
بسم الله الرحمن الرحيم

عنوان

تمرین سوم درس یادگیری ماشین (SVM)

دانشجو

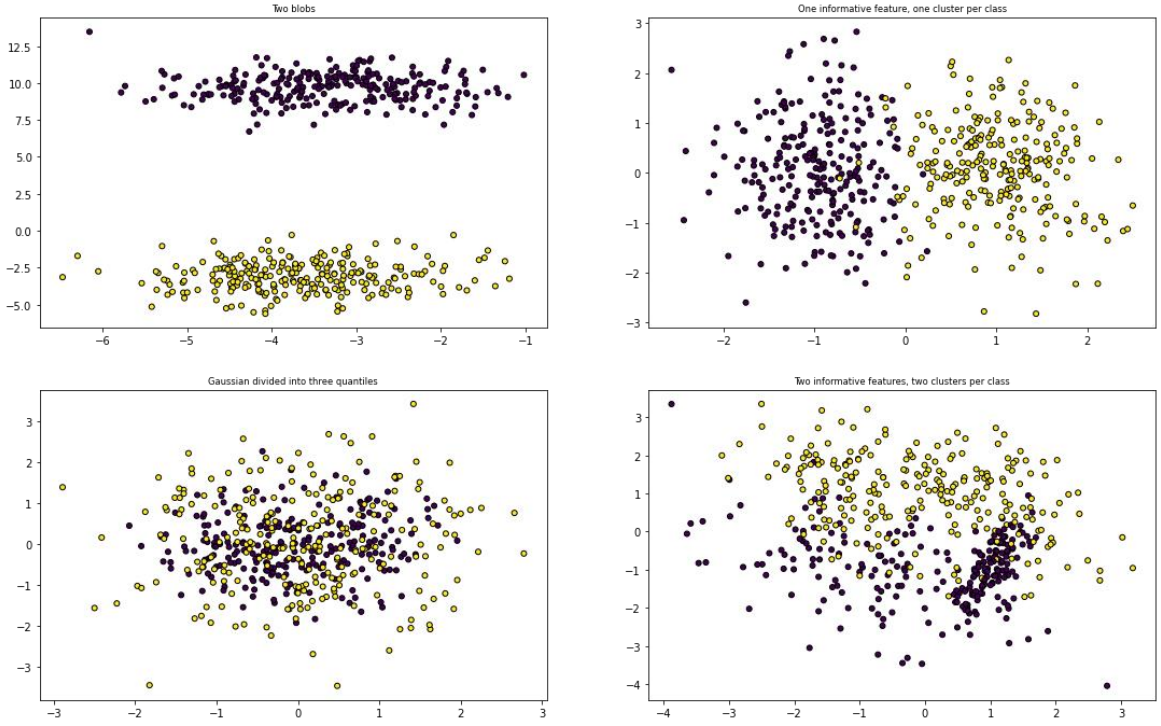
امیرحسین جراره – ۴۰۰۶۱۶۰۰۴

استاد درس

دکتر عبدی هجراندوست

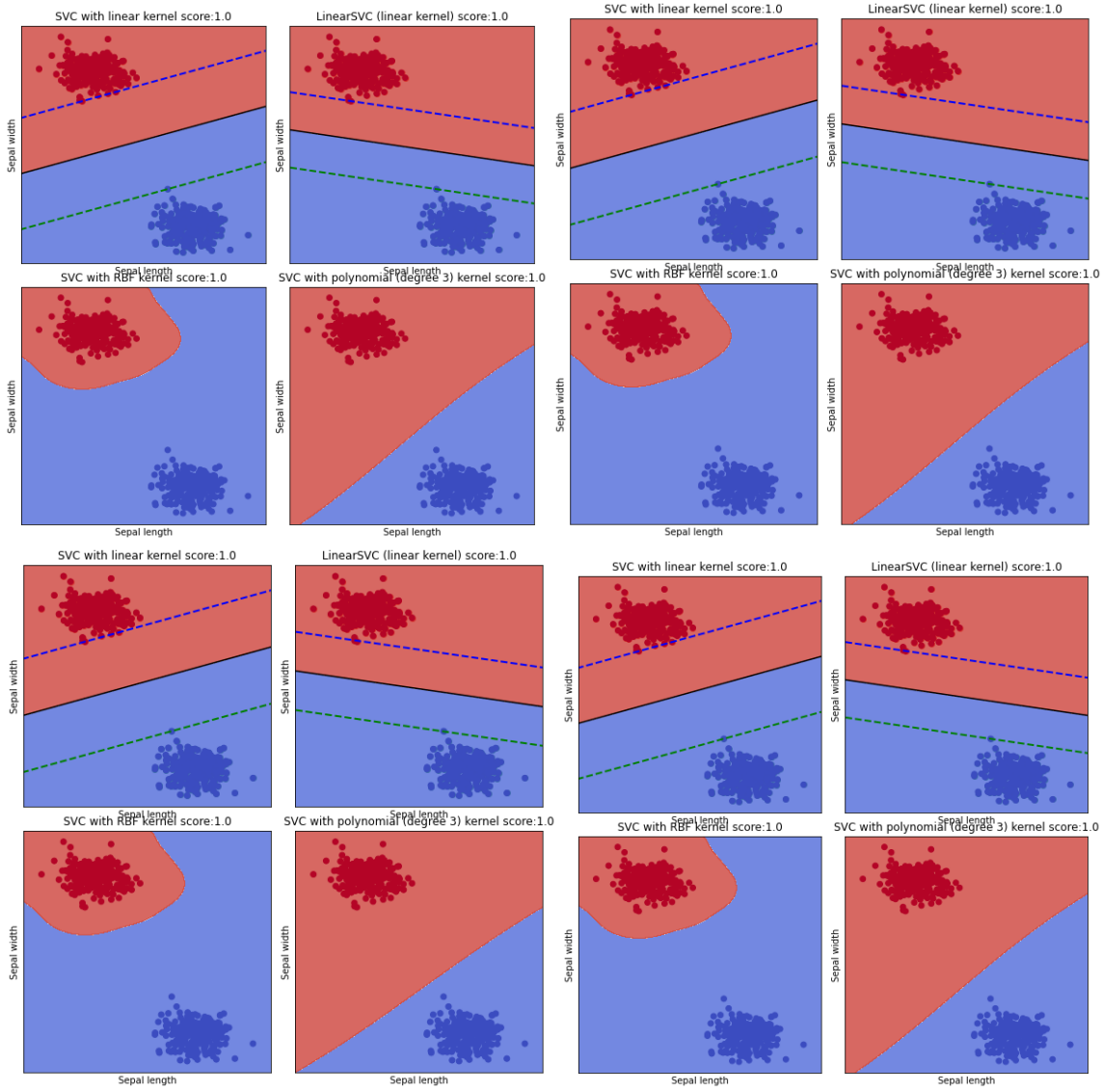
سوال ۱ : این فایل در سکشن SVM Dataset Analyzes در فایل ipynp قرار دارد

ابتدا با استفاده از ماژول دیتاست در کتابخانه sickit learn و همچنین به صورت دستی دیتاست هـای زیر را از ساده تا در هم تنیده و پیچیده با ۵۰۰ نقطه و در ۱۰ ویژگی تولید می نماییم.



