CS M146,Winter 2018
Problem Set 2

Xin Liu, 505037053

1 Perceptron

(a) AND

One valid perceptron:$w = \begin{bmatrix} 1 \\ 1 \\ -1.5 \end{bmatrix}$, and the corresponding hyperplane is $x_1 + x_2 - 1.5 = 0$

Another valid perceptron: $w = \begin{bmatrix} 1 \\ 2 \\ -2.5 \end{bmatrix}$, and the corresponding hyperplane is $x_1 + 2x_2 - 2.5 = 0$

(b) XOR

A valid perceptron does not exist, for XOR is not linearly separable.

2 Logistic Regression

$$\frac{\partial J}{\partial \theta_j} = -\frac{1}{\partial \theta_j} \sum_{n=1}^{N} [y_n log h_\theta(x_n) + (1 - y_n) log(1 - h_\theta(x_n))]$$
$$= -\sum_{n=1}^{N} \frac{1}{\partial \theta_j}[y_n log h_\theta(x_n) + (1 - y_n) log(1 - h_\theta(x_n))]$$
$$= -\sum_{n=1}^{N} [y_n \cdot \frac{1}{h_\theta(x_n)} h_\theta(x_n) \cdot (1 - h_\theta(x_n))x_n + (1 - y_n)\frac{1}{1-h_\theta(x_n)} \cdot (-1) \cdot h_\theta(x_n) \cdot (1 - h_\theta(x_n))x_n]$$
$$= -\sum_{n=1}^{N} [y_n \cdot (1 - h_\theta(x_n))x_n + (y_n - 1) \cdot h_\theta(x_n)x_n]$$
$$= -\sum_{n=1}^{N} [(y_n - h_\theta(x_n))x_n]$$
$$= \sum_{n=1}^{N} [(h_\theta(x_n) - y_n)x_n]$$

3 Locally Weighted Linear Regression

(a)

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

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

(b)
After setting each partial derivative to zero, we can obtain the following two equations.

$$\theta_0 \sum_{n=1}^{N} w_n + \theta_1 \sum_{n=1}^{N} w_n x_{n,1} - \sum_{n=1}^{N} w_n y_n = 0$$

$$\theta_0 \sum_{n=1}^{N} w_n x_{n,1} + \theta_1 \sum_{n=1}^{N} w_n x_{n,1}^2 - \sum_{n=1}^{N} w_n x_{n,1} y_n = 0$$

For convenience, use some variables to denote some expressions.

$$a = \sum_{n=1}^{N} w_n, b = \sum_{n=1}^{N} w_n x_{n,1}, c = \sum_{n=1}^{N} w_n x_{n,1}, d = \sum_{n=1}^{N} w_n x_{n,1}^2, e = \sum_{n=1}^{N} w_n y_n, f = \sum_{n=1}^{N} w_n x_{n,1} y_n$$

Subsequently, we can acquire the values of θ_0 and θ_1,

$$\theta_0 = \frac{de - bf}{ad - bc}$$

$$\theta_1 = \frac{af - ce}{ad - bc}$$

If we are allowed to represent the data in matrix form , we can obtain the analytical answer as well, which is more concise.

$$X = \begin{bmatrix} 1 & x_{1,1} \\ \cdots & \cdots \\ 1 & x_{n,1} \end{bmatrix}, y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, W = \begin{bmatrix} w_1 & 0 & 0 \\ 0 & \cdots & 0 \\ 0 & 0 & w_n \end{bmatrix}$$

Then,

$$\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = (X^T W X)^{-1} X^T W y$$

4 Understanding Linear Separability

(a) Considering the dataset is linearly separable, $\exists$ hyperplane $\overrightarrow{\vartheta}^T \vec{x} + \gamma$ such that,

$$\max_{(\overrightarrow{x}, y) \epsilon D, y=-1} (\overrightarrow{\vartheta}^T \vec{x} + \gamma) < 0 \leqslant \min_{(\overrightarrow{x}, y) \epsilon D, y=1} (\overrightarrow{\vartheta}^T \vec{x} + \gamma)$$

Use $\overrightarrow{x_i}$ to denote the positive data point which is closest to the hyperplane $\overrightarrow{\vartheta}^T \vec{x} + \gamma$, and use $\overrightarrow{x_j}$ to denote the negative data point which is closest to the hyperplane $\overrightarrow{\vartheta}^T \vec{x} + \gamma$.

