**دانشگاه صنعتی امیرکبیر**

**دانشکده مهندسی کامپیوتر**

# گزارش تکلیف سوم درس شناسائی آماری الگو

**دانشجو:**

**سید احمد نقوی نوزاد**

**ش-د: ٩٤١٣١٠٦٠**

**استاد:**

**دکتر رحمتی**

**پائیز ۹۵**

# جواب سوال ۱

$\{x_k\}, k=1,\dots,N \rightarrow$ independent training data samples from following densities:

$a.\quad f(x_k;\theta)=\dfrac{x_k}{\theta^2}\exp\left(-\dfrac{x_k^2}{2\theta^2}\right) \qquad x_k \ge 0 \qquad \theta>0 \qquad$ Reyleigh Density

$b.\quad f(x_k;\theta)=\sqrt{\theta}\,x_k^{\sqrt{\theta}-1} \qquad 0\le x_k \le 1 \qquad \theta>0 \qquad$ Beta Density

What is the Maximum Likelihood Estimate of $\theta$ in each case?

(قسمت الف)

$$a:\quad L_N(\theta)=\prod_{k=1}^{N} f(x_k;\theta)=\prod_{k=1}^{N}\frac{x_k}{\theta^2}\exp\left(-\frac{x_k^2}{2\theta^2}\right)=\frac{\prod_{k=1}^{N}x_k}{\theta^{2N}}\exp\left(-\frac{1}{2\theta^2}\sum_{k=1}^{N}x_k^2\right)\rightarrow$$

$$\ell_N(\theta)=\ln\left(\prod_{k=1}^{N}x_k\right)-\ln\left(\theta^{2N}\right)-\frac{1}{2\theta^2}\sum_{k=1}^{N}x_k^2=\sum_{k=1}^{N}\ln x_k-2N\ln\theta-\frac{1}{2\theta^2}\sum_{k=1}^{N}x_k^2\rightarrow$$

$$\frac{d}{d\theta}\ell_N(\theta)=-\frac{2N}{\theta}+\theta^{-3}\sum_{k=1}^{N}x_k^2=0\xrightarrow{\times\theta^3}-2N\theta^2+\sum_{k=1}^{N}x_k^2=0\Rightarrow\theta^2=\frac{\sum_{k=1}^{N}x_k^2}{2N}\Rightarrow$$

$$\theta=\sqrt{\frac{\sum_{k=1}^{N}x_k^2}{2N}}=\sqrt{\frac{\frac{N(N+1)(2N+1)}{6}}{2N}}\Rightarrow\theta=\sqrt{\frac{(N+1)(2N+1)}{12}}$$

(قسمت ب)

$$b:\quad L_N(\theta)=\prod_{k=1}^{N} f(x_k;\theta)=\prod_{k=1}^{N}\sqrt{\theta}\,x_k^{\sqrt{\theta}-1}=\theta^{N/2}\left(\prod_{k=1}^{N}x_k\right)^{\sqrt{\theta}-1}\rightarrow$$

$$\ell_N(\theta)=\frac{N}{2}\ln\theta+\left(\sqrt{\theta}-1\right)\sum_{k=1}^{N}\ln x_k\Rightarrow\frac{d}{d\theta}\ell_N(\theta)=\frac{N}{2\theta}+\frac{1}{2}\theta^{-1/2}\left(\sum_{k=1}^{N}\ln x_k\right)=0\xrightarrow{\times\theta}$$

$$\frac{N}{2}+\frac{1}{2}\theta^{1/2}\left(\sum_{k=1}^{N}\ln x_k\right)=0\Rightarrow\theta^{1/2}=\frac{-N}{\sum_{k=1}^{N}\ln x_k}\Rightarrow\theta=\frac{N^2}{\left(\sum_{k=1}^{N}\ln x_k\right)^2}$$

# جواب سوال ۲

$$f_x(x\mid\theta)\sim U(0,\theta)=\begin{cases}\dfrac{1}{\theta} & 0\le x\le\theta\\ 0 & \text{otherwise}\end{cases}$$

(قسمت الف)

$a.\quad D=\{x_1,x_2,\dots,x_n\}\rightarrow$ independently drawn samples according to $f_x(x\mid\theta)\rightarrow$
$MLE_\theta=\max[D]=X_{(n)}\,?$

