Exercise 1:

(1) The sequence is $x_i=(\cos i, \sin i)$, $i\in N^+=\{1,2,3,\cdots\}$, $y_i=\begin{cases}1, & \text{if } \sin i>0\\ -1 & \text{else;}\end{cases}$

a) prove the sequence that is linearly separable with $b=0$

$i\in N^+$, so i is positive rational number. $\sin i=0$ if $i=n\pi$, $\overset{n\in Z^+}{\text{so}}$ if $i\in\{1,2,\cdots\}$, $i>0$, $n\pi>0$, $n>0$, $\pi=\frac{i}{n}$, π is irrational, n, i rational

Contradiction, so $\sin i\neq 0$, so data is linearly separable by line $y=0$, where $b=0$

b) For $x_i=(\cos i, \sin i)$, $\|x_i\|_2=\sqrt{\cos^2 i+\sin^2 i}=1\leq 1$, so $\|x_i\|_2\leq 1$

(c) Suppose that modified algorithm only made limited mistakes, and after the mistakes are made, w', b' is in loop. $b'=y_{I_1}+y_{I_2}+\cdots, y_{I_t}=1$ or -1, b' is an integer.

Since for w', b', we won't made any mistake, find $(x_m, y_m), (x_n, y_n)$ where $y_m=1, y_n=-1$, so:

$w'^T x_m+b'>0$ $w'^T x_n+b'<0$, by the definition of sequence, $x_n\neq -x_m$ (If $x_n=-x_m$ $n=m+(2k+1)\pi$, $k\in Z$, contradiction, n is irrational)

So $x_m\cdot x_n=\|x_m\|_2\|x_n\|_2\cdot\cos\theta=\cos\theta>-1$ $(\cos\theta\neq -1)$ $1+x_m\cdot x_n>0$

$\exists k$, $k(1+x_m\cdot x_n)>-(w'^T x_n+b')$

We feed the algorithm with (x_m, y_m) for k times, then feed (x_n, y_n)

$w^*=w'+kx_m$ $b=b'+k$, so $(w'+kx_m)\cdot x_n+b'+k<0$ (no more mistakes)

$kx_m\cdot x_n+k<-(w'^T x_n+b')$, contradiction.

It will make unlimited errors

(2)

Exercise 2:

(1) $(2) = \min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2n} \left\| \begin{bmatrix} X & 1_n \\ \sqrt{2\lambda n} I_d & 0_d \end{bmatrix} \begin{bmatrix} w \\ b \end{bmatrix} - \begin{bmatrix} Y \\ 0_d \end{bmatrix} \right\|_2^2$

$= \min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2n} \left\| \begin{bmatrix} X\cdot w + b\cdot 1_n - Y \\ \sqrt{2\lambda n} I_d \cdot w \end{bmatrix} \right\|_2^2$ $\quad (a+b-c)^2 = a^2+b^2+c^2+2ab$

$= \min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2n} \left[\|X\cdot w + b\cdot 1_n - Y\|_2^2 + \|\sqrt{2\lambda n}\cdot w\|_2^2 \right]$

$= \min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2n} \|X\cdot w + b\cdot 1_n - Y\|_2^2 + 2\lambda n \cdot \|w\|_2^2 \times \frac{1}{2n}$

$= \min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2n} \|X\cdot w + b\cdot 1_n - Y\|_2^2 + \lambda \|w\|_2^2 = (1)$

QED

(2) By chain rule: $\frac{\partial}{\partial w} = \frac{1}{2n} \cdot 2 \cdot \frac{\partial}{\partial w} (X\cdot w + b\cdot 1_n - Y) \cdot (X\cdot w + b\cdot 1_n - Y) + \lambda \cdot 2w \cdot \frac{dw}{dw}$ ($X\cdot w$: $d\times 1$)

$$\frac{\partial (X\cdot w + b\cdot 1_n - Y)}{\partial w} = \left[\frac{\partial Y_1\cdot w + b - Y_1}{\partial w} \quad \cdots\cdots \quad \frac{Y_n\cdot w + b - Y_n}{\partial w} \right]$$ where Y_i is the i^{th} row of matrix X

$$= \begin{bmatrix} \frac{\partial a_{11}\cdot w_1 + b - Y_1}{\partial w_1} & \frac{\partial a_{21}\cdot w_1 + b - Y_2}{\partial w_1} & \cdots \\ \vdots & & \\ \frac{\partial a_{1d}\cdot w_d + b - Y_1}{\partial w_d} & \cdots & \end{bmatrix}$$ where a_{ij} is the entry of matrix X at i^{th} row j^{th} column

$$= \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ \vdots & & & \\ a_{1d} & \cdots & & a_{nd} \end{bmatrix}$$

$= X^T$.

put this back, so:

$$\frac{\partial}{\partial w} = \frac{1}{2n} \times 2 \times X^T (Xw + b\cdot 1_n - Y) + 2\lambda \cdot w$$

By chain rule, $\frac{\partial}{\partial b} = \frac{1}{2n} \times 2 \times \frac{\partial (X\cdot w + b\cdot 1_n - Y)}{\partial b} \cdot (X\cdot w + b\cdot 1_n - Y) + 2\lambda \cdot w \cdot \frac{dw}{db}$ ($\frac{dw}{db} = 0$ as w and b are independent)

$$\frac{\partial (X\cdot w + b\cdot 1_n - Y)}{\partial b} = \left[\frac{r_1\cdot w + b - Y_1}{\partial b} \quad \cdots\cdots \quad \frac{r_n\cdot w + b - Y_n}{\partial b} \right] = [1 \cdots 1] = 1_n^T$$

So $\frac{\partial}{\partial b} = \frac{1}{n} 1^T (X\cdot w + b\cdot 1_n - Y)$

(3)
(4)
(5)

$\lambda = 0$:

closed form:

Training Error : 0.0623

Training Loss : 19.062725

Test error : 2531.6381

Gradient descent:

Training Error : 0.0623

Training Loss : 19.062727

Test error : 2532.0434

$\lambda = 10$:

closed form:

Training Error : 0.4159

Training Loss : 127.3023

Test error : 3103.3522

Gradient descent:

Training Error : 0.0623

Training Loss : 25.7952

Test error : 2532.1109

I think the gradient descent algorithm is faster, the running time is $O(ncd)$, where we only need to iterate fixed number of times.

For closed form, the running time to solve linear equation is $O(nd^2)$
If we have a lot of data, it will be slow.

However, closed form is better, because it gives a precise answer, but descent gradient is just approximation.

Exercises:

The screen shot was in folder.

On the top right, the λ chosen and average mean squared error is computed.

5. (1) $\hat{w} = \arg\max_w \prod_{i=1}^{n} p(Y=y_i|x_i)$ due to independence

$= \arg\max_w \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} \cdot \exp(-\mu_i)$

$= \arg\max_w \sum_{i=1}^{n} \log\left(\frac{\mu_i^{y_i}}{y_i!}\right) - \mu_i$

$= \arg\max_w \sum_{i=1}^{n} y_i \log \mu_i - \mu_i$ y_i is independent, so $-\log y_i!$ term can be removed.

(2)

The maximum likelihood function is $\arg\max_w \sum_{i=1}^{n} y_i \log \mu_i - \mu_i$, where y_i is constant, $\log \mu_i$ is an increasing function on $\mathbb{R}^+$, and the range is $(0, +\infty)$

(3) $\log \mu_i =$