به نام خدا

تمرین سری 4 درس یادگیری ماشین

فردین عباسی 810199456

دانشکده مهندسی برق و کامپیوتر

دانشگاه تهران

بهار 1402

فهرست

سوال اول 3

1. 3

2. 3

3. 4

سوال دوم 5

1. 5

2. 5

3. 5

4. 5

5. 6

سوال سوم 7

سوال چهارم 8

1. 8

2. 9

3. 10

4. 11

5. 13

6. 13

سوال پنجم 14

1. 14

2. 14

3. 14

سوال ششم 15

1. 15

2. 15

3. 15

4. 15

سوال هفتم 16

1. 16

2. 18

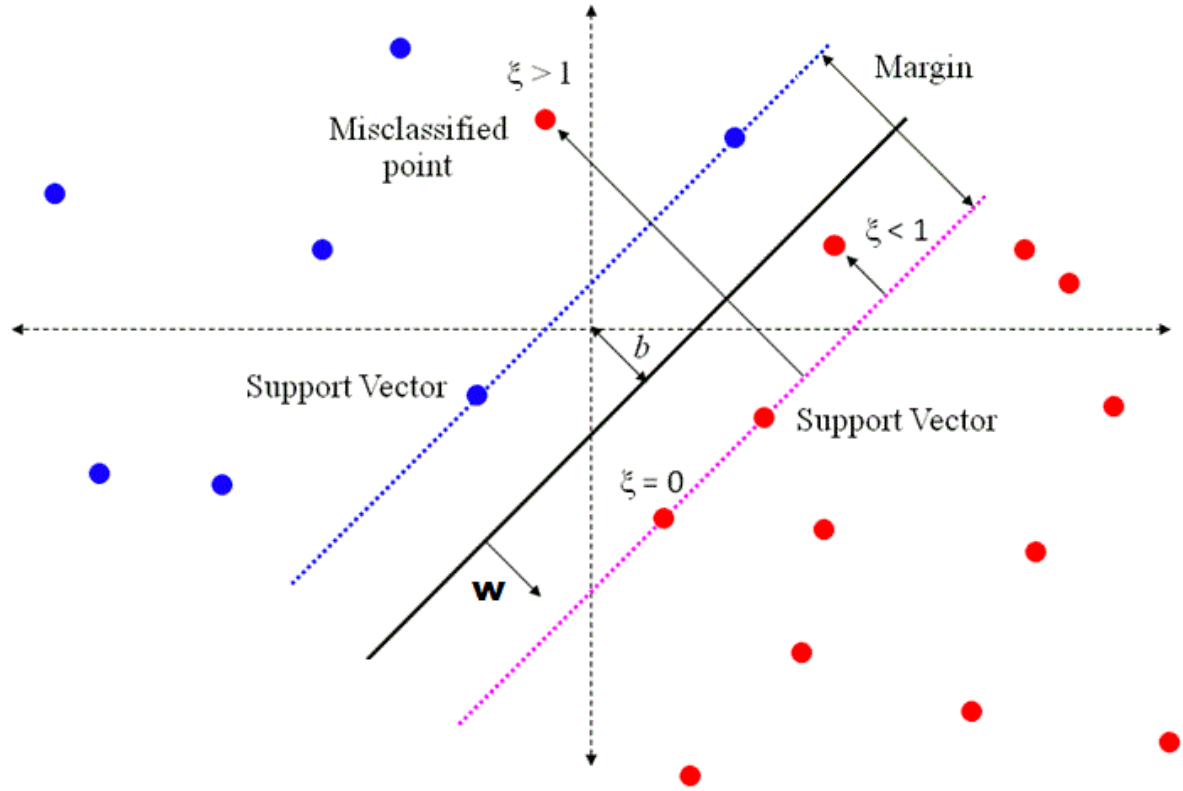
سوال اول

1.

اگر Hyperplane جدا کننده را به صورت $\omega^T x + b = 0$ در نظر بگیریم، جهت و مکان این Hyperplane بر اساس بردار ω تعریف می شود که عمود بر آن است. همچنین b این Hyperplane را در فضای مختصاتی جابه‌جا می‌کند.

2.

در SVM به طور کلی فرض بر این است که داده ها جدایی پذیر خطی هستند در نتیجه در صورت وجود نویز ممکن است Hard SVM کارا نباشد اماSoft SVMبه دلیل در نظر گرفتن خطا امکان جدا کردن داده ها را دارد.



$$Soft\ SVM: \min \frac{1}{2}||\omega||_2^2 + c\sum_{i=1}^{n} e_i$$

$$s.t.\, y_i(\omega^T x_i + b) \geq 1 - e_i\,, e_i \geq 0$$

که e_i اجازه می دهد نقاط نویزی داخل Margin و یا حتی اشتباه طبقه بندی شوند.

3.

در SVMممکن است داده ها در فضای اصلی جدایی پذیر خطی نباشند اما در feature space با استفاده از تابع نگاشت ϕ جدایی پذیر خطی باشند اما یافتن نگاشت $\phi(x_i)$ و سپس محاسبه ضرب داخلی آن هزینه محاسباتی بالایی دارد که برای حل این مشکل از Kernel Trick استفاده می کنیم.

$$K(x,y) = \phi(x)^T \phi(y)$$

تابع K(x,y)در اصل یک معیار شباهت در feature space است بدون آن که مجبور باشیم نگاشت را انجام دهیم.

$$d = 2: \phi_1(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}, \phi_2(x) = \begin{bmatrix} 2x \\ 2x^2 \end{bmatrix}$$

در $\phi_2(x)$ به دلیل اینکه در فضای feature space جدید فاصله بین sample ها بیشتر از $\phi_1(x)$ است، حاشیه بزرگتری ایجاد می کند.