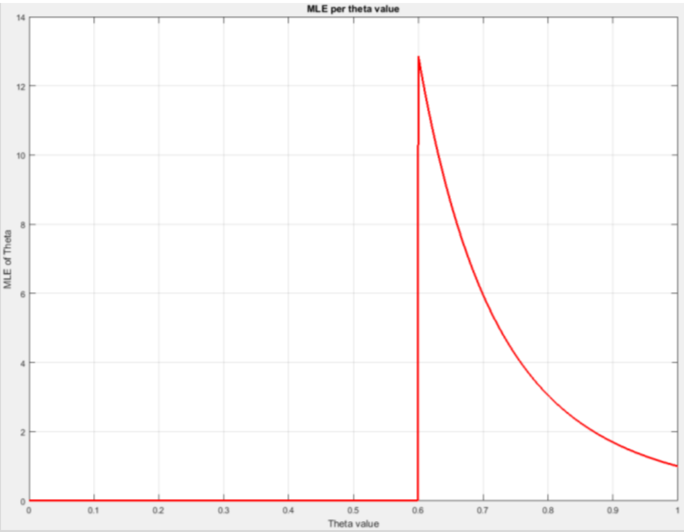
At first let's consider a fixed value of $\theta$. Then we have:

$$\mathrm{L}_n(\theta)=\prod_{k=1}^{n} f\left(x_k;\theta\right)=\begin{cases}\left(\dfrac{1}{\theta}\right)^n & \theta\ge X_{(n)}, \text{ because for every k: } f\left(x_k;\theta\right)=\dfrac{1}{\theta}\\ 0 & \theta<X_{(n)}, \text{ because for some k, } \mathrm{x}_k>\theta \text{ and then: } f\left(x_k;\theta\right)=0\end{cases}$$

Now $\mathrm{L}_n\left(\theta\right)$ is strictly decreasing over the interval $\left[X_{(n)},+\infty\right)$. Hence, $MLE_\theta=X_{(n)}$

قسمت ب)

*b*. Suppose that n=5 and $\mathrm{X}_{(n)}=.6$. Plot the likelihood function $f_x\left(D\mid\theta\right)$ in the range of $0\le\theta\le1$?




شکل ۱. نمودار MLE مربوط به پارامتر θ در بازه‌ی بین ۰ و ۱؛ همانطور که پیداست به ازای پارامتر θ برابر با ۰,۶ مقدار MLE بیشینه شده و بعد از آن با افزایش مقدار θ رو به افول می‌گذارد.

از آن‌جا که با توجه به قسمت الف، مقدار تخمین بیشینه‌ی درستنمائی[1] برابر با بیشترین مقدار موجود در نمونه‌هاست و به عبارتی در آن نقطه مقدار بیشینه‌ی خود را داراست، لذا جهت رسم نمودار MLE بر حسب پارامتر θ، دیگر نیازی به دانستن مقادیر کمتر از مقدار بیشینه نمی‌باشد.

# جواب سوال ۳

$$\text{Leave-One-Out Error}<\frac{\#SV_s}{l}$$

با توجه به این‌که مرز تصمیم‌گیری نهائی که با استفاده از مجموعه‌ی آموزشی حاصل می‌گردد، تنها به بردارهای پشتیبان وابسته[2] است، لذا در صورتی که در رویه‌ی LOO[3] یکی از داده‌های غیر بردار پشتیبان را از داده‌های آموزشی کنار گذاشته و دوباره مدل SVM را بر روی مابقی داده‌های موجود آموزش دهیم، مرز تصمیم‌گیری نهائی تغییری نکرده و البته که سایر داده‌های غیر بردار پشتیبان نیز به درستی دسته‌بندی می‌شوند. اما در صورتی که یکی از بردارهای پشتیبان را از مجموعه‌داده‌ی آموزشی حذف نمائیم و دوباره مدل SVM را آموزش دهیم، ممکن است مرز تصمیم‌گیری جدید حاصله متفاوت بوده و در نتیجه در لحظه‌ی آزمایش آن بردار پشتیبان دچار خطا شویم. به عبارت دیگر در رویه‌ی LOO، تنها در مورد بردارهای پشتیبان ممکن است به اشتباه تصمیم‌گیری نموده و در نتیجه نامعادله‌ی بالا ثابت می‌شود.

