باسمه تعالی

گزارش کار تمرین دوم درس یادگیری ماشین – جناب آقای دکتر باباعلی

ایمان کیانیان

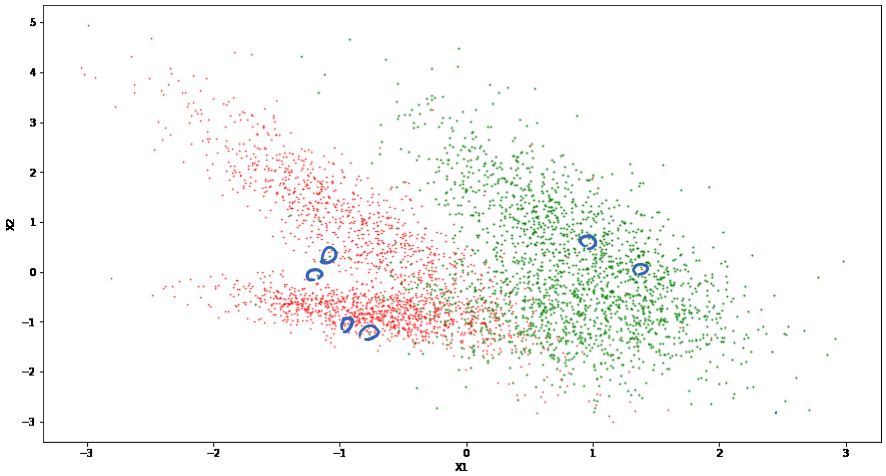
۱- مقدمه

در این تمرین ، دسته بندی خطی با استفاده از پرسپترون را انجام دادیم. داده های آموزشی ما شامل ۴۰۰۰ داده دارای ۲ فیچر یا مشخصه بود و یک ستون هم به عنوان خروجی یا label وجود داشت که دو مقداره بود (صفر یا یک). صفر نشانگر اصل بودن سکه و یک نشانه تقلبی بودن سکه بود. در واقع کار این دسته بند ، دسته بندی سکه ها به اصل یا تقلبی است. در ادامه برخی مشاهدات ، روش ها و تحلیل ها را ذکر خواهیم کرد.

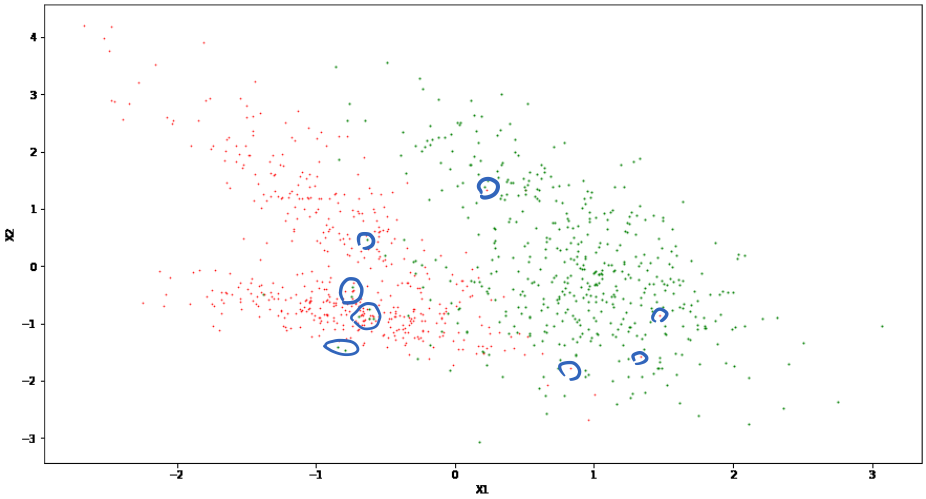
۲- داده ها

این دیتا ست شامل ۴۰۰۰ داده اموزشی و ۱۰۰۰ داده تست است که شامل ۲ فیچر هستند. همچنین هر داده آموزشی دارای یک label است که ۰ یا ۱ هستند. لیبل صفر به معنی اصل بودن سکه و ۱ به معنی تقلبی بودن سکه است.

