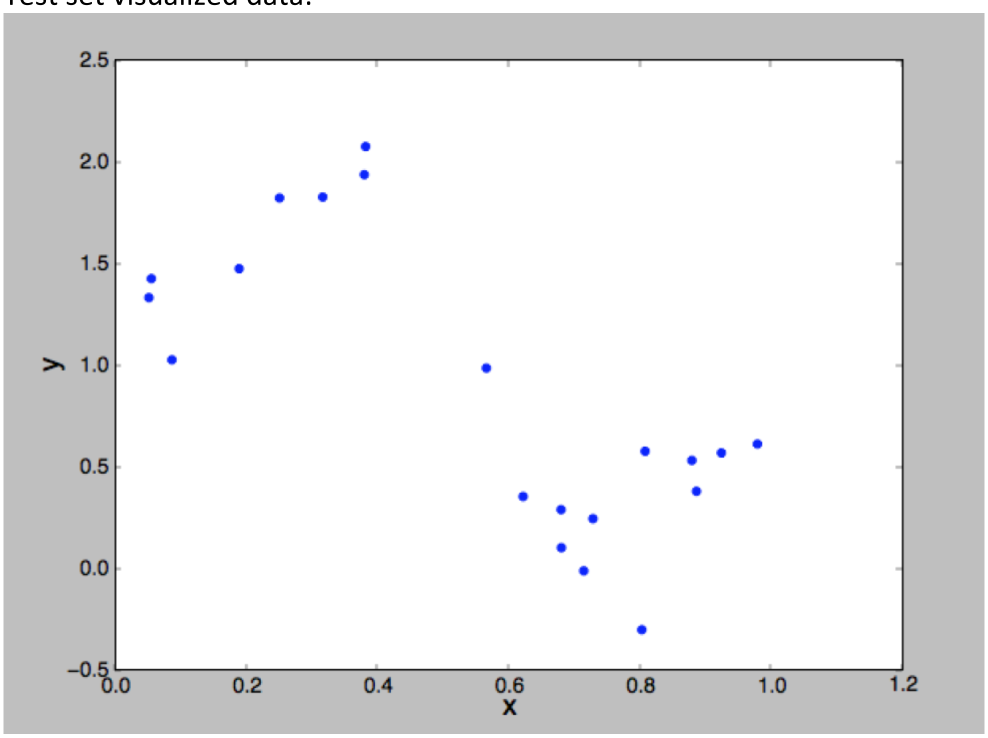
Problem 5:

a).

Train set visualized data:



Test set visualized data:



From the two plots, we can see that the data set could be linear separate easily. Although there are two “noisy” data in the middle of Train set plot, we could still separate the data set into positive and negative groups. Therefore, linear regression would be effective in predicting the data.

d).

η	Coefficient	Cost	Iterations	Time(s)
0.0001	[1.91573585 -1.74358989]	5. 4935655887	10000	0. 670198
0.001	[2.4463815 -2.81630184]	3.91257640947	10000	0.566816
0.01	[2.44640699 -2.81635338]	3.91257640579	1490	0.075016
0.0407	[2.44640706 -2.81635352]	3.91257640579	383	0.020842

Besides when η = 0.0001 which is not converge, the coefficient of rests is almost the same. η = 0.0001 and η = 0.001 reached the limit of 10000 iterations, and they took about the same time. η = 0.01 and η = 0.0407 took much less iterations and time.

e)

For closed-form, the resulting coefficient is: [2.44640709 -2.81635359], and the cost is: 3.91257640579, which is about same as those get from GD. The time spent is: 0.00278s, which is much faster than GD. However, the lower time cost on closed-form solution is because of the small size of the data set. If we run both methods on a very large data set, eventually GD would be faster than closed-form solution.

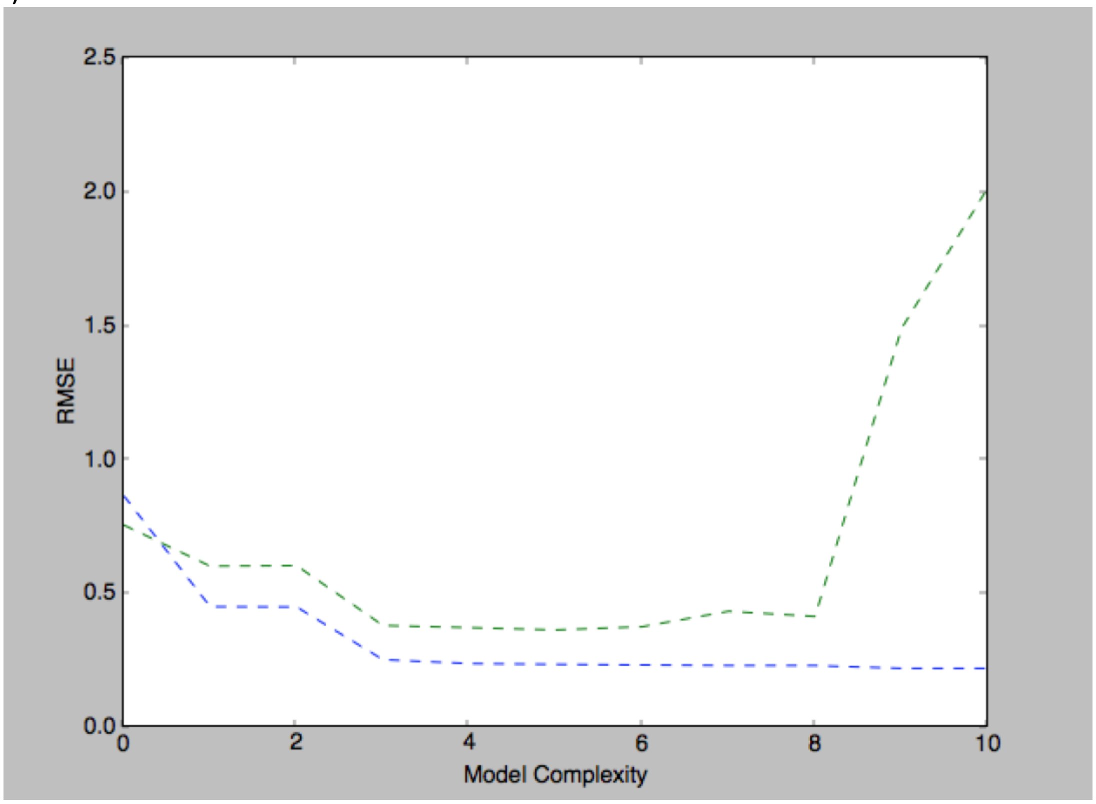
f)

The iteration is: 10000, time spent is: 0.6054s, and the cost is: 3.91257642432, which used up the limit of iterations. The coefficient I got is: [2.44634965 -2.81623746].

h).

Root-Mean-Square error (RMSE) represents the sample standard deviation of the differences between real data and predicted data, RMSE would normalizes the errors. $J(\theta)$ only measures the magnitude which doesn’t match. Therefore, RMSE is preferable to measure overfitting problem since it doesn’t rely on not-matched data only.

i).



*Green line indicates test error; Blue line indicates train error.

From the graph, we could know that degree 4 and 5 best fit the data since the error rate is the lowest and most steady. Test error after degree 6 is increasing very fast since overfitting happens.