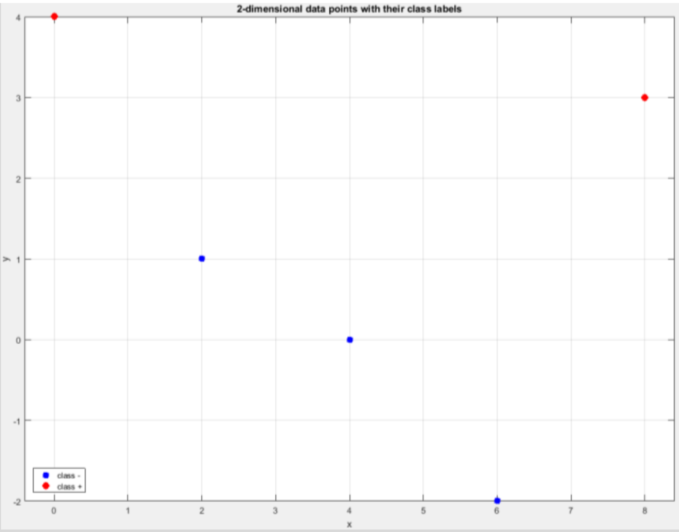
[1] Maximum Likelihood Estimate (MLE)
[2] Support Vectors
[3] Leave-One-Out

# جواب سوال ۴

داده‌های مسئله به قرار زیر می‌باشند:

| x | y | class |
|---|---|---|
| 0 | 4 | + |
| 8 | 3 | + |
| 6 | -2 | - |
| 4 | 0 | - |
| 2 | 1 | - |

## قسمت الف)




شکل ۲. داده‌های ترسیم‌شده مربوط به دو کلاس + و -.

**همان‌طور که قابل مشاهده است، داده‌های مربوط به دو کلاس به صورت خطی از یکدیگر جداپذیر می‌باشند و در نتیجه می‌توان با استفاده از یک شبکه‌ی پرسپترون خطی دوبُعدی آن‌ها را از یکدیگر جداسازی نمود.**

## قسمت ب)

**با توجه به شکل ۲، می‌توان مشاهده نمود که برای جداسازی داده‌های مربوط به دو کلاس به صورت خطی، نیاز به یک bias term نیز می‌باشد. لذا یک به هر یک از داده‌های دو کلاس یک ویژگی با مقدار ۱ اضافه می‌نمائیم، که به این کار اصطلاحاً augmentation می‌گویند. علاوه بر این کار، برای اجرای پرسپترون خطی می‌بایست داده‌های دو کلاس را به اصطلاح normalize نیز بنمائیم. به این صورت که تمامی مقادیر ویژگی‌های یکی از کلاس‌ها را قرینه نموده و به عبارتی در ۱- ضرب نمائیم.**

**در ادامه اجرای مرحله به مرحله‌ی پرسپترون را بر روی مجموعه‌داده‌ی فوق مشاهده می‌نمائیم:**

| x | y | class |
|---|---|---|
| 0 | 4 | + |
| 8 | 3 | + |
| 6 | -2 | - |
| 4 | 0 | - |
| 2 | 1 | - |

**نتیجه‌ی اعمال augmentation:**

| extra | x | y | class |
|---|---|---|---|
| 1 | 0 | 4 | + |
| 1 | 8 | 3 | + |
| 1 | 6 | -2 | - |
| 1 | 4 | 0 | - |
| 1 | 2 | 1 | - |

نتیجه‌ی اعمال **normalization**:

| # | extra | x | y | class |
|---|---|---|---|---|
| 1 | 1 | 0 | 4 | + |
| 2 | 1 | 8 | 3 | + |
| 3 | -1 | -6 | 2 | - |
| 4 | -1 | -4 | 0 | - |
| 5 | -1 | -2 | -1 | - |

مقادیر اولیه به قرار زیر می‌باشند:

$a^{(1)} = [1,1,1]$

sample is misclassified if: $\mathrm{a}^t y_i = \sum_{k=0}^{2} a_k y_i^{(k)} < 0$

gradient descent single sample rule: $\mathrm{a}^{(k+1)} = \mathrm{a}^{(k)} + \eta^{(k)} y$ ,

where $y \in Y_M$ and $Y_M$ is the the set of missclassified items and $\eta^{(k)}$=1.

