مبانی یادگیری ماشین

گزارش تمرین ۱ و ۲

عرشیا حسینمردی

شماره دانشجویی : 98222030

توجه شود که در گزارش بیشتر فقط نتیجه کد های زده شده نوشته شده و برای بررسی دقیقتر باید به کد مراجعه شود و در ساختار کد تلاش شده تا خواندن آن راحت باشد.

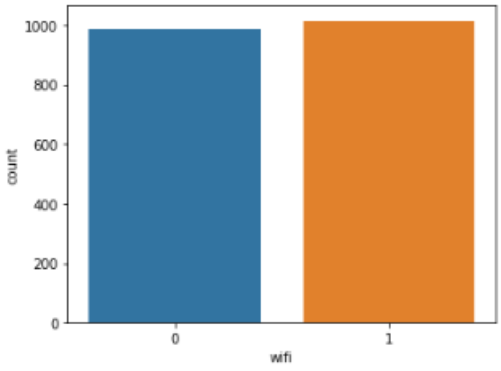
گزارش تمرین اول)

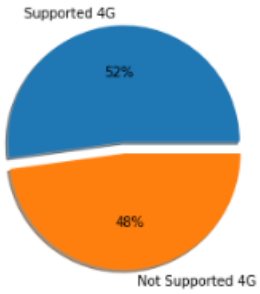
پاکسازی داده :

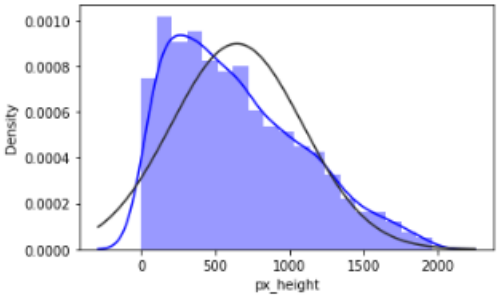
اول کاری که میکنیم این است که داده را بررسی میکنیم که داده های غیر موجود و پرت و تکراری را پیدا کنیم که دیتا این سوال نه داده ناموجود دارد و داده تکراری و ما از روش zscore داده های پرت را نیز پیدا کردیم و از داده ها خارج کردیم.

اطلاعات کلی داده ها :

ما داده ها را با نمودار های مختلف مثل pie chart , box plot و ... نشان داده ایم که نمونه هایی از آن در پایین آمده و برای دیدن کامل نمودار ها باید به کد مراجعه شود.







آزمون های فرض :

در این قسمت ما 5 آزمون فرض را با توجه به داده هایمان بررسی کردیم و که برای انجام آن از سه روش مختلف استفاده کردیم که کد آن و روش های آن در قسمت hypothesis testing تمرین اول مشخص است

آزمون اول :

ازمون اول در مورد رابطه بین داشتن وای فای و قیمت است که از یک توزیع هستند که از روش f_oneway یا anova استفاده کرده ایم.

آزمون دوم :

طول و عرض گوشی به همدیگر وابسته هستند و رابطه ای دارند که از روش pearson استفاده میکنیم و فرض ما درست است و به هم وابسته اند.

آزمون سوم :

در این آزمون فرض این است که داشتن 4G بر روی قیمت تاثیری ندارد که فرض درست است و از روش student t-test استفاده کردیم.

آزمون چهارم :

رزولوشن دوربین اصلی گوشی تاثیری روی قیمت ندارد که با انجام آزمون متوجه میشویم که درست است و تاثیری روی قیمت گوشی ندارد .

آزمون پنجم :

رابطه ای بین بلوتوث و وای فای وجود دارد که با انجام t-test دیدیم که بله این رابطه وجود دارد.

مدل :

مدل های کلاسیفایر در کد مشخص شده است و مدل logistic ما از OVA استفاده میکند.

و اینکه مدل ما برای کلاس های مختلف یکسان عمل نکرده (که گزارش آن در کد است) که دلیل آن میتواند این باشد که مدل ما logestic است تعداد کلاس های ما بیشتر از 2 است و 4 تا کلاس داریم.

