



# تمرین سری اول

پاسخ سوال ۱

امیر حسین محمدی

۹۹۲۰۱۰۸۱



مسئله ی دسته بندی دو دسته ای را در نظر بگیرید که داده ها متعلق به یک فضای ویژگی دو بعدی بوده و احتمال شرطی دسته ها از توزیع گوسی با ماتریس کوواریانس و میانگین های متفاوت بیابند.  $(p(x|C_i) = N(\mu_i | \Sigma_i))$  همچنین احتمال پیشین روی دسته اول برابر با  $\pi$  و روی دسته دو را برابر با  $1 - \pi$  در نظر بگیرید.  $(p(C_1) = \pi, p(C_2) = 1 - \pi)$



مرز دسته بند احتمالی بیز را بر حسب پارامترها و بردار ورودی  $X$  بیابید.

راه حل

جهت محاسبه ی مرز تصمیم در دسته  
بند های گوسی بیز داریم:

$$p(C_1|x) = p(C_2|x)$$



همچنین می دانیم:

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)}$$

طبق فرضیات  
مسئله داریم:

$$p(C_1) = \pi, \\ p(C_2) = 1 - \pi$$

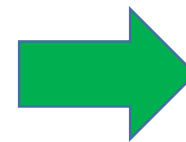
$$\bullet \quad p(C_1|x) = p(C_2|x) \Rightarrow \ln p(x|C_1) + \ln p(C_1) - \ln p(x) = \ln p(x|C_2) + \ln p(C_2) - \ln p(x)$$

$$\Rightarrow \ln p(x|C_1) + \ln p(C_1) = \ln p(x|C_2) + \ln p(C_2) \Rightarrow g(x) = \ln \frac{p(x|C_1)}{p(x|C_2)} + \ln \frac{p(C_1)}{p(C_2)} = \ln p(x|C_1) - \ln p(x|C_2) + \ln \frac{p(C_1)}{p(C_2)}$$

$$\ln p(x|C_1) = -\frac{d}{2} \ln 2\pi - \frac{1}{2} \ln \left| \Sigma_1 \right| - \frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1)$$

$$\ln p(x|C_2) = -\frac{d}{2} \ln 2\pi - \frac{1}{2} \ln \left| \Sigma_2 \right| - \frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2)$$

$$g(x) = -\frac{d}{2} \ln 2\pi - \frac{1}{2} \ln \left| \Sigma_1 \right| - \frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \frac{d}{2} \ln 2\pi + \frac{1}{2} \ln \left| \Sigma_2 \right| + \frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) + \ln \frac{\pi}{1 - \pi}$$



بنابراین پس از ساده سازی برای مرز تصمیم گیری بر اساس  
فرضیات مسئله خواهیم داشت:

$$g(x) = -\frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) + \ln \frac{\pi}{1 - \pi} - \frac{1}{2} \ln \left| \Sigma_1 \right| + \frac{1}{2} \ln \left| \Sigma_2 \right|$$



در صورتی که ماتریس کوواریانس دو دسته یکسان در نظر گرفته شود، نشان دهید مرز به صورت یک ابر صفحه در می آید. پارامترهای این ابر صفحه را بر حسب پارامترهای  $\pi$ ، بردار میانگین و ماتریس کوواریانس دسته ها بنویسید.



بر اساس محاسبات انجام شده در قسمت الف  
سوال ۱ داریم:

$$\bullet \quad p(C_1|x) = p(C_2|x) \quad \longrightarrow \quad \ln p(x|C_1) + \ln p(C_1) - \ln p(x) = \ln p(x|C_2) + \ln p(C_2) - \ln p(x)$$

$$\longrightarrow \ln p(x|C_1) + \ln p(C_1) = \ln p(x|C_2) + \ln p(C_2) \quad \longrightarrow$$

$$g(x) = \ln \frac{p(x|C_1)}{p(x|C_2)} + \ln \frac{p(C_1)}{p(C_2)} =$$
$$\ln p(x|C_1) - \ln p(x|C_2) + \ln \frac{p(C_1)}{p(C_2)}$$

$$\left\{ \begin{array}{l} \ln p(x|C_1) = -\frac{d}{2} \ln 2\pi - \frac{1}{2} \ln \left| \Sigma_1 \right| - \frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) \\ \ln p(x|C_2) = -\frac{d}{2} \ln 2\pi - \frac{1}{2} \ln \left| \Sigma_2 \right| - \frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \end{array} \right.$$

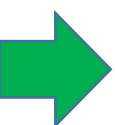


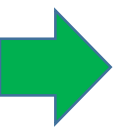

❖ بنابراین بر اساس فرضیات مسئله که  
کوواریانس دو کلاس را یکسان در نظر  
گرفته است، داریم:

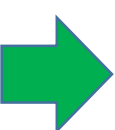


$$\Sigma_1 = \Sigma_2 = \Sigma$$

$$-\cancel{\frac{d}{2} \ln 2\pi} - \cancel{\frac{1}{2} \ln |\Sigma_1|} - \frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \ln \pi = -\cancel{\frac{d}{2} \ln 2\pi} - \cancel{\frac{1}{2} \ln |\Sigma_2|} - \frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) + \ln 1 - \pi$$


$$-\frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \ln \pi = -\frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) + \ln 1 - \pi$$


$$-\frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \ln \pi - \ln 1 - \pi = -\frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2)$$

$$C = \ln \pi - \ln 1 - \pi$$


$$-\frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + C = -\frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \xrightarrow{\times -2}$$

$$(x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + C = (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2)$$





$$\diamond (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) - 2C = (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \xrightarrow{\text{ساده سازی}}$$

$$\rightarrow x^T \Sigma^{-1} x - \mu_1^T \Sigma^{-1} x - x^T \Sigma^{-1} \mu_1 + \mu_1^T \Sigma^{-1} \mu_1 - 2C = x^T \Sigma^{-1} x - \mu_0^T \Sigma^{-1} x - x^T \Sigma^{-1} \mu_0 + \mu_0^T \Sigma^{-1} \mu_0$$

$$\rightarrow x^T \Sigma^{-1} x - 2\mu_1^T \Sigma^{-1} x + \mu_1^T \Sigma^{-1} \mu_1 - 2C = x^T \Sigma^{-1} x - 2\mu_0^T \Sigma^{-1} x + \mu_0^T \Sigma^{-1} \mu_0$$

$$\rightarrow -2\mu_1^T \Sigma^{-1} x + \mu_1^T \Sigma^{-1} \mu_1 - 2C = -2\mu_0^T \Sigma^{-1} x + \mu_0^T \Sigma^{-1} \mu_0$$

$$\underbrace{-2(\mu_1 - \mu_0)^T \Sigma^{-1} x}_w + \underbrace{(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0) - 2C}_b = 0 \rightarrow$$

$$wx + b = 0$$



بنابراین به معادله ی یک خط رسیدیم که در حالت کلی معادله ی یک ابر صفحه  
است



در قسمت ب، فرض کنید تابع هزینه به این صورت تعریف شود که اگر داده کلاس اول اشتباهها به داده کلاس دوم و داده کلاس دوم اشتباهها به داده کلاس اول نسبت داده شود به ترتیب هزینه های  $L_2, L_1$  را در بر بگیرد ( $L_2, L_1 > 0$ )، در خصوص تغییر مرز دسته بند با توجه به مقادیر مختلف  $L_2, L_1$  توضیح دهید.



❖ در دسته بند بیز می دانیم که طبق قاعده ی Bayes Decision Rule داریم:

If  $P(\mathcal{C}_1|\mathbf{x}) > P(\mathcal{C}_2|\mathbf{x})$  decide  $\mathcal{C}_1$   
otherwise decide  $\mathcal{C}_2$

❖ همچنین می دانیم برای بدست آوردن مقدار زیان در دسته بند بیز داریم :

$$p(error|\mathbf{x}) = \begin{cases} p(\mathcal{C}_2|\mathbf{x}) & \text{if we decide } \mathcal{C}_1 \\ P(\mathcal{C}_1|\mathbf{x}) & \text{if we decide } \mathcal{C}_2 \end{cases}$$

❖ بنابراین طبق دو رابطه ی بالا همواره در یک دسته بندی دو کلاسه طبق دسته بند بیز، مقدار error برابر با احتمال کلاسی است که کمترین مقدار احتمال را دارد و طبق رابطه ی زیر داریم:

$$P(error|\mathbf{x}) = \min\{P(\mathcal{C}_1|\mathbf{x}), P(\mathcal{C}_2|\mathbf{x})\}$$

بنابراین با توجه به روابط گفته باید به دنبال این باشیم که میزان error کمینه کنیم



► Decision regions:  $\mathcal{R}_k = \{x | \alpha(x) = k\}$

► All points in  $\mathcal{R}_k$  are assigned to class  $\mathcal{C}_k$

$$p(\text{error}) = E_{x,y}[I(\alpha(x) \neq y)]$$

$$= p(x \in \mathcal{R}_1, \mathcal{C}_2) + p(x \in \mathcal{R}_2, \mathcal{C}_1)$$

$$= \int_{\mathcal{R}_1} p(x, \mathcal{C}_2) dx + \int_{\mathcal{R}_2} p(x, \mathcal{C}_1) dx$$

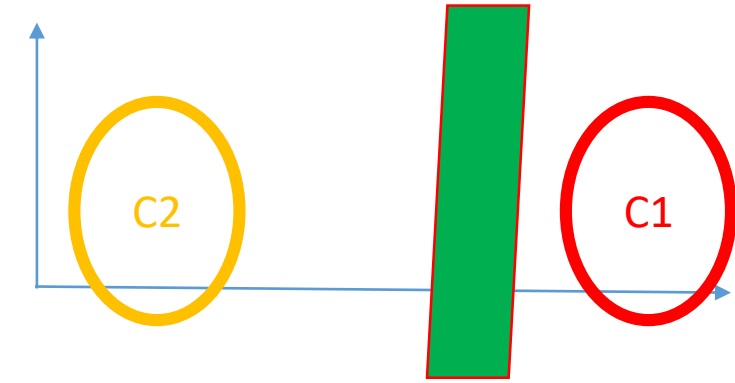
$$= \int_{\mathcal{R}_1} p(\mathcal{C}_2|x)p(x) dx + \int_{\mathcal{R}_2} p(\mathcal{C}_1|x)p(x) dx$$

Choose class with highest  $p(\mathcal{C}_k|x)$  as  $\alpha(x)$

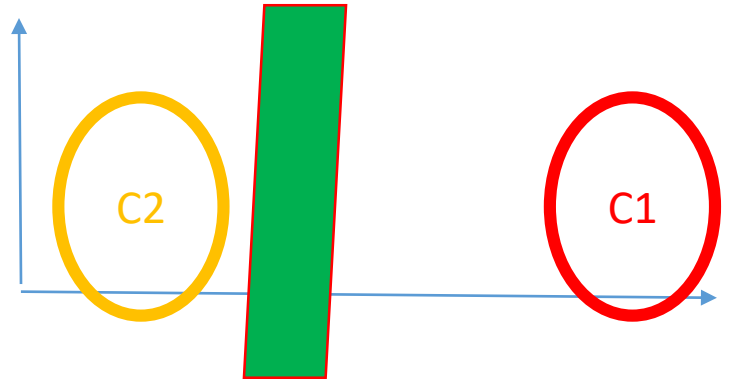
برای کمینه کردن میزان error طبق روابط روبه رو، باید بین حالاتی که ۱) در ناحیه ی R1 هستیم و به میزان  $p(\mathcal{C}_2|x)$  (یعنی به میزان تشخیص داده های کلاس ۲ به اشتباه به کلاس ۱ خطا داریم) یا ۲) در ناحیه R2 هستیم و به میزان  $p(\mathcal{C}_1|x)$  (یعنی به میزان تشخیص داده های کلاس ۱ به اشتباه به کلاس ۲ خطا داریم)، باید یکی را انتخاب کنیم که مینیمم است.