سوال دوم

1.

اگر بتوانیم ثابت کنیم که این کرنل حاصل از ضرب داخلی دو نگاشت است، آنگاه یک کرنل معتبراست:

$$K(x,y) = \langle \phi(x), \phi(y) \rangle$$

$$K(x,y) = \exp\left(-\frac{||x-y||_2^2}{2\sigma^2}\right) = \exp\left(-\frac{||x||_2^2}{2\sigma^2}\right)\exp\left(-\frac{||y||_2^2}{2\sigma^2}\right)\exp\left(\frac{x^T y}{\sigma^2}\right)$$

$$\phi(x) = \exp\left(-\frac{||x||_2^2}{2\sigma^2}\right)\exp\left(\frac{x}{\sigma^2}\right) \rightarrow K(x,y) = \langle \phi(x), \phi(y) \rangle \; Valid$$

2.

کرنل ها مثبت معین هستند در نتیجه:

$$K(x,y) = aK_1(x,y) + bK_2(x,y)$$

$$x^T K x = x^T(aK_1 + bK_2)x \xleftrightarrow{x^T K x > 0} x^T(aK_1 + bK_2)x \geq 0 \xleftrightarrow{x^T K_{1,2} x > 0} a, b > 0$$

کرنل K تنها در صورتی معتبر است که a,b مثبت باشد و به ازای دیگر مقادیر این گزاره صادق نیست.

3.

$$K(x,y) = f(x) \times f(y)$$

$$a^T K a = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, y_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j f(x_i) f(x_j) = \left(\sum_{i=1}^{n} a_i f(x_i)\right)^2 \geq 0 \overset{so}{\Rightarrow} K \; is \; Valid$$

4.

در ابتدا فرض می کنیم$\acute{K}$یک کرنل معتبر است و اگر بتوانیم ثابت کنیم Kنیز معتبر است آنگاه حکم ثابت است.

$$\acute{K}\big(g(x), g(y)\big) = K(x,y)$$

$$a^T K a = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, y_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \acute{K}\left(g(x_i), g(y_j)\right) \xleftrightarrow{b_i = a_i\sqrt{\acute{K}\left(g(x_i), g(y_j)\right)}}$$

$$a^T K a = \sum_{i=1}^{n}\sum_{j=1}^{n} b_i b_j \frac{\acute{K}\left(g(x_i), g(y_j)\right)}{\sqrt{\acute{K}\left(g(x_i), g(y_j)\right)}\sqrt{\acute{K}\left(g(x_i), g(y_j)\right)}} \xleftrightarrow{Quadratic \; Form}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} b_i b_j \acute{K}\left(\frac{g(x_i)}{\sqrt{\acute{K}\left(g(x_i), g(y_j)\right)}}, \frac{g(y_j)}{\sqrt{\acute{K}\left(g(x_i), g(y_j)\right)}}\right) \xleftrightarrow{\acute{K} \; is \; Valid}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} b_i b_j \langle \phi(g(x_i)), \phi\left(g(y_i)\right) \rangle \geq 0 \rightarrow a^T K a \geq 0$$

طبق روابط بازگشتی شرط برقرار و حکم ثابت است.

5.

$$K(x,y) = K_1(x,y) \times K_2(x,y)$$

$$K_1(x,y) = \sum_i \phi_i(x)\phi_i(y)\ , K_2(x,y) = \sum_i \acute{\phi}_i(x)\acute{\phi}_i(y)$$

$$K(x,y) = \left(\sum_i \phi_i(x)\phi_i(y)\right) \times \left(\sum_j \acute{\phi}_J(x)\acute{\phi}_J(y)\right) = \sum_{i,j} \phi_i(x)\acute{\phi}_J(x)\,\phi_i(y)\acute{\phi}_J(y) \xrightarrow{\phi_k(z)=\phi_i(z)\acute{\phi}_J(z)}$$

$$K(x,y) = \sum_k \phi_k(x)\phi_k(y) = \langle \phi_k(x), \phi_k(y) \rangle$$

سوال سوم

$$\min_{w,b} \frac{1}{2}||\omega||_2^2 + \frac{C}{2}\sum_{i=1}^{n} \epsilon_i^2$$

این رابطه تابع هزینه روش Soft Margin SVM است که در آن بر اساس رابطه زیر توانایی طبقه بندی کلاس ها را دارد:

$$\begin{cases} \omega^T x_i + b \geq 1 - \epsilon_i & y_i = 1 \\ \omega^T x_i + b \leq -1 + \epsilon_i & y_i = -1 \end{cases}$$

در Soft Margin SVM حتی اگر داده ها جدایی پذیر خطی نباشند با قبول کردن پارامتر ϵ_i به عنوان خطا امکان طبقه بندی کلاس ها را می دهد.