حال به ازای هر کدام از داده‌ها بررسی می‌نمائیم که آیا با مقادیر موجود برای بردار وزن a، به درستی دسته‌بندی شده‌اند یا نه، تا در صورت به اشتباه‌دسته‌بندی‌شدن مقدار بردار وزن را با توجه به قانون نزول در امتداد گرادیان به روز نمائیم. داریم:

| # | $a^t y$ | misclassified? |
|---|---|---|
| 1 | 1*1+1*0+1*4>0 | no |
| 2 | 1*1+1*8+1*3>0 | no |
| 3 | 1*(-1)+1*(-6)+1*2<0 | yes |

مقدار جدید بردار وزن به صورت زیر خواهد بود:

$$a^{(2)} = a^{(1)} + y_M = [1,1,1] + [-1,-6,2] = [0,-5,3]$$

در ادامه بررسی نمونه‌ها داریم:

| # | $a^t y$ | misclassified? |
|---|---|---|
| 4 | 0*(-1)+(-5)*(-4)+3*0>0 | no |
| 5 | 0*(-1)+(-5)*(-2)+3*-1>0 | no |

حال به ازای یک دور بررسی تمام داده‌های موجود، بردار وزن نهائی همان مقدار $a^{(2)}$ می‌باشد. در ادامه به ازای این مقدار بردار وزن، صحت دسته‌بندی را به ازای تمام داده‌ها دوباره حساب می‌کنیم:

| # | $a^t y$ | misclassified? |
|---|---|---|
| 1 | 0*(1)+(-5)*(0)+3*4>0 | no |
| 2 | 0*(1)+(-5)*(8)+3*3<0 | yes |
| 3 | 0*(-1)+(-5)*(-6)+3*2>0 | no |
| 4 | 0*(-1)+(-5)*(-4)+3*0>0 | no |
| 5 | 0*(-1)+(-5)*(-2)+3*(-1)>0 | no |

مشاهده می‌شود که به ازای داده‌ی دوم، خطا داریم. در نتیجه می‌بایست مجدداً قانون به‌روزرسانی پرسپترون برای بردار وزن را اعمال نمائیم. داریم:

$$a^{(3)} = a^{(2)} + y_M = [0,-5,3] + [1,8,3] = [1,3,6]$$

مجدداً صحت دسته‌بندی را به ازای تمامی داده‌ها به یکباره بررسی می‌نمائیم. داریم:

| # | $a^t y$ | misclassified? |
|---|---|---|
| 1 | 1*(1)+3*(0)+6*4>0 | no |
| 2 | 1*(1)+3*(8)+6*3>0 | no |
| 3 | 1*(-1)+3*(-6)+6*2<0 | yes |
| 4 | 1*(-1)+3*(-4)+6*0<0 | yes |
| 5 | 1*(-1)+3*(-2)+6*(-1)<0 | yes |

مشاهده می‌شود که رویه‌ی به‌روزرسانی هم‌چنان می‌بایست ادامه یابد. در نهایت پس از یک سری محاسبات نسبتاً طولانی به بردار وزن زیر دست خواهیم یافت:

$$a = [-10,-3,15]$$

مجدداً صحت این بردار وزن را به ازای تمامی نمونه‌های آموزشی موجود بررسی می‌نمائیم. داریم:

| # | $a^t y$ | misclassified? |
|---|---|---|
| 1 | -10*(1)-3*(0)+15*4>0 | no |
| 2 | -10*(1)-3*(8)+15*3>0 | no |
| 3 | -10*(-1)-3*(-6)+15*2>0 | no |
| 4 | -10*(-1)-3*(-4)+15*0>0 | no |
| 5 | -10*(-1)-3*(-2)+15*(-1)>0 | no |

### قسمت ج)

**معادله‌ی خط جداساز به صورت زیر می‌باشد:**

$$-10 + -3x + 15y = 0$$

**لازم به ذکر است که معادله‌ی خط مزبور در واقع یکی از معادلات خط جداساز ممکن می‌باشد که همگی در به اصطلاح «ناحیه‌ی پاسخ[4]» قرار دارند.**

### قسمت د)

**برای دسته‌بندی داده‌ی (5,2) می‌بایست مقدار آن را در معادله‌ی خط مذکور قرار داده و علامت آن را بررسی نمائیم. اگر علامت آن مثبت بود که به کلاس + و در غیر این صورت به کلاس – تعلق خواهد گرفت. داریم:**

$$-10-3\times5+15\times2=5>0 \Rightarrow (5,2)\in M_+$$

# جواب سوال ۵

کدهای مربوط به این سوال در فایل‌های با پیشوند ex05 قرار دارند.