❖ حال با توجه به فرضیات مسئله (قسمت ب) می دانیم که دسته بند ما یک ابر صفحه است و همچنین می دانیم که L1 میزان هزینه ای است که به ازای پیش بینی کلاس یک به اشتباه به کلاس دو منجر می شود و L2 میزان هزینه ای است که به ازای کلاس دو به اشتباه به کلاس یک منجر می شود.

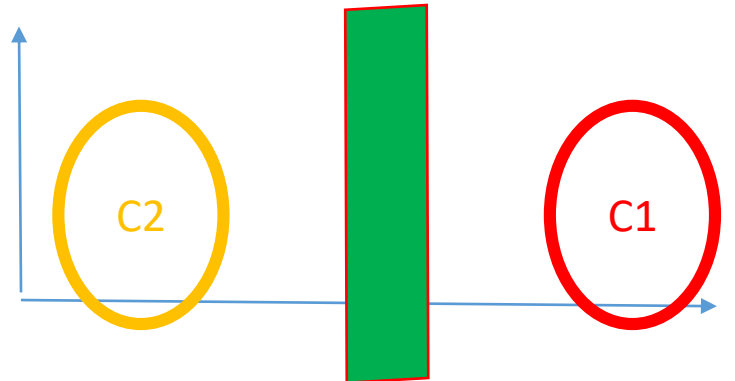
❖ اگر فرض کنیم که  $L1 > L2$  است:



❖ اگر فرض کنیم که  $L2 > L1$  است:



❖ اگر فرض کنیم که  $L2 = L1$  است:



در این حالت با توجه فرضیاتی که مطرح کردیم، خواهیم داشت  $p(C_2|x) > p(C_1|x)$ . که این به این معناست که فضای احتمالی  $p(C_2|x)$  بیشتر بوده است و  $error$  کمتری داشته است و این مسئله به جابجایی مرز دسته بندی به سمت کلاس یک می شود. بنابراین فضای مسئله به این صورت می شود که دو کلاس ۱ و ۲ به وسیله ی یک صفحه از یکدیگر جدا می شوند با این تفاوت که این صفحه به دسته ی یک نزدیک تر است.

در این حالت با توجه فرضیاتی که مطرح کردیم، خواهیم داشت  $p(C_2|x) < p(C_1|x)$ . که این به این معناست که فضای احتمالی  $p(C_1|x)$  بیشتر بوده است و  $error$  کمتری داشته است و این مسئله به جابجایی مرز دسته بندی به سمت کلاس دو می شود. بنابراین فضای مسئله به این صورت می شود که دو کلاس ۱ و ۲ به وسیله ی یک صفحه از یکدیگر جدا می شوند با این تفاوت که این صفحه به دسته ی دو نزدیک تر است.

در این حالت با توجه فرضیاتی که مطرح کردیم، خواهیم داشت  $p(C_2|x) = p(C_1|x)$ . که این به این معناست که فضای احتمالی  $p(C_2|x)$  و  $p(C_1|x)$  با یکدیگر برابر است و  $error$  یکسانی دارند. بنابراین فضای مسئله به این صورت می شود که دو کلاس ۱ و ۲ به وسیله ی یک صفحه از یکدیگر جدا می شوند با این تفاوت که این صفحه در جایی بین دو کلاس قرار می گیرد.



در حالت  $\Sigma_1 = \Sigma_2 = \sigma^2 I$  نشان دهید که دسته بند  
بیز برای هر داده، دسته ای را انتخاب می کند که میانگین آن به  
داده نزدیک تر است.



بر اساس محاسبات انجام شده در  
قسمت ب سوال ۱ داریم:

$$-\cancel{\frac{d}{2} \ln 2\pi} - \cancel{\frac{1}{2} \ln \left| \Sigma_1 \right|} - \frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \ln \pi = -\cancel{\frac{d}{2} \ln 2\pi} - \cancel{\frac{1}{2} \ln \left| \Sigma_2 \right|} - \frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) + \ln 1 - \pi$$

$$\rightarrow -\frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \ln \pi = -\frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) + \ln 1 - \pi$$

$$\rightarrow -\frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \ln \pi - \ln 1 - \pi = -\frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \rightarrow C = \ln \pi - \ln 1 - \pi$$

$$\rightarrow -\frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + C = -\frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \xrightarrow{\times -2}$$

$$\rightarrow \text{بنابراین در قسمت ب به رابطه ی مقابل رسیدیم.} \quad \blacklozenge (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) - 2C = (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2)$$



بنابر بر رابطه ی مقابل و فرضیات  
مطرح شده  $\Sigma_1 = \Sigma_2 = \sigma^2 I$  داریم:

$$(x - \mu_1)^T \sum_1^{-1} (x - \mu_1) - 2C = (x - \mu_2)^T \sum_2^{-1} (x - \mu_2)$$

$$(x - \mu_1)^T (\sigma^2)^{-1} (x - \mu_1) - 2C = (x - \mu_2)^T (\sigma^2)^{-1} (x - \mu_2) \rightarrow$$

$$\frac{(x - \mu_1)^T (x - \mu_1)}{\sigma^2} - 2C = \frac{(x - \mu_2)^T (x - \mu_2)}{\sigma^2} \rightarrow \text{طرفین ضرب در } \sigma^2$$

$$(x - \mu_1)^T (x - \mu_1) - 2C\sigma^2 = (x - \mu_2)^T (x - \mu_2) \rightarrow$$

$$g(x) = (x - \mu_2)^T (x - \mu_2) - (x - \mu_1)^T (x - \mu_1) + 2C\sigma^2 \rightarrow$$

❖ اگر احتمال prior هر دسته را با یکدیگر برابر در نظر بگیریم خواهیم داشت  $C=0$  زیرا  
 $\ln \frac{p(c)}{p(c)} = \ln 1 = 0$  بنابراین:

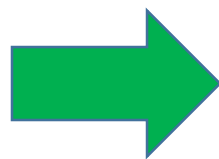
$$g(x) = (x - \mu_2)^T (x - \mu_2) - (x - \mu_1)^T (x - \mu_1)$$

$$(x - \mu_2)^T (x - \mu_2) - (x - \mu_1)^T (x - \mu_1) = 0 \rightarrow -x^2 - \mu_1^2 + 2x\mu_1 + x^2 + \mu_2^2 - 2x\mu_2 = 0$$



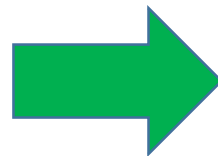


$$\diamond -x^2 - \mu_1^2 + 2x\mu_1 + x^2 + \mu_2^2 - 2x\mu_2 = 0$$

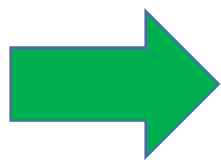


پس از ساده سازی داریم:

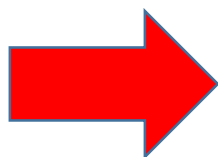
$$2x(\mu_1 - \mu_2) = \mu_1^2 - \mu_2^2$$



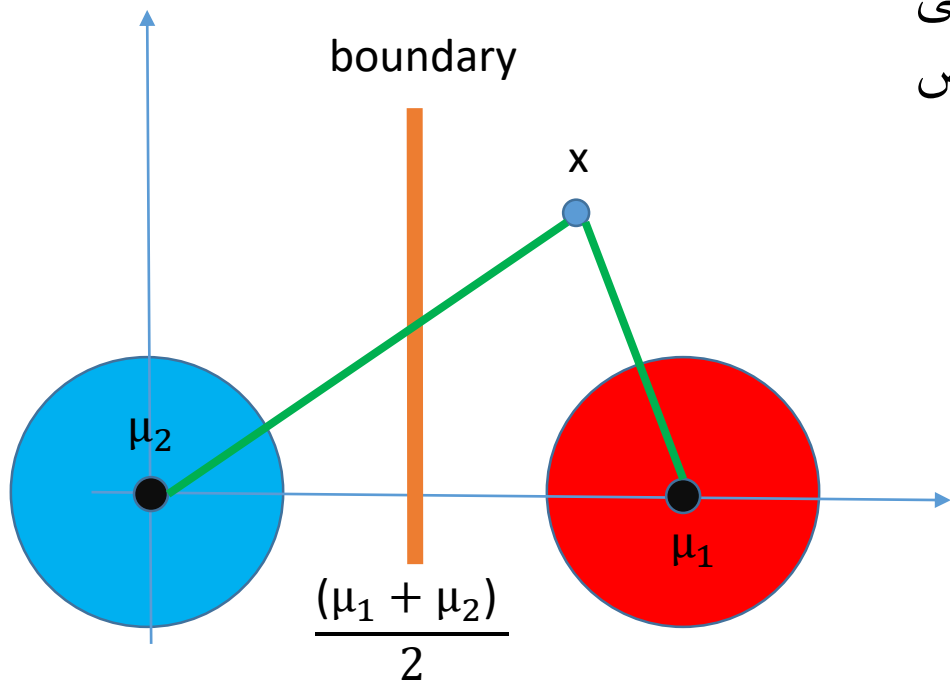
$$2x(\mu_1 - \mu_2) = (\mu_1 - \mu_2)(\mu_1 + \mu_2)$$



$$x = \frac{(\mu_1 + \mu_2)}{2}$$



بنابراین مرز دسته بندی در بین میانگین دو نقطه ی میانگین هر کلاس قرار می گیرد، از این رو می توان این گونه نتیجه گرفت که به ازای هر داده ی جدید، نزدیکی آن داده به نقطه ی میانگین هر کلاس، دسته ی آن کلاس را مشخص می کند(همانطور که در شکل زیر مشاهده می کنید) و فرض مسئله ی ما اثبات می شود.





ه) برای قسمت ب، با تشکیل توزیع  $P(x, y)$  و با استفاده از تخمین بیشینه درستنمایی و با فرض داشتن  $N$  داده آموزش  $\{(x^{(i)}, y^{(i)})\}$  پارامترهای  $\mu_1, \mu_2, \Sigma, \pi$  را به دست آورید. روابط قسمت الف چه تفاوتی با قسمت ب خواهند داشت؟ نیازی به نوشتن دقیق روابط قسمت الف نیست و توضیح تفاوت کافی است.

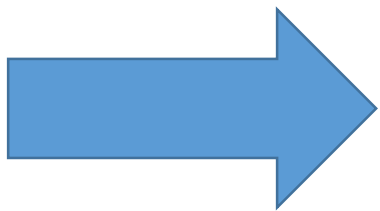


❖ با توجه به فرضیات مسئله خواهیم داشت:

$$\mu = \frac{1}{N} \sum_{n=1}^N x^{(n)}$$
$$\sigma^2 = \frac{1}{N} \sum_{n=1}^N (x^{(n)} - \mu)^2$$

$$\boldsymbol{\mu} = \frac{\sum_{n=1}^N \mathbf{x}^{(n)}}{N}$$

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}^{(n)} - \boldsymbol{\mu})(\mathbf{x}^{(n)} - \boldsymbol{\mu})^T$$



$$\bullet \quad p(x, y) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{-\frac{1}{2} (x - \mu_k)^T \boldsymbol{\Sigma}_k^{-1} (x - \mu_k)\right\}$$

$k=1,2$

$$\bullet \quad p(x, y) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_1|^{1/2}} \exp\left\{-\frac{1}{2} (x - \mu_1)^T \boldsymbol{\Sigma}_1^{-1} (x - \mu_1)\right\}$$

$k=1$

$$\bullet \quad p(x, y) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_2|^{1/2}} \exp\left\{-\frac{1}{2} (x - \mu_2)^T \boldsymbol{\Sigma}_2^{-1} (x - \mu_2)\right\}$$

$k=2$

$$Data = \{(x^{(n)}, y^{(n)})_{n=1}^N\}$$



❖ با توجه به فرضیات مسئله خواهیم داشت:

$$\pi = \frac{N_1}{N}$$

$$\mu_1 = \frac{\sum_{n=1}^N y^{(n)} x^{(n)}}{N_1}$$

$$\mu_2 = \frac{\sum_{n=1}^N (1 - y^{(n)}) x^{(n)}}{N_2}$$

$$N_1 = \sum_{n=1}^N y^{(n)}$$

$$N_2 = N - N_1$$

$$\sum_1 = \sum_2 = \sum \frac{1}{N_1} \sum_{n=1}^N y^{(n)} (x^{(n)} - \mu) (x^{(n)} - \mu)^T$$



❖ با توجه به قسمت الف پس از نوشتن روابط و ساده سازی ها به مرز دسته بندی زیر رسیدیم:

$$g(x) = -\frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \frac{1}{2}(x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) + \ln \frac{\pi}{1 - \pi} - \frac{1}{2} \ln \left| \Sigma_1 \right| + \frac{1}{2} \ln \left| \Sigma_2 \right|$$

❖ همانطور که در مرز دسته بندی بالامشاهده می کنیم، این مرز دسته بندی دارای جملات به شکل  $x^T \Sigma^{-1} x$  است که این جملات حاکی از آن است که مرز دسته بند به فرم درجه دو است. بنابراین در قسمت الف ما با یک مرز دسته بندی درجه دو مواجه هستیم.