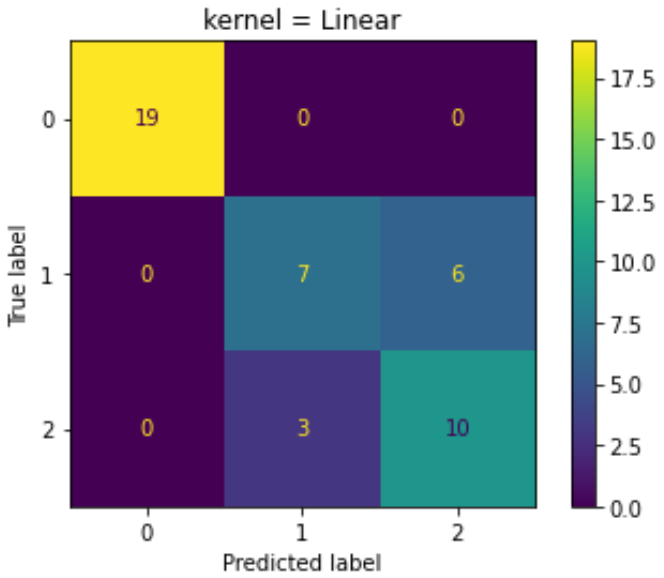
سوال چهارم

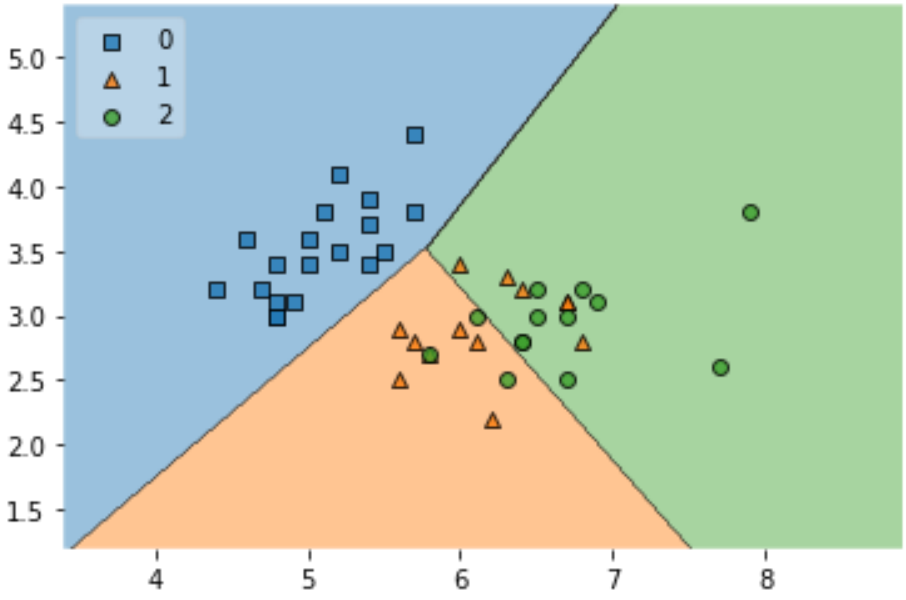
در ابتدا برای آشنایی بیشتر با دیتاست مشهورIris، این را مطالعه نمایید.

1.

نتایج طبقه بندی بر اساس فیچر های Sepal(کاسبرگ) بر اساس کرنل خطی به شرح زیر می باشد:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	19
1	0.70	0.54	0.61	13
2	0.62	0.77	0.69	13
accuracy			0.80	45
macro avg	0.78	0.77	0.77	45
weighted avg	0.81	0.80	0.80	45