### تمرین کامپیوتری ۱)

**متأسفانه حتی پس از صرف زمان زیاد جهت بررسی مسئله و البته تحقیق و تفحص فراوان، به دلیل عدم مشخص‌بودن تابع ضابطه[5] در صورت سؤال، قادر به پیاده‌سازی الگوریتم‌های مربوطه نگشتیم.**

### تمرین کامپیوتری ۲)

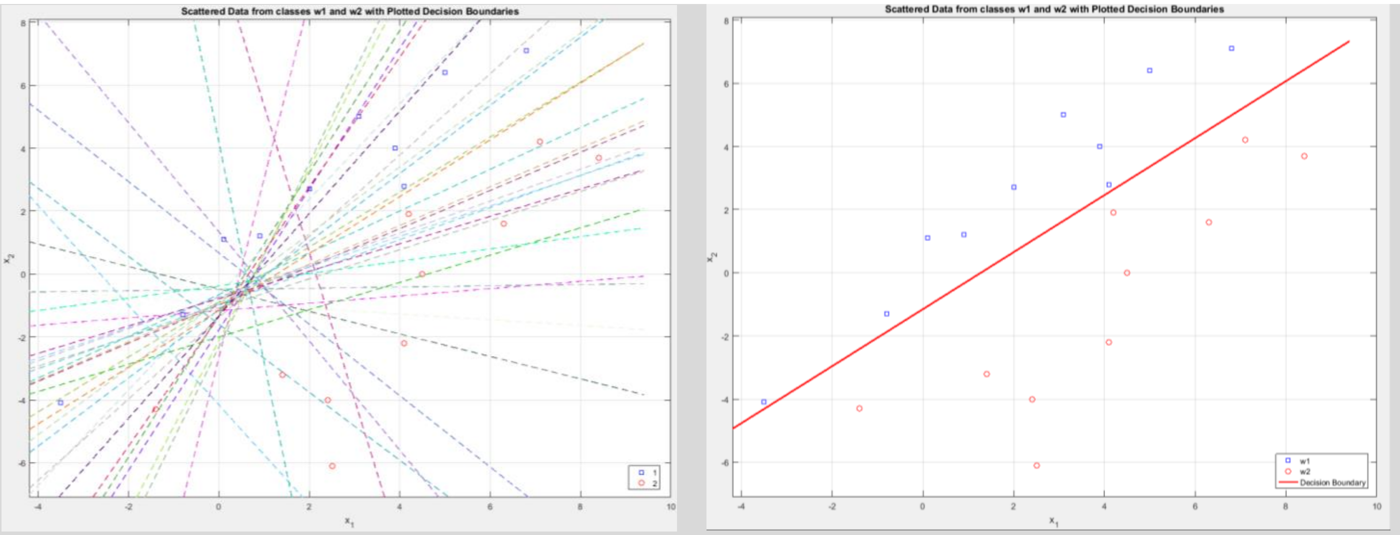
**در ادامه الگوریتم پرسپترون را بر روی مجموعه‌داده‌های خواسته‌شده پیاده‌سازی می‌نمائیم:**

#### قسمت الف)

**نتایج پیاده‌سازی بر روی مجموعه‌داده‌های w1 و w2 و با مقدار اولیه‌ی a=0 به صورت زیر می‌باشد:**

---

[4] Solution Region
[5] Criterion Function

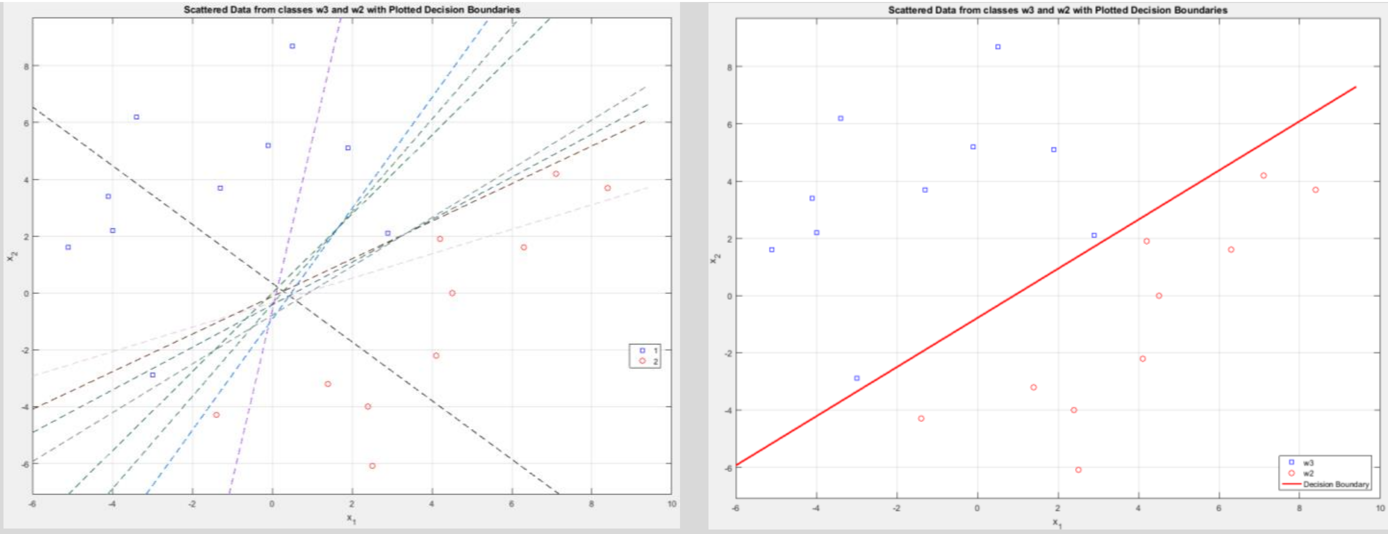



شکل ۳. سمت چپ: نتیجه‌ی اجرای مرحله به مرحله‌ی پرسپترون بر روی داده‌های موجود و البته به تعداد دفعات تکرار برابر با ۹ بار تا رسیدن به همگرائی؛ سمت راست: مرز تصمیم‌گیری نهائی حاصله از اجرای پرسپترون خطی بر روی داده‌های دو کلاس.

**طبق آزمایش انجام‌شده، تعداد دفعات لازم برای همگراشدن در مورد داده‌های دو کلاس w1 و w2 برابر ۹ می‌باشد.**

## قسمت ب)

**نتایج پیاده‌سازی بر روی مجموعه‌داده‌های w2 و w3 و با مقدار اولیه‌ی a=0 به صورت زیر می‌باشد:**




شکل ۴. سمت چپ: نتیجه‌ی اجرای مرحله به مرحله‌ی پرسپترون بر روی داده‌های موجود و البته به تعداد دفعات تکرار برابر با ۹ بار تا رسیدن به همگرائی؛ سمت راست: مرز تصمیم‌گیری نهائی حاصله از اجرای پرسپترون خطی بر روی داده‌های دو کلاس.

**طبق آزمایش انجام‌شده، تعداد دفعات لازم برای همگراشدن در مورد داده‌های دو کلاس w2 و w3 برابر ۵ می‌باشد.**

## قسمت ج)

**با توجه به این‌که مجموعه‌داده‌ی کلاس w2 در هر دو آزمایش وجود دارد، لذا علت وجود تفاوت در تعداد دفعات لازم برای همگراشدن را باید در وضعیت داده‌های کلاس w1 و w3 جستجو نمود. از شکل‌های ۳ و ٤ پیداست که پراکندگی داده‌های مربوط به کلاس w3 بیشتر از کلاس w1 می‌باشد. لذا در مورد آزمایش دوم، در همان دفعات تکرار اولیه، داده‌های دورتر و به اصطلاح پخش‌شده‌تر[6] به درستی دسته‌بندی شده و در دفعات بعدی تکرار الگوریتم، این داده‌های نزدیک به مرز تصمیم‌گیری هستند که با جابجائی‌های اندک مرز تصمیم‌گیری موجب خطا شده و سبب به‌روزرسانی‌های مکرر بردار وزن و البته تکرار الگوریتم پرسپترون می‌شوند. اما در مورد آزمایش اول، پراکندگی داده‌های کلاس w1 به مراتب کمتر بوده و در نتیجه با اندک تغییر مرز تصمیم‌گیری،**

[6] Scattered

نرخ خطای دسته‌بندی نیز به مراتب نسبت به آزمایش دوم بیشتر می‌باشد، و همین مسئله سبب می‌گردد تا تعداد دفعات تکرار الگوریتم پرسپترون نیز بیشتر از آزمایش دوم باشد.