❖ با توجه به قسمت ب پس از نوشتن روابط و

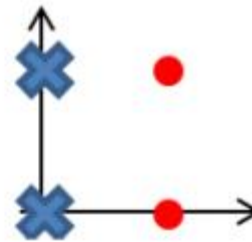
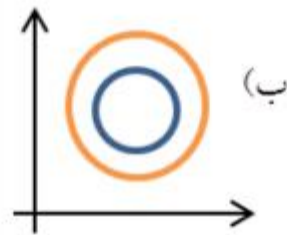
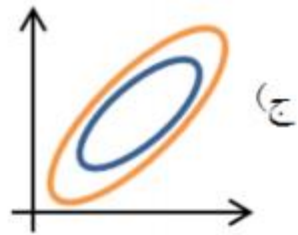
ساده سازی ها به مرز دسته بندی زیر رسیدیم:

$$g(x) = \underbrace{-2(\mu_1 - \mu_0)^T \Sigma^{-1}}_{wx} x + \underbrace{(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0) - 2C}_{b} = g(x)$$

❖ همانطور که در مرز دسته بندی بالامشاهده می کنیم، این مرز دسته بندی دارای جملات به شکل  $\Sigma^{-1} x$  است که این جمله حاکی از آن است که مرز دسته بند به فرم درجه یک است. بنابراین در قسمت ب ما با یک مرز دسته بندی درجه دو مواجه هستیم. بنابراین مرز دسته بندی قسمت الف و ب از نظر فرم معادله (قسمت الف درجه دو و قسمت ب درجه یک) متفاوت هستند.

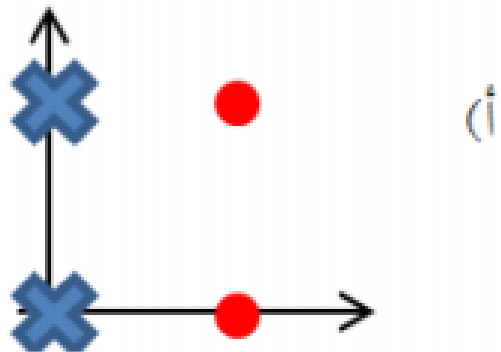


در شکل های زیر در حالت پیوسته منحنی های آبی و نارنجی، کانتورهای  $P(x|C_1) = 0.5$  و  $P(x|C_2)$  را نشان می دهند (همچنین احتمال پیشین دسته ها در شکل های ب و ج برابر در نظر گرفته می شود) و در حالت گسسته احتمال نقاط هم رنگ با هم برابر و احتمال نقاط آبی دو برابر نقاط قرمز است. با استدلال بررسی نمایید که آیا فرض استقلال شرطی که در دسته بند Naive Bayes استفاده می شود، در هر کدام از این سه مورد برقرار است یا نه؟ چنان چه این شرط برقرار است مرز تقریبی که توسط این دسته بند برای دو دسته بدست می آید را در شکل زیر مشخص نمایید. (محور افقی ویژگی  $x_1$  و محور عمودی ویژگی  $x_2$  است)





❖ فرض می کنیم که ویژگی های بر اساس مفروضات مسئله و شکل مقابل به صورت فرضی زیر باشد:



	X1	X2	target
0	0	0	C1
1	0	1	C1
2	1	0	C2
3	1	1	C2



ابتدا همبستگی داده ها را محاسبه می کنیم، بنابراین بر این کار ابتدا ماتریس کوواریانس داده ها را محاسبه می کنیم:

ماتریس کوواریانس

	C1	C2
C1	0.3333	0
C2	0	0.3333

بنابراین داده ها با یکدیگر همبستگی ندارند (روی یکدیگر تاثیری ندارند) و نسبت به یکدیگری استقلال دارند، بنابراین می توان نتیجه گرفت که می توان می توانیم فرض استقلال شرطی را برای دسته بند Naive Bayes در نظر بگیریم.

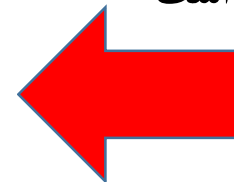


	C1	C2
C1	0.3333	0
C2	0	0.3333

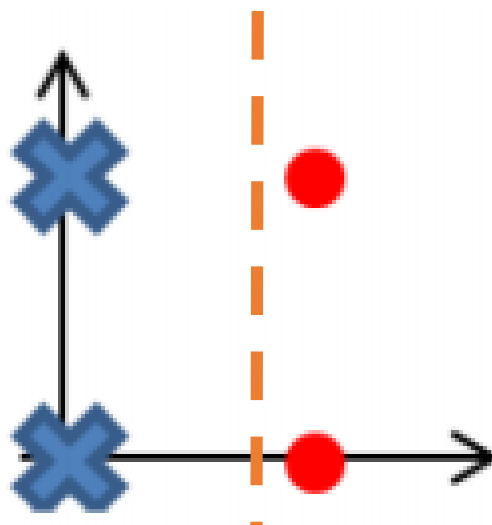
بنابراین داده ها با یکدیگر همبستگی ندارند (روی یکدیگر تاثیری ندارند) و نسبت به یکدیگری استقلال دارند، بنابراین می توان نتیجه گرفت که می توان می توانیم فرض استقلال شرطی را برای دسته بند Naive Bayes در نظر بگیریم.



همچنین بر اساس فرضیات مسئله می دانیم که احتمال دسته ی آبی دو برابر دسته ی قرمز است، بنابراین با توجه به این که ماتریس کوواریانس بدست آمده به فرم  $\Sigma_2 = \sigma^2 I$  است، طبق معادله ی بدست آمده در قسمت ب سوال یک، می دانیم که مرز دسته بند خطی (در صورت تعدد ویژگی ها صفحه و ابر صفحه) بین داده های دو کلاس C1 و C2 قرار می گیرد و چون احتمال داده های آبی دو برابر داده های قرمز است، بنابراین بنابراین مرز دسته بند به دسته ی با احتمال کمتر (داده های قرمز) نزدیک تر است



دسته بند

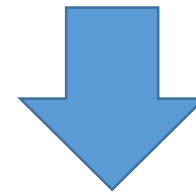
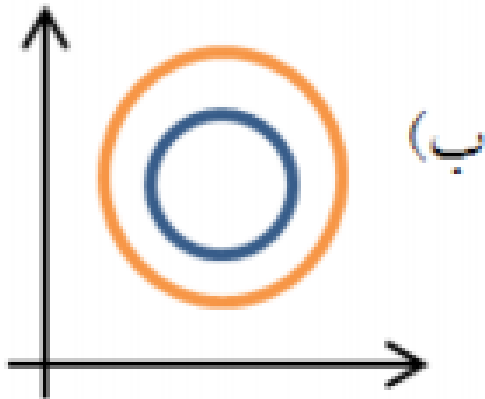


(f)

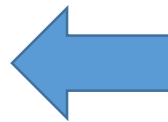




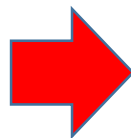
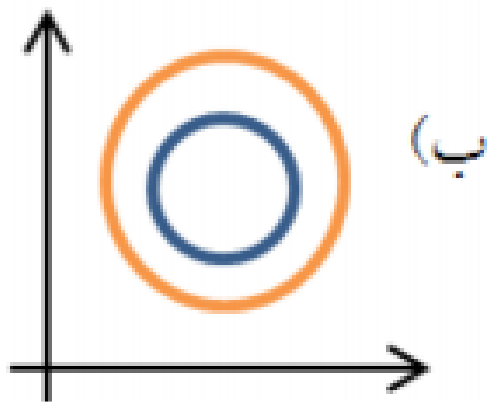
با توجه به توزیع پیوسته رو به رو می توانیم اینگونه توصیف کنیم که مرکز دو کلاس (میانگین دو کلاس) به صورت تقریبی بر روی یکدیگر قرار دارند با این تفاوت که مقدار واریانس داده ی نارنجی بزرگتر از داده ی آبی است، همچنین ماتریس کوواریانس به صورت زیر خواهد بود.



	C1	C2
C1	$\sigma_1^2$	0
C2	0	$\sigma_2^2$



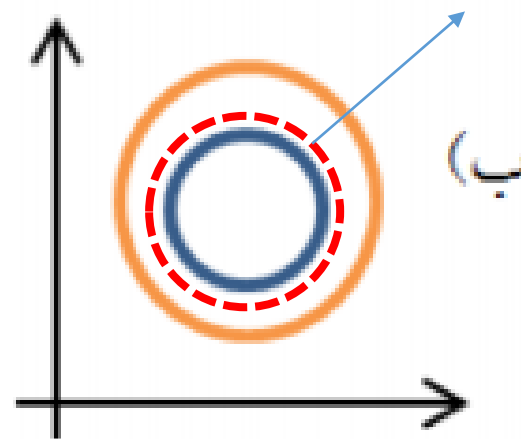
بنابراین بر اساس ماتریس کوواریانس مقابل، داده های موجود در دو کلاس با یکدیگر همبستگی ندارد و بر روی یکدیگر تاثیری ندارند. بنابراین داده های دو کلاس نسبت به یکدیگر مستقل هستند و می توانیم فرض استقلال شرطی را برای Naive Bayes در نظر بگیریم.

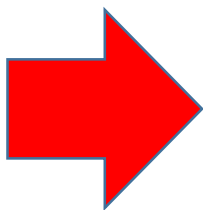
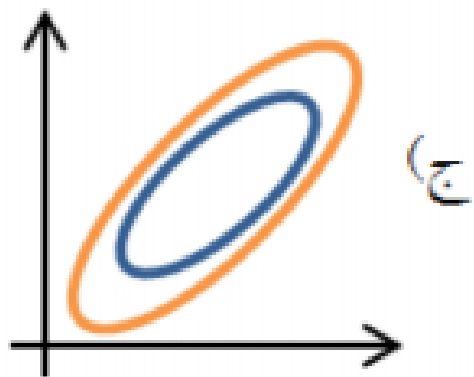


بنابراین بر اساس استدلال های مطرح شده و بر اساس  
شکل کانتوری مقابل، دسته بند Naive Bayes به  
صورت دایره وار (درجه دو) خواهد بود و چون طبق  
فرضیه های مسئله  $P(x|C_2) = P(x|C_1) = 0.5$   
است و احتمال پیشین نیز برابر است بنابراین مرز دسته  
ها به صورت تقریبی در جایی بین حاشیه ی کلاس ها  
قرار می گیرد:



دسته بند





با توجه به توزیع پیوسته رو به رو می توانیم اینگونه توصیف کنیم که مرکز دو کلاس (میانگین دو کلاس) به صورت تقریبی بر روی یکدیگر قرار دارند با این تفاوت که ویژگی  $X_1$  و  $X_2$  با یکدیگر همبستگی مثبت دارند، به این صورت که افزایش یکی بر روی دیگری تاثیر می گذارد و بالعکس. بنابراین قطر فرعی ماتریس کوواریانس مخالف صفر خواهد بود. بنابراین فرض استقلال را برای آنها نمی توان در نظر گرفت. بنابراین فرض استقلال شرطی برای دسته بند Naive Bayes برقرار نیست.



