علی ضیغمیان

۹۷۲۳۰۵۱

الف)

ستون RecordId از دیتاست حذف شد. چون از لحاظ منطقی نقشی در دسته بندی ندارد و صرفا باعث حواس پرتی مدل میشود.

ستون هایی با ویژگی غیرعددی کدگذاری شدند و سپس همه ی ستون ها نرمال شدند.

درنهایت ۳۰ درصد داده ها به عنوان تست جدا شد و داده ی اموزش به مدل svm در حالت خطی داده شد.

صحت مدل روی داده ی اموزش 97.6 % و صحت روی داده ی تست 96.8 % است.

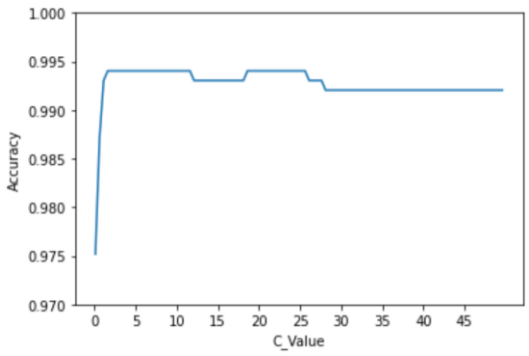
تعداد ساپورت وکتورهای هر دسته به شرح زیر است.

```
[ 41,  97, 233, 190]
```

ب)

یک ارایه برای c از 0.1 تا 50 با تغییر 0.5 ساخته شد و داخل یک حلقه هر بار با مقداری از c مدل ساخته شد و میزان صحت آن روی داده ی ارزیابی ذخیره شد.

درنهایت شکل تغییرات صحت بر حسب c به صورت زیر است:



مشاهده میشود که بهترین مقدار c در حدود 3 است. که صحتی در حدود 99.4 % روی داده های ارزیابی دارد.

صحت روی داده ی ترین 99.9 % و روی تست 99.5 % است.

تعداد ساپورت وکتورها:

```
[ 20,  52, 129, 106]
```

پ)

در حالت c=1 ، با کرنل rbf صحت روی داده ی ارزیابی 99.3 %

با کرنل sigmoid صحت 85.4 %

با کرنل poly صحت 99.3 %

بنظر کرنل rbf بهتر است.

صحت روی تست : 99.3 %

صحت روی اموزش : 99.8 %

تعداد ساپورت وکتورها

```
[ 30,  81, 191, 136]
```

ت) کرنل را روی rbf و c را برابر ۳ قرار میدهیم:

صحت روی ارزیابی % 99.4 است، روی اموزش % 99.9 و روی تست % 99.5 است.

که مشابه قسمت ب که کرنل بصورت پیش فرض روی rbf بود می باشد.

تعداد ساپورت وکتورها :

```
[ 20,  52, 129, 106]
```

با استفاده از svm غیرخطی توانستیم تا حد خوبی صحت را روی همه ی داده ها بالاببریم. غیرخطی کردن کرنل باعث شد تا تعداد ساپورت وکتورها نیز کم شود. همچنین با تنظیم پارامتر soft svm توانستیم از برخی داده هایی که پرت بودند صرف نظر کنیم و تا میشود حاشیه بین نواحی را زیاد نگه داریم. اما بصورت کلی صحت ها بطور عجیبی بالا بودند که برای من حس خوبی ایجاد نکرده چرا که تغییرات پارامترهای مدل تاثیر قابل حسی روی نتایج صحت نداشتند.