متوازن بودن داده ها:

بله داده ها متوازن است که در قیمت کد مشخص است و اگر متوازن نبود 3 راه زیر را میتوانیم انجام دهیم:

روش 1)

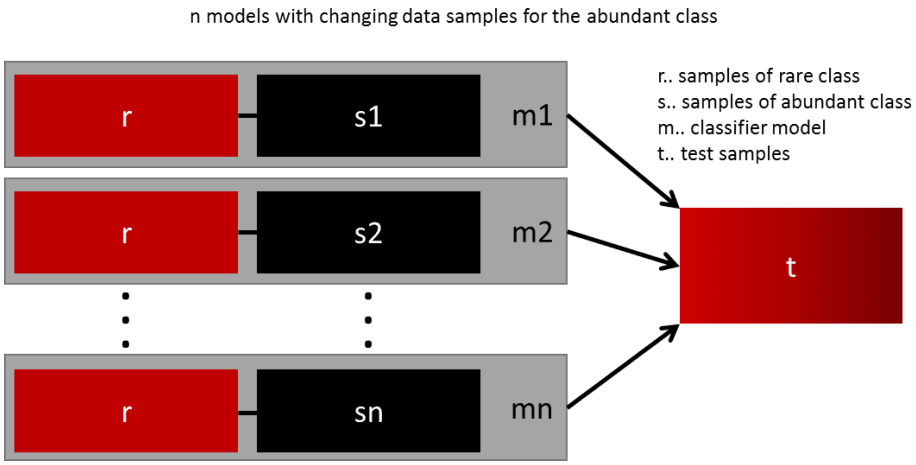
یک روش این است که تعداد داده ی کلاسی که زیاد است را کم کنیم تا تقریبا برابر با داده ی کلاسی که مقدار آن کم است شود که این روش زمانی کاربرد دارد که اندازه کلاس با داده کم به قدری باشد که برای مدل ما کار کند به عبارت دیگر؛ با نگه داشتن همه نمونه‌ها در کلاس نادر و انتخاب تصادفی تعداد مساوی از نمونه‌ها در کلاس فراوان، می‌توان یک مجموعه داده جدید متعادل را برای مدل‌سازی بیشتر بازیابی کرد.

روش 2)

یک روش دیگر این است که ما کلاس با داده کمیاب را زیاد کنیم تا داده ها متوازن شوند که یک راه میتواند این باشد که اگر کارفرما داده ها را تهیه کرده به او بگوییم که داده هایی که در کلاس کمیاب است . یا اینکه از روش هایی مثل bootstrapping یا SMOTE داده های جدید تولید کنیم.

روش 3)

یک راه ساده میتواند این باشد که مثلا اگر یک کلاس با 1000 داده و کلاس دیگری با 10000 داده داشتیم ، آنگاه 10000 نمونه کلاس زیاد به 10 چانک تقسیم میکنیم و 10 مدل مختلف را آموزش میدهیم



روش 4)

یک راه دیگر که علم زیاد تری نسبت به بقیه میخواهد این است که الگوریتم خود را عوض کنیم و از الگوریتمی استفاده کنیم که با داده های نامتنازن بهتر عمل کند یا اینکه مدل را عوض کنیم یا دید خود را نسبت به دیتا عوض کنیم تا شاید به ایده ای برسیم که مشکل داده نا متوازن ما را حل کند

اسکیلینگ

بر روی داده minmaxscaler و روشی که داده منهای میانگین و تقسیم بر std را انجام دادیم که دقت مدل مارا بسیار بالا برده و از حدو 67 به حدود 90 رساند که میتوانیم در کد به صورت کامل مشاهده کنیم.

همچنین برای تست داده ها را به نسبت ۸۰ به ۲۰ جدا شده است که در کد قابل مشاهده است.