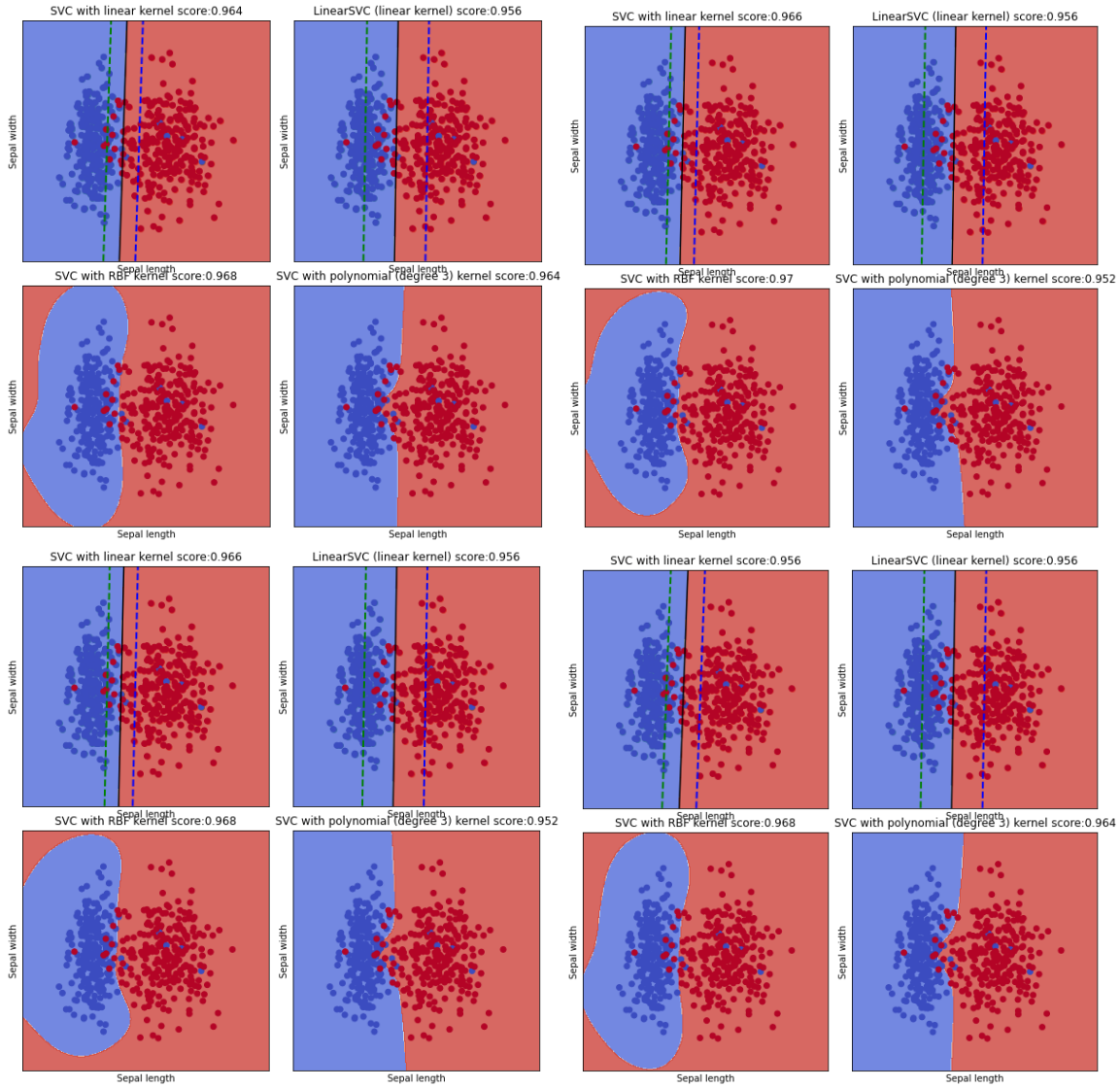
شکل (۱-۱) چهار دیتاست از ساده تا در هم تنیده ،بالا چپ دیتاست ۱: دو کلاس کاملا تفکیک شده ، بالا راست دیتاست ۲: دو کلاس تفکیک شده با مرکزیت نزدیک به هم و دارای ناحیه ی همپوشانی، پایین چپ دیتاست ۳: دو کلاس با پخش گوسی در صفحه پایین راست دیتاست ۴: دو کلاس تفکیک شده با مرکزیت نزدیک به هم و دارای ناحیه ی همپوشانی

سپس عملیات دسته بندی را برای ۴ دیتاست بالا و برای C های متفاوت ۰,۵ ، ۱ ، ۱۰ و ۲۰ تست مـی کنیم. خروجی به صورت زیر می باشد و چهار svm کرنل خطی ، خطی ، کرنل rbf و کرنـل چندجملـه ای تست می گیریم.



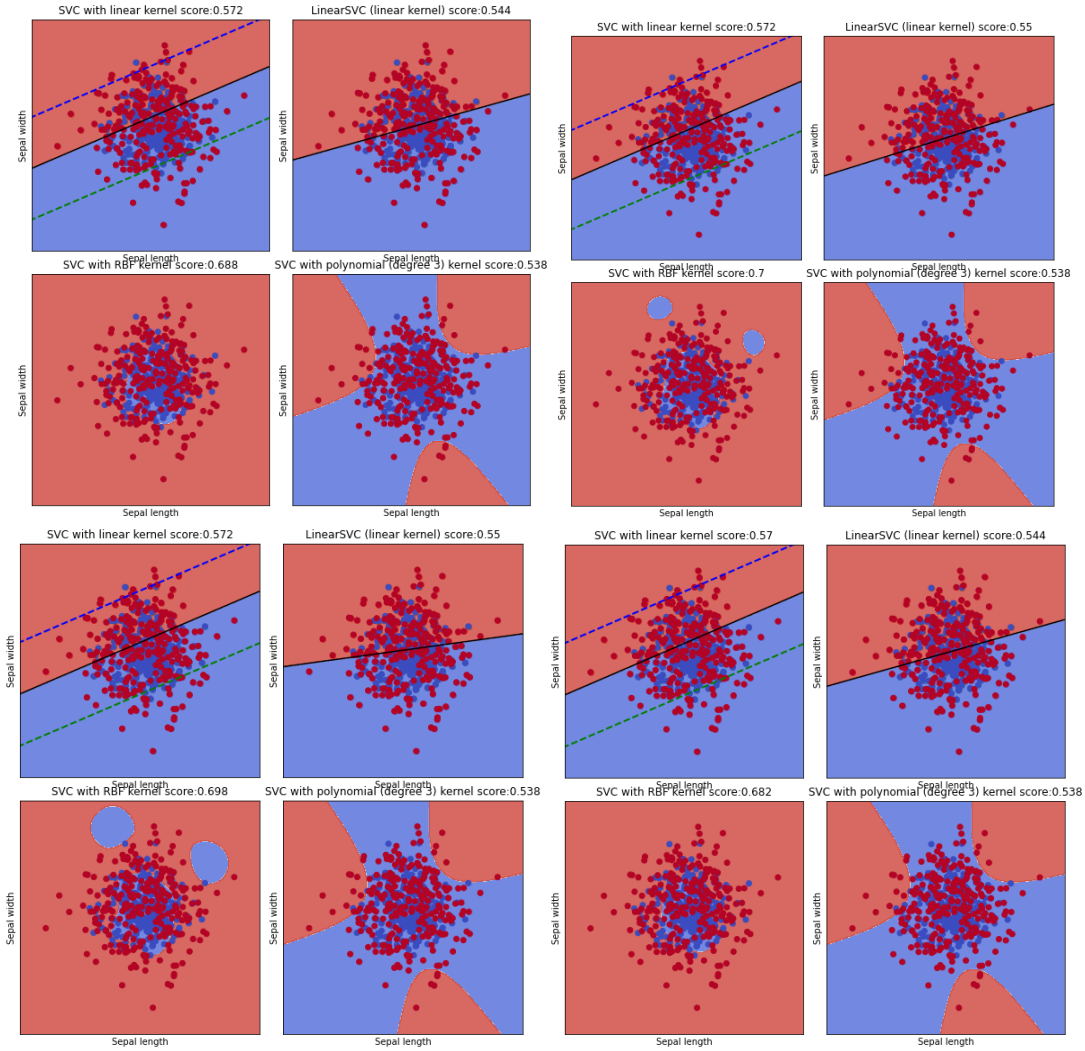
خروجی برای دیتاست ۱ برای C های ۰٫۵، ۱، ۱۰ و۲۰

در مجموعه دیتاست تفکیک پذیر همه ی دقت ها برابر ۱۰۰ درصد بود ولی کرنل خطی مقاوم تر است.