## تمرین کامپیوتری ۳)

**در این قسمت ابتدا به رویه‌ی الگوریتم Ho-Kashyap به طور مختصر اشاره نموده و سپس آن را بر روی داده‌های دو کلاس w1 و w3 اجرا می‌نمائیم و بعد از آن همین رویه را در مورد داده‌های دو کلاس w2 و w4 انجام خواهیم داد. داریم:**

Ho-Kashyap Procedure is as follows:

0) Start with arbitrary $a^{(1)}$ and $b^{(1)} > 0$, let k=1

Repeat steps (1) through (4)

1) $e^{(k)} = Ya^{(k)} - b^{(k)}$

2) Solve for $b^{(k+1)}$ using $a^{(k)}$ and $b^{(k)}$

$$b^{(k+1)} = b^{(k)} + \eta\left[e^{(k)} + \left|e^{(k)}\right|\right]$$

3) Solve for $a^{(k+1)}$ using $b^{(k+1)}$

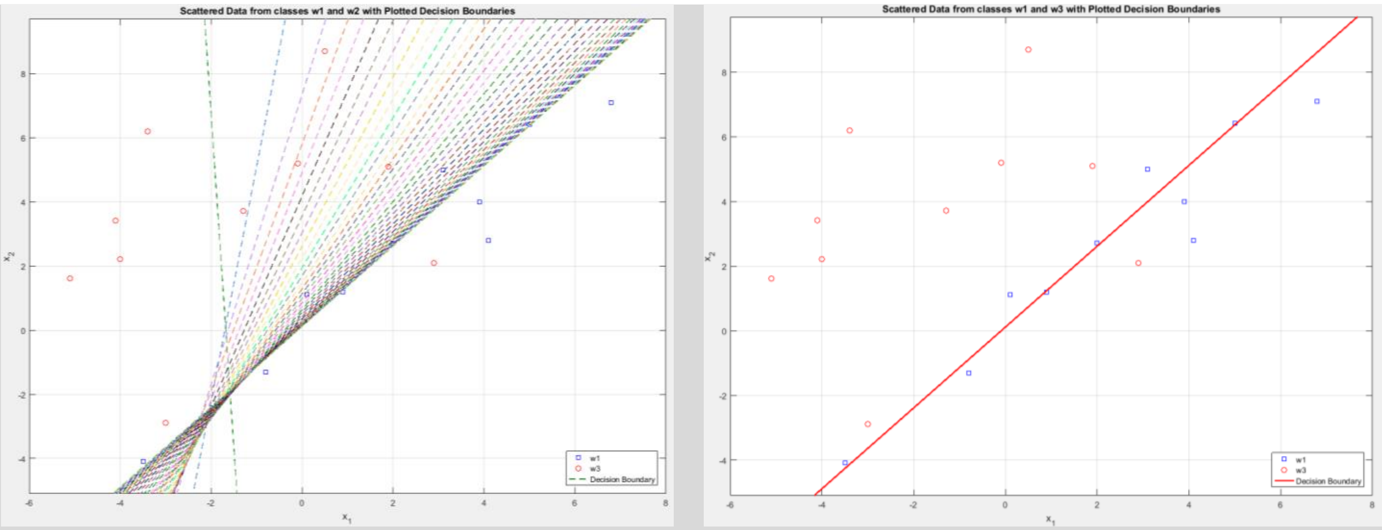
$$a^{(k+1)} = \left(Y^tY\right)^{-1}Y^tb^{(k+1)}$$