PCA

با اعمال pca های مختلف که در کد قابل مشاهده است ؛ در مدل logestic تغییر بسیاری نمیکند و در مدل رگرسیون خطی دقت آن کاهش میابد حالا شاید بپرسیم که چرا از این روش استفاده میکنیم ، از دلایل آن میتوان گفت : با Pca کار مصورسازی بسیار آسان تر میشود و اینکه کار با pca محاسبه را برای کامپیوتر بسیار آسان میکند. و برای کار با داده های با ابعاد بالا خیلی به کارمان میآید.

عدم توازن

کلاس داده ها را جوری تغییر دادیم که کلاس ها نامتوازن شد که در این حالت اگر از رگرسیون خطی استفاده کنیم دقت کاهش میابد و اگر از لوجستیک استفاده کنیم دقت بالا میرود . حالا ما از از روش smoth استفاده کردیم که داده های کلاس کمتر را از روشی اضافه میکند تا تعداد داده های کلاس کمتر به تعداد کلاس بیشتر برسد که از این روش دقت مدل ما کاهش پیدا کرد که دلیل آن میتواند این باشد که تعداد داده های کلاس کمتر ما بسیار کم است و چون از روی آنها داده های جدید میسازد داده های جدید نماینده خوبی برای کلاس کمتر نیستند و داده هایی که برای آموزش استفاده میکند خوب نیستند.

گزارش تمرین دوم)

پاکسازی داده :

در data set دوم ما اول یک سری از ستون هایی را که تاثیری در مدل ما ندارند مثل پلاک خانه یا کد پستی را حذف میکنیم.

داده های مثل houseNumber , streetPlain , street و scoutId مربوط به اطلاعات نسبتا unique برای هر خانه هستند که در این مدل به درد آموزش نمیخورد و همینطور facilities و description که در اصل توضیح هستند و اطلاعات آماری به ما نمیدهند را نیز حذف میکنیم.

بعد از آن داده های categorical که تعداد داده های منحصر به فرد آن بسیار است(مثلا محله که میتواند 1000 نوع محله را داشته باشد) را که فقط ممکن است مدل ما را خراب کند حذف میکنیم.

حالا سراغ بقیه فیچر ها میرویم و آنهایی که تعداد داده های null آن بیشتر از 50 درصد است را حذف میکنیم.

حالا سراغ پر کردن داده های null مانده میرویم :

داده های null دسته عددی را با میانگین پر میکنیم و داده های تهی کتگوریکال را با مد داده ها پر میکنیم

سپس داده های categorical که یک سری داده با درصد بسیار کم دارند را در یک دسته قرار میدهیم تا بار محاسباتی کم شود و در آخر بر روی داده های categorical ، one hot encoding یا به اصتلاح dummy encoding را انجام میدهیم.

در این بین خوب است بگوییم که ما داده های پرت را از روش zscore حذف کرده ایم و روی دیتا اسکیلینگ نیز انجام داده ایم و مهندسی ویژگی نیز برای بهبود مدل استفاده شده است .

پس از انجام مراحل بالا ، داده ما برای آموزش مدل آماده است.

Multiprocessing:

ما یک سری از بخش های پاکسازی را هم از روش ساده و هم از روش پردازش چندگانه رفتیم و زمان آن را ثبت کردیم که در بیشتر آن زمان پردازش چند گانه بیشتر بوده که شاید به خاطر این بوده که توابع ما کار پیچیده ای نمیکردند و اینکه تعداد داده های ما کم بوده که هزینه pool کردن بیشتر از پردازش داده روی تابع بوده است

مصور سازی داده :

مثل سوال قبل انواع داده های اینdata set را با نمودار های مختلف نمایش داده ایم که در قسمت کد مشخص است.

ما در این سوال از مدل رگرسیون خطی استفاده کرده ایم که نتیجه آن در قسمت کد مشخص است و بهتر است در آنجا دیده و بررسی شود.