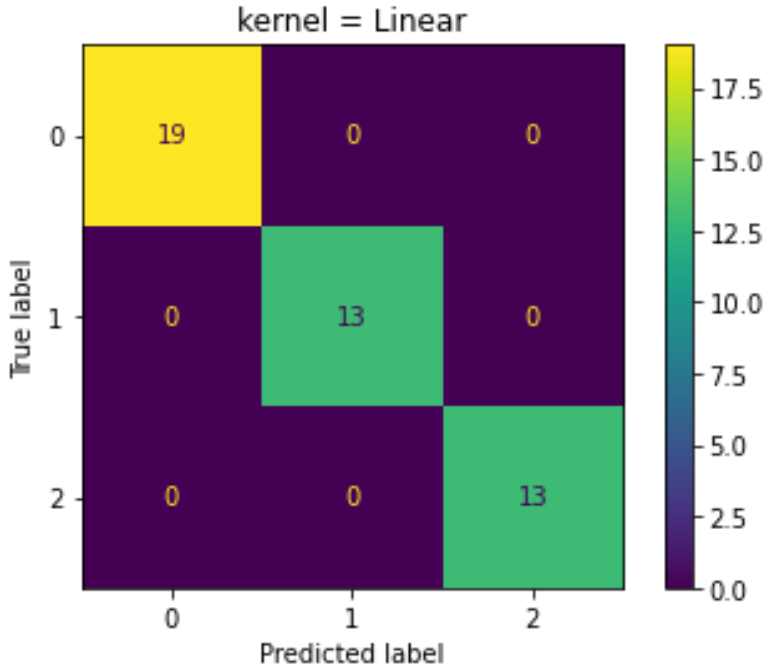


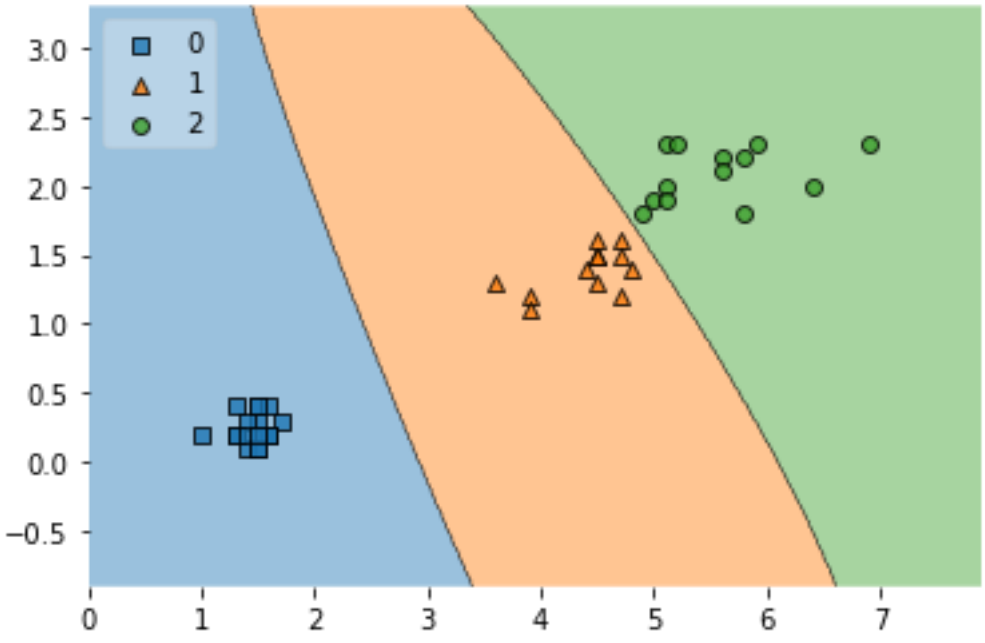


2.

نتایج طبقه بندی بر اساس فیچر های Petal(گلبرگ) بر اساس کرنل خطی به شرح زیر می باشد:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	19
1	1.00	1.00	1.00	13
2	1.00	1.00	1.00	13
accuracy			1.00	45
macro avg	1.00	1.00	1.00	45
weighted avg	1.00	1.00	1.00	45





همانطور که مشاهده می شود ویژگی گلبرگ به خوبی فضای خطی بین فیچر ها را جدا می کند و عملکرد ایده آلی دارد.

3.

$$Polynomial\ Kernel: K(x,y) = (x^T y + c)^d$$

این کرنل حالت کلی تری نسبت به کرنل خطی است اما به دلیل این که دقت و کارایی کمتری نسبت به دیگر کرنل ها دارد کم کاربرد تر است.

$$Radial\ Basis\ Function\ Kernel: K(x,y) = \exp\left(-\frac{||x-y||_2^2}{2\sigma^2}\right)$$

کرنل گوسی یکی از محبوب ترین کرنل ها است زیرا به دانش قبلی خاصی نیاز ندارد. این کرنل شباهت بین داده ها در ***feature space*** نامتناهی بررسی می کند و برای داده های غیرخطی عملکرد مناسبی دارد.

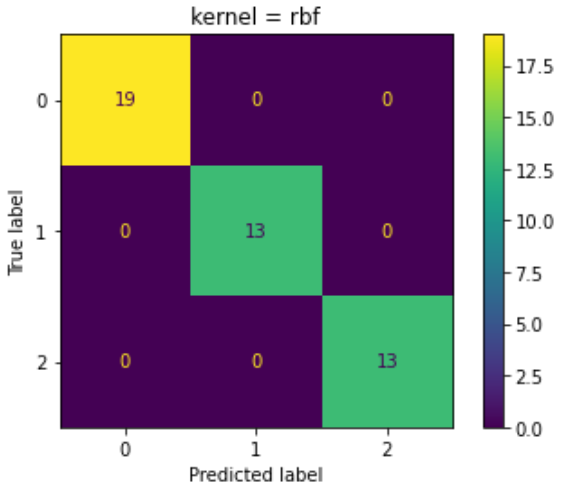
$$Linear\ Kernel: K(x,y) = xy$$

کرنل های خطی یکی از ساده ترین انواع کرنل ها است که ثابت می شود در فضای های با بعد بالا بهترین کرنل است و همچنین نسبت به دیگر کرنل ها سریع تر است. (مطالعه بیشتر)

نتایج کرنل خطی در بخش قبل بررسی شد و نتایج کرنل های چندجمله ای و گوسی به شرح زیر است:

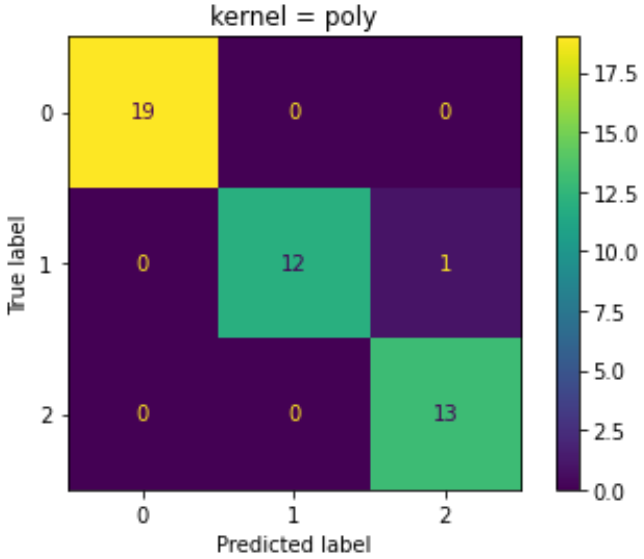
RBF Kernel:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	19
1	1.00	1.00	1.00	13
2	1.00	1.00	1.00	13
accuracy			1.00	45
macro avg	1.00	1.00	1.00	45
weighted avg	1.00	1.00	1.00	45



Polynomial Kernel:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	19
1	1.00	0.92	0.96	13
2	0.93	1.00	0.96	13
accuracy			0.98	45
macro avg	0.98	0.97	0.97	45
weighted avg	0.98	0.98	0.98	45



همانگونه که ملاحظه می شود در صورت استفاده از همه فیچر ها هر دو کرنل عملکرد خیلی خوبی دارند اما همانطور که پیش تر اشاره شد کرنل چندجمله ای عملکرد ضعیف تری دارد.

4.

پارامتر gamma، ضرایب مولفه های کرنل های RBF,Poly,sigmoid است که یا به صورت دستی تعین می شود و یا از یکی از دو حالت زیر پیروی می کند:

$$\begin{cases} auto: gamma = \dfrac{1}{n_{features}} \\ scale: gamma = \dfrac{1}{n_{features} \times Var_X} \end{cases}$$

در رابطه زیر پارامتر C نقش Regularization را ایفا می کند:

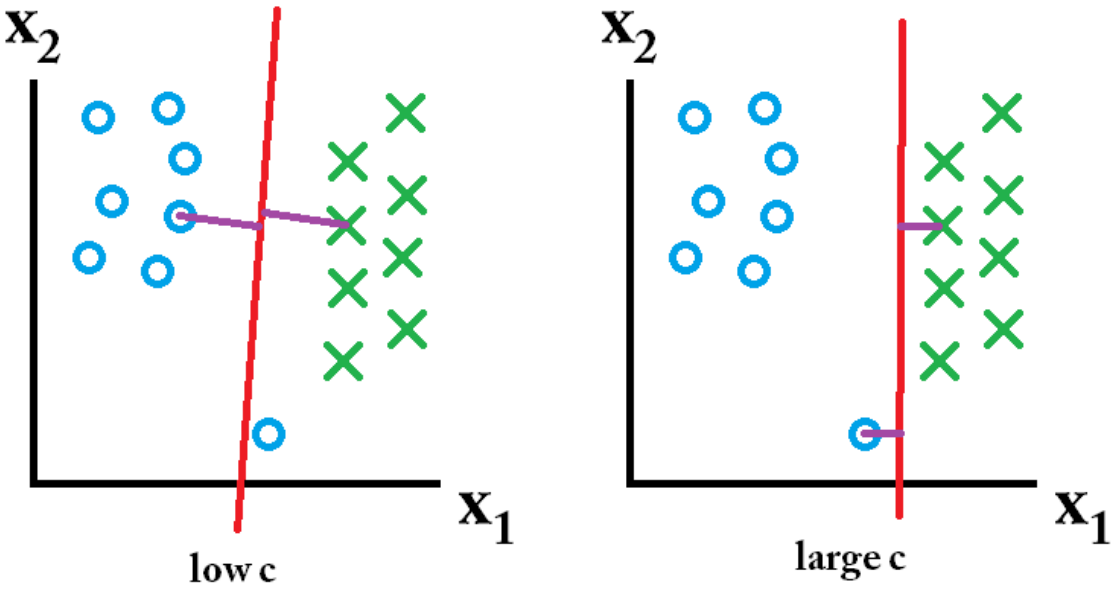
$$\min \frac{1}{2}||\omega||_2^2 + \boldsymbol{C}\sum_{i=1}^{N}\epsilon^2$$

در مسئله SVM دو هدف را دنبال می کنیم:

1- طبقه بندی کلاس ها با کمترین خطا

2- انتخاب Hyperplane با بیشترین Margin

اما همواره دست‌یابی به هر دو ممکن نیست و پارامترC در رابطه نقش Trade-Off میان این دو را بازی می کند که باید با توجه به دیتاست انتخاب شود. C بزرگتر به این معنی که دقت طبقه بند برای ما اهمیت بیشتری دارد و C کوچک تر یعنی حاشیه بزرگ تر برای ما اهمیت بیشتری دارد. (مطالعه بیشتر)



به صورت کلی ما صرفا باید C بهینه را انتخاب کنیم ولی در صورتی کرنل گوسی باشد انتخاب gamma نیز اهمیت پیدا می کند زیرا اگر gamma خیلی بزرگ باشد اثر C خیلی کم می شود و در صورتی که gamma خیلی کوچک باشد، تاثیر C همچون تاثیر بر کرنل خطی است.(مطالعه بیشتر)

5.

```
Optimal hyperparameters for Kernel = linear:{'C': 1, 'gamma': 0.1}
Accuracy on test set for Kernel = linear:1.0

Optimal hyperparameters for Kernel = poly:{'C': 0.1, 'gamma': 0.1}
Accuracy on test set for Kernel = poly:1.0

Optimal hyperparameters for Kernel = rbf:{'C': 100, 'gamma': 0.01}
Accuracy on test set for Kernel = rbf:1.0
```

مقادیر Hyperparameter بهینه برای طبقه بندی با استفاده از همه فیچر ها به شرح بالا است.

