باسمه تعالی

تمرین سری پنجم درس یادگیری ماشین

استاد درس: جناب آقای دکتر باباعلی

نام دانشجو: ایمان کیانیان

شماره دانشجویی: ۶۱۰۳۰۰۲۰۳

۱- مقدمه

در این تمرین قرار است به صورت کلی ابتدا پاکسازی و پیش پردازش داده ها با روش های مختلف را بررسی کنیم، سپس بعد از آماده شدن داده ها ، آنها را بررسی کرده و به دسته بند های بیز ساده و رگرسیون لجستیک بدهیم تا دسته بندی ها انجام شود و دقت ها در انواع حالت ها با هم مقایسه شود.

۲- سوال اول

در این سوال ابتدا از ما خواسته شده داده هایی که در کنار فایل توضیحات تمرین آپلود شده را لود کنیم. اینکار را با استفاده از pandas انجام میدهیم. داده ها شامل ۳۲۵۶۱ سمپل و ۱۵ ویژگی است . ستون آخر و متناظر با Income خروجی ای است که میخواهیم آموزش دهیم (ویژگی وابسته). بنابراین ۱۴ ویژگی و ۳۲۵۶۱ سمپل داریم.

حال طبق خواسته سوال ستون هایی که مقدار عددی دارند را به خروجی میبریم که شامل ستون های زیر است:

```
['Age', 'FinancialWeight', 'Education-num', 'CapitalGain', 'CapitalLoss', 'HourPerWeek']
```

سپس ستون هایی که مقدار غیر عددی (رشته) دارند را به خروجی میبریم:

```
['WorkClass', 'Education', 'MaritalStatus', 'Occupation', 'Relationship', 'Race', 'Sex', 'NativeCountry', 'Income']
```

سپس تعداد مقادیر خالی در هر ستون را به خروجی میبریم که به شکل زیر است:

```
Age                  0
WorkClass         1836
FinancialWeight      0
Education            0
Education-num        0
MaritalStatus        0
Occupation        1843
Relationship         0
Race                 0
Sex                  0
CapitalGain          0
CapitalLoss          0
HourPerWeek          0
NativeCountry      583
Income               0
```

بنابراین ستون های WorkClass، Occupation و NativeCountry مقادیر خالی و تعریف نشده دارند که نیاز به ویرایش دارند. باید این داده ها را اصلاح کنیم. روش های متنوعی وجود دارد مثلا حذف کامل آن نمونه ، استفاده از مقداری که بیشترین تکرار را دارد و ... . که ما از روش دوم استفاده میکنیم از آنجایی که ستون هایی که دارای مقادیر خالی هستند همگی اسمی (رشته) هستند پس میتوانیم پر تکرار ترین رشته بین سمپل ها را جایگزین مقادیر خالی کنیم.

در واقع ما از روش Impute missing values for categorical variable استفاده کردیم .

سپس مدلی درست میکنیم که برای همه ی ورودی ها 50=> Income را خروجی دهد (ثابت مستقل از اطلاعات ورودی). در اینصورت دقت روی این داده ها برابر 0.7591904425539756 است. یعنی تقریبا ۷۶ درصد داده ها در کلاس 50=> قرار دارند.

۳- سوال دو

در این سوال ابتدا باید داده ها را پیش پردازش کرده و برای انجام عمل Logistic Regression آماده کنیم. در این مرحله ابتدا ستون (فیچر) Education-num را حذف میکنیم چون معادل با ستون دیگری به نام Education است. با استفاده از LabelEncoder مقادیر کتگوریکال را با استفاده از اعداد صحیح لیبل گذاری میکنیم تا مقدار عددی به خود بگیرند. (۰ ، ۱ یا ...) . سپس با استفاده از Cross Validation با K=10 یک مدل Logistic Regression را آموزش میدهیم. دقت میانگین این ۱۰ بار اجرا برابر 0.7991155062805196 است. همانطور که میبینیم دقت این مدل کمی از دقت مدل BaseLine بیشتر است اما همچنان دقت خوبی نیست که نشان از نویزی بودن داده هاست و مدل ما نتوانسته ارتباط خوبی میان داده ها و خروجی ها پیدا کند.

۴- سوال سوم

در این سوال میخواهیم ابتدا داده ها را برای استفاده ی naïve bayes آماده کنیم. آماده سازی داده ها همانند سوال دو خواهد بود چیز اضافه ای نداریم. در این سوال روش های متفاوتی را برای این دسته بندی استفاده کردیم. یعنی با استفاده از توزیع های متنوعی داده ها را آموزش دادیم . که دقت هارا برای هر مورد ذکر میکنیم.

Method	Accuracy
Gaussian Naive Bayes	*0.7948774007071521*
Multinomial Naive Bayes	*0.782592680742819*
Bernoulli Naive Bayes	*0.7284171155832193*
Complement Naive Bayes	*0.782592680742819*

سپس با استفاده از یک مدل ترکیبی از توزیع multinomial برای داده های categorical و توزیع Gaussian برای داده های عددی و پیوسته ، و CV با K=10 داده ها را آموزش دادیم و دقت خوبی را نسبت به حالت های قبل گرفتیم . این دقت برابر 0.8222490989607968 است. که به نسبت حالت های قبل بهتر است. شاید با استفاده از پیش پردازش دیگر بتوانیم این دقت هارا بالا ببریم مثلا شاید استفاده از نرمال سازی و یا استاندارد سازی داده ها کمی دقت مارا افزایش دهد اما افزایش چشمگیری نداریم و در حد صدم است.

۵- سوال چهارم

برای مقایسه نتایج، ابتدا جدولی را تهیه کرده و نتیجه هر قسمت را در سطر های جدول می آوریم:

Method	Accuracy
BaseLine Model	0.7591904425539756
Linear Regression	0.7991155062805196
Naïve Bayes (Multinomial + Gaussian)	0.8222490989607968

علت دقت بالای *BaseLine* این است که داده های اولیه ما تعداد بیشتری برای کلاس *50=>* دارند. یعنی در این روش به دنبال این هستیم که هر نمونه ای که گرفتیم فارغ از اینکه داده های آن چه است کلاس *50=>* را پیش بینی کنیم.

همچنین مقایسه نتیجه *Linear Regression* و *Naïve Bayes* شاید بیانگر این باشد که داده های ما خاصیت *Bias* بیشتری نسبت به *Variance* دارند چون *Naïve Bayes* روشی است که برای زمانی که داده های ما *Bias* بیشتری دارند مناسب تر است. از دقتی که گرفتیم هم این قضیه پیداست. چون نتوانسته ایم دقت روی داده های آموزشی را به عدد خیلی بالایی برسانیم (با وجود تعداد نمونه های زیاد) پس میتوان نتیجه گرفت که احتمالا *Variance* داده ها کمتر از *Bias* داده هاست.

در کل با این نتایج میتوانیم کمی استنباط کنیم که روش خطی نمیتواند راهکار مناسبی باید و از طرفی *naïve bayes* هم خیلی نتیجه خوبی نمیدهد چون زمانی مناسب است که ویژگی های ما باهم *correlated* نباشند که در عمل خیلی بعید است همیشه مقداری ویژگی های ما *correlated* هستند. یا اگر بخواهیم بهتر توضیح دهیم شرط *independence* بودن در اکثر مسائل برقرار نیست و به همین خاطر *naïve bayes* به طور کلی دقت بالایی برای ما به همراه ندارد.