توضیحات کدها:

سوال یک: ابتدا ستون user id را حذف کرده چون برای هر سطری مقدار یکتا داشته و تاثیری در نتیجه نهایی و دست بندی حاصل نمیکند. از نمودارهای هیستوگرام و دایره ای متوجه میشویم که از لحاظ سنی بیشتر کاربران مربوط به بازه ۳۵ تا ۴۰ سال هستند و همچنین تعداد خانم ها بیشتر از آقایان است.

با استفاده از روش encoding تبدیل متغیر کتگوریکال جنسیت به متغیر عددی انجام شد. سپس داده ها را با روش standardscalerنرمال کرده و سپس به مجموعه داده آموزشی و آزمایشی تقسیم شد.

با توجه به اینکه برای مقدار بایاس همواره مقادیر داده ورودی را یک فرض میکنیم یک لیستی از مقادیر یک به اندازه تعداد داده ها ساخته شده و به ماتریس داده های ورودی چسبانده میشود. مقادیر اولیه وزن یک نیز در نظر گرفته میشود. توابع sigmoid cross Entropy , نیز در نظرگرفته شده است.

حال برای پیش بینی مقادیر خروجی اگر مقدار تابع sigmod بزرگتر از ۰.۵ باشد مقدار یک در نظر گرفته میشود و اگر کوچکتر از ۰.۵ باشد مقدار صفر به عنوان خورجی در نظرگرفته میشود. (چون مقدار تابع sigmoid نشان دهنده احتمال وقوع کلاس یک است)

Accuracyنشان دهنده تعداد مقادیر درست پیش بینی شده تقسیم بر تعداد کل داده هاست. یعنی مجموع مقادیری که برچسب صفر داشتند و مدل ما نیز مقدار صفر را به عنوان خروجی پیش بینی کرده است و مقادیری که برچسب یک داشتند و مدل ما نیز مقدار یک را به عنوان خروجی پیش بینی کرده است را تقسیم بر تعداد کل داده ها میکنیم.

سوال ۱_قسمت خ) تابع logistic regressionرا سه بار هربار با مقدار آلفای یکسان ۰.۰۱ و تعداد iterationهای متفاوت (۱۰۰ بار ، ۲۰۰۰ بار ، ۵۰۰۰ بار) اجرا کردیم. بنظر میرسد با تعداد دفعات ۵۰۰۰ بار مقدار خطا در نهایت کمتر شده و نتیجه بهتری دارد. از نظر دقت نیز مشاهده میکنیم روند صعودی دارد و هربار با کاهش خطای مورد نظر دقت افزایش میابد. در حالت اولیه با ۱۰۰ بار تکرار دقت در نهایت به مقدار ۶۸٪ رسیده است و با ۲۰۰۰ بار تکرار دقت نهایی به مقدار ۸۰٪ رسیده است. با ۵۰۰۰ بار تکرار نیز دقت به اندکی بالاتر از ۸۰٪ رسیده است.

سوال ۱_قسمت م) برای پیش بینی مقادیر تست نیز مقادیر ورودی مورد نیاز برای بایاس نیز که یک هستند به ماتریس ورودی چسبانده و اضافه میشود. مدل ساخته شده با مقادیر وزن نهایی بدست آمده در قسمت قبل روی داده های تست اجرا شده و میزان دقت نهایی ۶۱٪ بدست آمده است. مقدار خطای کراس آنتروپی آن نیز حدودا ۷۶٪ است.

سوال ۲: ابتدا ستون rownames که برای هر سطر مقدار یکتا دارد و تاثیری روی نتجیه نهایی و پیش بینی ما ندارد را حذف میکنیم. با استفاده از روش encoding متغیر کتگوریکال جنسیت را تبدیل به متغیر عددی میکنیم.

قسمت ب) بنظر میرسد در حالت نرمال شده مقادیر در فضای حالت دامنه کوچکتری داشته و مقادیر موجود در بازه کوتاهتری را شامل میشوند. قبل از نرمال سازی در بازه ۶ تا ۲۰ بودند اما بعد از نرمال سازی در بازه -۳ تا ۳ قرار گرفته.

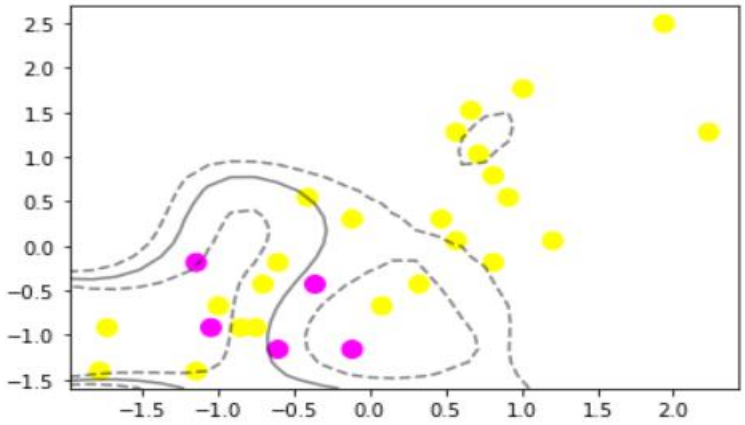
قسمت خ) بنظر میرسد در حالت خطی دقت روی داده های آموزشی ۷۸٪ ولی در حالت گاوسی دقت ۸۰٪ است و حالت گاوسی دقت و عملکرد بهتری دارد. چون داده ها به صورت خطی تفکیک پذیر نیستند منطقی است که کرنل گاوسی نتیجه

بهتری داشته باشد. با توجه به عدم تفکیک پذیری خطی داده ها باید به فضای دیگری مپ شود که کرنل گاوسی این کار را انجام میدهد تا در فضای جدید مپ شده داده ها بهتر تفکیک شوند.