همانگونه که مشاهده می شود این مسئله اهمیت انتخاب Hyperparameter ها را نشان می دهد که بسته به نوع کرنل باید انتخاب شود.

6.

$$\begin{cases} OVR: One\ Vs.\ Rest \\ OVO: One\ Vs.\ One \end{cases}$$

نتایج در کد موجود می باشد، با توجه به اینکه از HyperParameter های بهینه استفاده شد در هر دو حالت عملکرد 100% درست بود.

سوال پنجم

1.

$$K(x,y) = \exp\left(-\frac{||x-y||_2^2}{2}\right) = \theta(x) \times \theta(y)$$

$$\begin{cases} ||\theta(x)-\theta(y)||_2^2 = ||\theta(x)||_2^2 + ||\theta(y)||_2^2 - 2\theta(x)\theta(y) \\ ||\theta(x)||_2^2 = \langle\theta(x),\theta(x)\rangle = K(x,x) = \exp\left(-\frac{0}{2}\right) = 1 \end{cases} \overset{so}{\Rightarrow} ||\theta(x)-\theta(y)||_2^2 = 2\left(1 - e^{-||x-y||_2^2}\right) < 2$$

2.

$K(x,y) = \langle\phi(x),\phi(y)\rangle = (x^T y + 1)^2$ (مطالعه بیشتر)

$$K(x,y) = \left(\sum_{i=1}^{n} x_i y_i + 1\right)^2 = \sum_{i=1}^{n}(x_i^2)(y_i^2) + \sum_{i=2}^{n}\sum_{j=1}^{i-1}(\sqrt{2}x_i x_j)(\sqrt{2}y_i y_j) + \sum_{i=1}^{n}(\sqrt{2}x_i)(\sqrt{2}y_i) + 1$$

$$\xrightarrow{Quadratic\ Form} \phi(x) = \langle x_n^2, \dots, x_1^2, \sqrt{2}x_n x_{n-1}, \sqrt{2}x_n x_1, \sqrt{2}x_{n-1}x_{n-2}, \dots, \sqrt{2}x_{n-1}x_1, \dots, \sqrt{2}x_2 x_1, \sqrt{2}x_n, \dots, \sqrt{2}x_1, 1\rangle$$

$$Dimension_\phi = \binom{n+2}{2} = \frac{(n+2)(n+1)}{2}$$

3.

$Positive\ \&\ Negative\ Hyperplane: \omega^T x + b = \pm 1$

با توجه به اینکه دو مولفه نامعلوم ω, b را باید به دست آورد، در *Hard Linear SVM* به **دو** داده نیاز خواهیم داشت.

در صورتی که با اضافه شدن داده جدید هنوز فضا خطی جدا پذیر باشد باز هم به دو داده نیاز خواهیم داشت زیرا هنوز صرفا باید دو مقدار جدید برای ω, b مشخص کنیم.

سوال ششم

1.

$$N = 5\,, x = \frac{N+1}{2} = 3\,, p = 51\%\,, Percision_{total} = \sum_{i=x}^{N} \binom{N}{i} p^i (1-p)^{N-i}$$

$$Percision_{total} = \sum_{i=3}^{5} \binom{5}{i} 0.51^i (1-0.51)^{5-i} \approx 0.518$$

2.

$$N = 9\,, x = \frac{N+1}{2} = 5\,, p = 51\%\,, Percision_{total} = \sum_{i=x}^{N} \binom{N}{i} p^i (1-p)^{N-i}$$

$$Percision_{total} = \sum_{i=5}^{9} \binom{9}{i} 0.51^i (1-0.51)^{9-i} \approx 0.524$$

3.

$$N \to \infty\,, x \to \infty\,, p = 51\%\,, Percision_{total} = \sum_{i=x}^{N} \binom{N}{i} p^i (1-p)^{N-i} \to 1$$

به ازای Nهای خیلی بزرگ به این دقت میل می کند و به عبارتی سرعت همگرایی خیلی پایین است.

همچنین با توجه به محدودیت های سخت افزاری کنونی آموزش این تعداد طبقه بند در عمل ممکن نیست!