یک مسئله ی logistic regression را در نظر بگیرید. اگر فیچرهای موجود در مجموعه های داده یکسان باشد، چه اتفاقی برای این نوع دسته بند خواهد افتاد؟ به عنوان مثال ۲ مجموعه داده زیر را در نظر بگیرید:

$$DS_1 = (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(N)}, y^{(N)})$$

$$DS_2 = (x^{(1)}, x^{(1)}, y^{(1)}), (x^{(2)}, x^{(2)}, y^{(2)}), \dots, (x^{(N)}, x^{(N)}, y^{(N)})$$

در صورت آموزش روی مجموعه داده دوم به جای اول، عملکرد این نوع دسته بند چه تفاوتی خواهد کرد؟



برای پاسخ به این سوال در ابتدا باید به یک درک شهودی از فضای مسئله و چالش های موجود در دسته بند های خطی و logistic regression رسید. این چالش همواره در الگوریتم های یادگیری ماشین وجود دارد که وقتی داده ها و ویژگی های ورودی مسئله، همبستگی بالایی نسبت به یکدیگر دارند، این مسئله سبب کاهش کارایی الگوریتم های یادگیری ماشین می شود. یکی از این چالش ها، حافظه و زمان است، در واقع افزونگی داده ها و داده های با همبستگی زیاد می تواند سبب افزایش زمان اجرا و حافظه شود. اما این چالش به تنهایی سبب ایجاد مشکل در الگوریتم ها یادگیری ماشین نمی شود و مشکلی جدی و چالش برانگیز نیست. چالش اصلی و جدی در واقع کاهش دقت و عدم استفاده عمومی از مدل های یاد گرفته شده به علت همبستگی و عدم استقلال داده است. فرض کنید که ما یک مسئله ی تشخیص سرطان داریم و می خواهیم با استفاده از logistic regression بتوانیم سرطان خوش خیم و بد خیم را تشخیص دهیم. فرض می کنیم ویژگی  $X_1$  و  $X_2$  و  $X_3$  و... ویژگی هایی هستند که ما از آنها برای پیش بینی خوش خیمی و بد خیمی سرطان به ازای داده های جدید استفاده می کنیم. فرض می کنیم که ویژگی  $X_1$  و  $X_2$  به صورت خطی به یکدیگر وابسته هستند و رابطه خطی بین آنها به صورت  $X_1 = 1.2X_2$  برقرار است. حال فرض می کنیم مرز دسته بند به شکل زیر باشد:

$$\theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \dots$$



$$\theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \dots$$

با توجه به رابطه ی خطی که بین ویژگی های  $X_1$  و  $X_2$  وجود دارد، رابطه ی با صورت زیر قابل باز نویس است.

$$\theta_0 + (1.2\theta_1 + \theta_2)X_2 + \theta_3 X_3 + \dots$$

فرض کنیم ویژگی های  $X_1$  و  $X_2$  در بسیاری از موارد تاثیری در پیش بینی خوش خیم یا بدخیمی سرطان ندارند، بنابراین خواهیم داشت،  $\theta_1 = \theta_2 = 0$ . با این وجود به دلیل وابستگی خطی موجود بین  $X_1, X_2$ ، مدل یادگرفته شده تفاوت بین  $\theta_2 = -1.2$  و  $\theta_1 = 1$  یا  $\theta_2 = 12$  و  $\theta_1 = -10$  و یا هر ترکیب دیگری از  $\theta_1 = -\frac{\theta_2}{1.2}$  را تشخیص نمی دهد. بنابراین مدل ما ممکن است از نظر عددی ناپایدار شود. از طرف دیگر، دیگر نمی توانیم نتایج را به درستی تفسیر کنیم و این مسئله می تواند سبب افت دقت مدل شود. بنابراین با توجه به این ایده به فرضیات مسئله ی خود بر می گردیم:

$$\theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \dots$$



بنابراین با توجه به فرضیات، می دانیم که ویژگی ها مسئله دارای وابستگی خطی  $X1=X2$  هستند و مشابه آنچه توضیح داده شد از نظر عددی ناپایدار می شوند و این مسئله می تواند سبب افت دقت شود. بنابراین برای جلوگیری از بروز این اتفاق نیاز داریم تا در مرحله‌ی پیش پردازش داده های با همبستگی زیاد را حذف کنیم.



درستی یا نادرستی گزاره های زیر را با توضیح مختصر مشخص نمایید.

- در SVM حاشیه نرم، با افزایش پارامتر  $C$  حاشیه ی جداسازی دو دسته افزایش می یابد و خطای آموزش کاهش می یابد.
- هسته (Kernel) گاوسی با پارامتر  $\sigma$  را در نظر بگیرید. در حالت کلی از بین مقادیر مختلف پارامتر  $\sigma$  برای هسته گاوسی، مقداری که منجر به حاشیه ی بزرگتر می شود، مقدار بهتری است.





- در SVM حاشیه نرم، با افزایش پارامتر  $C$  حاشیه ی جداسازی دو دسته افزایش می یابد و خطای آموزش کاهش می یابد.

$$\min_{w, w_0, \{\xi_n\}_{n=1}^N} \frac{1}{2} \|w\|^2 + C \sum_{n=1}^N \xi_n$$

$$\text{s.t. } y^{(n)}(w^T x^{(n)} + w_0) \geq 1 - \xi_n \quad n = 1, \dots, N$$
$$\xi_n \geq 0$$

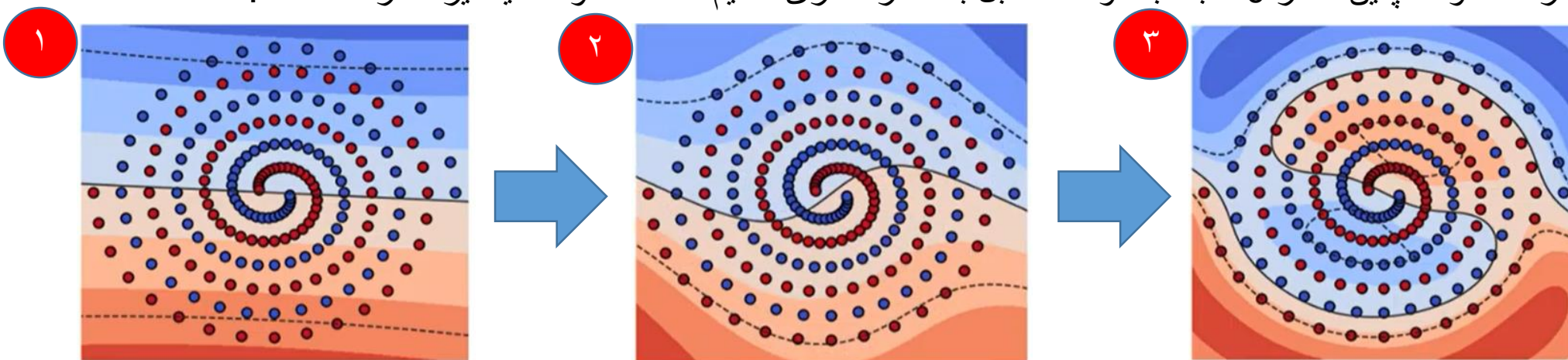
❖ می دانیم که رابطه ی SVM نرم به صورت مقابل است:

در عبارت مقابل هدف از جمله ی  $\frac{1}{2} \|W\|^2$  برای افزایش حاشیه مورد استفاده قرار می گیرد و مینیمم کردن این این جمله به افزایش حاشیه منجر می شود. می دانیم در SVM نرم علاوه بر افزایش حاشیه به دنبال کاهش یا افزایش میزان خطای آموزش هستیم برای این مسئله جمله  $C \sum_{n=1}^N \xi_n$  را اضافه می کنیم که در واقع با افزایش مقدار  $C$  به دنبال کاهش خطای آموزش می شویم. بنابراین در SVM نرم به دنبال دو چیز هستیم: ۱. افزایش حاشیه ۲. کاهش خطای آموزش. بنابراین با افزایش و کاهش مقدار  $C$  به افزایش یا کاهش حاشیه یا میزان خطای آزمایش اهمیت می دهیم به این صورت که اگر مقدار  $C$  کم باشد یعنی از خطای آموزش تا حدی صرف نظر می کنیم و به دنبال افزایش حاشیه هستیم و هرچه  $C$  بیشتر باشد، به دنبال کاهش خطای آموزش هستیم و از افزایش حاشیه تا حدی چشم پوشی می کنیم. اگر مقدار  $C$  بی نهایت شود در واقع ما به طور جدی از افزایش حاشیه چشم پوشی می کنیم. بنابراین طبق سوال این فرضیه که گفته شده است با افزایش  $C$  مقدار خطای آموزش کاهش می یابد درست است اما اینکه حاشیه ی جداسازی افزایش می یابد اشتباه است.



- هسته (Kernel) گاوسی با پارامتر  $\sigma$  را در نظر بگیرید. در حالت کلی از بین مقادیر مختلف پارامتر  $\sigma$  برای هسته گاوسی، مقداری که منجر به حاشیه ی بزرگتر می شود، مقدار بهتری است.

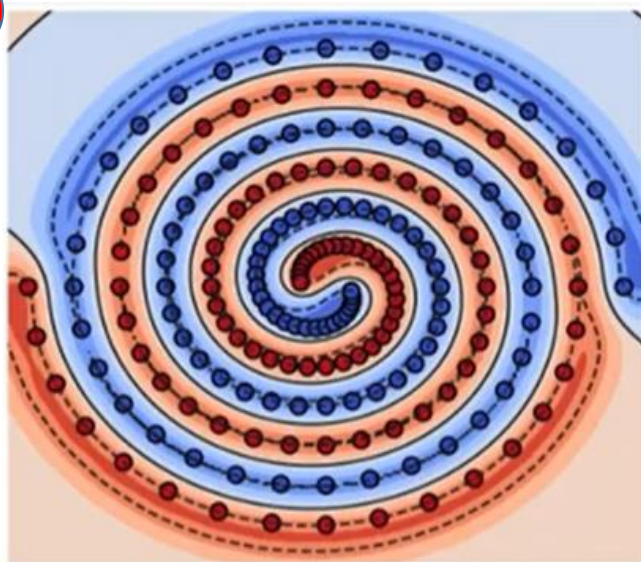
می دانیم که در مسائل یادگیری ماشین همواره، هایپرپارامترهایی وجود دارد که مشخص کردن مقدار آنها متناسب روند کلی ندارد و متناسب با هر مسئله می تواند این اعداد تغییر کند. از رایج ترین تکنیک های تعیین هایپر پارامترها Cross validation است. طبق فرضیات سوال این گونه مطرح شده است که برای تعیین هایپر پارامتر  $\sigma$  مقدار که به حاشیه ی بزرگتری منجر می شود بهتر است، اما همانطور که گفته شد برای تعیین مقدار هایپر پارامترها روند کلی وجود ندارد و می تواند مسئله به مسئله متفاوت باشد. برای مثال، شکل های زیر مجموعه ای از داده های حلزونی شکل است که شامل دو دسته است و الگوریتم SVM با مقادیر مختلف  $\sigma$  و مقدار ثابت C بر روی این داده ها اجرا شده است. شکل ها از سمت چپ به راست نشان دهنده دسته بندی SVM به ازای  $\sigma$  بزرگ تا کوچک است. همانطور که مشاهده می شود در شکل ۱ به دلیل بزرگ بودن مقدار  $\sigma$ ، مرز دسته بندی هموار تر شده است و همچنین حاشیه بزرگ تر است. همانطور که مشاهده می شود در شکل ۱ میزان خطای آموزش بسیار زیاد است، بنابراین مشخص می شود میزان پارامتر  $\sigma$  به درستی تنظیم نشده است. در شکل دو به دلیل کاهش مقدار  $\sigma$  کمی از همواری مرز دسته بندی کاسته شده است و همچنین مرز دسته بندی کوچکتر شده است اما همچنان خطای آموزش زیاد است. در شکل ۳ نیز وضعیتی مشابه شکل یک و دو وجود دارد، با این تفاوت که میزان خطای آموزش در شکل سه نسبت به شکل یک و دو کمتر است و همچنین مقدار  $\sigma$  نسبت به دو حالت قبل با مقدار کمتری تنظیم شده است و حاشیه نیز کمتر شده است.