دقت مدل گاوسی روی داده های تست نیز ۷۹٪ است و دقت نسبتا خوبی است که فاصله زیادی نیز با دقت روی داده های آموزشی نداشته در نتیجه overfiting نیز اتفاق نیفتاده است.

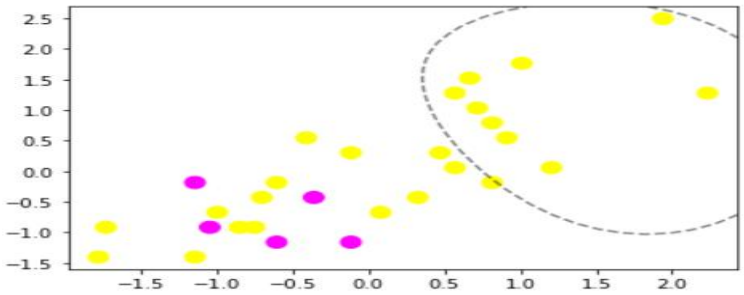
قبلتر مشاهده کردیم که حالت کرنل گاوسی مناسب تر از خطی است چون داده ها خطی تفکیک پذیر نیستند.

در حالت گاوسی اگر مقدار C را خیلی بزرگ ینی همان ۱۰ به توان ۵ در نظر بگیریم مدل overfit شده به گونه ای که دقت ترین ۸۰٪ اما دقت تست ۶۵٪ خواهد شد و مرز تصمیم ها نیز پیچیده خواهند شد و پیچ و خم زیادی خواهد داشت. که کاملا منطقی است چون میدانیم اگر مقدار C خیلی بزرگ باشد جریمه ترم دوم خطا در مسائل svm حتی وقتی مقار زیتا خیلی کوچک است زیاد شده در نتیجه به سمت کم کردن مقادیر زیتا میرویم تا جایی که به hard margin تبدیل شده و حتی یک نقطه نیز نباید در مارجین قرار گیرد در نتیجه نتایج پیچیده بدست می آید. و با توجه به اینکه داده ها تفکیک پذیر خطی نیستند سولوشنی پیدا نمیکند. مرز تصمیم به این شکل خواهد شد:

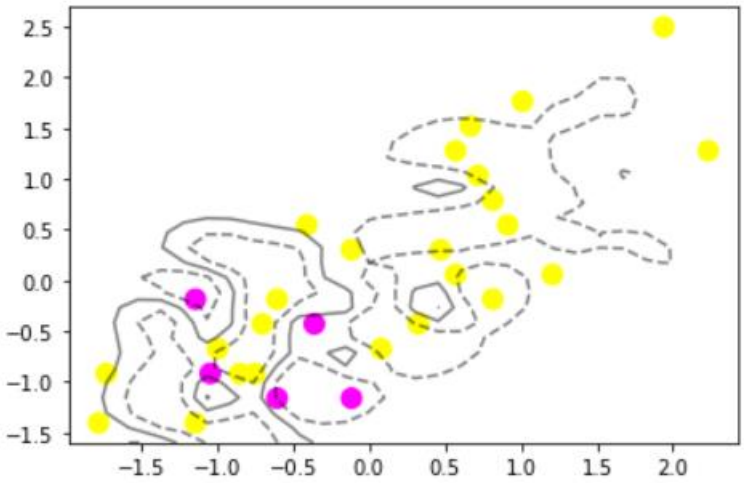


اما اگر در این حالت مقدار C را مقدار خیلی کوچکی در نظر بگیریم ینی همان ۱۰ به توان -۲ در نظر بگیریم دقت روی ترین نیز به ۶۳٪ رسیده که در مقایسه با حالت قبلی مقدار خیلی کمی است. که بنظر میرسد مدل اورفیت شده است و اصلا سولوشنی برنمیگرداند. عملکرد سولوشن حاصل روی دادهای تست به این شکل خواهد شد:

اصلا نتوانسته خوب عمل کند و داده ها را درست دسته بندی کند که خب منطقی است چون اگر مقدار سی خیلی کوچک باشد مقدار جریمه حاصل از ترم دوم مقدار کمی داشته پس به سمت بزرگ کردن زیتا ها پیش میرود در نتیجه تعداد زیادی داده میتوانند داخل مارجین قرار گرفته و دقت مناسبی نخواهد داشت.



از طرف دیگر مقدار سی را ثابت در نظر گرفته و مقدار گاما را تغییر میدهیم. اگر گاما را بزرگترین مقدار در این بازه در نظر بگیریم مشاهده میکنیم مدل اورفیت میشود چون دقت ترین ۹۳٪ در حالی که دقت تست ۵۵٪ خواهد شد. منطقی هم هست. چون مقدار گاما مشخص میکند نقاط تا چه اندازه فاصله برای تعیین مرز تصمیم در نظر گرفته شوند. اگر خیلی گاما زیاد باشد ینی تقاط با فاصله خیلی زیاد نیز در نظر گرفته میشوند اما اگر کوچک باشد تعداد اندکی نقطه در فاصله نزدیک در نظر گرفته میشوند. مدل روی داده های تست به این شکل خواهد شد:



اما اگر با مقدار ثابت سی مقدار گاما کوچکترین مقدار موجود در آن بازه در نظر گرفته شود مشاهده میکنیم مدل آندرفیت میشود. تا جایی که تبدیل به خط میشود که انگار سولوشن دقیق و درستی نیست و هیچ دسته بندی انجام نمیدهد. روی داده های تست به این شکل خواهد شد: