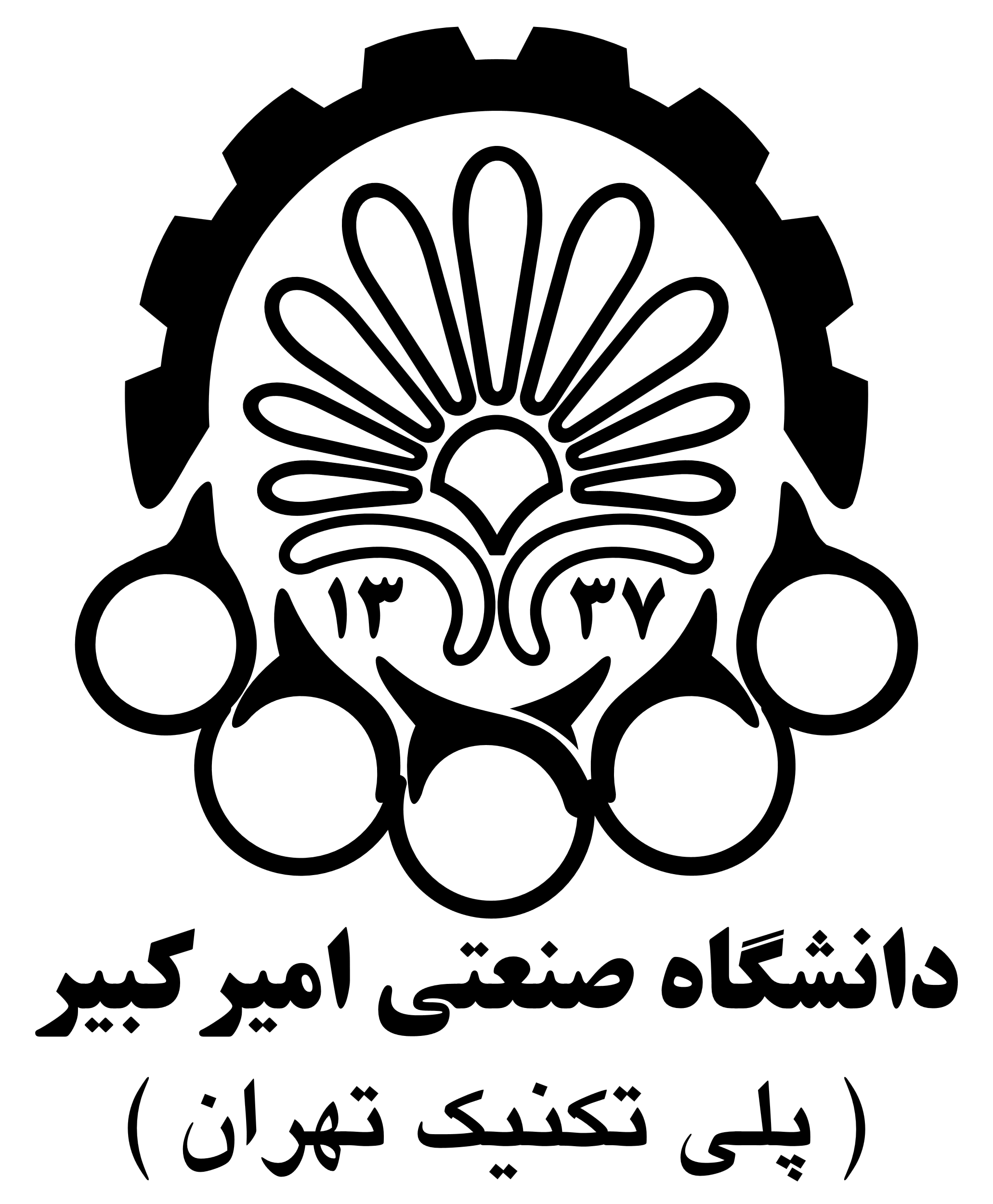
****

**دانشکده مهندسی کامپیوتر**

**تمر ین سوم درس یادگیری ماشین**

**دکتر ناظرفرد**

**غلامرضا دار 400131018**

**زمستان 1400**

فهرست مطالب

[بخش اول: پرسشهای تشریحی 3](#_Toc92497757)

[سوال 1) 3](#_Toc92497758)

[سوال 2) 4](#_Toc92497759)

[سوال 3) 6](#_Toc92497760)

[سوال 4) 8](#_Toc92497761)

[سوال 5) 9](#_Toc92497762)

[بخش دوم: پیاده سازی 10](#_Toc92497763)

[سوال 1) 10](#_Toc92497764)

[سوال 2) 15](#_Toc92497765)

# بخش اول: پرسشهای تشریحی

## سوال 1)

**الف )** ماشین های بردار پشتیبان با کرنل خطی پارامتریک هستند زیرا غیروابسته به داده آموزش، فرم کلی decision function آنها یک خط است. اما به عنوان مثال در مورد ماشین های بردار پشتیبان با کرنل RBF این قضیه درست نیست و با تغییر داده آموزش ممکن است فرم decision function تغییر کند.

**ب )** SVM های Hard Margin به شدت Overfit می‎شوند. پارامتر C در Soft Margin SVM و گاما در RBF برای مقابله با Overfitting استفاده می‎شوند. هرچه C کمتر باشد مدل دچار Regularization بیشتری می‎شود و از Overfitting کاسته می‎شود. پارامتر گاما نیز در Overfitting تاثیر دارد و پیدا کردن مقدار مناسب گاما بسیار مهم است. برای توضیحات بیشتر به قسمت د سوال اول پیاده سازی مراجعه کنید.

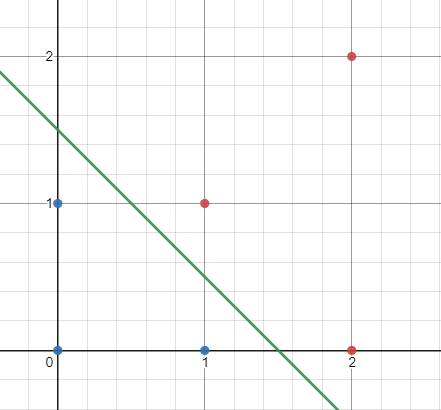
**ج )** در Hard Margin SVM هاDecision boundary فقط به بردار های پشتیبان بستگی دارد و داده های پرت و نویز تاثیر کمتری دارند. اما در Soft Margin SVM ها چه پارامتر C کمتر باشد داده های بیشتری

**د )** خیر. فقط در صورتی این اتفاق می‎افتد که داده ها به کمک یک ترکیب خطی از دسته بند های ضعیف قابل جداپذیری باشند.

**هـ )** بله. وزن های اختصاص داده شده به داده ها مانند احتمال هستند. مقادیر آنها بین 0 تا 1 است و مجموع همه وزن ها 1 می‎شود.

**و )** بله. الگوریتم Ada Boost به عنوان loss function از تابع نمایی استفاده میکند که باعث می‎شود داده های misclassify شده را بیشتر جریمه کند.

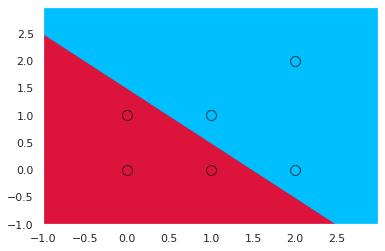
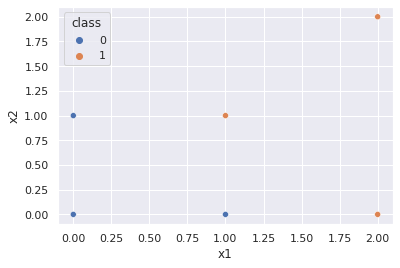
## سوال 2)

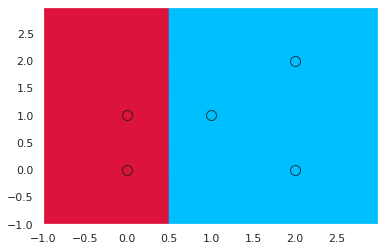
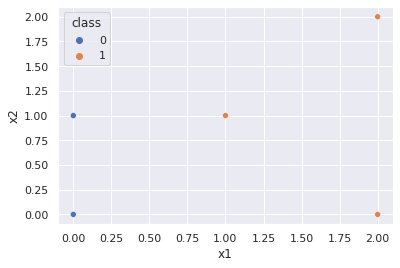
**الف )** بله جداپذیر خطی هستند.

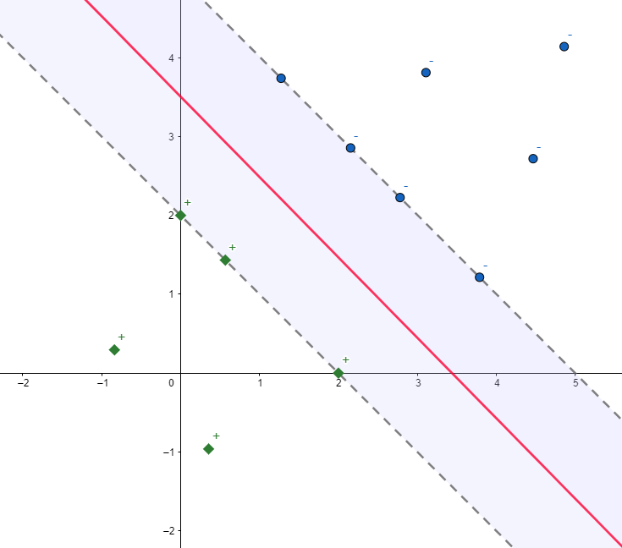
**ب)** محاسبه شیب خط و عرض از مبدا

معادله خط جدا کننده به شکل زیر است ()

**ج )** بله تغییر میکند. در مثالی که مشاهده میکنید نقطه [1,0] حذف شد و W از [2,2] به [2,0] تغییر یافت.



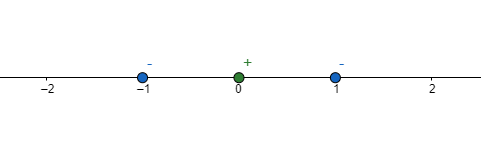


**د )** به شکل زیر توجه کنید. در SVM(Hard Margin) تضمین می‎شود در ناحیه هاشور زده(جاده) هیچ داده ای وجود نداشته باشد. چون در آن صورت آن نقطه به عنوان بردار پشتیبان انتخاب می‎شد و مارجین و در نتیجه ناحیه هاشور خورده تغییر میکرد. در این مثال اگر تمام بردار های پشتیبان هم حذف شوند تنها نقاطی که از خط جدا کننده دور تر هستند باقی خواهند ماند و نقش بردار پشتیبان را به خود میگیرند. در نتیجه این اتفاق **مارجین افزایش** می‎یابد. اگر فقط تعدادی از این بردار ها را حذف کنیم، سایر بردار های پشتیبان مارجین را **ثابت** نگه میدارند.

**نکته:**  تصاویر این سوال بدون استفاده از کد و به صورت دستی اما با نرم افزار رسم شده اند به جز قسمت (ج) که با کمک sklearn برای درک بهتر رسم شده.

## سوال 3)

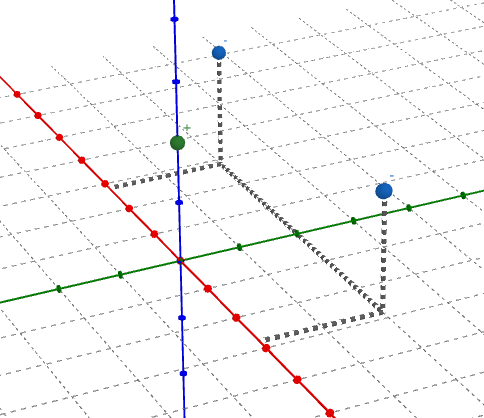
|  |  |
| --- | --- |
|  | Label |
| 0 | + |
| -1 | - |
| 1 | - |



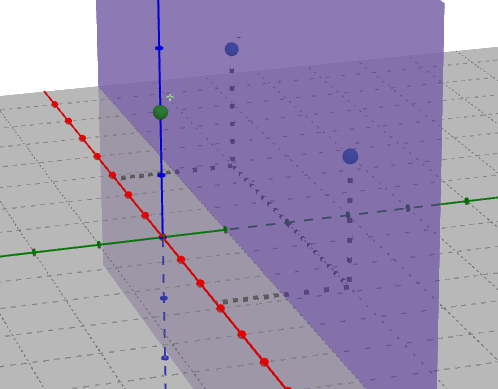
همانطور که مشاهده میکنید داده ها جداپذیر خطی نیستند. با استفاده از کرنل داده شده داده ها را به بعد بالاتر میبریم.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | Label |
| 1 | 0 | 0 | + |
| 1 | - | 1 | - |
| 1 |  | 1 | - |

پس از انتقال به بعد بالاتر داریم(هر واحد grid برابر 0.5):



به راحتی میتوان دید که در این حالت داده ها به کمک یک صفحه جدا می‎شوند. **معادله صفحه :**



حل سوال 3 با کمک روابط ریاضی :

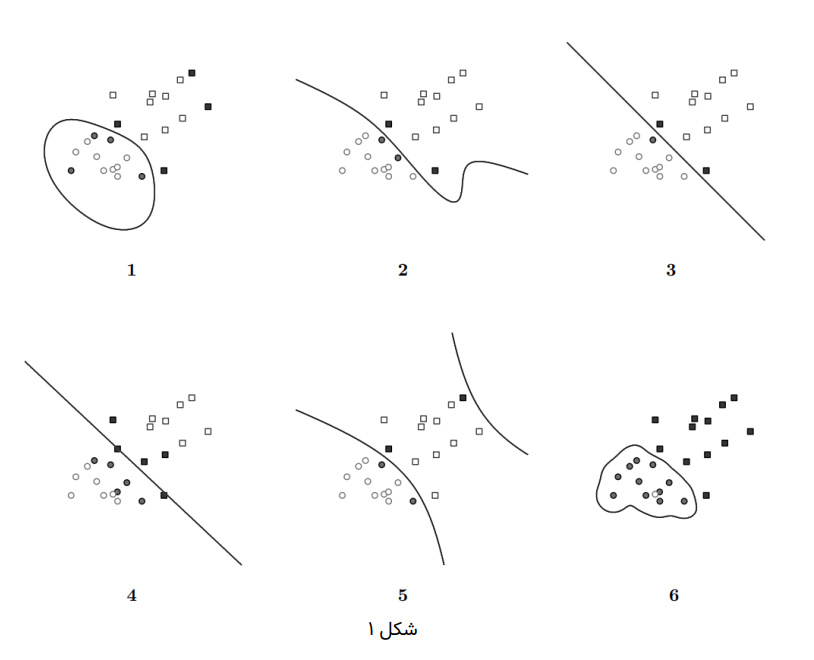
ابتدا مقادیر متناظر kernel را بدست می‎آوریم.

|  |  |  |
| --- | --- | --- |
| 1 | 1 | 1 |
| 0 | 4 | 1 |
| 4 | 0 | 1 |

معادله هایی که داریم را مینویسیم :

**معادله صفحه جداکننده :**

## سوال 4)



**ج**

**ب**

**د**

**الف**

**هـ**

----

شکل شماره 3و4 به طور واضح **خطی** هستند. از بین این دو، شکل 3 دارای **C بزرگتری** است زیرا به Hard Margin SVM نزدیک تر شده پس 3 همان **(ب)** است و 4 همان **(الف).**

شکل شماره 2 یک چند جمله ای با درجه بالاتر از 2 است (3 یا 4 یا بیشتر) پس نمیتواند گزینه مناسبی برای **(ج)** باشد.

به این دلیل است که کرنل 5 را برای شکل **(ج)** برگزیدیم.

شکل شماره 1و6 به احتمال زیاد decision boundary کرنل RBF هستند. هر چه مقدار بیشتر باشد، بردار های پشتیبان روی ناحیه کوچکتری اثر میگذارند و مدل Complex تر میشود و همچنین Decision boundary تیزتر(!) می‎شود. بنابراین 6 همان **(هـ)** است و 1 همان **(د)** است.

## سوال 5)

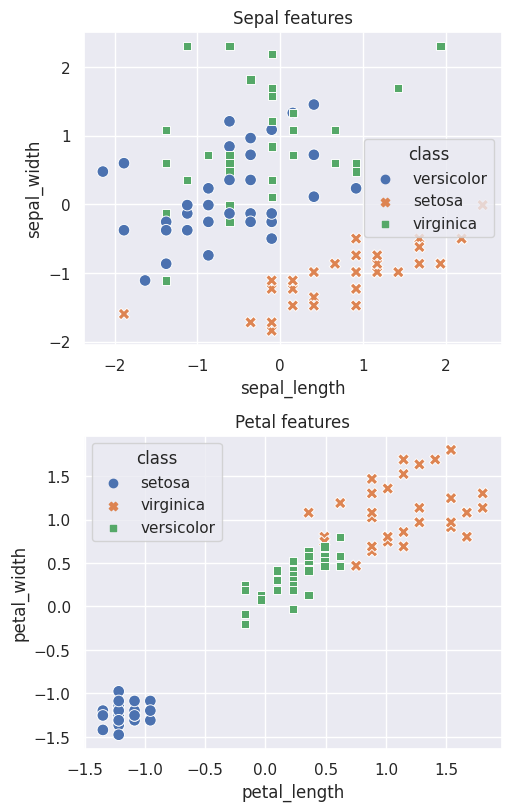
# بخش دوم: پیاده سازی

## سوال 1)

نوت بوک مربوط به این تمرین <https://colab.research.google.com/drive/1w78l4MI6mCXaGTqFnSwrd_hlBEdCkzfe?usp=sharing>

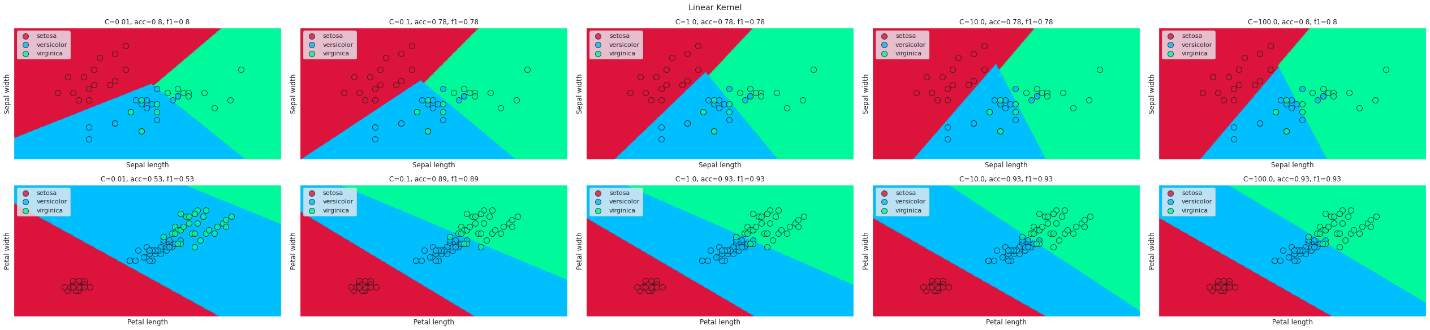
پس از لود کردن دیتاست Iris ، آنرا به بخش های Sepal features, Petal features تقسیم میکنیم (زیرا میخواهیم ویژگی ها را دو به دو بررسی کنیم تا قابل رسم باشد). سپس این بخش ها را به نسبت **30-70** به داده آموزش و ارزیابی تقسیم میکنیم.

**الف)** ابتدا به رسم داده ها میپردازیم. در این بخش دو نمودار خواهیم داشت. نمودار اول شامل ویژگی های Sepal Length, Sepal Width است و نمودار دوم شامل ویژگی های Petal Length, Petal Width.

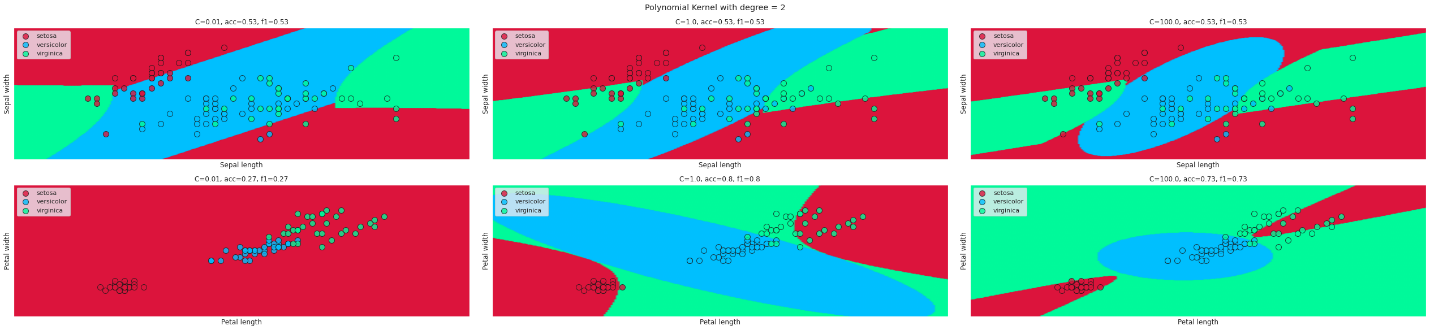


**ب)** در این بخش میخواهیم به ازای کرنل های مختلف، داده های تست این دیتاست را طبقه بندی کنیم و همچنین Decision boundary را رسم کنیم. این کار را به ازای پارامتر های مختلفی برای هر کرنل انجام میدهیم. در مجموع 64 تصویر خواهیم داشت. مقدار هر پارامتر به همراه Accuracy و F1-score متناظر در بالای هر نمودار قرار داده شده است. **توصیه می‎شود برای دیدن بهتر نمودار ها فایل های تصویری که در پوشه این تمرین وجود دارد را ببینید.**

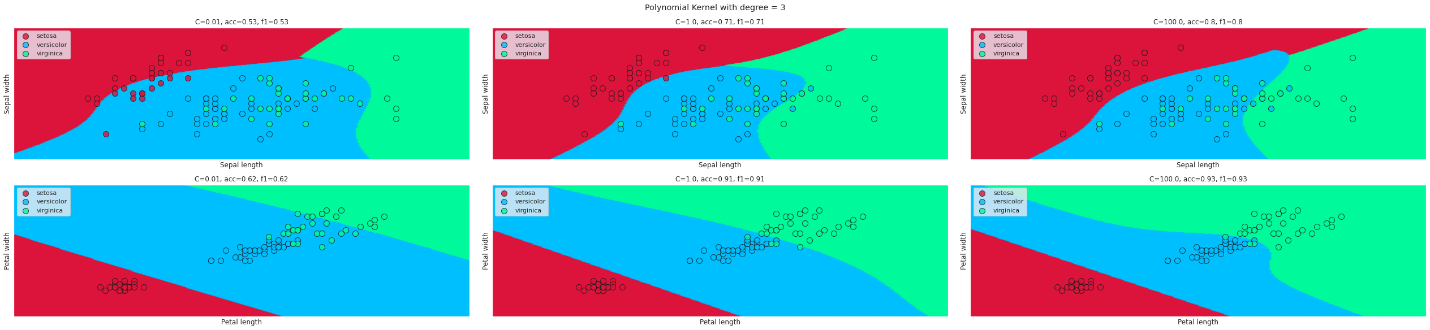
**Linear Kernel ;**



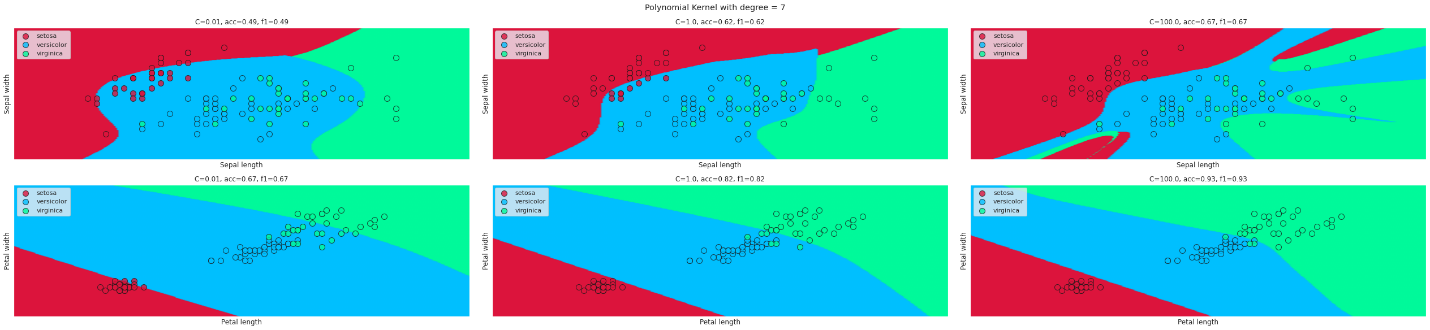
**Polynomial Kernel** ;



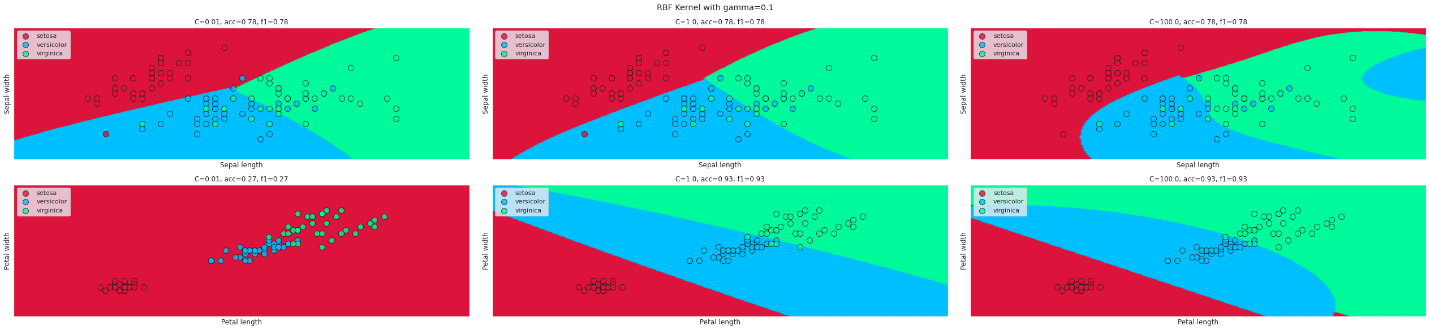
**Polynomial Kernel** ;



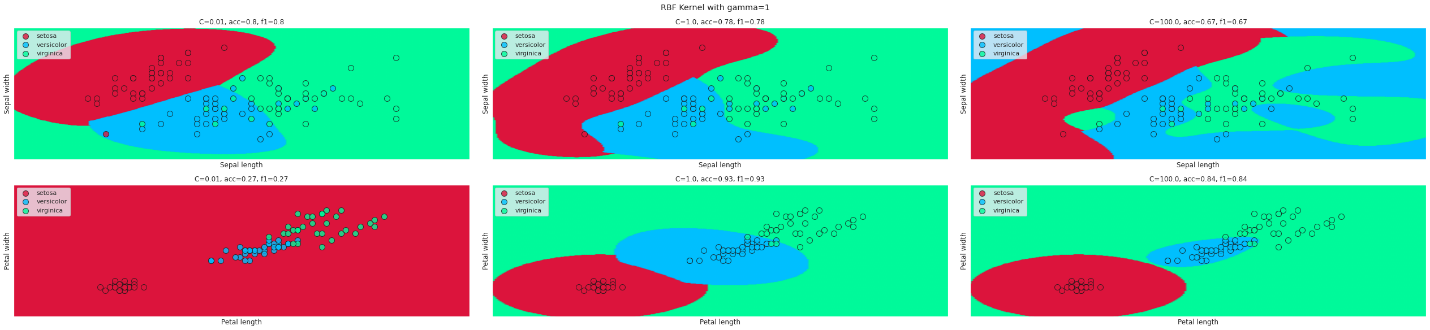
**Polynomial Kernel** ;



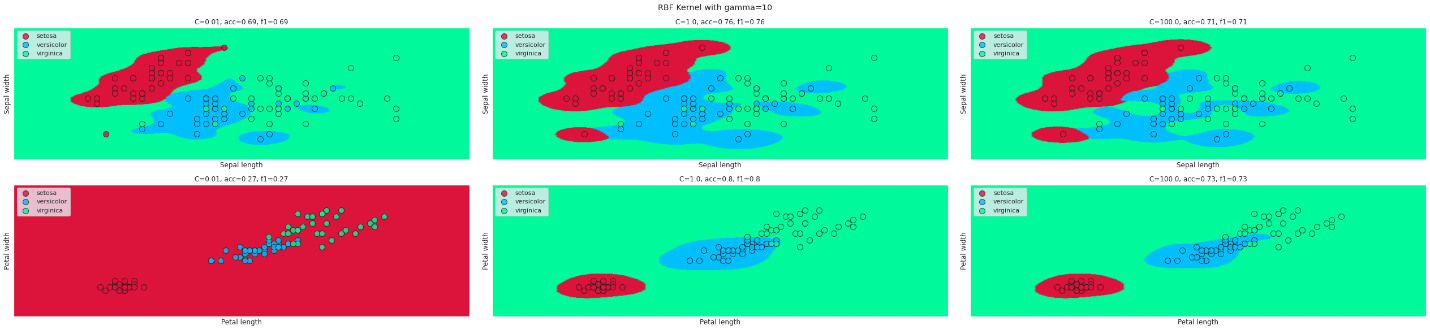
**RBF Kernel** ;