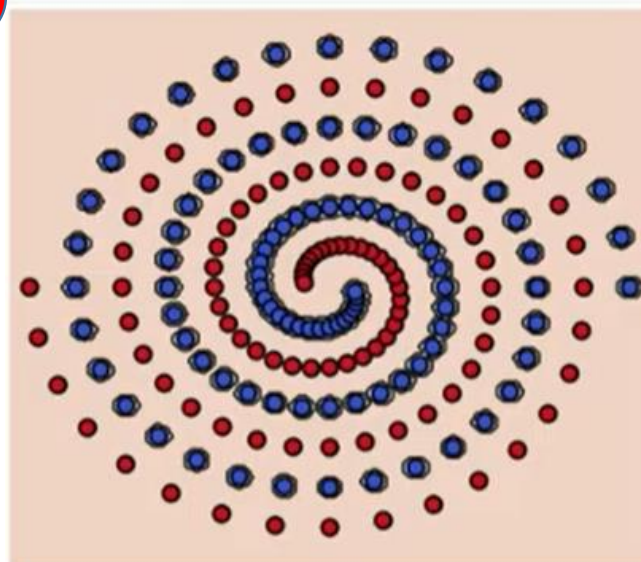


همانطور که مشاهده می کنید، در شکل دسته بند بسیار پیچیده تر از حالات قبلی است و همچنین میزان حاشیه نسبت به حالات قبل کاهش پیدا کرده است اما نکته‌ی قابل توجه این است که میزان خطای آموزش به کمترین مقدار خود نسبت به حالت قبل رسیده است و همه ی این مسائل به این دلیل است که میزان  $\sigma$  نسبت به حالات قبل کمتر است. بنابراین تا همین جا می توان نتیجه گرفت که تعیین میزان هاپیر پارامتر  $\sigma$  متاثر افزایش میزان حاشیه نیست. همچنین برای اینکه نشان دهیم تعیین میزان  $\sigma$  متاثر از کاهش میزان حاشیه نیست می توانیم به شکل ۵ نگاه کنیم، در این شکل میزان  $\sigma$  نسبت به حالات قبل بیشتر کاهش یافته و این مسئله سبب پیچیده تر شدن مرز تصمیم گیری شده است به گونه ای که هر مرز و حاشیه آن به دور داده افتاده است و این مسئله باعث می شود بیش برآزش رخ دهد و خط بر روی داده های اعتبار سنجی به طور چشم گیری می تواند افزایش یابد. بنابراین همانطور که مشاهده شد تعیین میزان پارامتر  $\sigma$  به طور کلی قانونی ندارد و نمی توان نتیجه گیری کرد که به طور کلی میزانی از  $\sigma$  که میزان حاشیه بیشتری را ایجاد می کند بهتر است. همانطور که در این مسئله داده های حلزونی بررسی کردیم.

۴



۵





❖ ثابت کنید در svm خطی در حالتی که داده ها به صورت خطی جدا پذیر هستند، عبارت زیر همواره برقرار است. دقت کنید که از حالت Hard margin استفاده کنید.



❖ می دانیم که تابع هدف در Hard margin SVM به شکل مقابل تعریف می شود:

$$\begin{aligned} & \min \frac{1}{2} ||W||^2 \\ s.t \quad & (w^T x^t + b) \geq +1 \quad \text{if } y^t = +1 \\ & (w^T x^t + b) \leq -1 \quad \text{if } y^t = -1 \end{aligned}$$

❖ و پس از ساده سازی خواهیم داشت:

$$\begin{aligned} & \min \frac{1}{2} ||W||^2 \\ s.t \quad & (w^T x^t + b) \geq +1 \end{aligned}$$

حل مسئله و دخیل کردن شرط با استفاده  
از ضریب لاگرانژ:

$$\begin{aligned} L_p &= \frac{1}{2} ||W||^2 - \sum_{t=1}^m \alpha^t [y^t (W^T X^t + b) - 1] \\ &= \frac{1}{2} ||W||^2 - \sum_{t=1}^m \alpha^t y^t (W^T X^t + b) + \sum_{t=1}^m \alpha^t \end{aligned}$$



برای پیدا کردن مرز تصمیم گیری داریم:

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow w = \sum_{t=1}^m \dot{\alpha}^t y^t x^t$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{t=1}^m \dot{\alpha}^t y^t = 0$$

بنابراین با توجه به روابط بدست آمده در بالا برای تعریف تابع هدف به شکل دوگان داریم:

$$L_d = \frac{1}{2}(W^T W) - W^T \sum_{t=1}^m \dot{\alpha}^t y^t x^t - b \sum_{t=1}^m \dot{\alpha}^t y^t + \sum_{t=1}^m \dot{\alpha}^t \quad \longrightarrow \quad L_d = \frac{1}{2}(W^T W) - W^T \sum_{t=1}^m \overset{W}{\cancel{\dot{\alpha}^t y^t x^t}} - b \sum_{t=1}^m \overset{0}{\cancel{\dot{\alpha}^t y^t}} + \sum_{t=1}^m \dot{\alpha}^t$$

$$L_d = \frac{1}{2}(W^T W) - W^T W + \sum_{t=1}^m \dot{\alpha}^t = -\frac{1}{2}(W^T W) + \sum_{t=1}^m \dot{\alpha}^t = -\frac{1}{2}||W||^2 + \sum_{t=1}^m \dot{\alpha}^t \quad \longrightarrow \quad \text{بنابراین می خواهیم این عبارت مینیمم شود و بنابر Karush-Kuhn-Tucker (KKT) conditions داریم:}$$

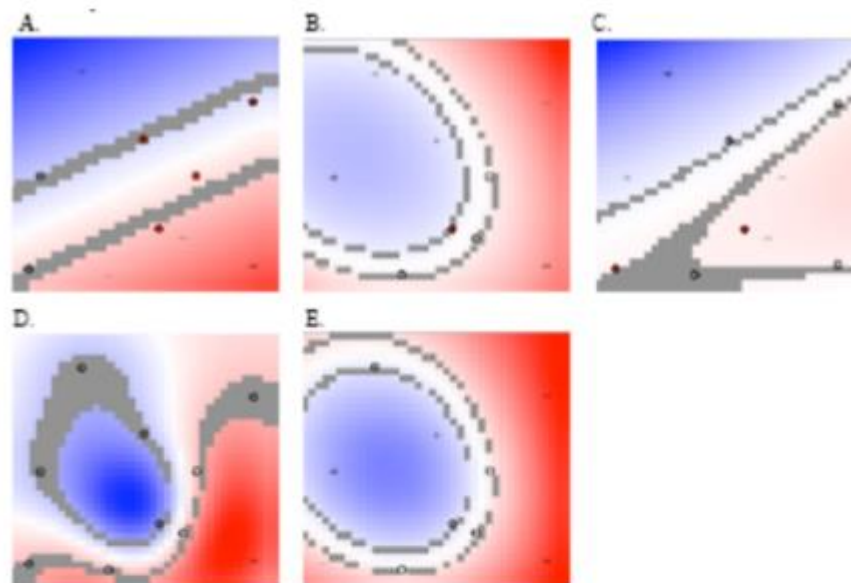
$$-\frac{1}{2}||W||^2 + \sum_{t=1}^m \dot{\alpha}^t = 0 \quad \longrightarrow \quad \frac{1}{2}||W||^2 = \sum_{t=1}^m \dot{\alpha}^t \quad \longrightarrow \quad ||W||^2 = \sum_{t=1}^m 2\dot{\alpha}^t \quad \longrightarrow \quad \text{می دانیم بسیاری از ضرایب آلفا برابر صفر است و تنها تعداد اندکی از ضرایب دارای مقدار بزرگ تر از صفر هستند بنابراین خواهیم داشت:}$$

$$\longrightarrow ||W||^2 = \sum_{t=1}^m \alpha^t$$

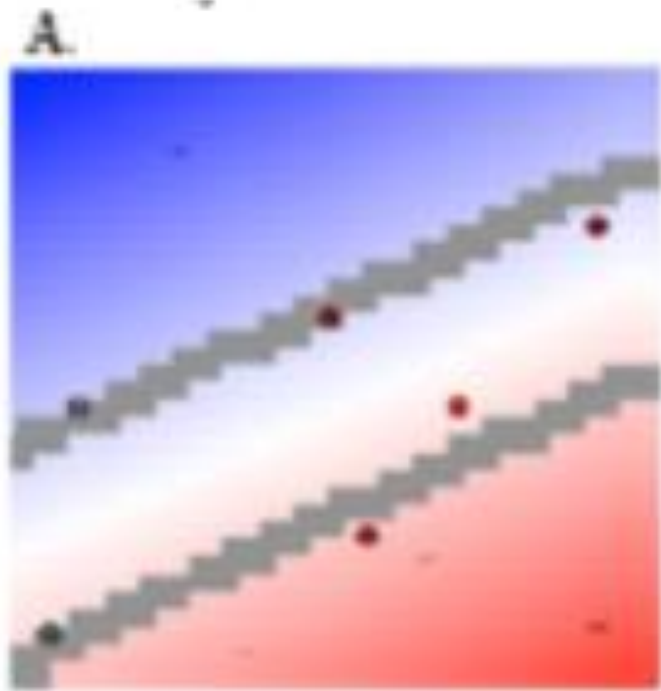




برای کی مجموعه داده ی مشخص پنج دیاگرام حاصل از SVM با کرنل های مختلف در شکل زیر را در نظر بگیرید. با یک توضیح مختصر در جدول مربوطه مشخص کنید که هر دیاگرام می تواند توسط کدام یک از کرنل های زیر ایجاد شود.



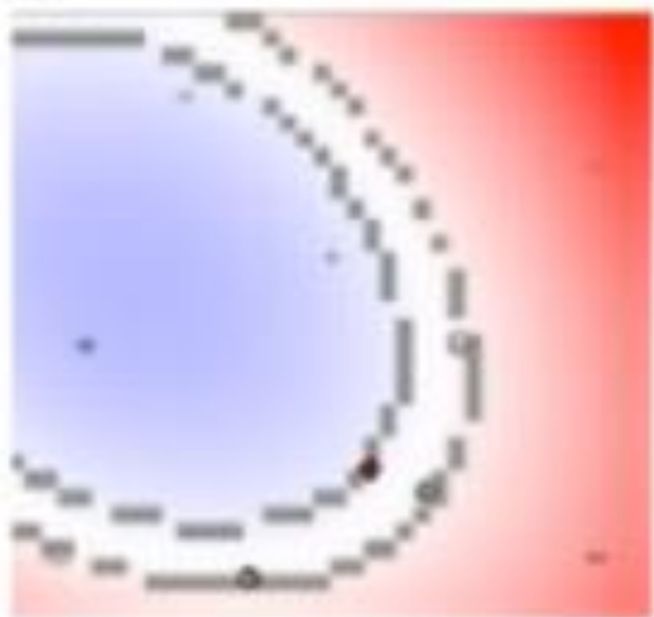
RBF, $\sigma = 0.1$	
RBF, $\sigma = 0.5$	
RBF, $\sigma = 2$	
Linear	
Second Order Polynomial	



شکل رو نشان دهنده ی یک دسته بند با کرنل خطی (Linear) است، این مسئله را به چند دلیل می توان نتیجه گرفت. یکی از این دلایل مرز هموار و خطی مشخص شده در شکل است و همچنین حاشیه های هموار و خطی، از دیگر دلایل که می توان به آن اشاره کرد این است که در کرنل خطی هر چقدر که از حاشیه ها دور می شویم احتمال تخصیص یک داده به آن کلاس افزایش می یابد و این مسئله نیز در شکل کاملاً مشخص به صورتی که به صورت متوازن با دور شدن از حاشیه ها رنگ آبی و قرمز پر رنگ تر شده است.