داده های آموزشی را به صورت تصویری در عکس زیر مشاهده میکنید. نقطه های o سبز رنگ، سکه های تقلبی و نقطه های x قرمز رنگ ، سکه های اصل هستند. انتظار داریم برای نمونه نقاطی که با دایره مشخص شده اند ، دچار خطا شوند که بعدا خواهیم دید اینطور خواهد شد.



و داده های تست به صورت زیر هستند :



هدف ما این است خطی رسم کنید که داده های این دو کلاس را از همه جدا کند. البته واضح است که یک خط ساده نمیتواند این داده هارا به طور کامل جدا کند. بنابراین اگر از روش پرسپترون استفاده کنیم قطعا دچار خطا خواهیم بود. البته یکی از روش های ما استفاده از کرنل است که ممکن است خطا را کم یا حتی صفر کند.

۳- پرسپترون

ابتدا تابعی به اسم sign_function پیاده سازی کردیم که عملکرد زیر را دارد:

$$f(x) = \begin{cases} 0 & if\ x \leq 0 \\ 1 & if\ x > 0 \end{cases}$$

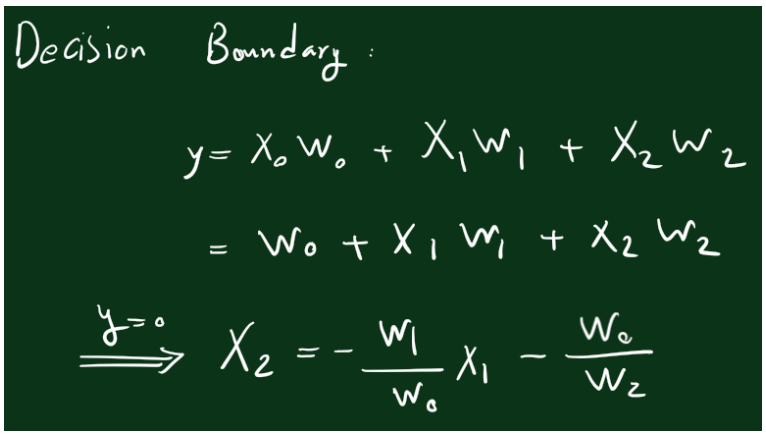
سپس تابعی نوشتیم به نام Preprocess که یک ستون شامل ستون ۱ به داده های ما اضافه کرد.

تابع perceptron که در واقع X ، Y، epochs و Learning Rate را گرفته و مقدار W بهینه را حساب میکند. همچنین تعداد misclassified را در هر epoch ذخیره کرده و به خروجی میدهد. طبیعتا کمترین misclassified بهترین حالت W برای ما خواهد بود.

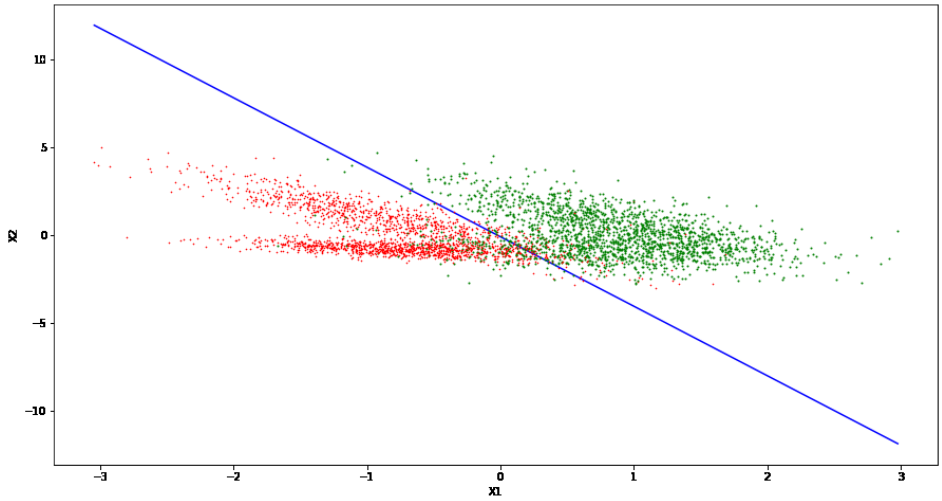
ما داده ها را با epoch = 1000 و lr = 1e-3 به تابع پرسپترون دادیم و خروجی هارا دریافت کردیم. W ی بدست آمده هدف ما بود حال مرز تصمیم گیری را باید بر اساس W بدست آمده رسم کنیم. در قسمت بعدی توضیحات در خصوص رسم مرز تصمیم داده خواهد شد.