So,

$$d_+ = \min_{(\overrightarrow{x}, y) \epsilon D, y=1} (\overrightarrow{\vartheta}^T \vec{x} + \gamma) = \overrightarrow{\vartheta}^T \vec{x_i} + \gamma$$

$$d_- = \max_{(\overrightarrow{x}, y) \epsilon D, y=-1} (\overrightarrow{\vartheta}^T \vec{x} + \gamma) = \overrightarrow{\vartheta}^T \vec{x_j} + \gamma$$

Obviously,
$d_- < 0 \leqslant d_+$, and $\exists \zeta \geqslant 0$ such that $d_- - \zeta < 0 \leqslant d_+ - \zeta$. Therefore, $\overrightarrow{\vartheta}^T \vec{x} + \gamma$ -ζ= 0 can also linearly separate D.

Now, our goal is to find such a ζ that $\overrightarrow{x_i}$ and $\overrightarrow{x_j}$ have the same distance from the hyperplane $\overrightarrow{\vartheta}^T \vec{x} + \gamma$ -ζ= 0.

Since their have equal distances,

$\frac{|\overrightarrow{v}^T\vec{x_i}+\gamma-\zeta|}{\|\overrightarrow{v}\|} = \frac{|\overrightarrow{v}^T\vec{x_j}+\gamma-\zeta|}{\|\overrightarrow{v}\|}$

$\overrightarrow{v}^T\vec{x}_{\mathrm{i}} + \gamma - \zeta = -(\overrightarrow{v}^T\vec{x}_j + \gamma - \zeta)$
$d_+ - \zeta = -d_- + \zeta$
Finally, we obtain that $\zeta = (d_+ + d_-)/2$.

Now, our new hyperplane is $\overrightarrow{v}^T\vec{x} + \gamma$ -ζ= 0 ($\zeta = (d_+ + d_-)/2$).

$\underset{(\overrightarrow{x},y)\epsilon D,y=1}{min}(\overrightarrow{v}^T\vec{x} + \gamma - \zeta) = \underset{(\overrightarrow{x},y)\epsilon D,y=1}{min}(\overrightarrow{v}^T\vec{x} + \gamma) - \zeta = d_+ - \zeta = (d_+ - d_-)/2.$
$\underset{(\overrightarrow{x},y)\epsilon D,y=-1}{max}(\overrightarrow{v}^T\vec{x} + \gamma - \zeta) = \underset{(\overrightarrow{x},y)\epsilon D,y=-1}{max}(\overrightarrow{v}^T\vec{x} + \gamma) - \zeta = d_- - \zeta = (d_- - d_+)/2.$

Hence,
$y(\overrightarrow{v}^T\vec{x} + \gamma - \zeta) \geqslant (d_+ - d_-)/2,\ \forall(\overrightarrow{x}, y)\epsilon D$

Let $\overrightarrow{v}^T = \zeta\overrightarrow{w}^T, \gamma - \zeta = \zeta\theta,$

$y(\zeta\overrightarrow{w}^T\overrightarrow{x} + \zeta\theta) \geqslant \frac{d_+ - d_-}{2} > \frac{d_+ + d_-}{2} = \zeta(\zeta \geqslant 0)$

$y(\overrightarrow{w}^T\overrightarrow{x} + \theta) \geqslant 1, \forall(\overrightarrow{x}, y)\epsilon D$

In this case, $\delta = 0$.
Thus, an optimal solution to the linear program (2) is $\delta = 0$.

(b) If there is an optimal solution with $\delta = 0$, it means that $\exists$ hyperplane $\overrightarrow{w}^T\vec{x} + \theta$,
$y(\overrightarrow{w}^T\vec{x} + \theta) >= 1, \forall(\vec{x}, y)\epsilon D$
Obviously,

for $\forall(\vec{x}, y)\epsilon D$ and $y = 1$,
$(\overrightarrow{w}^T\vec{x} + \theta) \geqslant 1 > 0$
for $\forall(\vec{x}, y)\epsilon D$ and $y = -1$,
$(\overrightarrow{w}^T\vec{x} + \theta) \leqslant -1 < 0$

According to the description of (1), we can determine that D is linear separable.

(c) The situation varies with the value of δ.
If $0 < \delta < 1$, the situation is similar to the question (b), so the dataset D is still linearly separable.
If $\delta \geqslant 1$, we can not guarantee the dataset D is linearly separable.
However, if the $\delta_{min}$$\geqslant 1$, we are sure that the dataset D is **not** linearly separable.

(d) The optimal solution is this alternative LP formulation is $\delta = 0$, $\overrightarrow{w} = \vec{0}$, $\theta = 0$
As is shown above, the downside of this formulation is that regardless of the actual dataset, the optimal solution can always be a trivial and meaningless solution ($\delta = 0$, $\overrightarrow{w} = \vec{0}$, $\theta = 0$), which is undoubtedly not appropriate to be a reasonable hyperplane.

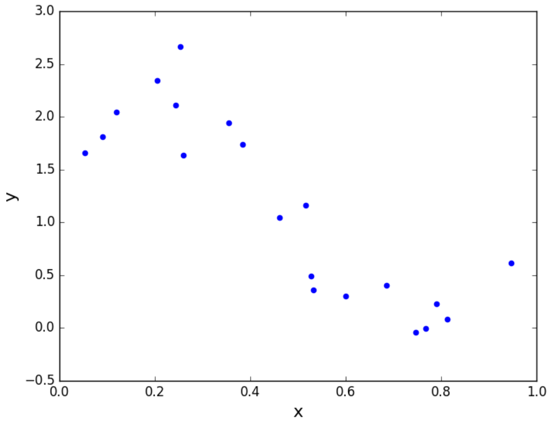
(e) After applying two data points into the constraints of (2), we can get the following two inequalities.
$w_1 + w_2 + w_3 + \theta \geqslant 1 - \delta$
$-(-w_1 - w_2 - w_3 + \theta) \geqslant 1 - \delta$

Since the dateset only includes two points, we can determine that dataset D is linearly separable, based on question(a).
Thus, $\delta = 0$, and
$w_1 + w_2 + w_3 + \theta \geqslant 1$
$-(-w_1 - w_2 - w_3 + \theta) \geqslant 1$
After simple transformations, we can obtain that
$w_1 + w_2 + w_3 \geqslant 1 - \theta$
$w_1 + w_2 + w_3 \geqslant 1 + \theta$
To select the intersection of these two constraints, we can conslude the following constraint is valid,
$w_1 + w_2 + w_3 \geqslant 1 + |\theta|$
To sum up, the possible optimal solutions are $w_1 + w_2 + w_3 \geqslant 1 + |\theta|$ plus $\delta = 0$.

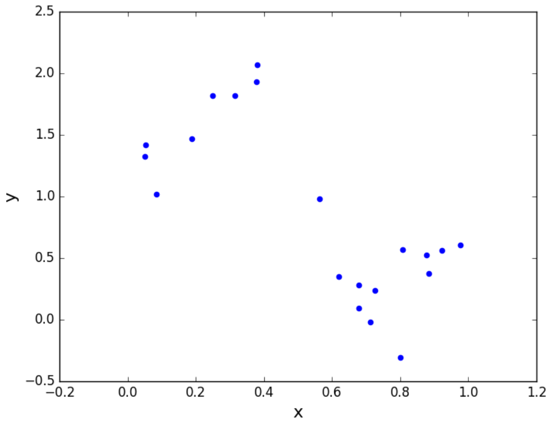
5 Implementation: Polynomial Regression

(a)

Train Data



Test Data



I observe that the data distributions of train set and test set are somewhat different, and both are not purely linear.Therefore, I guess the simple linear regression approach can not predict data well.

(d)

	Number of Iterations	Final Value of Loss Function	Coefficients
$\eta = 0.01$	764	3.912576	$\begin{bmatrix} 2.4464 & -2.8164 \end{bmatrix}$
$\eta = 0.001$	7020	3.912576	$\begin{bmatrix} 2.4464 & -2.8164 \end{bmatrix}$
$\eta = 0.0001$	10000	4.086397	$\begin{bmatrix} 2.2704 & -2.4606 \end{bmatrix}$
$\eta = 0.0407$	Can not converge	Very very large	$\begin{bmatrix} -9.4047 * 10^{18} & -4.6523 * 10^{18} \end{bmatrix}$

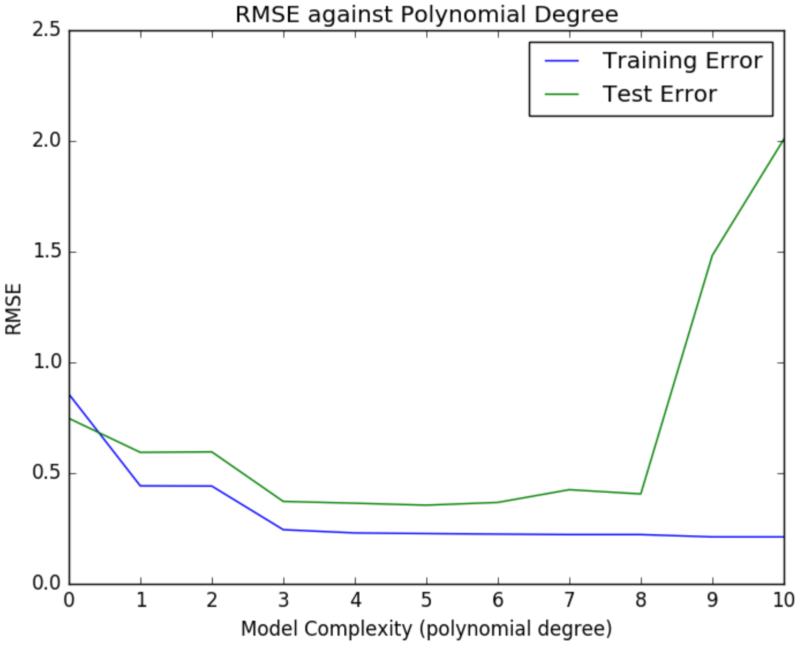
Regarding the coefficients, $\eta = 0.01$ and $\eta = 0.001$ achieve the exactly same one because the GD has converged under both conditions. The coefficient of $\eta = 0.0001$ is slightly different, for the algorithm has not converged to the minimum yet. Since $\eta = 0.0407$ can not enable GD to converge eventually, the coefficient is weird and meaningless to discusss further.
As is shown in the table above,the GD converges fastest when $\eta = 0.01$ and slowest when $\eta = 0.0001$. What's more, the GD will not converge when $\eta = 0.0407$ because the learning rate is too large.

(e) The closed-form solution is $\begin{bmatrix} 2.4464 & -2.8164 \end{bmatrix}$, which is same as the one obtained via Gradient Descent(converged). In terms of the speed, the closed-form solution runs faster for this particular problem.

(f) After settting learning rate η as a function of the number of iterations, it takes 1678 times of iterations for GD algorithm to converge.

(h) RMSE represents the average error of all data points. By contrast, $J(\theta)$ becomes bigger and bigger as the size of dataset increases. Smaller $J(\theta)$ does not necessarily indicate high precision because the dataset's small size may be the cause.

(i)



I think polynomial of degree 5 would best fit for this particular data, for the test error is minimal when the degree equals to 5.

The phenomenon of overfitting is significant when the degree is greater than 8. As is depicted in the plot, the test error increases a lot when the model is comparatively complex, although the training error is still diminishing.

The phenomenon of underfitting is evident when the degree is relatively small, for example, 0 and 1. When underfitting occurs, both training error and test error are very large, which has been demonstrated in the plot.