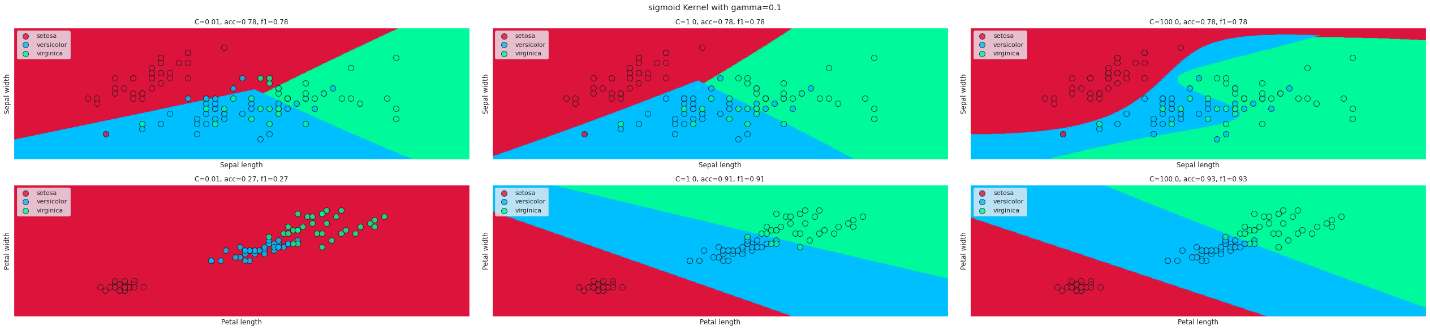
**RBF Kernel** ;



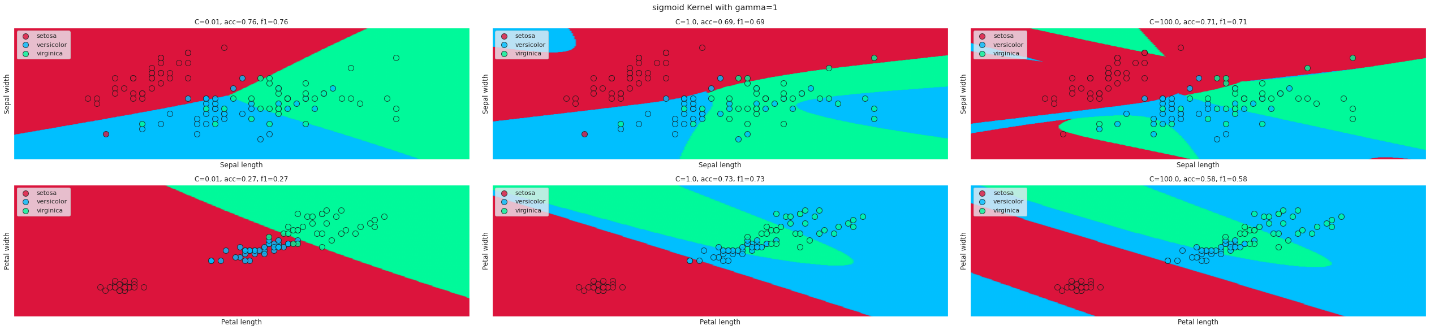
**RBF Kernel** ;



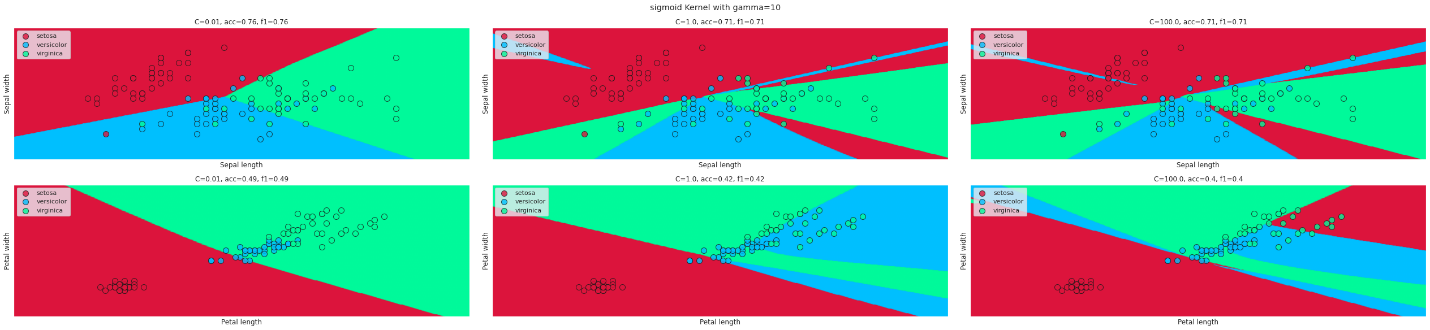
**Sigmoid Kernel** ;



**Sigmoid Kernel** ;



**Sigmoid Kernel** ;



**ج)** معیار Accuracy, F1-measure در تصاویر بالا موجود هستند اما جهت بهتر دیده شدن اعداد به پوشه این تمرین، یا notebook مراجعه کنید.