4.

$$N = 5\,, x = \frac{N+1}{2} = 3\,, p = 50\%\,, Percision_{total} = \sum_{i=x}^{N} \binom{N}{i} p^i (1-p)^{N-i}$$

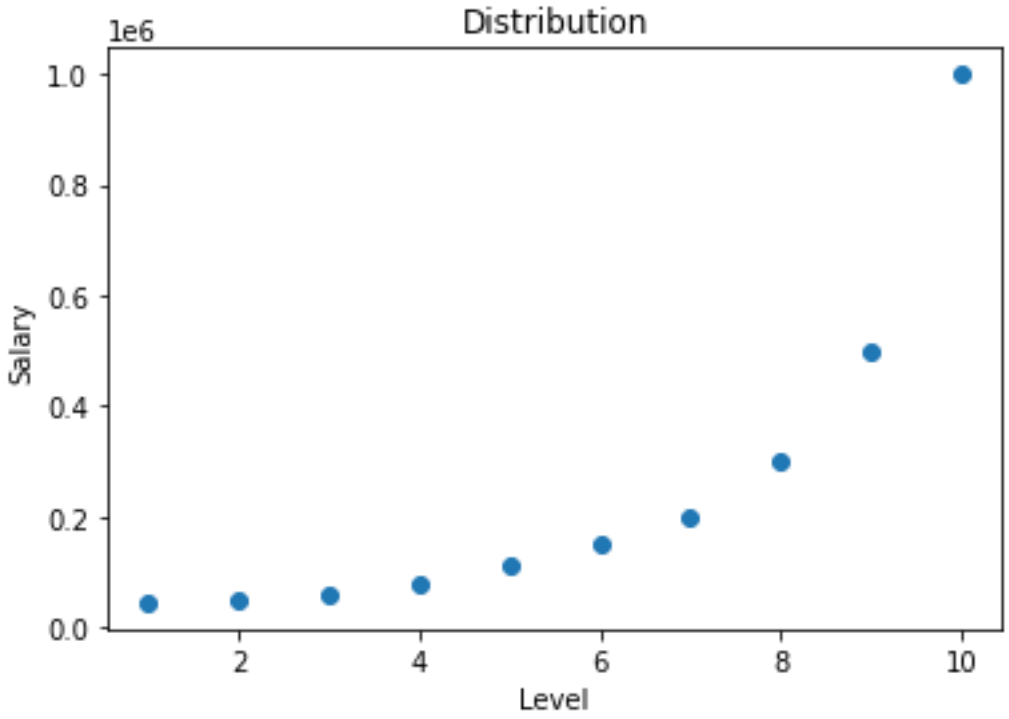
$$Percision_{total} = \sum_{i=3}^{5} \binom{5}{i} 0.5^i (1-0.5)^{5-i} = 0.5$$

به طور کلی در Ensemble Learning، از کنار هم قرار گرفتن Weak Learner ها دقت بالاتری نسبت به عملکرد انفرادی آنها به دست می آوریم، اما باید به این موضوع توجه کرد که اگر عملا طبقه بند ها به صورت شانسی طبقه بندی کنند، از کنار هم قرار گرفتن آنها نتیجه ای حاصل نمی شود!

سوال هفتم

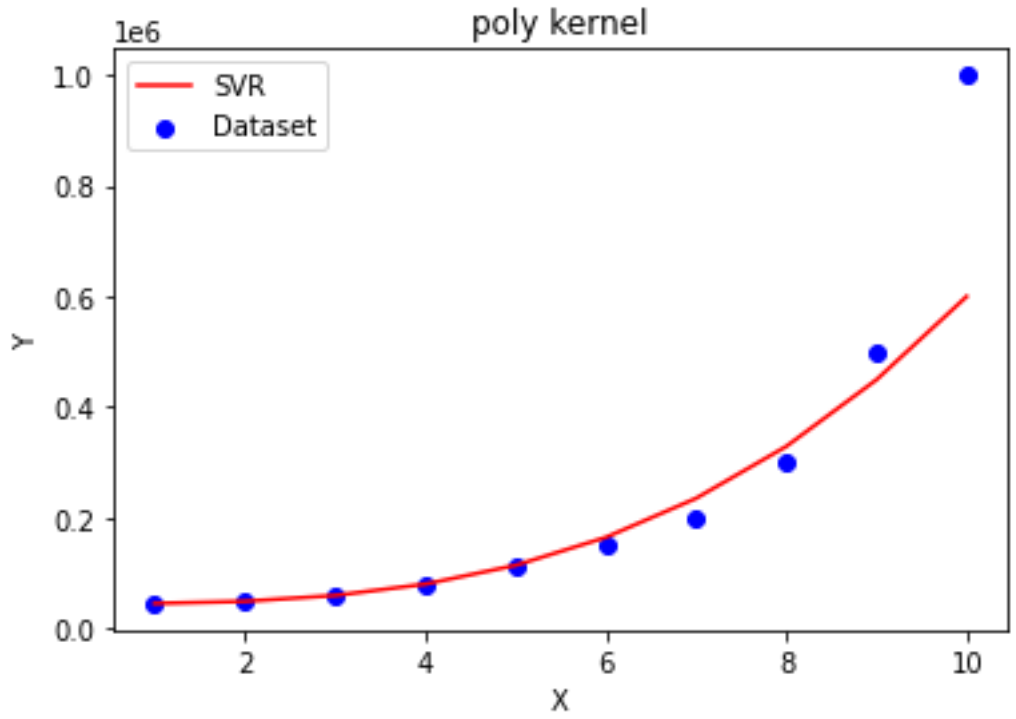
1.