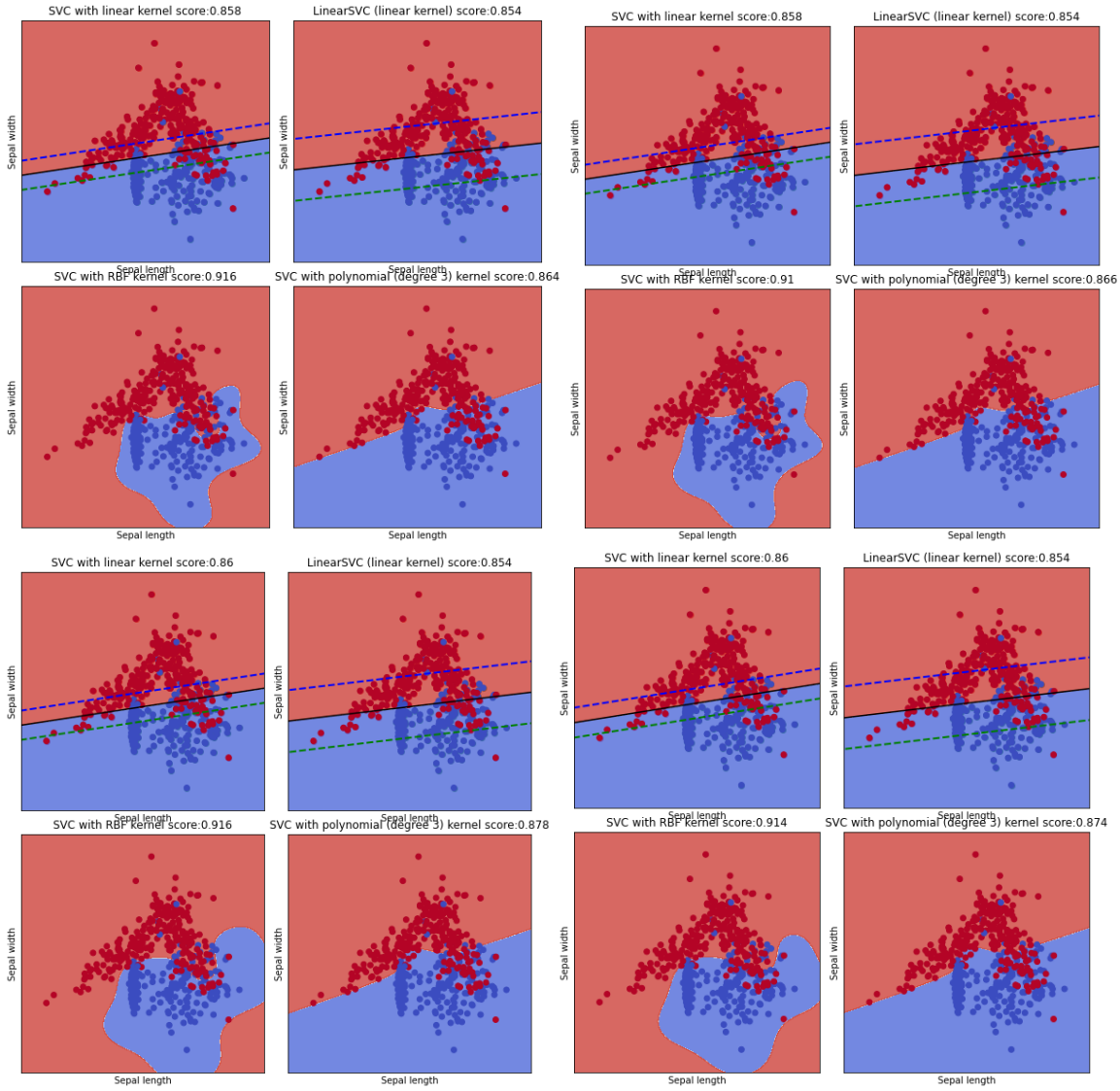
خروجی برای دیتاست ۲ برای C های ۰٫۵، ۱، ۱۰ و ۲۰

برای این دیتاست نیز با وجود نزدیکی جواب ها ولی C با مقدار ۲۰ و کرنل rbf بهترین جواب را محقق کرده است.



خروجی برای دیتاست ۳ برای C های ۰,۵، ۱، ۱۰ و۲۰

برای دیتاست ۳ با توزیع گوسی نیز با وجود نزدیکی جواب ها ولی C با مقدار ۱۰ و کرنل rbf بهترین جواب را محقق کرده است.



خروجی برای دیتاست ۴ برای C های ۰٫۵، ۱، ۱۰ و ۲۰

برای دیتاست ۳ نیز با وجود نزدیکی جواب ها ولی C با مقدار ۱ و کرنل rbf بهترین جواب را محقق کرده است.

همچنین خط مارجین را نیز برای کرنل های خطی رسم می کنیم.هر چه پیچیدگی داده بیشتر باشد دقت کمتر می شود داده های غیر خطی نمی توانند تفکیک پذیری را انجام دهند. به طور کلی اگر پخش داده ها خطی باشد کرنل خطی و ساپورت وکتور خطی عملکرد بهتر و مقاوم تری را دارند ولی برای داده های غیر تفکیک پذیر خطی کرنل غیر حطی مناسب تراست و برای فضاهای محدب کرنل rbf بهترین عمل کرد را دارد. همانند مجموعه داده قبل این مجموعه داده نیز rbf به دلیل کرنل محدب بهترین عملکرد را دارد.البته برای هر داده باز هم به طور حتمی نمی توان بهترین کرنل را به صورت کلی گفت.