4) k = k + 1

Until $e^{(k)} \geq 0$ or $k > k_{\max}$ or $b^{(k+1)} = b^{(k)}$

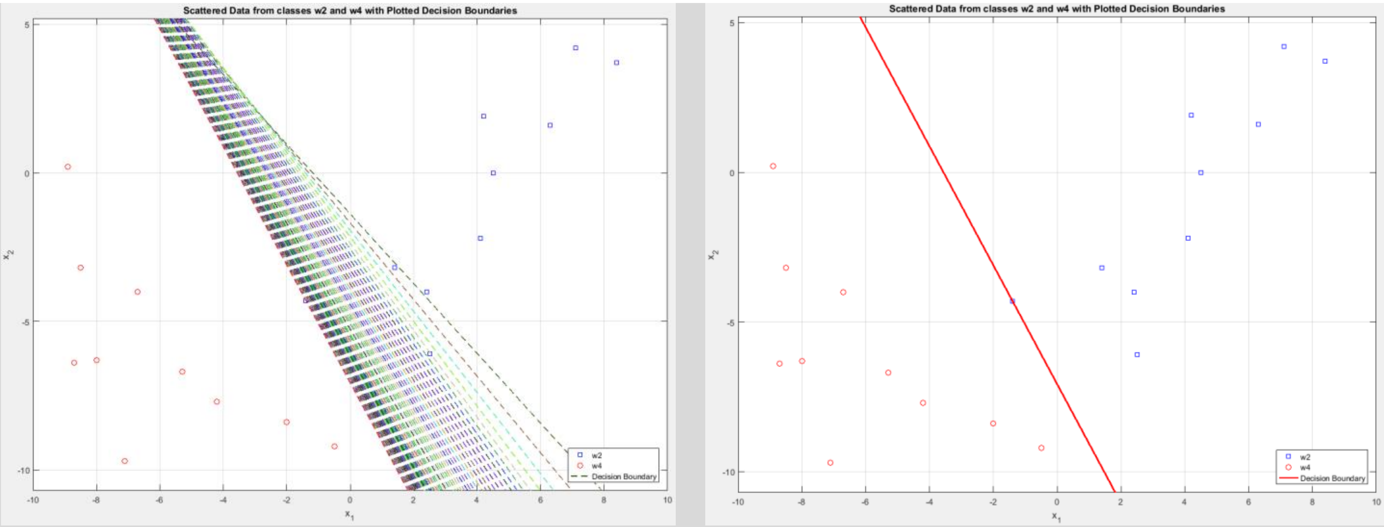
Note: For convergence, learning rate should be fixed between 0 and 1.

**نتایج اجرای الگوریتم مربوطه بر روی داده‌های دو کلاس w1 و w3 به شرح زیر می‌باشند (توضیحات مکفی در ذیل جدول آمده است):**




شکل ۵. سمت چپ: نتیجه‌ی اجرای مرحله به مرحله‌ی الگوریتم **Ho-Kashyap** بر روی داده‌های دو کلاس **w1** و **w3**؛ سمت راست: مرز تصمیم‌گیری نهائی حاصله از اجرای الگوریتم **Ho-Kashyap** بر روی داده‌های دو کلاس مزبور؛ لازم به ذکر است که مقدار نرخ یادگیری مورد استفاده در این آزمایش برابر ۰,۹ می‌باشد. همانطور که قابل مشاهده است، به دلیل اینکه داده‌های دو کلاس به صورت خطی جداپذیر نمی‌باشند، لذا الگوریتم **Ho-Kashyap** تلاش می‌کند تا آن‌جا که ممکن است مرز تصمیم‌گیری بهینه را رسم نماید و در نهایت به دلیل تجاوز تعداد دفعات تکرار الگوریتم از حد مجاز (در اینجا $k_{max} = 200$) متوقف می‌گردد.

**نتایج اجرای الگوریتم مربوطه بر روی داده‌های دو کلاس w2 و w4 به شرح زیر می‌باشند (توضیحات مکفی در ذیل جدول آمده است):**




شکل ۵. سمت چپ: نتیجه‌ی اجرای مرحله به مرحله‌ی الگوریتم **Ho-Kashyap** بر روی داده‌های دو کلاس **w2** و **w4** ؛ سمت راست: مرز تصمیم‌گیری نهائی حاصله از اجرای الگوریتم **Ho-Kashyap** بر روی داده‌های دو کلاس مزبور؛ لازم به ذکر است که مقدار نرخ یادگیری مورد استفاده در این آزمایش برابر ۰٫۹ می‌باشد. همانطور که قابل مشاهده است، با توجه به اینکه داده‌های دو کلاس به صورت خطی جداپذیر می‌باشند، در نهایت الگوریتم **Ho-Kashyap** پس از تعداد دفعات تکرار فراوان (**k=1000**) موفق به کسب یک مرز تصمیم‌گیری نسبتاً مناسب می‌شود.