در ابتدا نحوه توزیع حقوق بر اساس سطح موقعیت شغلی را مطابق تصویر زیر بررسی می کنیم:

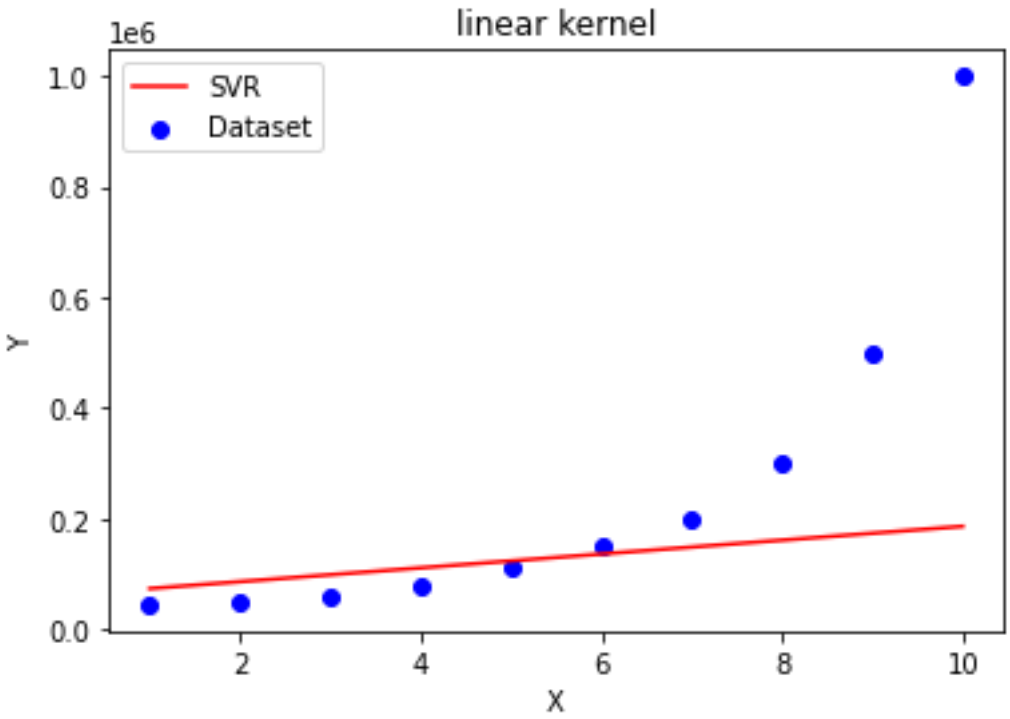


نتایج استفاده از SVR به ازای کرنل های مختلف به شرح زیر است:

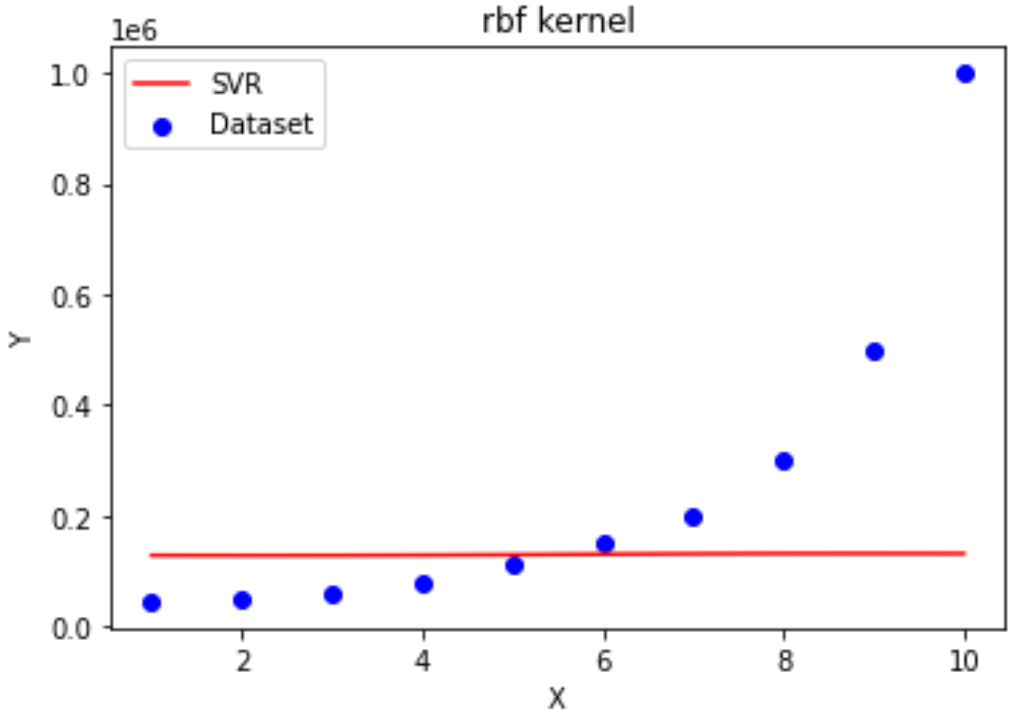
Kernel = Poly:



Kernel = Linear:



Kernel = RBF:



2.

در این بخش باید داده های Categorical با روش های مختلف مانند One hot Encoding و یا Label Encoding تغیر یابد و همچنین فیچر های عددی نرمالایز شود.

```
print('MAE=',mean_absolute_error(y_test, y_pred))

results_df = pd.DataFrame()
results_df['Real'] = y_test
results_df['Predicted'] = y_pred
results_df['Difference'] = y_pred-y_test

# Save the results dataframe to a new csv file
results_df.to_csv('results.csv', index=False)
```

```
MAE= 37.17776174455382
```