بخش یک

سوال ۱)

ابتدا دیتاست پاکسازی شده ی سوال اول را آپلود شده و سپس داده ها به ۲بخش آموزشی و تست تقسیم می‌شود تا مدل ها را ابتدا با داده های آموزشی فیت کنیم سپس با داده های تست به بررسی عملکرد مدل بپردازیم؛ سپس کلاس رگرسیو خطی را پیاده سازی شده و از هیچ پکیجی استفاده نکردیم همینطور کراس ولیدیشن در ادامه پیاده سازی می‌شود.

حالت ۱ :حال با استفاده از کلاس نوشته شده مدل رگرسیون خود را بر روی ستون 'livingSpaceRange' می سازیم و تارگت را ستون متراژ خانه قرار می‌دهیم.

حالت ۲: مانند حالت ۱ عمل میکنیم با این تفاوت که از پکیج آماده کتابخانه sklearn استفاده میکنیم و همان داده های حالت یک را به آن می‌دهیم.

حالت ۳:مانند حالت ۲ با پکیج ها شروع به ساختن رگرسیون خطی میکنیم با این تفاوت که ورودی آن ۴ فیچر(۲فیچر با کمترین و ۲ فیچر با بیشترین کورریلیشن نسبت به متغیر هدف) است؛ این فیچر ها عبارت اند از:

```
['livingSpaceRange', 'pricetrend', 'thermalChar', 'floor']
```

حالت ۴: با استفاده از پکیج ها مدلی برروی کل داده(همه ی ستون ها) زده شده است که نتایج آن ها در کد قابل مشاهده می‌باشد.

حالت ۵ ،۶: با استفاده از پکیج های کتابخانه sklearn رگرسیون ridge و lasso پیاده سازی می‌شود اطلاعات کراس ولیدیشن در کد قابل مشاهده می‌باشد.

سوال ۲)نتایج داده های تست در فولد های مختلف به شرح زیر است:

برای رگرسیون ridge :

```
[-1.20018641e+02 -2.19416678e+05 -9.94773279e+02 -6.14467954e+05
 -1.29273385e+02 -3.73481680e+03 -1.33324155e+02 -1.33148380e+02
                                  -2.60635011e+02 -1.22763347e+02]
```

برای رگرسیون lasso :

```
[-8.64977508e+01 -2.19417812e+05 -9.73699754e+02 -6.14438116e+05
 -1.01990843e+02 -3.71868310e+03 -1.07178888e+02 -1.02551562e+02
                                  -2.31626666e+02 -8.99023678e+01]
```

لازم به ذکر است برای رگرسیون لاسو و ریدج داده ها مقیاس بندی شده اند که نیاز به این کار حتما وجود داد و پس از تست های مختلف هایپر پارامتر ها ۰.۱ برای ۲مدل انتخاب شده اند

بخش دو

سوال ۱)

روش Ridge

در رگرسیون خطی، از روش کمترین مربعات خطا برای تخمین ضرایب استفاده می‌کریدم و به دنبال یافتن ضرایبی بودیم که خطای زیر را کمینه کند:

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2$$

روش رگرسیون ridge بسیار مشابه روند بالاست اما یک تفاوتی دارد. در این روش به دنبال یافتن ضرایبی هستیم که مقدار خطای زیر را کمینه کند:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

که $\lambda \geq 0$ یک پارامتر تنظیم‌کننده است. در واقع با این تابع خطا هم به دنبال کمینه کردن خطای آموزشی و هم به دنبال کم کردن واریانس ضرایب و فشرده کردن آن‌ها به سمت صفر هستیم. هر چه ضریب λ بزرگتر باشد، فشرده‌سازی بیشتر صورت می‌گیرد و ضرایب بسیار به صفر نزدیک می‌شوند. به ازای هر مقدار λ دسته ضرایب $\hat{\beta}_R^{\lambda}$ متفاوتی خواهیم داشت و باید مقدار λ با CV به درستی انتخاب شود.

روش Lasso

روش ridge یک عیب دارد و آن این است که در این روش، تمام متغیرها در مدل باقی می‌مانند و این باعث می‌شود تفسیرپذیری مدل دشوار باشد. در مثال اعتبار کارت‌های بانکی، پارامترهای income, limit, rating, student مهم‌ترین پارامترها هستند و ما انتظار داریم که مدل نهایی تنها شامل این پارامترها باشد، اما در ridge تمام پارامترهای دیگر هم در مدل باقی می‌ماند و به ندرت پارامتری از مدل حذف می‌شود.

روش Lasso یک جایگزین جدید برای روش ridge است که در آن هدف تخمین ضریب به نحوی ات که تابع خطای زیر کمینه شود:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

این تابع خطا بسیار شبیه به تابع خطای ridge است اما با این تفاوت که در بخش فشرده‌سازی به جای نرم ℓ_2 از نرم ℓ_1 استفاده شده است و همین تفاوت منجر به صفر شدن بسیاری از ضرایب و تولید مدل‌های اسپارس می‌شود. بنابراین با روش Lasso به نوعی انتخاب زیرمجموعه هم انجام می‌شود و این باعث می‌شود تفسیرپذیری مدل آسان‌تر شود.

در رگرسیون ریدج عامل رگورالیزشین باعث بزرگ نشدن زیاد وزن ها میشود و بعضی اوقات کمک میکند که وزن های زیاد مقداری نزدیک ۰ پیدا کنند اما هیچ وقت صفر نمیشوند اما در لاسو بخاطر وجود نرم ۱ مقداری زیادی از وزن ها صفر میشود و میتوان به این مدل به دید فیچر سلکشن هم نگاه کرد چراکه به سادگی مدل کمک زیادی میکند اما در خیلی از دیتاست ها رگرسیون ریدج بهتر عمل میکند.

سوال ۲)

راهکار های متفاوتی برای این موضوع وجود دارد برای مثال اما بهترین آنها انتخاب چند مقدار و چک کردن آنها به کمک کراس ولیدیشن می‌باشد.

گرچه اگر گزینه های انتخابی زیاد باشد هزینه ی محاسباتی برنامه سنگین و زیاد میشود

از روش گرید سرچ نیز برای انتخاب استفاده می‌شود که نحوه ی عملکرد آن شبیه به روش بالاست.

سوال ۳)

افزایش فلد ها کمک میکند تا بایاس و اریب مدل ما کم شود و پیش بینی ما نسب به داده های تست بهتر باشد اما یک مشکل وجود دارد و این است که اگر تعداد فولد ها زیاد شود هزینه ی محاسباتی بالا می‌رود و برای داده های حجیم امکان پذیر نیست، پس تقریبا ترید آفی بین افزایش فولد ها و دقت پیش بینی وجود دارد که نسبت به اندازه ی داده تعداد فولد ها را در نظر میگیریم.

سوال ۴)

در این روش به جای اینکه دسته ای از داده ها در قسمت ولیدیشن قرار بگیرند فقط یک مشاهده یا داده را به عنوان ولیدیشن قرار می‌دهیم؛ نهایتا روش LOOCV خطای زیر را به عنوان تخمینی از نرخ خطای تست گزارش می‌کند:

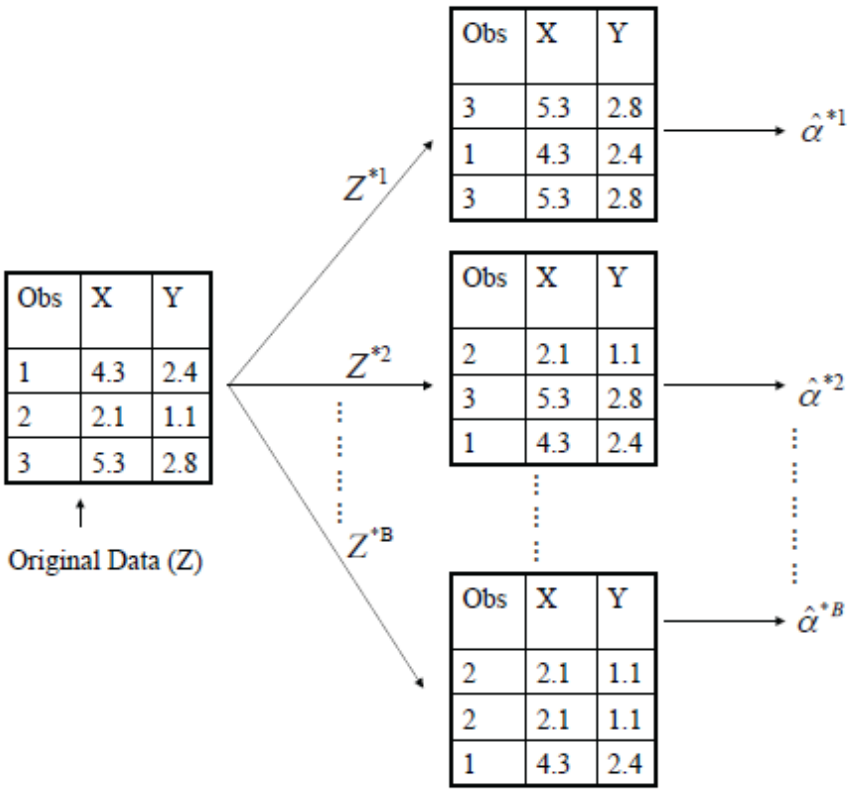
$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

روش LOOCV مزایای زیادی نسبت به روش ارزیابی بخش قبل دارد. این روش اریب bias کمتری دارد. زیرا به جای آنکه مدل را بر اساس بخشی از داده‌های آموزشی بدست آورد، تقریبا بر اساس تمام داده‌های آموزشی (همه به جز یکی) بدست می‌آورد. بنابراین به بخشی از داده‌های آموزشی وابسته نیست و ناریب‌تر است. هم‌چنین میزان خطای ارزیابی که بدست می‌آورد بیشتر از نرخ خطای تست نیست چون تقریبا از تمام داده‌های آموزشی استفاده می‌کند. علاوه بر این، خطایی که با این روش بدست می‌آید، پس از اعمال مجدد روش، تغییر نمی کند چون هیچ بخشی از آن تصادفی نیست.

سوال ۵)

روش bootstrap روش کارآمدی برای محاسبه میزان دقت و خطای استاندارد (standard error) متغیر تخمین زده شده است.به صورت تصادفی n عضو با جایگذاری انتخاب می‌کنیم و مجموعه داده جدید Z_1^* را می‌سازیم. منظور از انتخاب با جایگذاری آن است که یک داده می‌تواند چندین بار انتخاب شود. به عنوان مثال در Z_1^*، داده سوم دوبار انتخاب شده است، داده اول یک بار انتخاب شده است و داده دوم اصلا انتخاب نشده است. لازم به ذکر است که وقتی داده‌ای انتخاب می‌شود، هم X و هم Y انتخاب می‌شود. با این روش مجموعه داده Z_1^* را ساخته و تخمین جدیدی از α به نام $\hat{\alpha}_1^*$ بدست می‌آوریم. این روند را B بار تکرار می‌کنیم و B مجموعه داده $Z_1^*.Z_2^*.\ldots.Z_B^*$ ساخته و تخمین‌های $\hat{\alpha}_1^*.\hat{\alpha}_2^*.\ldots.\hat{\alpha}_B^*$ را محاسبه می‌کنیم. سپس خطای استاندارد $\hat{\alpha}$ از طریق فرمول زیر حساب می‌شود:

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}_r^* - \frac{1}{\mathrm{B}}\sum_{r'=1}^{B}\hat{\alpha}_{r'}^*\right)^2}$$



سوال ۶)

cross fold 5x2 به این معناست که 2 fold-را برای 5 مرتبه تکرار کنیم. به این دلیل از 2 foldاستفاده میشود و به این دلیل استفاده میکنیم که داده train و val فقط یکبار در مدل تاثیر داشته باشند

بخش سه

سوال ۱)

رگرسیون لاجیستیک را در دیتاست فوق پیاده کردیم و معیار های سنجش این مدل به در کد نشان داده شده است.

سوال ۲)

پس از بررسی مشخص شد که دیتا کاملا بالانس است و هر کتگوری در ستون قیمت ۵۰۰ نمونه دارد و با توجه به این که۴ کلاس داریم ۲۰۰۰ نمونه وجود دارد.

سوال۳)

عملیات مربوط به این بخش در کد انجام شده است

سوال۴)

گزارش فعالیت های خواسته شده در کد نشان داده شده است.

سوال ۵)

پس از انجام تغییر کلاس ها به صفر و یک، با داده های ناهمگن مواجه هستیم، ۳روش پیشنهادی برای رفع این مشکل وجود دارد.

روش اول همان بوتس ترپ هست اما کاربرد آن بیشتر در حدس زدن پارامتر های یک توزیع یا داده می‌باشد که در این مورد هم میتوان استفاده شود اما ترجیح روش دیگریست.

در روش بعدی که آپسمپلینگ نام دارد سعی بر این است که مقدار کلاسی که دارای نمومه ی کپتر است را با روش های آماری و جایگزینی زیاد کرد تا دیتا بالانس شود که در این جا ازین روش استفاده کردیم

روش دیگر نیز داون سمپلینگ است که برعکس حالت بالاست اما چون مقدار دیتا کم میشود ترجیح داده شد ازین روش استفاده نکنیم.

سوال ۶ ،۷) با استفاده کدهای پیاده سازی شده در بخش ۶ مدل را فیچر های انتخاب شده آموزش می‌دهیم که نتایج آن در کد نشان داده شده است.

سوال ۸، ۹) نتیاج بعد از اعمال pca به در فایل کد وجود دارد.

سوال ۱۰) به شکل سوال ۶ پیاده سازی شده است

سوال ۱۱)در کد انجام شده است.

بخش چهار

سوال ۱)

گاهی ممکن است به جای دو کلاس، داده‌ها در چند کلاس قرار بگیرند. به عنوان مثال مسئله بیماران اورژانسی که سه حالت سکته، مصرف بیش از حد مواد مخدر و تشنج را داشتند، در این دسته مسایل قرار می‌گیرند. در چنین شرایطی ما باید احتمالات $\Pr(Y =$ stroke$|X)$ و $\Pr(Y =$drug overdose$|X)$, و $\Pr(Y =$ epileptic seizure$|X) = 1 - \Pr(Y =$ stroke$|X) - \Pr(Y =$ drug overdose$|X)$ را مدل کنیم. می توان مدل لاجستیک در بخش قبل را به حالتی با بیش از دو کلاس تعمیم داد، اما معمولا برای حالت چندکلاسه از کلاس بند discriminant analysis استفاده می شود.

سوال ۲)

بله نتایج در حالت همگن بهتر می‌باشد

سوال۳)

بله نتایج ۶ در فولد های مختلف بهتر است چراکه بعضی متغیر ها تاثیری در مدل ندارد و فقط بعد را زیاد میکنند همچنین در حالت ۶ مدل تفسیر پذیر تر است

سوال ۴)

LDA و مدل های آماری

سوال ۶)

روش LDA یک روش یادگریی نظارت شده است که برای تفکی ک کالس ها به کار مریود . درواقع هدف اصیل LDA آن است که م کند تا بتوان به راحیت آن ها را دسته بندی کرد .LDA سیع دارد تا با پیدا ریان تفکیک پذیر ی میان کالس ها را ز یاد را باال ببید. یار های اصیل برای ریان تفکیک پذیر ی مع کردن یک خط جدید)محور(و تصو یر کردن داده ها بر رو ی آن، این م اینکار، میانگ ری داده ها و همچن ری ی پراکندگ (فراوای(آن ها میباشد . این روش با کم کردن م ریان ی پراکندگ داده ها و افزایش فاصله میانگ ری آنها، داده ها را بر رو ی آن محور مورد نظر، که توسط احتماالت ب ریی پیدا میشود، تصو یر میکند و پس از و به راحیت میتوان آن ها را کالس داده ها به راحیت قابل تفکیک هستند ز یرا کالس های مختلف آنها کامال جدا میشود اینکار، بندی کرد.

سوال ۸)

سنجش درسیت مدل است، مانند بقیه معیار ها که قبال استفاده میکردیم و درسیت و دقت 8. این معیار یگ از معیار های درسیت بری ان مدل را با آن بدست میاوردیم، مانند معیار accuracy. این معیار درواقع م ری داده predict شده و م ریان واقیع یی داده را به خو نشان میدهد و بی ی شیبرای کالس بندی های بای یی کاربرد دارد.این معیار توسط تست squared chi بدست ت ما کامال درست بوده، اگر -1 یم آید و اعدادی به عنوان خروجی میدهد . اگر این عدد 1 باشد ه این مع ت است که پیشبی ت باشد به این ت مع است که پیشبی ت ما کامال غلط است)کالس دیگر پیشبی شده(و اگر 0 باشد به این مع ت است که یک ت پیشبی رندم داشتیم و اصال خوب نیست. این معیار توسط فرمول ز یر بدست یم آید:

$$|\mathrm{MCC}| = \sqrt{\frac{\chi^2}{n}}$$