تمرین سری ۲ واحد درسی داده کاوی

جناب آقای دکتر فراهانی
دستیاران آموزشی : نوید کاشی ، علی شریفی

۲۷ اسفند ۱۳۹۹

پیشاپیش سال نو بر شما مبارک باد. با امید سالی سرشار از سلامتی و برکت و شادی.
یا مقلب القلوب و الابصار، یا مدبر اللیل و النهار، یا محول الحول و الاحوال، حول حالنا الی احسن الحال
ای تغییر دهنده دلها و دیده ها، ای مدبر شب و روز، ای گرداننده سال و حالت ها، بگردان حال ما را به نیکوترین حال

۱ بخش ۱

در بخش ۱ تمرین ۲ از شما خواسته شده است که با استفاده از دیتاست تمرین ۱ که به پیش پردازش و پاکسازی آن پرداخته اید (میتوانید دوباره روی دیتاست از اول کار کنید اما مشکلی نخواهد داشت اگر از حاصل کار های خود در تمرین ۱ بهره ببرید .) به انجام تسک های زیر بپردازید .

دیتاست تمرین ۱

۱. در تمرین ۱ شما دو معیار جهت بررسی حاصل مدل های رگرسیونی شما بر روی داده های تست مطرح شد که عبارت بودند از MSE و Accuracy . برای هر یک از مدل های رگرسیونی زیر با توجه به دو معیار فوق ، 5-fold Cross Validation و 10-fold Cross Validation را اجرا کنید .

- حالت (۱) : با استفاده از کدهای پیاده سازی شده در تمرین ۱ برای مدل رگرسیونی ، یک مدل رگرسیون بسازید (بدون استفاده از پکیج ها) که ورودی آن یک فیچر است که بیشترین کورولیشن را با فیچر متراژ خانه [1] دارد (در صورتی که بیش از یک فیچر بیشترین کورولیشن را دارد ، یکی را به دلخواه انتخاب کنید .) و تارگت آن متراژ خانه باشد .
- حالت (۲) : مشابه حالت (۱) عمل کنید با این تفاوت که از پکیج ها جهت ساخت مدل خود استفاده کنید . یعنی با استفاده از پکیج ها یک مدل رگرسیون بسازید که ورودی آن فیچر با بیشترین کورولیشن است و تارگت ما متراژ خانه میباشد .
- حالت (۳) : با استفاده از پکیج ها یک مدل بسازید که ورودی آن ۲ فیچر با بیشترین مقدار و ۲ فیچر با کمترین مقدار کورولیشن میباشند . (در مجموع ۴ فیچر) و تارگت متراژ خانه
- حالت (۴) : استفاده از پکیج برای ساخت مدل رگرسیون روی فیچرهای دلخواه و تارگت متراژ خانه
- حالت (۵) : استفاده از پکیج برای ساخت رگرسیون Ridge رو فیچر های دلخواه و تارگت متراژ خانه
- حالت (۶) : استفاده از پکیج برای ساخت رگرسیون Lasso رو فیچرهای دلخواه و تارگت متراژ خانه

۲. آیا به ازای فیچرهای ورودی یکسان و فولد های یکسان تفاوتی بین رگرسیون و رگرسیون Lasso ، رگرسیون Ridge مشاهده شد . در هنگام پیاده سازی خروجی فولد های یکسان را برای سه مدل نمایش دهید .

[1] LivingSpace

۲ بخش ۲

در این بخش لزومی به پیاده سازی کد نیست و با درج منبع در کتاب های رفرنس یا مقالات به پرسش های زیر پاسخ دهید .

۱. رگرسیون LASSO و رگرسیون RIDGE را هر کدام در یک پاراگراف توضیح دهید و سپس با هم مقایسه کنید . کاربرد های هر کدام را بیان کنید .

۲. راهکار هایی مناسب جهت انتخاب پارامتر مناسب برای ضریب ترم regularization در رگرسیون Ridge و رگرسیون Lasso پیشنهاد دهید .

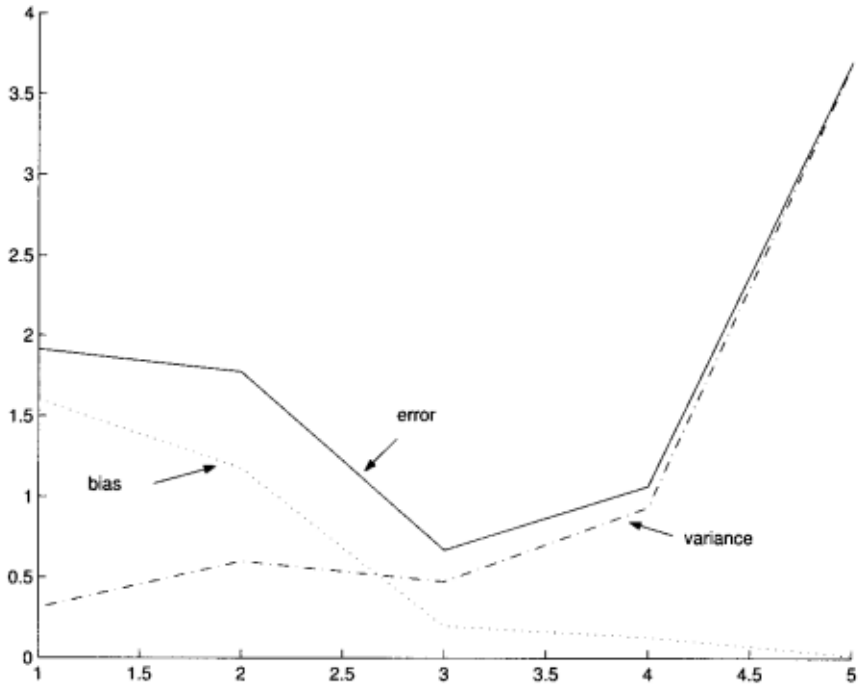
۳. بررسی کنید که به صورت کلی افزایش تعداد فولدها چه تاثیراتی بر مدل ها دارند و اصلا چرا ما باید گاها به دنبال افزایش تعداد فولد ها باشیم .

۴. Leave one out چیست و در چه جاهایی از آن استفاده میکنیم ؟

۵. (بخش امتیازی) Bootstrapping چیست و چه تفاوتی با Cross Validation دارد؟ در کجا ها از Bootstrapping استفاده میشود .

۶. 5x2 cross validation را در یک پاراگراف توضیح دهید سپس بیان کنید در چه جاهایی استفاده از این روش میتواند مفید باشد .

۷. (بخش امتیازی) آیا میتوان با استفاده از روش Elbow با استفاده نموداری مشابه نمودار زیر که نمایان گر بایاس ، واریانس و مرتبه مدل است . بهترین مرتبه مدل برای پیچیدگی مدل را یافت ؟
به طور مثال با استفاده از روش elbow میتوان در نظر گرفت که بر روی دیتاست ، مدلی از مرتبه ۳ جواب خوبی به ما میدهد . آیا همواره در تمامی مسائل و **نه صرفا بحث تحلیلی** میتوان اینگونه قضاوت کرد و مرتبه مناسب را به دست آورد ؟(راهنمایی برای پاسخ به این سوال توجه به مفهوم بایاس میتواند کمک کننده باشد.)
در شکل (۱) خطا از حاصل جمع توان بایاس و واریانس به دست می آید .



شکل ۱: نمودار بایاس و واریانس بنا بر مرتبه های مختلف مدل

۳ بخش ۳

در این تمرین با استفاده از دیتاست بخش ۳ به بررسی مسئله کلاس بندی میپردازیم .
توجه کنید فقط در سوال ۴ و ۵ میباشد که مسئله شما شامل دو کلاس است و در سایر سوالات شما با دیتاست اصلی که ۴ کلاسه است کار میکنید .
پیش پردازش ها و پاکسازی مناسب را همانطور که در تمرین ۱ توضیح داده شده بود برای این دیتاست هم اعمال کنید .

۱. با استفاده از پکیج و تمامی فیچرهای دیتاست یک رگرسیون لجستیگ پیاده سازی کنید . تارگت در این مسئله ستون price-range می باشد .معیار های precision ، recall ، f1-score را گزارش دهید .

۲. در این تمرین ما ۴ کلاس برای قیمت داریم که توسط ستون price-range مشخص میگردد . بررسی بکنید آیا تعداد نمونه های هر کلاس متوازن است یعنی تعداد نمونه ها در هر کلاس با هم برابر است (یا نزدیک به برابر است)

۳. یک کلاس بندی جدید برای نمونه ها تعریف کنید . تمامی نمونه هایی که دارای کلاس ۱و۲و۳ هستند را لیبل کلاس آنها را ۱ و نمونه ها با لیبل کلاس ۰ را دست نزنید . بعد از انجام این کار شما دو کلاس ۰ و ۱ را خواهید داشت .

۴. با استفاده از پکیج بر روی داده هایی که در حالت قبل تبدیل به دو کلاس ۰ و ۱ شده اند یک رگرسیون لجستیک پیاده سازی کنید که تمامی فیچرهای را شامل میشود و معیار های precision ، recall ، f1-score را گزارش دهید .

۵. در صورتی که داده های هر کلاس متوازن نباشد اصطلاحا Imbalanced data داشته باشیم ممکن چه مشکلاتی برای ما ایجاد کند . سه راه حل برای رفع این مشکل را هر کدام در یک پاراگراف شرح دهید . به دلخواه یکی از این سه روش را بر روی داده های بخش ۳ اعمال کنید و سپس یک مدل رگرسیون لجستیک با استفاده از پکیج روی حاصل داده ها اجرا کنید .

۶. روش انتخاب ویژگی Forward Selection را پیاده سازی کنید . برای معیار انتخاب فیچر جدید در هر مرحله از AUC استفاده کنید .
در روش انتخاب پیشرو ما از یک مجموعه تهی شروع کرده و در هرگام سعی داریم فیچر را به مجموعه فیچرهای انتخابی اضافه کنیم که AUC را افرایش دهد .

۷. با استفاده از کد پیاده سازی شده در بخش قبل به انتخاب ویژگی ها از فیچر ها بپردازید و سپس مدل لجستیک (با استفاده از پکیج) را بر روی فیچرهای انتخاب شده اجرا کنید و معیار های precision ، recall ، f1-score را گزارش کنید .

۸. با استفاده از الگوریتم PCA در حالتی که تعداد Component ها با تعداد فیچرهای انتخابی حاصل روش انتخاب ویژگی پیشرو برابر باشد (یعنی اگر در سوال ۷ شما با استفاده از انتخاب ویژگی پیشرو به طور مثال ۵ فیچر را انتخاب کردید در الگوریتم PCA هم به عنوان آرگومان ورودی تعداد Component را ۵ درج کنید .) دیتاست را تغییر دهید .

۹. با استفاده از دیتاست تغییر یافته در سوال ۸ و به کمک پکیج ها یک رگرسیون لجستیک را پیاده سازی کنید و معیار های precision ، recall ، f1-score را گزارش کنید .

۱۰. (بخش امتیازی) روش انتخاب ویژگی Backward Selection را پیاده سازی کنید و با استفاده از فیچرهای انتخاب شده و کمک پکیج یک رگرسیون لجستیک را پیاده سازی کنید . معیار های precision ، recall ، f1-score را گزارش کنید و نتایج را با سوال (۷) بخش ۳ مقایسه کنید .

۱۱. با استفاده از 5-Fold Cross Validation و 10-Fold Cross Validation روی کلیه فیچر ها و به کمک پکیج رگرسیون لجستیک پیاده سازی کنید و نتایج را گزارش دهید .

۴ بخش ۴

در این بخش به پیاده سازی نیازی نیست و با استفاده از نتایج خروجی بخش ۳ و کتاب ها و مقالات به پرسش های زیر پاسخ دهید .

۱. مسئله رگرسیون لجستیک را چگونه برای حالت چند کلاسه Multi Class یعنی بیش از دو کلاس حل میکنیم . به صورت مختصر توضیح دهید .

۲. نتایج سوال (۴) و سوال (۵) بخش ۳ را با هم مقایسه کنید . آیا تفاوت محسوسی مشاهده میشود .

۳. نتایج سوال (۱) و سوال (۶) بخش ۳ را با هم مقایسه کنید . آیا تفاوت محسوسی مشاهده میشود . به صورت کلی چرا ما به دنبال Feature Selection می باشیم .

۴. ۲ روش دیگر Feature Selection را معرفی و هر کدام را حداقل در یک پاراگراف توضیح دهید .

۵. روش های انتخاب ویژگی های پیشرو و پسرو چه معایبی دارند و چگونه میتوان این معایب را حل نمود ؟

۶. روش LDA(Linear discriminant analysis) را به صورت کامل توضیح دهید . با استفاده از مقالات و کتاب های مرجع بررسی کنید ، در مسائل یکسان LDA کارمندتر است و یا PCA .

۷. (بخش امتیازی) چگونه میتوان با استفاده از statistical significance tests به مقایسه مدل ها پرداخت ؟ (توضیح کامل) [۲]

۸. معیار Matthews Correlation Coefficient(MCC) چیست و در چه جاهایی استفاده میشود .

[۲] راهنمایی : Statistical Significance Tests for Comparing Models را جستجو کنید .