۴- رسم مرز تصمیم

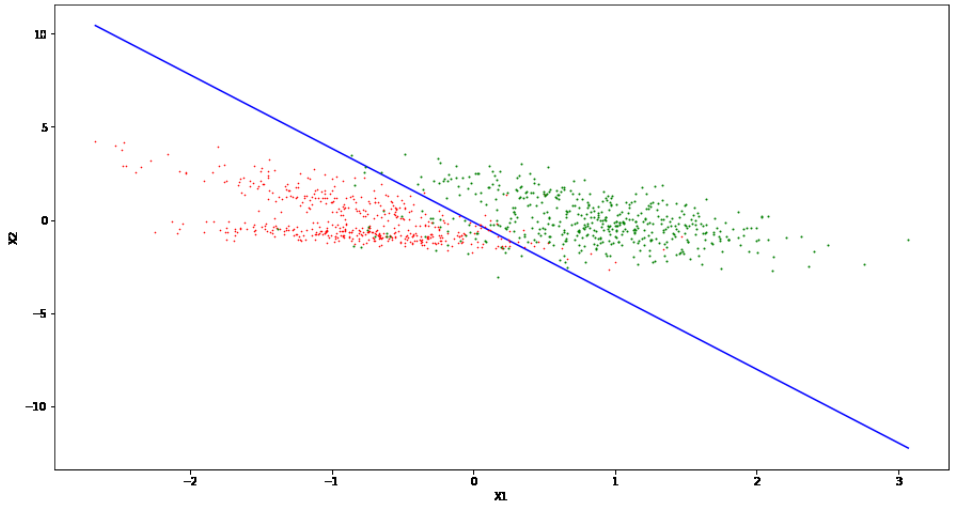
رسم مرز تصمیم با استفاده از W بدست آمده و داده ها چندان سخت نخواهد بود . کافی است رابطه ریاضی زیر را در نظر گرفته و تابعی برای رسم مرز تصمیم بنویسیم.



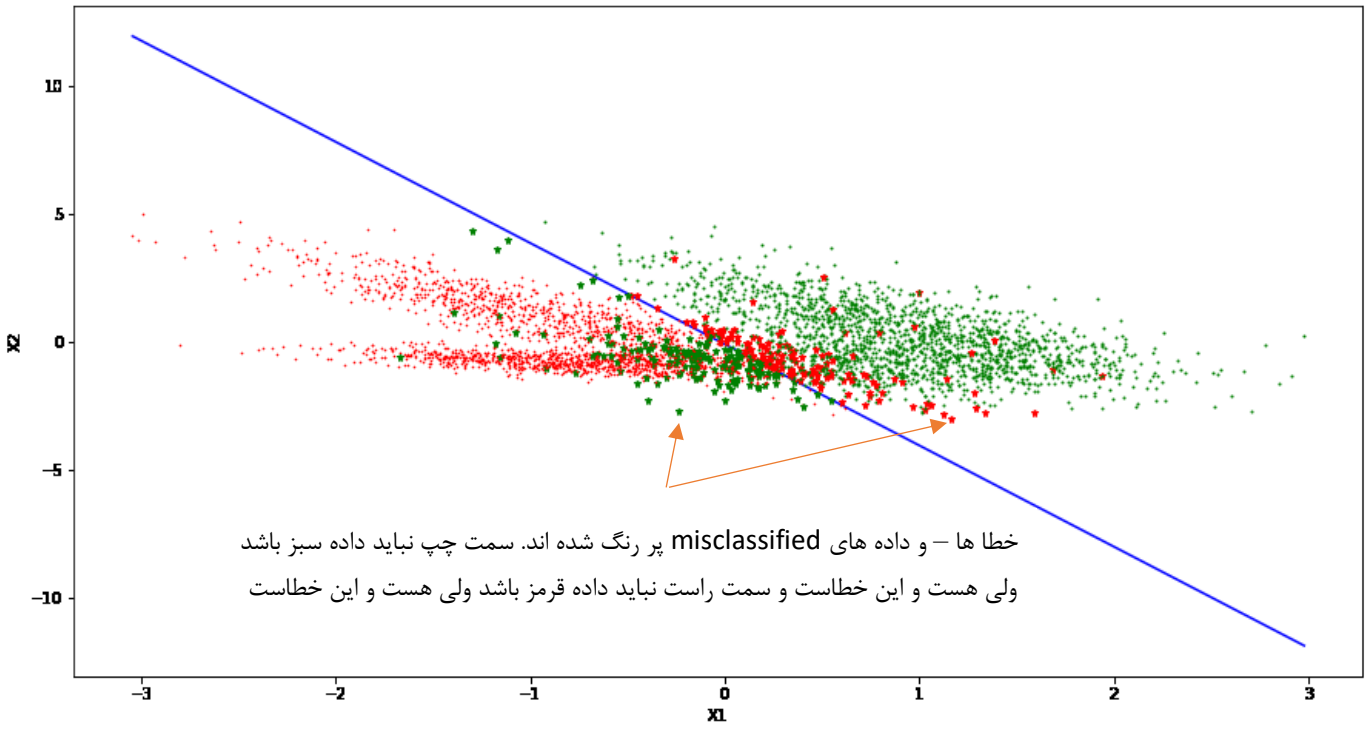
همچنین با استفاده از این و رسم مرز ها به شکل زیر برای داده های آموزشی میرسیم:

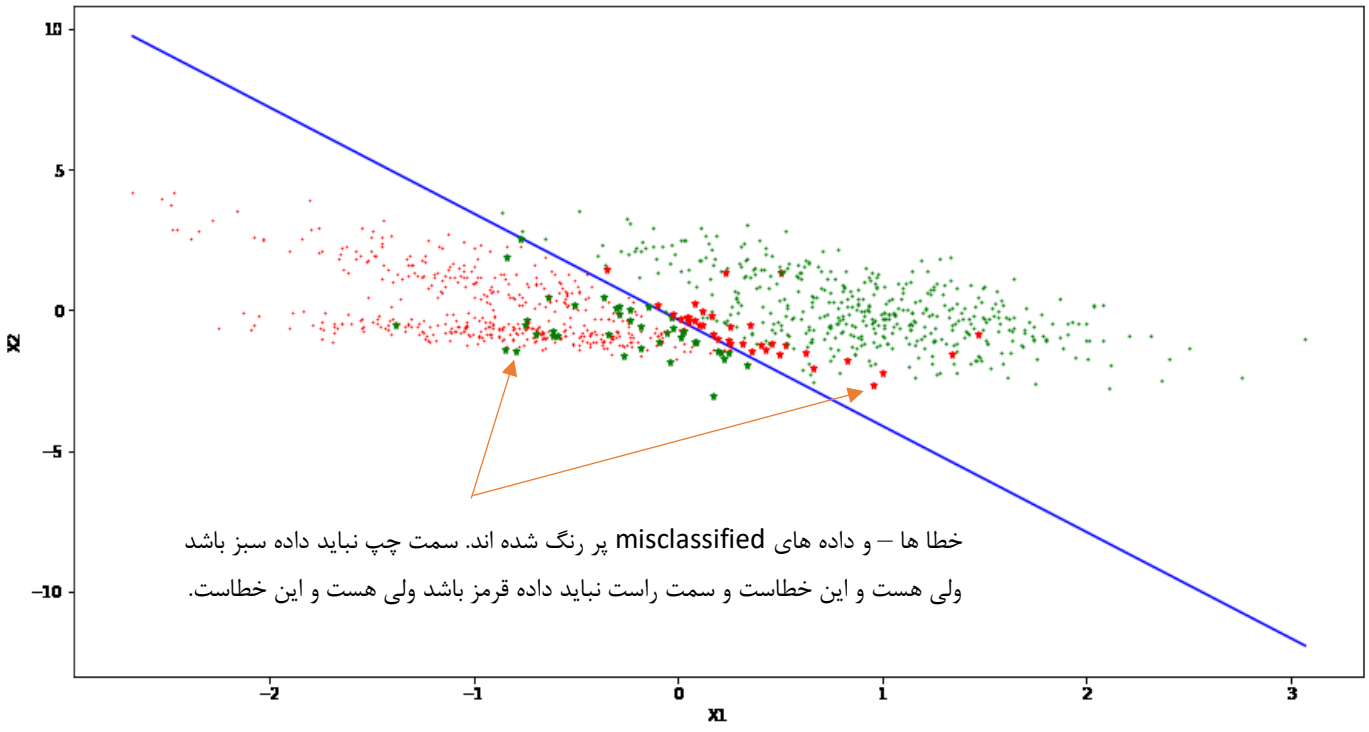


و همچنین برای داده های تست :



واضح است که در هر دو شکل مرز تصمیم گیری در جای مناسبی قرار دارد . اما طبیعتا چون داده ها کاملا قابل جدا سازی توسط یک خط نیستند و باهم تداخل دارند، مقداری خطا داریم و آن داده هایی است که باید در سمت راست خط باشد ولی در سمت چپ خط است (برخی نقطه های سبز) و بلعکس. برای نشان دادن داده هایی که دچار خطا شده اند و misclassify میشوند آن نقاط را پر رنگ کردیم و شکل نمودار برای داده های آموزشی و تست به شکل زیر در آمد:





۵- محاسبه خطا

در قسمت ما با استفاده از جدول Confusion Matrix ، مقادیر TP ، TN، FP و FN را برای داده های آموزشی بدست آوردیم. جدول ما به صورت زیر است:

N = 4000	Predicted : 0	Predicted : 1	
Actual : 0	TN = 1833	FP = 174	**2007**
Actual : 1	FN = 149	TP = 1844	**1993**
	1982	***2018***	

همچنین برای داده های تست :

N = 1000	Predicted : 0	Predicted : 1	
Actual : 0	TN = 468	FP = 35	**503**
Actual : 1	FN = 40	TP = 457	**497**
	508	***492***	

به راحتی میتوانیم دقت خودمان را در هر حالت بدست آوریم با استفاده از فرمول زیر:

$$Error = \frac{TN + TP}{TN + TP + FN + FP}$$

همچنین توسط این نمودار میتوان تحلیل های دیگری نیز ارائه کرد که در این تمرین به آن نیاز نیست.

که مقدار دقت برای داده های آموزشی برابر ۰/۹۱۹۲۵ و برای داده های تست برابر ۰/۹۲۵ است. به وضوح دقت روی داده های تست بهتر است.علت وجود خطا هم به دلیل این است که داده های ما به صورت خطی قابل جدا سازی نیستند وگرنه روش پرسپترون اگر داده ها به صورت خطی قابل جدا شدن باشند باید دقت ۱۰۰ درصد را برای ما تولید کند.