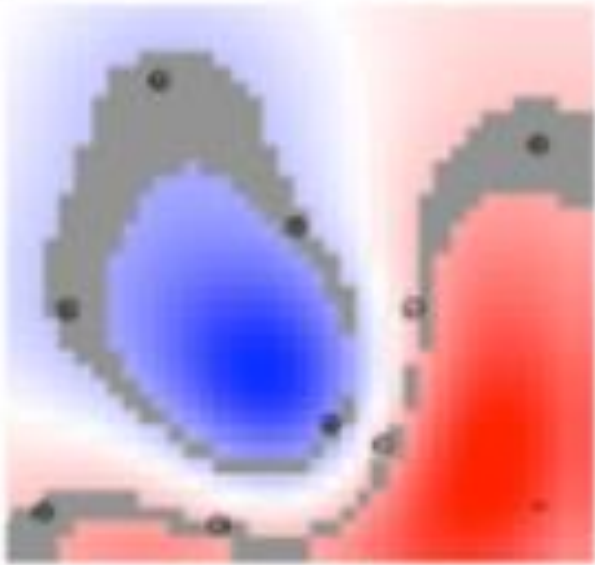


B.



شکل رو به رو نشان دهنده ی یک دسته بند با کرنل چند جمله ای از درجه ی دو است. این مسئله از مرز درجه دو که در شکل مقابل تقریباً شبیه یک نیم دایره است کاملاً مشخص است.

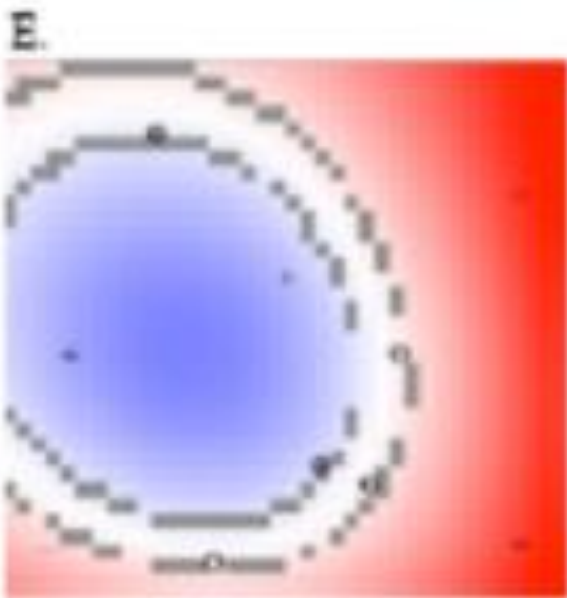
D.

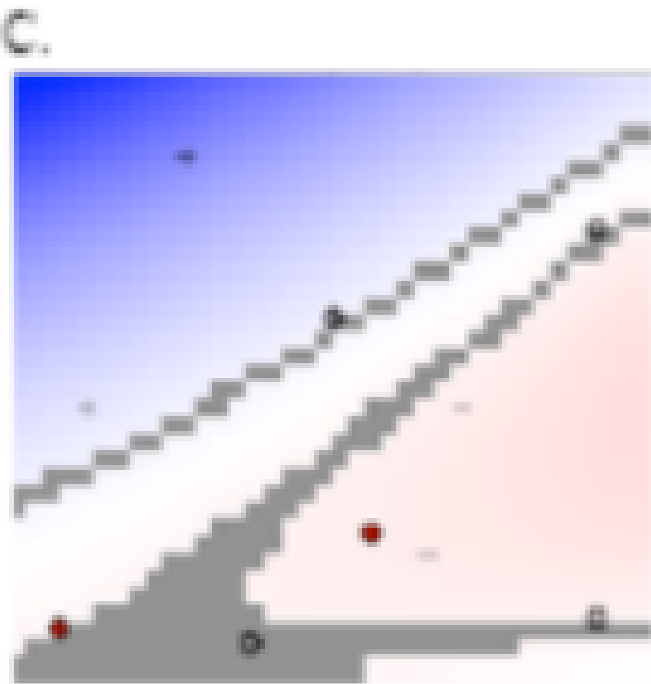


شکل رو به رو نشان دهنده ی یک دسته بند با کرنل گوسی با مقدار  $RBF=0.08$  است. می دانیم که هر چقدر مقدار  $\sigma$  کوچکتر شود از همواری مرز دسته بندی کرنل گوسی کاسته می شود و مرز تصمیم گیری پیچیده تر می شود. بنابراین با توجه به مقادیر مختلف گفته شد و پیچیدگی مرز ها در شکل ها، شکل مقابل دارای بیشترین پیچیدگی است بنابراین کمترین مقدار  $RBF$  یعنی  $0.08$  را به خود اختصاص می دهد.



شکل رو به رو نشان دهنده ی یک دسته بند با کرنل گوسی با مقدار  $RBF=0.5$  است. مرز هموار تر نسبت به شکل قبل به دلیل مقدار بزرگتر  $\sigma$  است. می دانیم هر چقدر  $\sigma$  بزرگتر شود مرز هموار تر می شود. بنابراین با توجه به مقادیر مختلف گفته شد و پیچیدگی مرز ها در شکل ها، شکل مقابل دارای بیشترین پیچیدگی  $D$  است. بنابراین مقدار  $RBF$  کمتر پس از  $0.08$  یعنی  $0.5$  را به خود اختصاص می دهد.





شکل رو به رو نشان دهنده ی یک دسته بند با کرنل گوسی با مقدار  $RBF=2$  است. همانطور که گفته شد هر چقدر مقدار  $\sigma$  بیشتر باشد مرز هموار تر می شود و از پیچیدگی آن کاسته می شود. مرز بسیار هموار تر نسبت به دو شکل قبل (D,E) به دلیل مقدار بزرگتر  $\sigma$  است. از آنجایی که مقدار  $RBF$  در این شکل بسیار بزرگ تر است (نسبت به دو شکل قبلی) مرز موجود در شکل مقابل به صورت تقریبی به یک مرز خطی تمایل پیدا کرده است.



فرض کنید  $S$  مجموعه‌ی رشته‌هایی با طول حداکثر ۱۰۰ باشند که هر حرف در هر رشته از الفبای محدود  $A$  انتخاب شده باشد. برای هر  $s \in S$  که  $s = a_1, \dots, a_{100}$  داریم  $a_j \in A$ . کرنل  $\mathcal{K} : S \times S \rightarrow \mathcal{R}$  را برای هر دو رشته در  $S$  به شکل  $\mathcal{K}(s_1, s_2)$  نشان می‌دهیم و آن را تعداد زیر رشته‌های منحصر به فرد مشترک در  $s_1$  و  $s_2$  معرفی می‌کنیم. به طور مثال اگر فرض کنیم  $A = \{a, e, i, o, u\}$  و  $s_1 = auue$  و  $s_2 = aaueee$  آنگاه خواهیم داشت:  $\mathcal{K}(s_1, s_2) = 9$  چرا که زیر رشته‌های زیر مشترک هستند:  $a, u, e, uu, ue, uue, au, auu, auue$  حالا اثبات کنید این تعریف یک تعریف معتبر برای تابع کرنل است.



با توجه به فرضیات مسئله، در ابتدا به طراحی کرنل  
مورد نیاز برای این مسئله می پردازیم :

❖ می دانیم که طبق تعریف، یک رشته (String, Substring) مجموعه ای محدود از حروف الفباست که به ترتیب در کنار  
یک دیگر قرار دارند و به مجموعه ی رشته ها با طول های مختلف را می توان صورت زیر تعریف می شود:

$$\Sigma^* = \bigcup_{n=0}^{\infty} \Sigma^n.$$

که در این رابطه  $\Sigma^n$ ، به معنای تمامی رشته ها به طول  $n$  است و  $\Sigma^*$ ، به معنای تمامی رشته ها با طول  
های مختلف است.

❖ می دانیم که رشته  $S1$  زیر رشته ی  $S2$  به شرطی که رابطه ی زیر برقرار باشد:

$$S2 = uS1v$$

که متغیرهای  $u, v$  می تواند رشته هایی با طول مختلف باشند و حتی می توانند تهی باشند.



❖ می دانیم که طبق خواسته ی مسئله می خواهیم تمامی زیر رشته های منحصر به فرد و مشترک بین دو رشته S1 و S2 را بدست آوریم، بنابراین برای بدست آوردن تابع کرنل این مسئله ابتدا باید روش مپینگ ویژگی ها به فضایی با ابعاد بیشتر را طراحی کنیم، مدل Mapping به صورت زیر خواهد بود و آن با سمبل  $\Phi$  تعریف می کنیم:

$$\Phi: s \rightarrow (\Phi_u(s))_{u \in I}$$

که در این رابطه  $I$  مجموعه ای از تمام رشته های به طول  $p$  است ( $\Sigma^p$ ) که یک فضای برداری با ابعاد  $|\Sigma|^p$  را ارائه می دهد.

❖ بنابراین بر اساس این مپینگ و تعاریف ارائه شده می توانیم دو رشته S1 و S2 را ه  $\kappa_p(s, t) = \langle \phi^p(s), \phi^p(t) \rangle = \sum_{u \in \Sigma^p} \phi_u^p(s) \phi_u^p(t)$  .  
دهیم که به صورت زیر تعریف می شود:

$$\Phi_u^p(s1) = |\{(v1, v2): s1 = v1uv2\}|, u \in \sum^p$$

$$\Phi_u^p(s2) = |\{(v1, v2): s2 = v1uv2\}|, u \in \sum^p$$

که در واقع  $p$  طول زیر رشته ها را مشخص می کند.



پس از تفکیک هر رشته به زیر رشته هایی با طول  $p$  و بردن به ویژگی ها به ابعاد بالاتر، برای بدست آوردن زیر رشته های مشترک و منحصر به فرد به طول  $p$  بین دو رشته  $S1$  و  $S2$  داریم می توانم از ضرب داخلی مپینگ دو رشته استفاده کنیم زیرا می دانیم این مسئله سبب یافتن تعداد جملات مشترک به طول  $p$  بین دو رشته می شود :

$$\sum_{u \in \Sigma^p} \Phi_u^p(s1) \Phi_u^p(s2)$$

❖ که در واقع شکل بالا در واقع همان تابع کرنل مورد نظر ماست که زیر رشته های مشترک ، و منحصر به فرد به طول  $p$  بین دو رشته ی  $S1$  و  $S2$  را پیدا می کند و به صورت زیر می توانیم باز نویسی کنیم:

$$k_p(s1, s2) = \langle \Phi_u^p(s1), \Phi_u^p(s2) \rangle = \sum_{u \in \Sigma^p} \Phi_u^p(s1) \Phi_u^p(s2)$$

اما همانطور که گفته شد، کرنل بالا برای یافتن زیر رشته های مشترک منحصر به فرد به طول  $p$  بین دو رشته ی  $S1$  و  $S2$  است. برای یافتن تمام زیر رشته های مشترک و منحصر به فرد بین  $S1$  و  $S2$  با طول های مختلف داریم:

$$k_1(s1, s2) + k_2(s1, s2) + k_3(s1, s2) + \dots + k_{\max(\text{length}(S1), \text{Length}(S2))}(s1, s2)$$

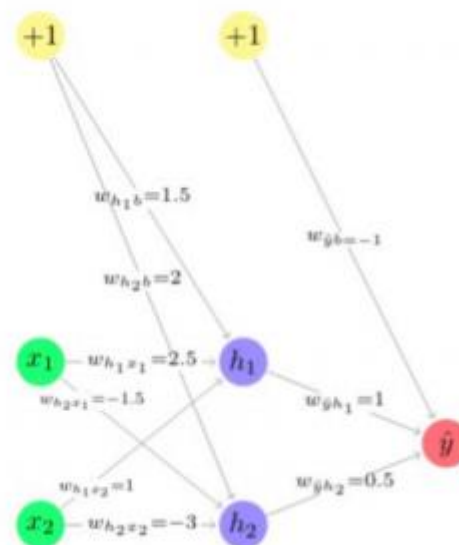
از خواص کرنل ها می دانیم که جمع کرنل ها نیز منجر به یک کرنل معتبر می شود بنابراین کرنل بالا یک کرنل معتبر است.

$$\sum_{p=1}^{\max(\text{length}(S1), \text{Length}(S2))} k_p(s1, s2) = \sum_{p=1}^{\max(\text{length}(S1), \text{Length}(S2))} \langle \Phi_u^p(s1), \Phi_u^p(s2) \rangle = \sum_{p=1}^{\max(\text{length}(S1), \text{Length}(S2))} \sum_{u \in \Sigma^p} \Phi_u^p(s1) \Phi_u^p(s2)$$





شکل زیر یک شبکه عصبی دو لایه با دو گره  $x_1$  و  $x_2$  در لایه ورودی، دو گره در لایه پنهان و یک گره در لایه خروجی را نشان می‌دهد. هر گره دارای یک ورودی بایوس با مقدار یک می‌باشد. فرض کنید از تابع سیگموئید به عنوان تابع فعالسازی در گره‌های لایه پنهان و خروجی استفاده می‌شود. سیگموئید تابعی است به فرم  $g(z) = \frac{1}{1 + e^{-z}}$  به طوری که  $z = \sum_{i=1}^n w_i x_i$  (همچنین نرخ یادگیری را برابر با ۰/۱ در نظر بگیرید)



- فرض کنید  $x_1$  و  $x_2$  به ترتیب برابر با صفر و یک باشند. خروجی مقادیر گره‌های  $h_1, h_2$  و  $\hat{y}$  را بدست آورید
- فرض کنید  $x_1, x_2$  و مقدار واقعی  $y$  به ترتیب برابر با صفر، یک و یک باشند. محاسبات مربوط به الگوریتم back-propagation را تنها برای یک گام انجام دهید. همچنین تابع لاس logistic regression را در نظر بگیرید.



- فرض کنید  $x_1$  و  $x_2$  به ترتیب برابر با صفر و یک باشند. خروجی مقادیر گره‌های  $h_1, h_2$  و  $\hat{y}$  را بدست آورید



$$h1 = X_1W_{h1x1} + X_2W_{h1x2} + bW_{h1b} = 0 * 2.5 + 1 * 1 + 1 * 1.5 = 2.5$$

$$h2 = X_1W_{h2x1} + X_2W_{h2x2} + bW_{h2b} = 0 * -1.5 + 1 * -3 + 1 * 2 = -1$$

$$out\ h1 = \frac{1}{(1 + e^{-h1})} = \frac{1}{(1 + e^{-2.5})} = 0.92414181997$$

$$out\ h2 = \frac{1}{(1 + e^{-h2})} = \frac{1}{(1 + e^1)} = 0.26894142137$$

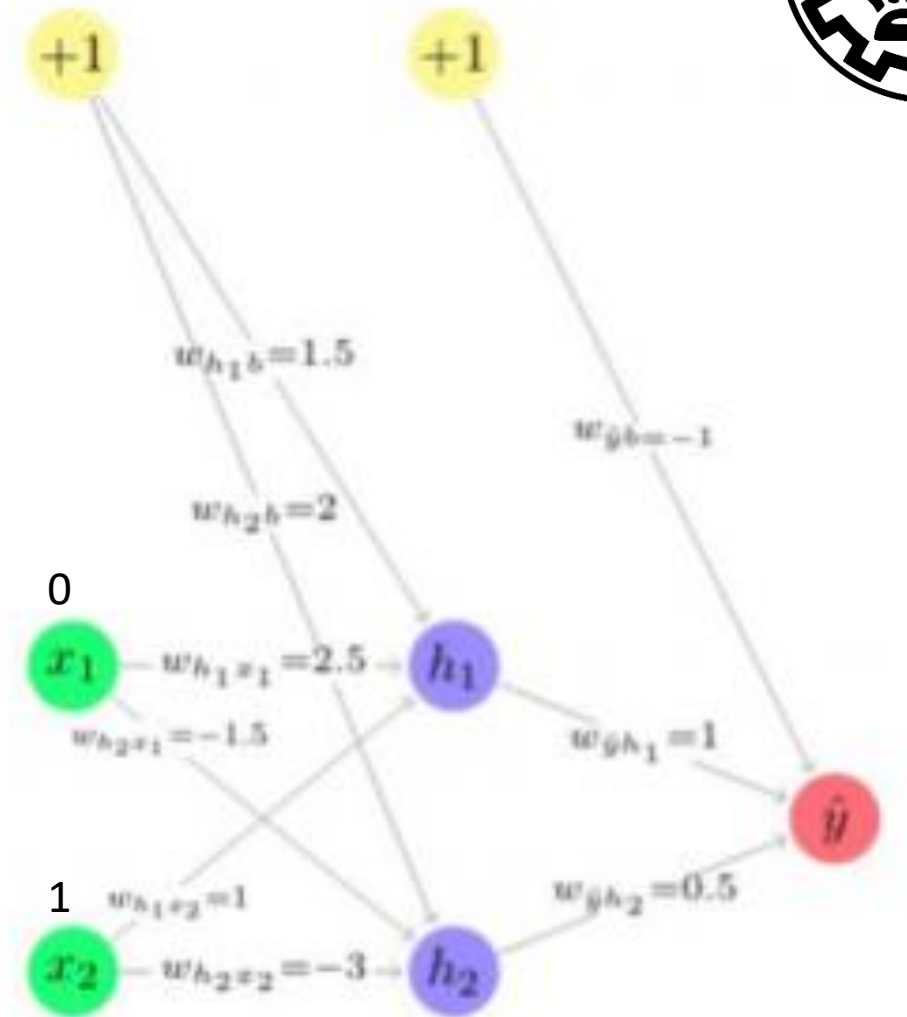
$$\begin{aligned} ypred &= out\ h1 * W_{ypredh1} + out\ h2 * W_{ypredh2} + b * W_{ypredb} \\ &= 0.92414181997 * 1 + 0.26894142137 * 0.5 + 1 * -1 = 0.05861253065 \end{aligned}$$

$$out\ ypred = \frac{1}{(1 + e^{-ypred})} = \frac{1}{(1 + e^{-0.05861253065})} = 0.51464893912$$

برای تابع لاس داریم:

$$\begin{aligned} J(\mathbf{w}) &= - \sum_{i=1}^n \log p(y^{(i)} | \mathbf{w}, \mathbf{x}^{(i)}) \\ &= \sum_{i=1}^n -y^{(i)} \log(f(\mathbf{x}^{(i)}; \mathbf{w})) - (1 - y^{(i)}) \log(1 - f(\mathbf{x}^{(i)}; \mathbf{w})) \end{aligned}$$

$$E = J(\mathbf{w}) = -y * \log(out\ ypred) - 0 = -1 * \log(0.51464893912) = 0.28848$$



## Back propagation

فرض می کنیم  $\hat{y} = y_{pred}$  ←



$$E = J(w) = -y * \log(\text{out } y_{pred}) - (1 - y) * \log(1 - \text{out } y_{pred})$$

به روز رسانی  $w_{ypred h1}, w_{ypred h2}, w_{ypred b}$  با استفاده از الگوریتم پس انتشار

$$\diamond \frac{\partial E}{\partial w_{ypred h1}} = \frac{\partial E}{\partial \text{out } y_{pred}} * \frac{\partial \text{out } y_{pred}}{\partial y_{pred}} * \frac{\partial y_{pred}}{\partial w_{ypred h1}}$$

$$\frac{\partial E}{\partial \text{out } y_{pred}} = -\frac{y}{\text{out } y_{pred}} + \frac{(1 - y)}{(1 - \text{out } y_{pred})}$$

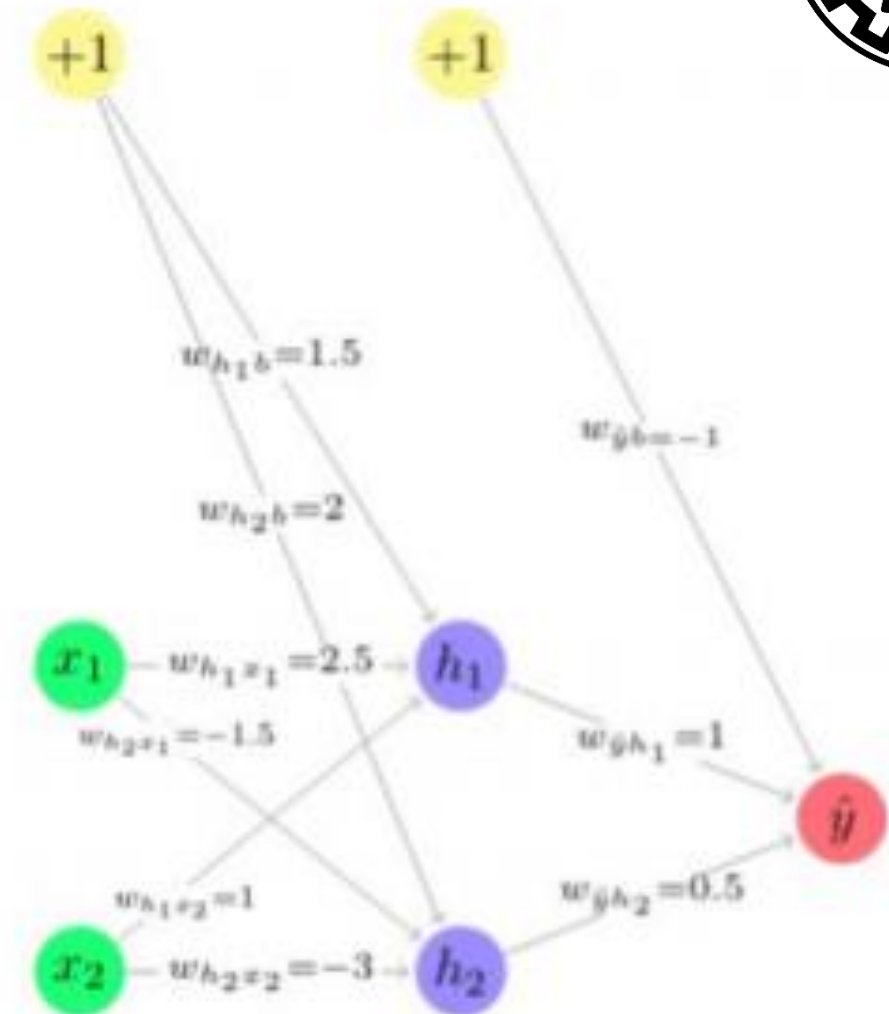
$$\frac{\partial \text{out } y_{pred}}{\partial y_{pred}} = \text{out } y_{pred} * (1 - \text{out } y_{pred})$$

$$\frac{\partial E}{\partial \text{out } y_{pred}} * \frac{\partial \text{out } y_{pred}}{\partial y_{pred}} = \text{out } y_{pred} - y_{pred}$$

$$\frac{\partial y_{pred}}{\partial w_{ypred h1}} = \text{out } h1$$

$$\frac{\partial E}{\partial w_{ypred h1}} = (\text{out } y_{pred} - y_{pred}) * \text{out } h1 =$$

$$(0.51464893912 - 1) * 0.92414181997 = -0.44853321272$$



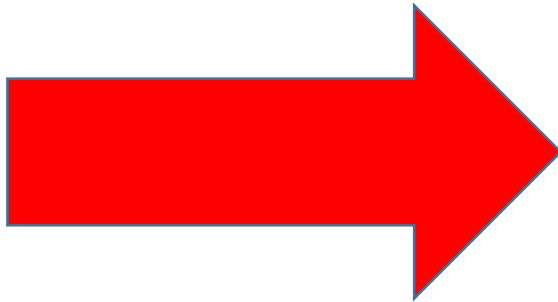


$$\diamond \frac{\partial E}{\partial w_{ypred h2}} = \frac{\partial E}{\partial out_{ypred}} * \frac{\partial out_{ypred}}{\partial ypred} * \frac{\partial ypred}{\partial w_{ypred h2}}$$

$$\frac{\partial E}{\partial w_{ypred h2}} = (out_{ypred} - ypred) * out_{h2} = (0.51464893912 - 1) * 0.26894142137 = -0.13053100417$$

$$\diamond \frac{\partial E}{\partial w_{ypred b}} = \frac{\partial E}{\partial out_{ypred}} * \frac{\partial out_{ypred}}{\partial ypred} * \frac{\partial ypred}{\partial w_{ypred b}}$$

$$\frac{\partial E}{\partial w_{ypred b}} = (out_{ypred} - ypred) * b = (0.51464893912 - 1) * 1 = -0.48535106088$$



$$\frac{\partial E}{\partial w_{ypred h1}} = -0.44853321272$$

$$\frac{\partial E}{\partial w_{ypred h2}} = -0.13053100417$$

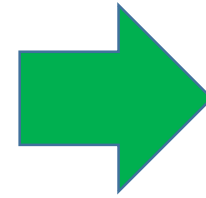
$$\frac{\partial E}{\partial w_{ypred b}} = -0.48535106088$$

$$\diamond \frac{\partial E}{\partial w_{h1x1}} = \frac{\partial E}{\partial out\ h1} * \frac{\partial out\ h1}{\partial h1} * \frac{\partial h1}{\partial w_{h1x1}}$$

$$\frac{\partial E}{\partial out\ h1} = \frac{\partial E}{\partial ypred} * \frac{\partial ypred}{\partial out\ h1}$$

$$\frac{\partial E}{\partial ypred} = \frac{\partial E}{\partial out\ ypred} * \frac{\partial out\ ypred}{\partial ypred} = out\ ypred - ypred = 0.51464893912 - 1$$

$$\frac{\partial ypred}{\partial out\ h1} = W_{ypred\ h1} = 1$$



$$\frac{\partial E}{\partial w_{h1x1}} = \frac{\partial E}{\partial out\ h1} * \frac{\partial out\ h1}{\partial h1} * \frac{\partial h1}{\partial w_{h1x1}} = 0$$

$$\frac{\partial out\ h1}{\partial h1} = out\ h1 * (1 - out\ h1) = 0.92414181997 * (1 - 0.92414181997)$$

$$\frac{\partial h1}{\partial w_{h1x1}} = X1 = 0$$

$$\diamond \frac{\partial E}{\partial w_{h1x2}} = \frac{\partial E}{\partial out\ h1} * \frac{\partial out\ h1}{\partial h1} * \frac{\partial h1}{\partial w_{h1x2}}$$

$$\frac{\partial E}{\partial out\ h1} = \frac{\partial E}{\partial ypred} * \frac{\partial ypred}{\partial out\ h1}$$

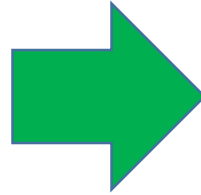
$$\frac{\partial E}{\partial ypred} = \frac{\partial E}{\partial out\ ypred} * \frac{\partial out\ ypred}{\partial ypred} = out\ ypred - ypred$$

$$= 0.51464893912 - 1 = -0.48535106088$$

$$\frac{\partial ypred}{\partial out\ h1} = W_{ypred\ h1} = 1$$

$$\frac{\partial out\ h1}{\partial h1} = out\ h1 * (1 - out\ h1) = 0.92414181997 * (1 - 0.92414181997) = 0.84828363994$$

$$\frac{\partial h1}{\partial w_{h1x2}} = X2 = 1$$



$$\frac{\partial E}{\partial w_{h1x2}} = \frac{\partial E}{\partial out\ h1} * \frac{\partial out\ h1}{\partial h1} * \frac{\partial h1}{\partial w_{h1x2}} = -0.4117153638$$

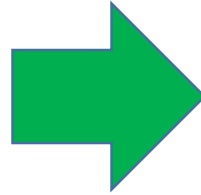
$$\diamond \frac{\partial E}{\partial w_{h1b}} = \frac{\partial E}{\partial out\ h1} * \frac{\partial out\ h1}{\partial h1} * \frac{\partial h1}{\partial w_{h1b}}$$

$$\frac{\partial E}{\partial out\ h1} = \frac{\partial E}{\partial ypred} * \frac{\partial ypred}{\partial out\ h1}$$

$$\frac{\partial E}{\partial ypred} = \frac{\partial E}{\partial out\ ypred} * \frac{\partial out\ ypred}{\partial ypred} = out\ ypred - ypred$$

$$= 0.51464893912 - 1 = -0.48535106088$$

$$\frac{\partial ypred}{\partial out\ h1} = W_{ypred\ h1} = 1$$



$$\frac{\partial E}{\partial w_{h1b}} = \frac{\partial E}{\partial out\ h1} * \frac{\partial out\ h1}{\partial h1} * \frac{\partial h1}{\partial w_{h1b}} = -0.4117153638$$

$$\frac{\partial out\ h1}{\partial h1} = out\ h1 * (1 - out\ h1) = 0.92414181997 * (1 - 0.92414181997) = 0.84828363994$$

$$\frac{\partial h1}{\partial w_{h1b}} = b = 1$$



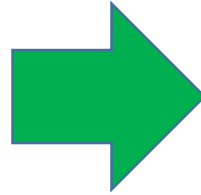
$$\diamond \frac{\partial E}{\partial w_{h2x1}} = \frac{\partial E}{\partial out\ h2} * \frac{\partial out\ h2}{\partial h2} * \frac{\partial h1}{\partial w_{h2x1}}$$

$$\frac{\partial E}{\partial out\ h2} = \frac{\partial E}{\partial ypred} * \frac{\partial ypred}{\partial out\ h2}$$

$$\frac{\partial E}{\partial ypred} = \frac{\partial E}{\partial out\ ypred} * \frac{\partial out\ ypred}{\partial ypred} = out\ ypred - ypred$$

$$= 0.51464893912 - 1 = -0.48535106088$$

$$\frac{\partial ypred}{\partial out\ h2} = W_{ypred h2} = 0.5$$



$$\frac{\partial E}{\partial w_{h2x1}} = \frac{\partial E}{\partial out\ h2} * \frac{\partial out\ h2}{\partial h2} * \frac{\partial h2}{\partial w_{h2x1}} = 0$$

$$\frac{\partial out\ h2}{\partial h2} = out\ h2 * (1 - out\ h2) = 0.26894142137 * (1 - 0.26894142137)$$

$$\frac{\partial h2}{\partial w_{h2x1}} = X1 = 0$$

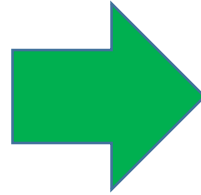
$$\diamond \frac{\partial E}{\partial w_{h2x2}} = \frac{\partial E}{\partial out\ h2} * \frac{\partial out\ h2}{\partial h2} * \frac{\partial h1}{\partial w_{h2x2}}$$

$$\frac{\partial E}{\partial out\ h2} = \frac{\partial E}{\partial ypred} * \frac{\partial ypred}{\partial out\ h2}$$

$$\frac{\partial E}{\partial ypred} = \frac{\partial E}{\partial out\ ypred} * \frac{\partial out\ ypred}{\partial ypred} = out\ ypred - ypred$$

$$= 0.51464893912 - 1 = -0.48535106088$$

$$\frac{\partial ypred}{\partial out\ h2} = W_{ypred h2} = 0.5$$



$$\frac{\partial E}{\partial w_{h2x2}} = \frac{\partial E}{\partial out\ h2} * \frac{\partial out\ h2}{\partial h2} * \frac{\partial h2}{\partial w_{h2x2}} = -0.047712904$$

$$\frac{\partial out\ h2}{\partial h2} = out\ h2 * (1 - out\ h2) = 0.26894142137 * (1 - 0.26894142137)$$

$$\frac{\partial h2}{\partial w_{h2x2}} = X2 = 1$$

$$\diamond \frac{\partial E}{\partial w_{h2b}} = \frac{\partial E}{\partial out\ h2} * \frac{\partial out\ h2}{\partial h2} * \frac{\partial h1}{\partial w_{h2b}}$$

$$\frac{\partial E}{\partial out\ h2} = \frac{\partial E}{\partial ypred} * \frac{\partial ypred}{\partial out\ h2}$$

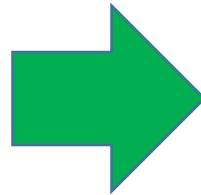
$$\frac{\partial E}{\partial ypred} = \frac{\partial E}{\partial out\ ypred} * \frac{\partial out\ ypred}{\partial ypred} = out\ ypred - ypred$$

$$= 0.51464893912 - 1 = -0.48535106088$$

$$\frac{\partial ypred}{\partial out\ h2} = W_{ypred\ h2} = 0.5$$

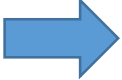
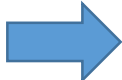




$$\frac{\partial out\ h2}{\partial h2} = out\ h2 * (1 - out\ h2) = 0.26894142137 * (1 - 0.26894142137)$$

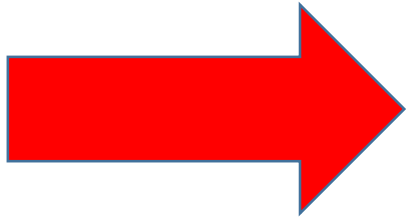
$$\frac{\partial h2}{\partial w_{h2b}} = b = 1$$



$$\frac{\partial E}{\partial w_{h2b}} = \frac{\partial E}{\partial out\ h2} * \frac{\partial out\ h2}{\partial h2} * \frac{\partial h2}{\partial w_{h2b}} = -0.047712904$$

## Update Weights

- $\frac{\partial E}{\partial w_{ypred\ h1}} = -0.44853321272$    $w_{ypred\ h1} = w_{ypred\ h1} - 0.1 * \frac{\partial E}{\partial w_{ypred\ h1}}$  
- $\frac{\partial E}{\partial w_{ypred\ h2}} = -0.13053100417$    $w_{ypred\ h2} = w_{ypred\ h2} - 0.1 * \frac{\partial E}{\partial w_{ypred\ h2}}$  
- $\frac{\partial E}{\partial w_{ypred\ b}} = -0.48535106088$    $w_{ypred\ b} = w_{ypred\ b} - 0.1 * \frac{\partial E}{\partial w_{ypred\ b}}$  



$$w_{ypred\ h1} = 1 - 0.1 * -0.448533 = 1.044853321$$

$$w_{ypred\ h2} = 0.5 - 0.1 * -0.130531 = 0.5130531$$

$$w_{ypred\ b} = -1 - 0.1 * -0.4853510 = -0.9514649$$

## Update Weights

$$\frac{\partial E}{\partial w_{h1x1}} = \frac{\partial E}{\partial out\ h1} * \frac{\partial out\ h1}{\partial h1} * \frac{\partial h1}{\partial w_{h1x1}} = 0$$



$$w_{h1x1} = 2.5 - 0.1 * 0 = 2.5$$

$$\frac{\partial E}{\partial w_{h1x2}} = \frac{\partial E}{\partial out\ h1} * \frac{\partial out\ h1}{\partial h1} * \frac{\partial h1}{\partial w_{h1x2}} = -4117153638$$



$$w_{h1x2} = 1 - 0.1 * -0.4117153638 = 1.041171$$

$$\frac{\partial E}{\partial w_{h1b}} = \frac{\partial E}{\partial out\ h1} * \frac{\partial out\ h1}{\partial h1} * \frac{\partial h1}{\partial w_{h1b}} = -4117153638$$



$$w_{h1b} = 1.5 - 0.1 * -0.4117153638 = 1.541171$$

$$\frac{\partial E}{\partial w_{h2x1}} = \frac{\partial E}{\partial out\ h2} * \frac{\partial out\ h2}{\partial h2} * \frac{\partial h2}{\partial w_{h2x1}} = 0$$



$$w_{h2x1} = -1.5 - 0.1 * 0 = -1.5$$

$$\frac{\partial E}{\partial w_{h2x2}} = \frac{\partial E}{\partial out\ h2} * \frac{\partial out\ h2}{\partial h2} * \frac{\partial h2}{\partial w_{h2x2}} = -0.047712904$$



$$w_{h2x2} = -3 - 0.1 * -0.047712904 = -2.99522$$

$$\frac{\partial E}{\partial w_{h2b}} = \frac{\partial E}{\partial out\ h2} * \frac{\partial out\ h2}{\partial h2} * \frac{\partial h2}{\partial w_{h2b}} = -0.047712904$$



$$w_{h2b} = 2 - 0.1 * -0.047712904 = 1.9952286$$