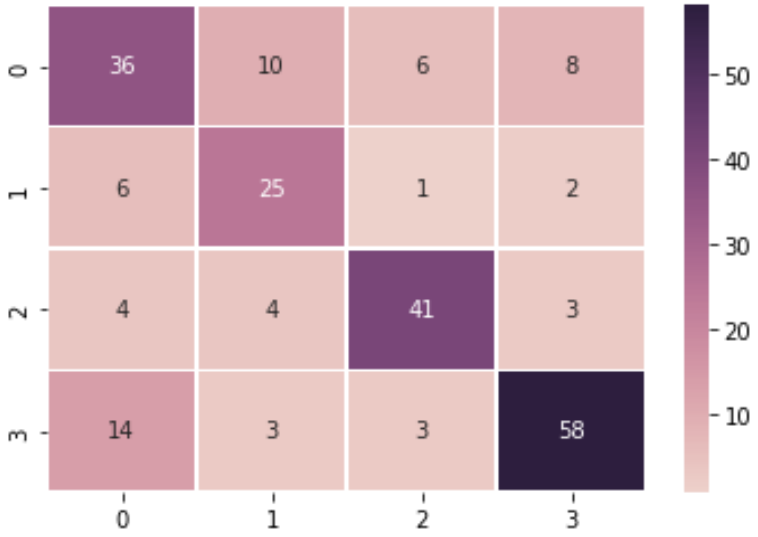
سوال ۲: این فایل در سکشن SVM Weather Classification در فایل ipynp قرار دارد

برای بخش دوم مسئله و دسته بندی Multi-class Weather Dataset for Image Classification پس از دریافت با اااستفاده از کتابخه opencv تصاویر را خوانده ، خاکستری کرده و سایز تصویر را به ۶۴ * ۶۴ ریسایز می نماییم.کلاس ها را نیز با استفاده از دیکشنری زیر عددی می کنیم.

```
label_dict = {
    "cloudy"  :0,
    "rain"    :1,
    "shine"   :2,
    "sunrise" :3
}
```

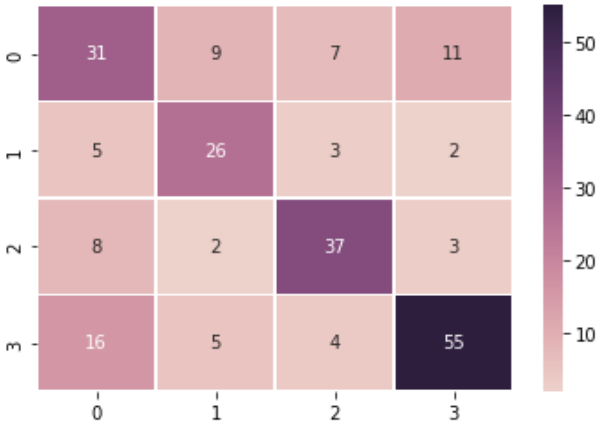
خروجی برای کرنل rbf و C برابر ۲۰ به صورت زیر می باشد.

	precision	recall	f1-score	support
0	0.60	0.60	0.60	60
1	0.60	0.74	0.66	34
2	0.80	0.79	0.80	52
3	0.82	0.74	0.78	78
accuracy			0.71	224
macro avg	0.70	0.72	0.71	224
weighted avg	0.72	0.71	0.72	224



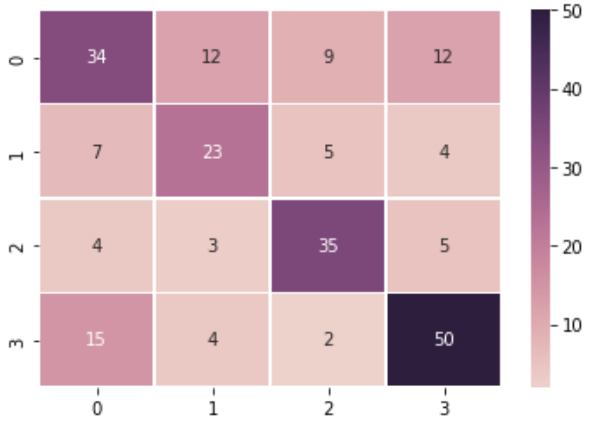
برای کرنل rbf و مقدار C برابر ۱ خروجی به صورت زیر می باشد. همانطور که می بینید دقت کمی کاهش یافته است.