<https://colab.research.google.com/drive/1w78l4MI6mCXaGTqFnSwrd_hlBEdCkzfe?usp=sharing>

**د)** در مورد پارامتر های SVM میتوانیم نکات زیر را بیان کنیم.

**C** **:** این پارامتر یک Tradeoff بین Hard Margin بودن SVM و پهنای جاده آن است. هر چه مقدار این پارامتر بیشتر باشد به Hard Margin SVM نزدیک تر می‎شویم و تعداد Misclassification ها کاهش می‎یابد. برای مقابله با Overfitting بهتر است مقدار C را کم کنیم.

**Degree :** درجه چند جمله ای در کرنل چند جمله ای مشخص میکند منحنی جدا کننده کلاس ها چه مقدار پیچیدگی داشته باشد. اگر درجه 2 باشد داده ها با یک خط یا یک سهمی از هم جدا میشوند، اگر درجه بالاتر باشد ممکن است منحنی هایی با پیچیدگی بیشترdecision boundary را تعیین کنند. در نتیجه هر چهdegree بیشتر باشد احتمال Overfitting بالاتر می‎رود.

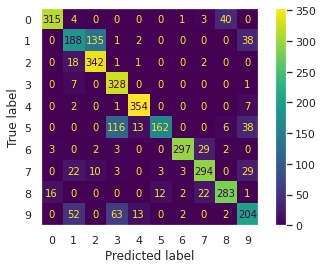
**:** این پارامتر که در کرنل های RBF و Sigmoid مورد استفاده قرار میگیرد، تعیین میکند بردارهای پشتیبان علاوه بر خودشان به چه تعداد داده دیگر تاثیر میگذارند. اگر این پارامتر بسیار زیاد باشد، بردارهای پشتیبان فقط به خودشان اثر میگذارند و Overfitting رخ میدهد. اگر مقدار این پارامتر خیلی کم باشد مدل به Linear SVM تبدیل میشود و در واقع هر بردار پشتیبان روی تمام نقاط دیتاست تاثیر میگذارند. مقادیر میانی این پارامتر بهترین نتیجه را میدهند.

## سوال 2)

**الف) Random Forest**

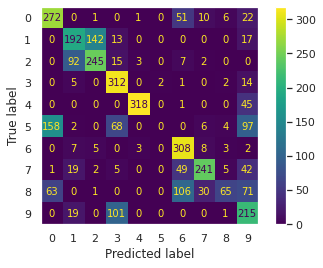
نوت بوک مربوط به این تمرین <https://colab.research.google.com/drive/11EkT4FEim1jgfc2BrmBACjDkohlY1an5?usp=sharing>





**ب) Ada Boost**

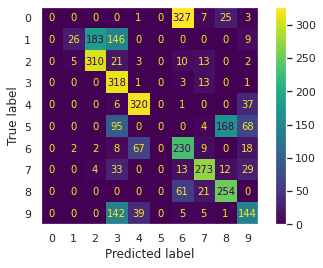
****

****

**ج) Ada Boost با مقادیر n\_estimator متفاوت**

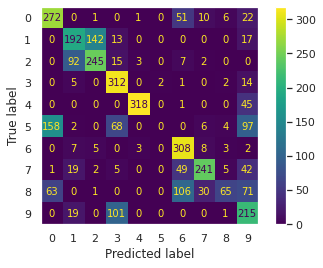
**n\_estimator = 5**

****

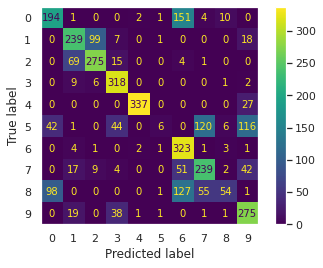
****

**n\_estimator = 10**

****

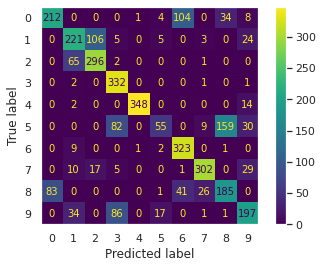
****

**n\_estimator = 20**

**n\_estimator = 50**





**د)** **XGBoost**

ابتدا با استفاده از مقادیر پیشفرض XGBoost دسته بند را آموزش میدهیم.



**Test Accuracy : 0.961**

**F1-Score : 0.966**

سپس با در نظر گرفتن مقادیر خاصی برای پارامتر های XGBoost به نتابج زیر دست یافتیم.



**Test Accuracy : 0.967**

**F1-Score : 0.967**