	precision	recall	f1-score	support
0	0.52	0.53	0.53	58
1	0.62	0.72	0.67	36
2	0.73	0.74	0.73	50
3	0.77	0.69	0.73	80
accuracy			0.67	224
macro avg	0.66	0.67	0.66	224
weighted avg	0.67	0.67	0.67	224



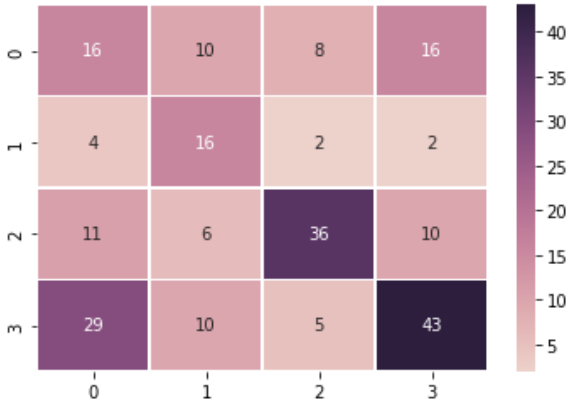
برای کرنل خطی و مقدار C برابر ۲۰ خروجی به صورت زیر می باشد. همانطور که می بینید دقت بسیار زیاد کاهش یافته است.

	precision	recall	f1-score	support
0	0.57	0.51	0.54	67
1	0.55	0.59	0.57	39
2	0.69	0.74	0.71	47
3	0.70	0.70	0.70	71
accuracy			0.63	224
macro avg	0.63	0.64	0.63	224
weighted avg	0.63	0.63	0.63	224



برای کرنل سیگموید و مقدار C برابر ۲۰ خروجی به صورت زیر می باشد. همانطور که می بینید دقت از حالت خطی نیز کمتر شده است.

	precision	recall	f1-score	support
0	0.27	0.32	0.29	50
1	0.38	0.67	0.48	24
2	0.71	0.57	0.63	63
3	0.61	0.49	0.54	87
accuracy			0.50	224
macro avg	0.49	0.51	0.49	224
weighted avg	0.53	0.50	0.51	224



برای کرنل چندجمله ای و مقدار C برابر ۲۰ خروجی به صورت زیر می باشد. همانطور که می بینید دقت از حالت کرنل خطی و سیگموئید بهتر است ولی هنوز rbf کمتر می باشد.

	precision	recall	f1-score	support
0	0.57	0.44	0.50	77
1	0.57	0.60	0.59	40
2	0.59	0.86	0.70	35
3	0.70	0.69	0.70	72
accuracy			0.62	224
macro avg	0.61	0.65	0.62	224
weighted avg	0.62	0.62	0.61	224

در مجموع بهترین کرنل برای این سیستم کرنل rbf و مقدار C برابر ۲۰ می باشد.

سوال ۳:

در مقابل سایر روش ها با توجه به ساپورت وکتـور هـا و عـدم در نظـر گـرفتن سـایر داده هـا ی دور از ساپورت وکتورها الگوریتم SVM به نسبت درخت تصمیم ، KNN و حتی شبکه عصبی نسبت بـه ایـن امر بسیار مقاوم تر می باشد و اگر توزیع داده عوض نشود تقریبا تاثیری ندارد ولی باز هم عدم تعـادل در داده ها باعث نزدیک شدن اثر ساپورت وکتور ها به داده با تعداد بیشـتر اسـت زیـرا نـویز آن هـا بیشـتر الگوریتم را تحت تاثیر قرار می دهد.

برای حل چالش کمبود داده می توان از مباحث دیتا آگمنتیشن استفاده کرد.این کار باعث افـزایش داده از روی سایر داده ها می شود.

چالش ها:

در مجموع در حل این مسئله چالش های زیر مطرح بودند.

- رسم مارجین برای ساپورت وکتور ها
- رسم و کاهش ابعا از فضای ۱۰ بعدی به ۲ بعدی و مقایسه روش ها TSNE ، PCA و اسـتفاده از ۲ ویژگی رندوم
- تولید دیتاست از مجموعه تصویر و تولید برچسب از نام آن ها و حذف داده های خراب
- کار و پردازش تصویر با استفاده از کتابخانه opencv که کد لینک کمک شایانی بـرای ایـن امـر می‌کند.